A MATHEMATICAL NOTATIONS

Notation	**Meaning**
$X \in \mathcal{X}$	Input example
$Y \in \mathcal{Y}$	The ground-truth label
f	The soft classifier
C	Number of classes
$\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$	Joint space of input-output pairs
$\mathcal{P}_{X,Y}$	Data distribution of $\mathcal{Z}$
$\mathcal{D}_{\text{tr}}$	Training data
$\mathcal{D}_{\text{cal}}$	Calibration data
$\mathcal{D}_{\text{test}}$	Test data
ϵ_c	The top-k error for class c
$r_f(x, y)$	The rank of y in prediction $f(x)$
$\widehat{\mathcal{C}}(X)$	Prediction set for input X
$V((X,Y) \in \mathcal{Z})$	Non-conformity scoring function
α	Target mis-coverage rate

Table 2: **Key notations used in this paper.**

B TECHNICAL PROOFS OF THEORETICAL RESULTS

B.1 PROOF OF PROPOSITION 1

Proposition 2. *(Proposition 1 restated, class-conditional over- and under-coverage of MCP) Given* α*, assume* $|Rob(\alpha)| < |\mathcal{Y}|$*. If there exist* $\xi, \xi' > 0$ *such that for* $y \in Rob(\alpha), y' \notin Rob(\alpha)$*:*

$$\mathbb{P}\{V(X,Y) \leq Q^{class}_{1-\alpha}(Y)|Y = y \in Rob(\alpha)\} - \mathbb{P}\{V(X,Y) \leq Q^{MCP}_{1-\alpha}|Y = y \in Rob(\alpha)\} \leq -\xi,$$
$$\mathbb{P}\{V(X,Y) \leq Q^{class}_{1-\alpha}(Y)|Y = y' \notin Rob(\alpha)\} - \mathbb{P}\{V(X,Y) \leq Q^{MCP}_{1-\alpha}|Y = y' \notin Rob(\alpha)\}$$
$$\geq \frac{1}{n_{y'}} + \xi'.$$

Then class y *and* y' *are over- and under-covered, respectively:*

$$\mathbb{P}\{V(X,Y) \leq Q^{MCP}_{1-\alpha}|Y = y\} \geq 1 - \alpha + \xi, \quad \mathbb{P}\{V(X,Y) \leq Q^{MCP}_{1-\alpha}|Y = y'\} \leq 1 - \alpha - \xi'.$$

Proof. (of Proposition 1)

(1) Class y is over-covered. We start with the lower bound of the class-conditional coverage using class-conditional quantile for class $y \in \text{Rob}(\alpha)$, i.e., Theorem 1 in Romano *et al.* (2020), as follows.

$$\begin{aligned}
1 - \alpha \leq\, & \mathbb{P}\{V(X,Y) \leq Q^{\text{class}}_{1-\alpha}(y)|Y = y \in \text{Rob}(\alpha)\} \\
=& \mathbb{P}\{V(X,Y) \leq Q^{\text{MCP}}_{1-\alpha}|Y = y \in \text{Rob}(\alpha)\} \\
& + \mathbb{1}[y \in \text{Rob}(\alpha)] \cdot \Bigg(\mathbb{P}\Big\{V(X,Y) \leq Q^{\text{class}}_{1-\alpha}(y)]\Big| Y = y \in \text{Rob}(\alpha)\Big\} \\
& \qquad\qquad - \mathbb{P}\Big\{V(X,Y) \leq Q^{\text{MCP}}_{1-\alpha}\Big| Y = c \in \text{Rob}(\alpha)\Big\}\Bigg) \\
& + \mathbb{1}[y \notin \text{Rob}(\alpha)] \cdot \Bigg(\mathbb{P}\Big\{V(X,Y) \leq Q^{\text{class}}_{1-\alpha}(y)\Big| Y = c \notin \text{Rob}(\alpha)\Big\} \\
& \qquad\qquad - \mathbb{P}\Big\{V(X,Y) \leq Q^{\text{MCP}}_{1-\alpha}\Big| Y = c \notin \text{Rob}(\alpha)\Big\}\Bigg).
\end{aligned} \tag{13}$$

