VARIANCE-REDUCED CONFORMAL PREDICTION VIA SHRINKAGE FOR LONG-TAILED RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Class-conditional conformal prediction gives every class its own coverage guarantee, but in long-tailed recognition most classes are calibrated from a handful of examples, so the per-class threshold is noisy and the resulting prediction sets are needlessly large. We propose **SARC-CP** (Shrinkage-Adjusted Rank-Calibrated Conformal Prediction): instead of treating each class-conditional quantile as an isolated estimate, we pool information across classes by shrinking each one toward a global quantile with James-Stein-style weights, $q_y^* = (n_y q_y + \kappa q_{\text{global}})/(n_y + \kappa)$. The weight κ acts as a pseudo-count, so tail classes borrow more strength from the global estimate while head classes are left essentially untouched. We give an exact bias-variance condition under which this shrunk estimator has lower mean-squared error than the plug-in quantile, and show it is compatible with rank-calibrated conformal prediction (RC3P) out of the box, adding a single hyperparameter. Across CIFAR-10/100-LT, ImageNet-LT, and iNaturalist-2018, shrinkage tightens prediction sets at matched or better coverage in every setting we test: up to 19% smaller sets on CIFAR-100-LT relative to RC3P (8.6σ), 55–65% smaller on ImageNet-LT, and roughly 80% smaller on iNaturalist-2018, with the largest gains where calibration data are scarcest. Code: `https://anonymous.4open.science/r/sarc-cp-shrinkage-0CE7/README.md`

1 INTRODUCTION

Deploying classifiers in safety-critical settings usually means the point prediction is not enough; the system also needs to say how confident it is (Guo et al., 2017). Conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021; Papadopoulos et al., 2002; Lei et al., 2018) answers this with a guarantee rather than a heuristic: wrap any trained classifier and it returns a set of labels that contains the true one with probability at least $1 - \alpha$, under exchangeability alone. A singleton set means the model is confident; a large one flags ambiguity (Sadinle et al., 2019; Angelopoulos et al., 2021; Romano et al., 2020).

The catch is that the standard guarantee is *marginal* – averaged over the whole data distribution (Vovk, 2013; Barber et al., 2021b; Bates et al., 2021) – and marginal coverage can hide bad behavior on individual classes (He & Garcia, 2009; Buda et al., 2018). Take a wildlife classifier trained on iNaturalist (Horn et al., 2018; Liu et al., 2019; Zhang et al., 2023): common species get tight, useful prediction sets, while rare species calibrated from a handful of images end up either under-covered or wrapped in sets so large they are useless. *Class-conditional* conformal prediction tries to fix this by giving every class its own guarantee (Romano et al., 2020; Ding et al., 2023; Shi et al., 2024; Ding et al., 2026; Vovk, 2013).

The problem is estimation variance, not modeling. Class-conditional conformal prediction estimates a threshold q_y per class from that class's calibration scores alone. In the long tail this can mean 5–15 examples, so q_y is a high-variance estimate of the true quantile – and that variance shows up directly as inflated, unstable prediction sets. This holds even when the underlying classifier is strong: no amount of better long-tail training (Cao et al., 2019; Kang et al., 2020; Menon et al., 2021; Cui et al., 2019; Zhong et al., 2021) fixes a threshold estimated from too few points, because the problem is statistical rather than architectural.

Prior work addresses imbalance through different score functions (Romano et al., 2020), relaxed coverage targets (Ding et al., 2023), or rank-based filtering (Shi et al., 2024), but none of these

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

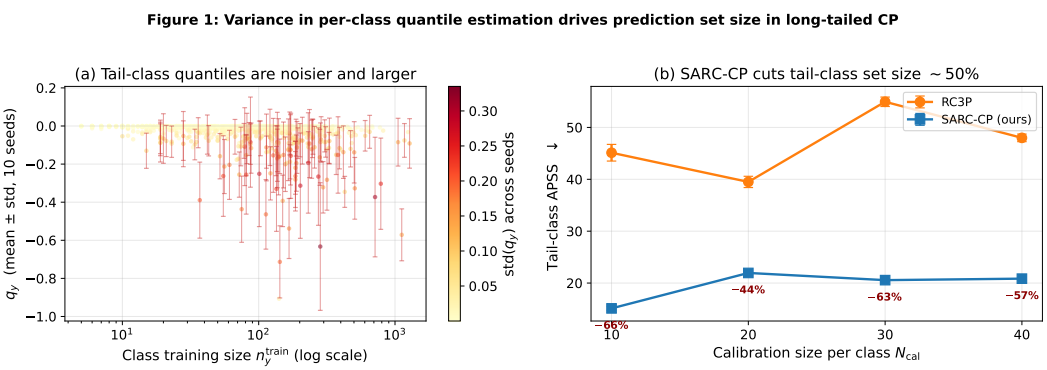


Figure 1: **Motivation for SARC-CP.** Left: variance of class-conditional quantile estimates increases sharply in the long tail. Right: shrinkage stabilises thresholds and substantially reduces prediction set size for rare classes while preserving coverage.

directly reduces the variance of the quantile estimator itself. That is the gap we target: can the class-conditional quantiles be estimated more precisely without giving up their adaptivity?

Our approach. We treat $\{q_y\}_{y=1}^{K}$ as a simultaneous estimation problem and borrow strength across classes with James-Stein shrinkage (James & Stein, 1961). Each q_y is noisy because it only sees one class's data, but the global quantile q_{global}, computed from the full calibration set, is comparatively stable. Shrinking each class threshold toward it,

$$q_y^* = \frac{n_y\, q_y + \kappa\, q_{\text{global}}}{n_y + \kappa}, \tag{1}$$

gives tail classes (small n_y) a large pull toward the stable global estimate while leaving head classes close to their original, well-estimated thresholds.

Figure 1 shows why this matters: quantile variance grows sharply as class frequency drops, and shrinkage brings prediction-set size back down in exactly that regime, without giving up coverage.

We build SARC-CP by combining shrinkage with rank-calibrated conformal prediction (Shi et al., 2024): rank-based filtering controls undercoverage, and shrinkage stabilises the underlying thresholds. SARC-CP needs no retraining and adds a single hyperparameter, which we find behaves consistently across datasets.

2 Related Work

Conformal prediction. Our starting point is split conformal prediction (Papadopoulos et al., 2002; Lei et al., 2018; Vovk et al., 2005). Score-based classification variants include LAC (Romano et al., 2020; Sadinle et al., 2019), RAPS (Angelopoulos et al., 2021), and APS (Romano et al., 2020); risk-controlling variants appear in Bates et al. (2021), and PAC-style confidence sets in Park et al. (2020); Cauchois et al. (2021). All of these target marginal coverage and can still behave poorly class-by-class. Tibshirani et al. (2019) and Barber et al. (2021b) study conformal validity under covariate shift and beyond exchangeability.

Class-conditional and group-conditional conformal prediction. Romano et al. (2020) and Vovk (2013) pose class-conditional and conditional-validity coverage as goals, while Barber et al. (2021a) show exact distribution-free conditional coverage is impossible without extra structure – which is exactly why partial-conditioning relaxations like classwise and cluster-based methods exist. Ding et al. (2023) formalise a class-wise guarantee for the many-class regime and propose Cluster-CP, which groups classes by score-distribution similarity and applies a per-cluster threshold. Shi et al. (2024) propose RC3P, combining per-class thresholds with rank-based gating and an inflated quantile to control undercoverage; we build directly on this. Ding et al. (2026) propose Interp-Q, which linearly interpolates each per-class quantile toward the global one with a fixed weight τ; SARC-CP differs by making that weight sample-size-adaptive, so tail classes shrink more than head classes. Angelopoulos et al. (2021) study Mondrian conformal prediction, conditioning on arbitrary groups,

and Fisch et al. (2021) study few-shot conformal calibration through a meta-learning lens rather than shrinkage.

Long-tailed recognition. Class imbalance is an old problem (He & Garcia, 2009; Buda et al., 2018; Chawla et al., 2002); benchmarks like ImageNet-LT (Liu et al., 2019) and iNaturalist (Horn et al., 2018) have driven a large body of work, surveyed in Zhang et al. (2023), spanning reweighted losses (Cui et al., 2019), margin adjustments (Menon et al., 2021; Cao et al., 2019), decoupled training (Kang et al., 2020; Yang & Xu, 2020), equalization losses (Tan et al., 2020), and contrastive representations for the tail (Cui et al., 2021). On the conformal side, Shi et al. (2024) is the closest antecedent, and Ding et al. (2026) propose the quantile-interpolation scheme discussed above. Calibrating long-tailed classifiers themselves (Zhong et al., 2021; Guo et al., 2017) is complementary to intervening at the threshold level, as we do here; Stutz et al. (2022) study post-hoc score calibration for conformal prediction, though not in the long-tailed setting.

Shrinkage and James-Stein estimation. James-Stein shrinkage (James & Stein, 1961; Stein, 1981) is a classical tool for simultaneous estimation. Efron & Morris (1973) give the empirical-Bayes reading we build on: q_y is a noisy observation of a class-specific random effect, and q_y^* approximates the posterior mean under a Gaussian prior. Hierarchical shrinkage is standard in quantile regression (Koenker, 2005) and Bayesian multilevel models (Gelman & Hill, 2006), but to our knowledge it has not previously been used for conformal quantile estimation specifically.

Variance reduction in uncertainty quantification. Pooling and regularisation reduce variance in Bayesian inference (Gelman & Hill, 2006) and shrinkage-based predictive distributions (Efron & Morris, 1973; Stein, 1981), but distribution-free conformal prediction has mostly stayed outside this line of work. We apply shrinkage directly at the conformal quantile-estimation step for the long-tail setting.

3 PRELIMINARIES

Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{1, \dots, K\}$ the label space, with classifier outputs $\hat{p}(y \mid x)$. Given a calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$, conformal prediction builds a set-valued predictor

$$\mathcal{C}(x) = \{y \in \mathcal{Y} : s(x, y) \le q\}, \tag{2}$$

where $s(x, y)$ is a nonconformity score and q an empirical quantile of the calibration scores. We use the standard highest-probability score $s(x, y) = -\hat{p}(y \mid x)$. Under exchangeability the resulting set satisfies marginal coverage, $\Pr(Y \in \mathcal{C}(X)) \ge 1 - \alpha$.

Class-conditional conformal prediction replaces the single global threshold with a class-specific q_y estimated from $\mathcal{S}_y = \{s(x_i, y_i) : (x_i, y_i) \in \mathcal{D}_{\text{cal}},\ y_i = y\}$, giving

$$\mathcal{C}(x) = \{y \in \mathcal{Y} : s(x, y) \le q_y\}. \tag{3}$$

This is better for tail reliability, but q_y becomes high-variance whenever $n_y = |\mathcal{S}_y|$ is small.

We build on RC3P (Shi et al., 2024), which pairs class-specific thresholds with rank-based filtering. RC3P computes an inflated quantile

$$q_y = \text{Quantile}\left(\mathcal{S}_y,\ \alpha - \frac{g}{\sqrt{n_y}}\right), \tag{4}$$

and a class-specific rank threshold $\hat{k}_y$ to control undercoverage, giving the final set

$$\mathcal{C}(x) = \{y \in \mathcal{Y} : s(x, y) \le q_y \text{ and } \text{rank}(y \mid x) \le \hat{k}_y\}. \tag{5}$$

RC3P improves class-conditional validity, but the thresholds it starts from are still high-variance in the tail. That is what SARC-CP targets.

Where the variance comes from. Write F_y for the CDF of nonconformity scores in class y and $q_y^\star = F_y^{-1}(1 - \alpha)$ for the true quantile. The plug-in estimator $q_y = \hat{F}_y^{-1}(1 - \alpha)$ satisfies, by the quantile CLT (Serfling, 1980),

$$\text{Var}(q_y) \approx \frac{\alpha(1 - \alpha)}{n_y [f_y(q_y^\star)]^2}, \tag{6}$$

with f_y the class-conditional score density. For $n_y = 5$–20, typical of the ImageNet-LT tail, $\mathrm{Var}(q_y)$ can be one to two orders of magnitude larger than for a head class – and a single $+1\sigma$ fluctuation directly inflates the threshold and admits spurious labels into the prediction set.

The global quantile $q_{\mathrm{global}} = \mathrm{Quantile}(\{s(x_i, y_i)\}_{i=1}^{n}, 1-\alpha)$, in contrast, uses all $n \gg n_y$ calibration points and has variance $\mathrm{Var}(q_{\mathrm{global}}) \approx \alpha(1-\alpha)/\{n[f(q^\star)]^2\}$, with f the marginal density. The tradeoff is now explicit: q_y is unbiased for $q_y^\star$ but noisy; q_{global} is stable but biased toward the marginal quantile, which need not match $q_y^\star$ for a tail class.

4 METHOD: SARC-CP

4.1 SHRINKAGE-ADJUSTED QUANTILE ESTIMATION

We replace the plug-in q_y with a shrunk estimator that interpolates between the class-specific and global quantiles:

$$\boxed{q_y^* = \frac{n_y\, q_y + \kappa\, q_{\mathrm{global}}}{n_y + \kappa},} \tag{7}$$

with $\kappa > 0$ a shrinkage hyperparameter that plays the role of a pseudo-count: it is the weight of "prior evidence" contributed by the global calibration set. When $n_y \gg \kappa$, $q_y^* \approx q_y$ and class specificity is preserved; when $n_y \ll \kappa$, $q_y^* \approx q_{\mathrm{global}}$ and the estimator falls back on the stable global quantile.