By assumption, for $y \in \text{Rob}(\alpha)$ and $\xi > 0$ such that the following inequality holds:

$$\mathbb{P}\{V(X,Y) \leq Q^{\text{class}}_{1-\alpha}(y)|Y = y \in \text{Rob}(\alpha)\} - \mathbb{P}\{V(X,Y) \leq Q^{\text{MCP}}_{1-\alpha}|Y = y \in \text{Rob}(\alpha)\} \leq -\xi, \tag{14}$$

by plugging inequality (14) into inequality (13), we derive the class-conditional over-coverage of CP on class y:

$$\mathbb{P}\{V(X,Y) \leq Q^{\text{MCP}}_{1-\alpha}|Y = y \in Rob(\alpha)\} \geq 1 - \alpha + \xi.$$

(2) Class c' is under-covered. We start with the upper bound of the class-conditional coverage using class-conditional quantile for class $y' \notin \text{Rob}(\alpha)$, i.e., Theorem 1 in Romano *et al.* (2020), as follows.

$$\begin{aligned}
&1 - \alpha + \frac{1}{n_{y'}} \geq \mathbb{P}\{V(X,Y) \leq Q^{\text{class}}_{1-\alpha}(y')|Y = y' \notin \text{Rob}(\alpha)\} \\
=&\mathbb{P}\{V(X,Y) \leq Q^{\text{MCP}}_{1-\alpha}|Y = y' \notin \text{Rob}(\alpha)\} \\
&+ \mathbb{1}[y' \in \text{Rob}(\alpha)] \cdot \Bigg(\mathbb{P}\{V(X,Y) \leq Q^{\text{class}}_{1-\alpha}(y')|Y = y' \notin \text{Rob}(\alpha)\} \\
&\qquad\qquad - \mathbb{P}\{V(X,Y) \leq Q^{\text{MCP}}_{1-\alpha}|Y = y' \notin \text{Rob}(\alpha)\}\Bigg) \\
&+ \mathbb{1}[y' \notin \text{Rob}(\alpha)] \cdot \Bigg(\mathbb{P}\{V(X,Y) \leq Q^{\text{class}}_{1-\alpha}(y')|Y = y' \notin \text{Rob}(\alpha)\} \\
&\qquad\qquad - \mathbb{P}\{V(X,Y) \leq Q^{\text{MCP}}_{1-\alpha}|Y = y' \notin \text{Rob}(\alpha)\}\Bigg).
\end{aligned} \tag{15}$$

By assumption, for $y' \notin \text{Rob}(\alpha)$ and $\xi' > 0$ such that the following inequality holds:

$$\begin{aligned}
\mathbb{P}\{V(X,Y) \leq Q^{\text{class}}_{1-\alpha}(y')|Y = y' \notin \text{Rob}(\alpha)\} - \mathbb{P}\{V(X,Y) \leq Q^{\text{MCP}}_{1-\alpha}|Y = y' \notin \text{Rob}(\alpha)\} \\
\geq \frac{1}{n_{y'}} + \xi',
\end{aligned} \tag{16}$$

by plugging inequality (16) into inequality (15), we derive the class-conditional under-coverage of CP on class y':

$$\mathbb{P}\{V(X,Y) \leq Q^{\text{MCP}}_{1-\alpha}|Y = c' \notin \text{Rob}(\alpha)\} \leq 1 - \alpha - \xi'.$$

The above two results of over-coverage of class y and under-coverage of class y' show that the class-conditional coverage can easily deviate from the marginal coverage as long as there exists a margin (i.e., ξ, ξ') for class-conditional coverage between using marginal quantile (i.e., $Q^{\text{MCP}}_{1-\alpha}$) and class-conditional quantile (i.e., $Q^{\text{class}}_{1-\alpha}(y), Q^{\text{class}}_{1-\alpha}(y')$), as in (14) and (16). □

B.2 PROOF OF THEOREM 1

Theorem 3. *(Theorem 1 restated, k-CCP guarantees class-conditional coverage) Suppose that selecting $\{k(y)\}_{y\in\mathcal{Y}}$ results in class-wise top-$k(y)$ error $\{\epsilon_y\}_{y\in\mathcal{Y}}$. If the nominated mis-coverage probability $\tilde{\alpha}_y$ of CCP for class y is set as*

$$\tilde{\alpha}_y \leq \alpha - \varepsilon_{n_y} - \delta - \epsilon_y, \quad \textit{for } 0 < \delta < 1, \;\; \varepsilon_{n_y} = \sqrt{(3(1-\alpha)\log(2/\delta))/n_y},$$

then k-CCP can achieve the class-conditional coverage as defined in equation (1).

Before proving Theorem 1, we introduce the following technical lemma.