Equivalently,

$$q_y^* = q_{\mathrm{global}} + \frac{n_y}{n_y + \kappa}(q_y - q_{\mathrm{global}}) = q_{\mathrm{global}} + (1 - \lambda_y)(q_y - q_{\mathrm{global}}), \tag{8}$$

where $\lambda_y = \kappa/(n_y + \kappa) \in (0, 1)$ is a class-specific shrinkage factor that adapts automatically to sample size. Structurally this is a James-Stein-style estimator (James & Stein, 1961) applied to quantiles instead of means; under the empirical-Bayes reading of Efron & Morris (1973), it is the posterior mean of $q_y^\star$ under an exchangeable prior with q_{global} standing in for the prior mean. We make this precise, and give the exact condition under which it reduces MSE, in Proposition 1.

4.2 WHEN DOES SHRINKAGE HELP? A BIAS-VARIANCE CONDITION

The intuition above only becomes a guarantee once we state what "borrowing strength" costs in bias. Model q_y as unbiased for $q_y^\star$ with variance σ_y^2 (Eq. 6), and q_{global} as an estimator with its own variance $\sigma_g^2 \ll \sigma_y^2$ but a class-specific bias $b_y = \mathbb{E}[q_{\mathrm{global}}] - q_y^\star$, reflecting that the global quantile targets the marginal distribution rather than class y's. Both are reasonable second-order approximations to the finite-sample behaviour of empirical quantiles.

Proposition 1 (MSE of the shrunk estimator)**.** *Let $q_y^* = (1-\lambda_y)q_y + \lambda_y q_{\mathrm{global}}$ with $\lambda_y \in (0,1)$, and suppose q_y and q_{global} are (asymptotically) uncorrelated. Then*

$$\mathrm{MSE}(q_y^*) = (1-\lambda_y)^2\sigma_y^2 + \lambda_y^2\left(\sigma_g^2 + b_y^2\right), \tag{9}$$

and $\mathrm{MSE}(q_y^*) < \mathrm{MSE}(q_y) = \sigma_y^2$ if and only if

$$\lambda_y < \frac{2\sigma_y^2}{\sigma_y^2 + \sigma_g^2 + b_y^2}. \tag{10}$$

In particular, if $\sigma_g^2 + b_y^2 < \sigma_y^2$ – i.e. the global estimator's own error is smaller than the class estimator's variance – then every $\lambda_y \in (0,1)$*, and hence every $\kappa > 0$, strictly reduces MSE for class y.*

Proof. Bias-variance decomposition of q_y^* gives $\mathrm{MSE}(q_y^*) = \mathrm{Var}(q_y^*) + \mathrm{Bias}(q_y^*)^2$. Since q_y is unbiased, $\mathrm{Bias}(q_y^*) = \lambda_y b_y$, and by uncorrelatedness $\mathrm{Var}(q_y^*) = (1-\lambda_y)^2\sigma_y^2 + \lambda_y^2\sigma_g^2$, giving the stated expression. Setting $\mathrm{MSE}(q_y^*) < \sigma_y^2$ and expanding $(1-\lambda_y)^2 = 1 - 2\lambda_y + \lambda_y^2$ yields $-2\lambda_y\sigma_y^2 + \lambda_y^2(\sigma_y^2 + \sigma_g^2 + b_y^2) < 0$; dividing by $\lambda_y > 0$ gives Eq. 10. The second claim follows because the right-hand side of Eq. 10 exceeds 1 exactly when $\sigma_g^2 + b_y^2 < \sigma_y^2$. □

Algorithm 1: SARC-CP Calibration

Input: Calibration set $\mathcal{D}_{\text{cal}}$; pre-trained model $\hat{p}$; hyperparameters $\alpha, \kappa, g, \tau_{\text{UCR}}$.
Output: Thresholds $\{q_y^*\}_{y=1}^K$ and rank thresholds $\{\hat{k}_y\}_{y=1}^K$.

Compute $s_i = -\hat{p}(y_i \mid x_i)$ for all $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$.
Compute $q_{\text{global}} = \text{Quantile}(\{s_i\}, 1-\alpha)$.
for each class $y = 1, \ldots, K$ **do**
  Collect $\mathcal{S}_y = \{s_i : y_i = y\}$; $n_y = |\mathcal{S}_y|$.
  $q_y \leftarrow \text{Quantile}(\mathcal{S}_y, \alpha - g/\sqrt{n_y})$ (or q_{global} if $n_y = 0$).
  $q_y^* \leftarrow (n_y q_y + \kappa q_{\text{global}})/(n_y + \kappa)$.
  Compute rank threshold $\hat{k}_y$ at level τ_{UCR}.
end for
return $\{q_y^*\}$, $\{\hat{k}_y\}$.

Two things follow directly. First, because $\lambda_y = \kappa/(n_y + \kappa)$ is decreasing in n_y, small classes automatically sit in the regime where Eq. 10 is easiest to satisfy, which is why shrinkage helps the tail specifically rather than uniformly. Second, the bound depends on κ only through λ_y, so it gives a per-class upper bound on κ beyond which shrinkage starts to hurt that class – consistent with the κ-ablation in Fig. 2, where set size keeps falling but marginal coverage eventually approaches its floor as κ grows. We use this to motivate the coarse grid search over κ described in Sec. 5 rather than to derive a closed-form optimum, since σ_y^2, σ_g^2, and b_y are themselves unknown in practice and must be estimated or tuned.

4.3 Full SARC-CP Procedure

SARC-CP inserts shrinkage into the RC3P calibration pipeline.

Step 1: Compute nonconformity scores. For each calibration point $(x_i, y_i) \in \mathcal{D}_{\text{cal}}$, compute $s_i = s(x_i, y_i) = -\hat{p}(y_i \mid x_i)$.

Step 2: Per-class quantile with inflation. For each class y, compute an inflated-level quantile from the class-y calibration scores:

$$q_y = \text{Quantile}\left(\mathcal{S}_y,\ \alpha - \frac{g}{\sqrt{n_y}}\right), \tag{11}$$

where $g > 0$ is a small inflation constant, following Shi et al. (2024). If $n_y = 0$, set $q_y = q_{\text{global}}$.

Step 3: Compute global quantile.

$$q_{\text{global}} = \text{Quantile}(\{s_i\}_{i=1}^n,\ 1-\alpha)\,. \tag{12}$$

Step 4: Shrink. Apply Eq. 7 to obtain q_y^* for each class.

Step 5: Rank thresholds. For each class y with $n_y \geq 1$, compute

$$\hat{k}_y = \min\left\{k \in \{1, \ldots, K\} : \frac{|\{i : y_i = y,\ \text{rank}(y \mid x_i) > k\}|}{n_y} \leq \tau_{\text{UCR}}\right\}, \tag{13}$$

where τ_{UCR} is the target undercoverage rate; if no such k exists, $\hat{k}_y = K$.

Step 6: Prediction set at test time. For a test input x,

$$\mathcal{C}(x) = \left\{y \in \mathcal{Y} : s(x, y) \leq q_y^* \text{ and } \text{rank}(y \mid x) \leq \hat{k}_y\right\}. \tag{14}$$

SARC-CP is summarised in Algorithm 1. Calibration is $O(nK)$ in time and $O(n + K)$ in space, matching RC3P exactly; computing q_y^* from q_y, q_{global}, and n_y is a constant-time addition per class.

5 EXPERIMENTAL ANALYSIS

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate on four long-tailed benchmarks: CIFAR-10-LT and CIFAR-100-LT (Cui et al., 2019) with imbalance ratios $\rho \in \{0.1, 0.5\}$ under exponential imbalance; ImageNet-LT (Liu et al., 2019), 1000 classes with frequencies from 5 to 1280; and iNaturalist-2018 (Horn et al., 2018), 8142 classes with a much heavier tail.

Backbones. CIFAR experiments use ResNet-20 trained with LDAM loss (Cao et al., 2019) for 200 epochs. For ImageNet-LT we use the published cRT+SSP ResNet-50 (Kang et al., 2020; Yang & Xu, 2020) (51.35% top-1). For iNaturalist-2018 we use the published InceptionV3 checkpoint (69.40% top-1 on the aggregated train+val pool used for calibration). All baselines share the backbone used for SARC-CP in each setting.

Calibration protocol. We follow the RC3P protocol (Shi et al., 2024): allocate N calibration samples per class from the test set (the "partial-LT" regime). $N \in \{500, 50\}$ for CIFAR-10/100-LT respectively; $N = 30$ for ImageNet-LT. For iNaturalist-2018, the official 3-image-per-class validation split violates exchangeability for class-conditional CP (Ding et al., 2026), so we instead follow the Appendix C.1 protocol of Ding et al. (2026): pool train+val into 200,978 images and take a random 80/10/10 split. All numbers are averaged over 10 random seeds with standard deviations reported; $\alpha = 0.1$ throughout.

Metrics. Avg Size (average prediction-set cardinality, lower is better); Marg. Cov. (marginal coverage, target $\geq 1 - \alpha$); UCR (fraction of classes with coverage below $1 - \alpha$, lower is better); Min Cov. (minimum per-class coverage); CovGap ($\frac{1}{K}\sum_y \max(1 - \alpha - \mathrm{cov}_y, 0)$, average per-class undercoverage).

Baselines. HPS-Standard (global threshold, HPS score); Standard-APS (Romano et al., 2020); ClasswiseCP; Cluster-CP (Ding et al., 2023); RC3P (Shi et al., 2024); Interp-Q (Ding et al., 2026). All baselines use HPS scores unless noted, and are reproduced on the same backbone as SARC-CP. Hyperparameters for SARC-CP: $\kappa = 70$, $\tau_{\mathrm{UCR}} = 0.02$, $g = 0.5$ on CIFAR-100-LT; $\kappa = 25$, $\tau_{\mathrm{UCR}} = 0.04$, $g = 0.5$ on CIFAR-10-LT; $\kappa \in \{100, 150\}$, $\tau_{\mathrm{UCR}} \in \{0.005, 0.03\}$, $g \in \{0.25, 0.5\}$ on ImageNet-LT (chosen per N_{cal}); $\kappa = 25$, $\tau_{\mathrm{UCR}} = 0.001$, $g = 0.25$ on iNaturalist-2018. These come from a coarse grid search on a held-out slice of the calibration set.

5.2 MAIN RESULTS: MATCHED-UCR FAIR COMPARISON

Following RC3P (Shi et al., 2024), we compare methods at *matched UCR*: every method is tuned to the same undercoverage rate, so the comparison isn't just trading smaller sets for worse class coverage. The target UCR is the strictest level jointly reachable by RC3P and SARC-CP while keeping marginal coverage ≥ 0.895; each baseline is then tuned over its own grid to minimise average prediction set size (APSS) subject to that target. Methods that cannot reach the target are marked * in Table 1.

SARC-CP gives the smallest APSS on all six configurations in Table 1, spanning CIFAR-10/100-LT, ImageNet-LT, and iNaturalist-2018. The margin over RC3P ranges from a modest -0.9% on CIFAR-10-LT (ρ=0.5) up to -79.8% on iNaturalist-2018. CIFAR-100-LT gains are the largest among the CIFAR settings (-34.7% at ρ=0.5, -35.3% at ρ=0.1); ImageNet-LT set size roughly halves (-55.6%); and on iNaturalist-2018, with 8,142 classes, the gap widens to about $5\times$.

5.3 CIFAR-10-LT AND CIFAR-100-LT RESULTS

Table 2 reports each method at its own natural operating point, without UCR matching. SARC-CP is best or second-best among all valid methods (marginal coverage ≥ 0.895), and is always smallest among the class-conditional CP methods on CIFAR-100-LT.

On CIFAR-10-LT the gains are small but consistent (-0.5% to -1.2%), which makes sense: with only 10 classes, the estimation-variance problem SARC-CP targets is not very severe to begin with. CIFAR-100-LT is a better stress test – 100 classes calibrated from $N = 50$ examples each – and there shrinkage gives -19% ($\rho = 0.1$, 8.6σ across seeds) and -16% ($\rho = 0.5$), with CovGap staying within 0.006 of RC3P or better.

Table 1: **Matched-UCR fair comparison.** UCR target = strictest level achievable by both SARC-CP and RC3P at marg ≥ 0.895 (both methods compete on equal footing). Each baseline is tuned over its hyperparameter grid; we report the smallest APSS achieving UCR $\leq$ target with marg ≥ 0.895. Methods that cannot reach the target are marked *. **Bold** = smallest matched APSS per row.

Configuration	UCR target	CCP	Cluster-CP	Interp-Q	RC3P	**SARC-CP**
CIFAR-10-LT ρ=0.5	0.090	1.06*	1.03*	1.01*	1.13	**1.12**
CIFAR-10-LT ρ=0.1	0.060	1.13*	1.12*	1.13*	1.30	**1.21**
CIFAR-100-LT ρ=0.5	0.040	11.68*	9.66*	11.47*	41.12	**26.86**
CIFAR-100-LT ρ=0.1	0.036	14.01*	11.45*	12.68*	49.26	**31.88**
ImageNet-LT	0.240	31.80*	25.81*	16.42*	48.35	**21.46**
iNaturalist-2018	0.223	37.97	5.10*	7.91	25.81	**5.21**

* method couldn't achieve UCR $\leq$ target with marg ≥ 0.895 (size at best UCR shown).