Lemma 1. *(Concentration inequalities for quantiles) Define $\varepsilon_n = \sqrt{3(1-\alpha)\log(2/\delta)/n}$. Let $Q_{1-\alpha} = max\{t : \mathbb{P}_V\{V \leq t\} \geq 1-\alpha\}$ be the true quantile of a random variable V given α, and $\widehat{Q}_{1-\alpha} = V_{(\lceil n(1-\alpha)\rceil)}$ be the empirical quantile estimated by n randomly sampled set $\{V_1, ..., V_n\}_{i=1}^n$. Then with probability at least $1-\delta$, we have $\widehat{Q}_{1-\alpha-\varepsilon_n-1/n} \leq Q_{1-\alpha} \leq \widehat{Q}_{1-\alpha+\varepsilon_n}$ where $\tilde{O}$ hides the logarithmic factor.*

Lemma 1 has been studied in a previous paper (see Proposition 2(a) in Vovk (2012)). To make the proof and conclusion of the Lemma 1 complete with the Theorem 1, we prove it again at the end of this subsection. Now we begin to prove Theorem 1.

Proof. (of Theorem 1)

Let $y \in \mathcal{Y}$ denote any class label. With the lower bound of the coverage on class y (Theorem 1 in Romano *et al.* (2020)), we have

$$\begin{aligned}
&1 - (\tilde{\alpha} + \varepsilon_{n_y}) \\
\leq&\mathbb{P}\{Y_{n+1} \in \mathcal{C}^{\text{CCP}}_{1-\tilde{\alpha}-\varepsilon_{n_y}}(X_{n+1}) | Y = y\} = \mathbb{P}\{V(X_{n+1}, Y_{n+1}) \leq Q^{\text{class}}_{1-\tilde{\alpha}-\varepsilon_{n_y}} | Y = y\} \\
=&(\mathbb{P}\{\text{Lemma 1 holds}\} + \mathbb{P}\{\text{Lemma 1 not holds}\}) \cdot \mathbb{P}\{V(X_{n+1}, Y_{n+1}) \leq Q^{\text{class}}_{1-\tilde{\alpha}-\varepsilon_{n_y}} | Y = y\} \\
\leq&1 \cdot \mathbb{P}\{V(X_{n+1}, Y_{n+1}) \leq \widehat{Q}^{\text{class}}_{1-\tilde{\alpha}} | Y = y\} + \delta \cdot 1 \\
=&\mathbb{P}\{V(X_{n+1}, Y_{n+1}) \leq \widehat{Q}^{\text{class}}_{1-\tilde{\alpha}}, r_f(X_{n+1}, Y_{n+1}) \leq \widehat{k}(y) | Y = y\} \\
&+ \mathbb{P}\{V(X_{n+1}, Y_{n+1}) \leq \widehat{Q}^{\text{class}}_{1-\tilde{\alpha}}, r_f(X_{n+1}, Y_{n+1}) > \widehat{k}(y) | Y = y\} + \delta \\
\leq&\mathbb{P}\{V(X_{n+1}, Y_{n+1}) \leq \widehat{Q}^{\text{class}}_{1-\tilde{\alpha}}, r_f(X_{n+1}, Y_{n+1}) \leq \widehat{k}(y) | Y = y\} \\
&+ \underbrace{\mathbb{P}\{r_f(X_{n+1}, Y_{n+1}) > \widehat{k}(y) | Y = y\}}_{\leq \epsilon_y} + \delta \\
\leq&\mathbb{P}\{Y_{n+1} \in \widehat{\mathcal{C}}^{k\text{-CCP}}_{1-\tilde{\alpha}}(y) | Y = y\} + \epsilon_y + \delta,
\end{aligned}$$

where the second inequality is due to Lemma 1, and the last inequality is due to the definition of ϵ_y, i.e., $\mathbb{P}\{r_f(X_{n+1}, Y_{n+1}) \leq \widehat{k}(y) | Y = y\} \geq 1 - \epsilon_y$.