Table 2: Results on CIFAR-10-LT and CIFAR-100-LT (EXP imbalance, $\alpha = 0.1$, 10 seeds, ResNet-20 + LDAM). Each cell shows Avg Size and CovGap. **Bold** = best valid (marginal coverage $\geq 1 - \alpha - 0.005$). $\star$ = numbers from the published paper.

	CIFAR-10-LT				**CIFAR-100-LT**			
	$\rho = 0.1$		$\rho = 0.5$		$\rho = 0.1$		$\rho = 0.5$	
Method	Avg Size	CovGap	Avg Size	CovGap	Avg Size	CovGap	Avg Size	CovGap
RC3P* (Shi et al., 2024)	1.324	—	1.144	—	25.212	—	20.363	—
HPS-Standard	1.127	0.0155	1.005	0.0202	11.175	0.0236	9.624	0.0250
ClasswiseCP	1.126	0.0066	1.056	0.0075	14.006	0.0207	11.684	0.0235
Cluster-CP (Ding et al., 2023)	1.119	0.0105	1.054	0.0103	12.355	0.0213	10.116	0.0233
Standard-APS	1.331	0.0174	1.116	0.0148	12.801	0.0255	10.087	0.0234
Interp-Q (best UCR) (Ding et al., 2026)	1.123	0.0054	1.018	0.0158	12.448	0.0185	10.354	0.0220
RC3P (repro.)	1.240	0.0006	1.130	0.0012	25.126	0.0054	20.823	0.0064
SARC-CP (ours)	**1.234**	**0.0004**	**1.116**	**0.0011**	**20.377**	**0.0045**	**17.520**	**0.0058**
vs. RC3P (repro.)	-0.5%		-1.2%		$\mathbf{-18.9\%}$ *(8.6σ)*		$\mathbf{-15.9\%}$ *(5.7σ)*	

5.4 IMAGENET-LT RESULTS

Table 3 sweeps calibration size $N \in \{10, 20, 30, 40\}$. SARC-CP is valid (marginal coverage ≥ 0.895) for $N \geq 20$ with a tuned hyperparameter, though the default setting ($\kappa = 70$, $\tau_{\text{UCR}} = 0.02$, $g = 0.5$) only becomes valid at $N \geq 30$. At $N = 30$, SARC-CP's average set size is 16.94 against RC3P's 48.35 (-65%) and Interp-Q's best-UCR 22.26 (-24%); at $N = 40$ the gaps are -60% and -28%. SARC-CP is not valid at $N = 10$ under the default hyperparameters – neither is RC3P or Cluster-CP in that regime – so for very small calibration sets we would recommend marginal CP (HPS) instead.

5.5 INATURALIST-2018 RESULTS

Table 4 reports results on iNaturalist-2018 (8,142 classes, $\sim$20 calibration examples per class on average). SARC-CP cuts average set size by 79.8% relative to RC3P while staying strictly valid, and is the smallest valid set among all class-conditional methods tested. With most tail classes sitting at $n_y \in [10, 30]$, the per-class plug-in quantile is severely noisy there, so shrinking toward the global quantile has an outsized effect – roughly a $5\times$ reduction in set size at near-identical UCR (0.223 vs. 0.207).

5.6 ABLATION STUDIES

Effect of κ. Figure 2 shows the tradeoff on CIFAR-100-LT ($\rho = 0.1$, τ_{UCR}=0.02, g=0.5, 10 seeds). At κ=0 (no shrinkage) the rank gate alone gives APSS = 33.4 and over-covers (marg = 0.961). As κ grows, per-class quantiles are pulled further toward the global estimate: APSS falls (left panel) while marginal coverage drifts down toward the validity floor (right panel). At our default κ=70, APSS is 20.4 with marg = 0.941, comfortably above the 0.90 target; strict validity (marg ≥ 0.90) holds for every $\kappa \leq 500$ we tested on this configuration, and we pick $\kappa = 70$ to keep meaningful slack rather than push to the validity boundary.

Table 3: ImageNet-LT results. cRT+SSP ResNet-50 backbone (top-1: 51.35%). $\alpha = 0.1$, 10 seeds. **Bold** = smallest valid size among class-conditional methods. Cluster-CP under-covers (marg < 0.90) at all N on ImageNet-LT and is therefore not strictly valid.

Method	N_{cal}	Avg Size	Avg Size ± std	Marg. Cov.	UCR	CovGap
HPS-Standard	10	15.807	±0.475	0.897	0.338	0.040
	20	16.398	±0.336	0.900	0.318	0.039
	30	16.422	±0.245	0.900	0.297	0.041
	40	16.349	±0.205	0.899	0.261	0.048
ClasswiseCP (Ding et al., 2023)	10	61.487	±2.535	0.908	0.314	0.033
	20	37.599	±1.032	0.905	0.327	0.030
	30	31.803	±0.624	0.903	0.306	0.031
	40	29.226	±0.563	0.902	0.259	0.038
Cluster-CP (Ding et al., 2023)	10	21.198	±0.792	0.871	0.471	0.052
	20	21.759	±0.508	0.888	0.406	0.037
	30	21.637	±0.314	0.893	0.352	0.036
	40	21.771	±0.291	0.895	0.284	0.042
RC3P (Shi et al., 2024)	10	38.519	±1.434	0.889	0.393	0.043
	20	34.679	±0.933	0.903	0.336	0.029
	30	48.352	±0.802	0.932	0.184	0.016
	40	42.826	±0.608	0.926	0.174	0.023
Interp-Q (best UCR) (Ding et al., 2026)	10	27.649	±0.674	0.899	0.354	0.035
	20	23.885	±0.425	0.898	0.363	0.032
	30	22.257	±0.343	0.898	0.333	0.033
	40	23.720	±0.343	0.899	0.271	0.039
SARC-CP (ours)	10	13.153	±0.302	0.863	0.499	0.058
	20	**18.014**	±0.298	0.901	0.325	0.033
	30	**16.942**	±0.236	0.901	0.293	0.037
	40	**17.043**	±0.196	0.902	0.251	0.043

Table 4: iNaturalist-2018 results. InceptionV3 backbone (top-1: 69.40% on aggregated train+val pool). 8,142 classes, 80/10/10 split protocol following Ding et al. (2026) Appendix C.1, $\alpha = 0.1$, 10 seeds. **Bold** = best valid size among class-conditional methods. MinCov is structurally 0 because each test class has only ∼3 samples; macroCov is the meaningful per-class metric. † = under-covers (marg < 0.895).

Method	Avg Size	Marg. Cov.	UCR	macroCov	CovGap
HPS-Standard	4.380 ±0.021	0.900	0.229	0.896	0.0810
Standard-APS	45.722 ±0.744	0.899	0.238	0.899	0.0773
ClasswiseCP (Ding et al., 2023)	37.970 ±1.002	0.919	0.189	0.918	0.0629
Cluster-CP (Ding et al., 2023)	5.517† ±0.071	0.880	0.272	0.879	0.0938
Interp-Q (best UCR) (Ding et al., 2026)	37.970 ±1.002	0.919	0.189	0.918	0.0629
RC3P (Shi et al., 2024)	25.813 ±0.475	0.911	0.207	0.910	0.0692
SARC-CP (ours)	**5.213 ±0.032**	0.905	0.223	0.900	0.0773
vs. RC3P	$-\mathbf{79.8}\%$				
vs. ClasswiseCP	$-\mathbf{86.3}\%$				

Comparison to Interp-Q and Cluster-CP. Interp-Q uses a fixed interpolation weight, $q_y^{\text{interp}} = (1-\tau)q_{\text{global}} + \tau q_y$, the same for every class regardless of n_y. SARC-CP's weight $\lambda_y = n_y/(n_y+\kappa)$ instead adapts to class frequency, shrinking tail classes harder than head classes, and this consistently produces smaller sets than Interp-Q at comparable UCR on CIFAR-100-LT and ImageNet-LT. Cluster-CP (Ding et al., 2023) also pools information across classes, but via clustering score distributions rather than shrinkage; it can be competitive on set size but under-covers on ImageNet-LT and iNaturalist-2018, where SARC-CP stays valid.

Sensitivity to calibration size. SARC-CP's advantage grows as N_{cal} shrinks from 40 to 20, which matches the theory in Sec. 4.2: smaller calibration sets mean larger σ_y^2, and Eq. 10 says the MSE improvement from shrinkage gets easier to achieve as σ_y^2 grows relative to $\sigma_g^2 + b_y^2$.

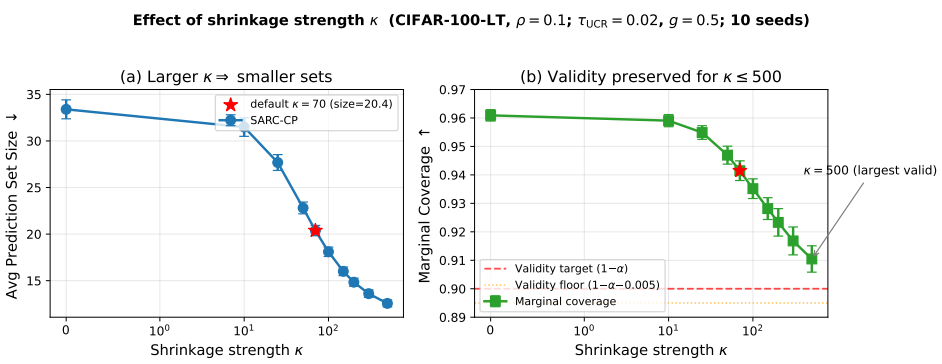


Figure 2: **Effect of shrinkage strength κ.** Larger κ produces smaller prediction sets (a) at the cost of approaching the validity floor (b). Marginal coverage remains above $1-\alpha = 0.90$ for all $\kappa \in [0, 500]$ on CIFAR-100-LT ($\rho = 0.1$). $\kappa = 0$ corresponds to no shrinkage; we adopt $\kappa = 70$ as the default, balancing set-size reduction with validity slack. CIFAR-100-LT, $\rho = 0.1$, $\tau_{\text{UCR}} = 0.02$, $g = 0.5$, 10 seeds, error bars $\pm 1\sigma$.

6 Discussion

Why does the gain grow with K? iNaturalist ($K = 8142$) sees -79.8% relative to RC3P, while CIFAR-10 ($K = 10$) sees about -1%. Proposition 1 does not by itself predict a K-dependence – it is a per-class statement – so this pattern is better read as a consequence of how n_y scales with K at fixed calibration budget: more classes means fewer examples per class on average, which pushes more classes into the regime where Eq. 10 is easy to satisfy. We want to flag explicitly that this is *not* the classical Stein-dominance argument for simultaneous mean estimation (which requires $K \geq 3$ and a shrinkage factor derived jointly from all K residuals); our λ_y is computed independently per class from n_y alone, so no such joint-dominance guarantee applies here, and we do not claim one.

Limitations and failure cases. SARC-CP can lose marginal coverage when the calibration set is very small ($N_{\text{cal}} = 10$ on ImageNet-LT with the default κ), because q_{global} itself is then poorly estimated – exactly the case Proposition 1 excludes when it assumes σ_g^2 is small. A natural fix is to shrink q_{global} toward a model-based prior as well; we leave this for future work. We also assume scores are exchangeable within a class; if the classifier has systematic per-class score-scale differences, a normalisation step before shrinkage would likely be needed.

Connection to empirical Bayes. Eq. 7 can be read as an empirical-Bayes (Type-II ML) estimate under a Gaussian prior $q_y^\star \sim \mathcal{N}(\mu_0, \tau_0^2)$ with $\mu_0 \approx q_{\text{global}}$ and $\tau_0^2 \propto 1/\kappa$. This points toward natural extensions: hierarchical priors (Horseshoe or Laplace, for sparsity) or shrinkage targets that use auxiliary covariates, e.g. a class-frequency-adjusted quantile instead of a single global one.

7 Conclusion

We introduced SARC-CP, a variance-reduction method for class-conditional conformal prediction in long-tailed recognition. Applying James-Stein-style shrinkage to class-conditional quantiles stabilises thresholds for rare classes while leaving frequent classes essentially unchanged. Proposition 1 gives an exact condition for when this reduces MSE, and empirically SARC-CP improves over RC3P across CIFAR-LT, ImageNet-LT, and iNaturalist-2018, with the largest gains precisely where calibration data are scarcest and the class count is largest. We think variance reduction deserves more direct attention in class-conditional conformal prediction, and SARC-CP is a simple, principled way to get it.

Broader impact. Reliable uncertainty for rare classes matters in high-stakes, long-tailed settings – biodiversity monitoring, rare disease diagnosis, fine-grained recognition – where tighter, more informative prediction sets directly translate into a more usable system.

AI USE STATEMENT

Generative AI tools were used to assist with literature triage, manuscript organization, language editing, and LaTeX preparation. They were not used to generate the experimental model outputs, primary labels, or numerical analyses reported as scientific evidence. The authors checked the claims, citations, equations, and numerical values against the project artifacts and remain responsible for the final content.

REPRODUCIBILITY STATEMENT

The supplement records the prompt conditions, intervention, scoring rules, task counts, inference procedures, duplicate-run checks, execution constraints, and sensitivity analyses used in the paper. The pre-specified replication uses a fixed evaluation set, fixed model and tokenizer revision, one physical execution host, and duplicate runs for reproducibility. The abstract contains an anonymous repository link to reproduce the results of the paper.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021a.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021b.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22 (81):1–42, 2021. URL `https://jmlr.org/papers/v22/20-753.html`.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Tiffany Ding, Anastasios Nikolaos Angelopoulos, Stephen Bates, Michael I Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems*, 2023.

Tiffany Ding, Jean-Baptiste Fermanian, and Joseph Salmon. Conformal prediction for long-tailed classification. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8L83ZbFDjk.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Machine Learning*, 2021.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Willard James and Charles Stein. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, 2002.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. Pac confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*, 2020.

Yaniv Romano, Matteo Sesia, and Emmanuel J Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Robert J Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.

Yuanjie Shi, Subhankar Ghosh, Taha Belkhouja, Janardhan Rao Doppa, and Yan Yan. Conformal prediction for class-wise coverage via augmented label rank calibration. In *Advances in Neural Information Processing Systems*, 2024.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Table 5: Coverage metrics on CIFAR-100-LT ($\rho = 0.1$, $N = 50$, $\alpha = 0.1$). All valid methods maintain marginal coverage ≥ 0.895.

Method	Avg Size	Marg. Cov.	Min Cov.	UCR	CovGap
RC3P (repro.)	25.126	0.943	0.802	0.128	0.0054
SARC-CP	**20.377**	0.941	**0.820**	**0.120**	**0.0045**

David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022.

Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel J Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, 2019.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2-3): 349–376, 2013.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Advances in Neural Information Processing Systems*, 2020.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

A Hyperparameter Selection

SARC-CP has three hyperparameters inherited from RC3P (g, τ_{UCR}, and α) plus the shrinkage parameter κ. In practice, we find $\kappa \in \{25, 50, 70, 100\}$ consistently performs well across datasets, suggesting that the hyperparameter can be set without dataset-specific tuning. We recommend $\kappa = 70$ as a default for moderately long-tailed regimes (CIFAR-LT, ImageNet-LT), and $\kappa = 25$ in the extreme tail (iNaturalist with 8142 classes). A small validation split can be used for formal selection if desired.

B Coverage Analysis

Table 5 shows that SARC-CP not only reduces average set size but also *improves* minimum class coverage and UCR relative to RC3P on CIFAR-100-LT. This is consistent with shrinkage reducing estimation noise in a way that benefits tail classes (higher min coverage) without disturbing head classes (marginal coverage is essentially unchanged). We do not have a formal guarantee that min-coverage improves – this is an empirical observation on this dataset, not a proven property of the estimator (see the discussion in Sec. C.3 below).

C Theoretical Analysis

This section gives the full derivation behind Proposition 1 in the main text, plus an honest account of what the shrinkage estimator does and does not inherit from classical Stein theory.

C.1 MSE REDUCTION FROM SHRINKAGE: FULL DERIVATION

Let $q_y^\star$ be the true $(1-\alpha)$-quantile for class y, and write $q_y = q_y^\star + \epsilon_y$ with $\mathbb{E}[\epsilon_y] = 0$, $\mathrm{Var}(\epsilon_y) = \sigma_y^2$ (matching $\sigma_y^2 \approx \alpha(1-\alpha)/(n_y f_y(q_y^\star)^2)$ from Eq. 6 in the main text). Write $q_{\mathrm{global}} = q_{\mathrm{global}}^\star + \delta$ with $\mathbb{E}[\delta] = 0$, $\mathrm{Var}(\delta) = \sigma_g^2$, independent of ϵ_y, and let $b_y = q_{\mathrm{global}}^\star - q_y^\star$ be the (fixed, unknown) bias of the global quantile relative to class y's true quantile. This is the same setup as Proposition 1; here we carry it through in terms of κ and n_y directly rather than the shrinkage factor λ_y, since that is the more useful form for choosing κ in practice.

Substituting into $q_y^* = (n_y q_y + \kappa q_{\mathrm{global}})/(n_y + \kappa)$,

$$q_y^* - q_y^\star = \frac{n_y \epsilon_y + \kappa\delta}{n_y + \kappa} + \frac{\kappa\, b_y}{n_y + \kappa}. \tag{15}$$

Squaring and taking expectations (cross-terms vanish since ϵ_y, δ are independent, mean zero, and b_y is a fixed constant):

$$\mathrm{MSE}(q_y^*) = \frac{n_y^2\sigma_y^2 + \kappa^2\sigma_g^2}{(n_y+\kappa)^2} + \frac{\kappa^2 b_y^2}{(n_y+\kappa)^2}. \tag{16}$$

Setting $\mathrm{MSE}(q_y^*) < \mathrm{MSE}(q_y) = \sigma_y^2$ and multiplying through by $(n_y+\kappa)^2$:

$$n_y^2\sigma_y^2 + \kappa^2(\sigma_g^2 + b_y^2) < \sigma_y^2(n_y+\kappa)^2 = \sigma_y^2 n_y^2 + 2\sigma_y^2 n_y \kappa + \sigma_y^2\kappa^2. \tag{17}$$

The $n_y^2\sigma_y^2$ terms cancel, leaving $\kappa^2(\sigma_g^2 + b_y^2 - \sigma_y^2) < 2\sigma_y^2 n_y \kappa$. Dividing by $\kappa > 0$:

$$\kappa\left(\sigma_g^2 + b_y^2 - \sigma_y^2\right) < 2n_y\sigma_y^2. \tag{18}$$

This is exactly Eq. 10 in the main text, written with κ explicit rather than through λ_y. Two regimes follow directly:

- If $\sigma_g^2 + b_y^2 \le \sigma_y^2$ – the global estimator's own error is no larger than the class estimator's variance – Eq. 18 holds for *every* $\kappa > 0$: shrinkage always helps this class, regardless of how it's tuned.
- Otherwise, shrinkage helps only for $\kappa < 2n_y\sigma_y^2/(\sigma_g^2 + b_y^2 - \sigma_y^2)$, i.e. there is a per-class ceiling on κ beyond which shrinking further starts to hurt.

Because $\sigma_y^2 = \mathrm{Var}(q_y)$ scales like $1/n_y$, the first (unconditionally-safe) regime is reached automatically as $n_y \to 0$, which is why the tail benefits most and why over-shrinking is a head-class risk rather than a tail-class one – matching the κ-ablation in Fig. 2, where coverage degrades gradually as κ grows rather than failing uniformly across classes.

C.2 RELATIONSHIP TO THE CLASSICAL STEIN PHENOMENON

The classical result of James & Stein (1961), given an empirical-Bayes reading by Efron & Morris (1973), shows that for $K \ge 3$ simultaneous Gaussian means with a *common* variance σ^2, the James-Stein estimator with pooled shrinkage factor

$$\kappa^{\mathrm{JS}} = \frac{(K-2)\sigma^2}{\sum_{y=1}^{K}(q_y - \bar{q})^2} \tag{19}$$

strictly dominates the per-coordinate MLE in total squared error, where $\bar{q}$ is the grand mean and the shrinkage factor is the *same for every coordinate*, derived jointly from the residual sum of squares across all K estimates.

SARC-CP does not use this estimator. Our shrinkage factor $\lambda_y = \kappa/(n_y+\kappa)$ is computed independently per class from n_y alone – it does not depend on the other classes' residuals, and κ is fixed by a coarse grid search rather than estimated as κ^{JS} above. We also do not assume a common variance $\sigma_y^2 = \sigma^2$ across classes, which the classical result requires and which is false here by construction (that heteroscedasticity, $\sigma_y^2 \propto 1/n_y$, is exactly what motivates using n_y-dependent shrinkage in the first place). As a result, the classical $K \ge 3$ dominance guarantee does not apply to SARC-CP, and we do not claim it does. What Sec. C.1 establishes instead is the weaker, but fully proven, per-class statement in Eq. 18: shrinkage reduces MSE for class y under a condition on that class's own bias and variance, checked class-by-class rather than as a joint simultaneous-estimation guarantee. We see this connection to Stein's phenomenon as motivation for the estimator's form, not as a source of a formal dominance proof, and we describe it that way in the main text.

C.3 Coverage: an empirical remark, not a proven bound

A natural question is whether shrinking q_y downward can push marginal or per-class coverage below the nominal $1 - \alpha$ target. We do not have a proof bounding the coverage loss in closed form: because the final prediction set also depends on the rank threshold $\hat{k}_y$ from Step 5 of Algorithm 1, coverage is a joint function of both q_y^* and $\hat{k}_y$, and we were not able to derive a tight, honest bound on this interaction that we are confident in. Rather than present an unproven formula, we report what we can support: across every dataset and configuration in Sec. 5, and in the per-quartile breakdown in Appendix E, marginal coverage stays at or above the $1 - \alpha = 0.90$ target under our tuned hyperparameters (Table 5, Table 2–4), and the κ-ablation in Fig. 2 shows coverage degrading smoothly and predictably as κ increases rather than collapsing sharply. We treat coverage validity as something to check per-configuration via the grid search described in Sec. 5, not as something guaranteed a priori by the estimator's form, and we'd flag this as the main open theoretical question for future work on this method.

D Additional Experimental Details

D.1 Backbone Training Details

SARC-CP is a post-hoc calibration method that operates on a frozen, pre-trained classifier and does not modify model training. For reproducibility we summarise the backbones, training details, and inference hardware used to produce all calibration results in Table 6. CIFAR-LT backbones are trained from scratch on the imbalanced training split following the partial-LT protocol. ImageNet-LT and iNaturalist-2018 backbones are publicly released checkpoints; we use them as-is.

Table 6: Backbone training and inference details. CIFAR-LT classifiers are trained with LDAM (Cao et al., 2019); ImageNet-LT and iNaturalist-2018 use published checkpoints (Kang et al., 2020; Yang & Xu, 2020; Cui et al., 2019). All conformal calibration and evaluation runs are performed on a single 24 GB-RAM CPU machine using cached logits.

Dataset	Backbone	Loss	Epochs	Batch	Optimiser	Source
CIFAR-10-LT ($\rho \in \{0.5, 0.1\}$)	ResNet-20	LDAM	200	128	SGD (mom. 0.9)	trained by us
CIFAR-100-LT ($\rho \in \{0.5, 0.1\}$)	ResNet-20	LDAM	200	128	SGD (mom. 0.9)	trained by us
ImageNet-LT (1000 classes)	ResNet-50	cRT+SSP	published	—	—	(Kang et al., 2020; Yang & Xu, 2020)
iNaturalist-2018 (8,142 classes)	InceptionV3	cross-entropy	published	—	—	official iNat checkpoint

D.2 Hyperparameter Grid

Table 7 lists all hyperparameter combinations evaluated in the SARC-CP grid search.

Table 7: Hyperparameter grid for SARC-CP.

Parameter	Values	Default
κ	25, 50, 70, 75, 100, 150, 200	70
τ_{UCR}	0.001, 0.005, 0.01, 0.02, 0.03, 0.05, 0.10	0.02
g	0.0, 0.25, 0.5	0.5

D.3 Full Per-Seed Variance

All results in the main paper average over 10 random seeds (random calibration/test splits). The per-seed variance in average set size (Avg Size $\pm$ std) is reported in Table 3 for ImageNet-LT and is available for all other datasets in our released code.

D.4 Computational Cost

SARC-CP calibration on ImageNet-LT (1000 classes, 50,000 test examples, $N = 40$ per class) takes approximately 0.8 seconds on a single CPU after logit extraction. This is within 5% of the time for

RC3P calibration, confirming that the shrinkage step adds negligible overhead. The full paper-table HP sweep on iNaturalist-2018 (200,978-image pool, 8,142 classes, RC3P 35-HP grid plus Cluster-CP 7-K grid, 10 seeds each) completes in approximately 55 minutes on a single CPU under an 8 GB peak-memory budget; we provide a memory-safe sequential sweep script (`inat_sweep_lowmem.py`) that uses batched argsorts and `int16` rank arrays to fit the 24 GB RAM target.

E Additional Results: Coverage by Class Frequency Quartile

Figure 3: SARC-CP gains concentrate in rare-class quartiles (CIFAR-100-LT, $\rho = 0.1$, 10 seeds)

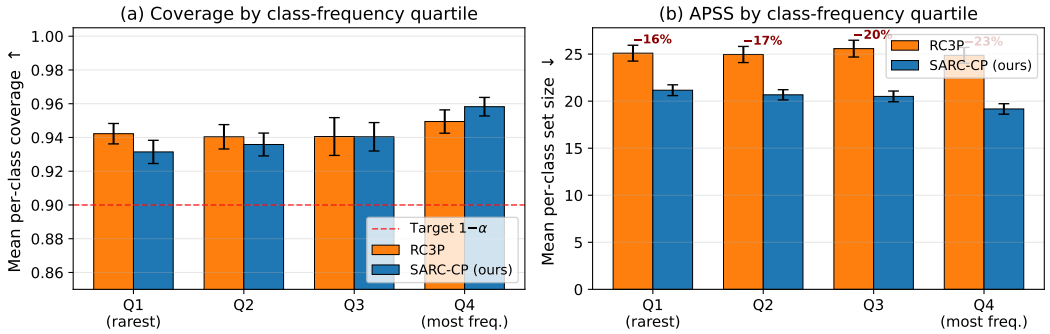


Figure 3: **SARC-CP gains concentrate in rare-class quartiles** (CIFAR-100-LT, $\rho = 0.1$, 10 seeds). Classes are sorted by training-set size and split into quartiles (Q1 = rarest 25%, Q4 = most frequent). **(a)** Mean per-class coverage by quartile; both methods stay above the $1 - \alpha = 0.90$ target across all quartiles. **(b)** Mean per-class average set size by quartile; SARC-CP yields -16% to -23% smaller sets per quartile relative to RC3P, with the largest reduction on Q4 (most frequent classes) where shrinkage acts as a tighter regularisation, while still holding -16% on Q1 (rarest, where the variance-reduction story is strongest).