Re-arranging the above inequality, we have

$$\mathbb{P}\{Y_{n+1} \in \widehat{\mathcal{C}}^{k\text{-CCP}}_{1-\tilde{\alpha}}(y) | Y = y\} \geq 1 - \tilde{\alpha} - \varepsilon_{n_y} - \delta - \epsilon_y \geq 1 - \alpha,$$

where the last inequality is due to $\tilde{\alpha}_y \leq \alpha - \varepsilon_{n_y} - \delta - \epsilon_y$. This implies that k-CCP guarantees the class-conditional coverage on any class y. ☐

After proving Theorem 1, now we show the deferred proof of lemma 1:

Proof. (of Lemma 1)

Define $Z_i = \mathbb{1}[V_i \leq Q_{1-\alpha}]$ where $1 \leq i \leq n$ and $\mathbb{1}[\cdot]$ is an indicator function. Then Z_i is a Bernoulli random variable with $\mathbb{P}\{Z_i = 1\} = 1 - \alpha$ and $\mathbb{P}\{Z_i = 0\} = \alpha$ from the definition of $Q(\alpha)$. Let $\widehat{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\mathbb{E}[\widehat{Z}] = 1 - \alpha$.

According to Chernoff bound, we know

$$\mathbb{P}\Bigg\{\Bigg|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[\widehat{Z}]\Bigg| \geq \varepsilon\mathbb{E}[\widehat{Z}]\Bigg\} \leq 2\exp\Bigg(-n\mathbb{E}[\widehat{Z}]\varepsilon^2/3\Bigg) = 2\exp\Bigg(-n(1-\alpha)\varepsilon^2/3\Bigg).$$

By setting $\delta = 2\exp(-n(1-\alpha)\varepsilon^2/3)$, i.e., $\varepsilon = \frac{\varepsilon_n}{1-\alpha} = \sqrt{(3\log(2/\delta))/((1-\alpha)n)}$, we have with probability at least $1 - \delta$:

$$\Bigg|\frac{1}{n}\sum_{i=1}^n \mathbb{1}[V_i \leq Q_{1-\alpha}] - (1-\alpha)\Bigg| \leq (1-\alpha)\varepsilon = \frac{\varepsilon_n}{1-\alpha}(1-\alpha) = \varepsilon_n = \sqrt{(3(1-\alpha)\log(2/\delta))/n}. \tag{17}$$

Recall the definition of the empirical quantile $\widehat{Q}_{1-\alpha}$ given α:

$$\widehat{Q}_{1-\alpha} = \max\Bigg\{t : \frac{1}{n}\sum_{i=1}^n \mathbb{1}[V_i \leq t] \geq 1 - \alpha\Bigg\}.$$

Then we know the following upper bound and lower bound for $1-\alpha$:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[V_i \leq \widehat{Q}_{1-\alpha-1/n}] \leq (1-\alpha) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[V_i \leq \widehat{Q}_{1-\alpha}]. \tag{18}$$

Re-arranging (17) and using the above upper/lower bounds, with probability at least $1-\delta$, we have

$$\begin{aligned}
&(1-\alpha)(1-\varepsilon) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[V_i \leq Q_{1-\alpha}] \leq (1-\alpha)(1+\varepsilon) \\
\Leftrightarrow\quad & 1-(\underbrace{1-(1-\alpha)(1-\varepsilon)}_{\triangleq\alpha'}) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[V_i \leq Q_{1-\alpha}] \leq 1-(\underbrace{1-(1-\alpha)(1+\varepsilon)}_{\triangleq\alpha''}) \\
\overset{(18)}{\Rightarrow}\quad & \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[V_i \leq \widehat{Q}_{1-\alpha'-1/n}] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[V_i \leq Q_{1-\alpha}] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[V_i \leq \widehat{Q}_{1-\alpha''}] \\
\Leftrightarrow\quad & \widehat{Q}_{1-\alpha'-1/n} \leq Q_{1-\alpha} \leq \widehat{Q}_{1-\alpha''}.
\end{aligned} \tag{19}$$

Finally, we simplify α' and α'' as follows

$$\begin{aligned}
\alpha' &= 1-(1-\alpha)(1-\varepsilon) = \alpha+\varepsilon(1-\alpha) = \alpha+\sqrt{3(1-\alpha)\log(2/\delta)/n} = \alpha+\varepsilon_n, \\
\alpha'' &= 1-(1-\alpha)(1+\varepsilon) = \alpha-\varepsilon(1-\alpha) = \alpha-\sqrt{3(1-\alpha)\log(2/\delta)/n} = \alpha-\varepsilon_n.
\end{aligned} \tag{20}$$

Therefore, plugging (20) into (19), we have

$$\widehat{Q}_{1-\alpha-\varepsilon_n-1/n} \leq Q_{1-\alpha} \leq \widehat{Q}_{1-\alpha+\varepsilon_n}.$$

□

B.3 PROOF OF THEOREM 2

Theorem 4. *(Theorem 2 restated, k-CCP produces smaller prediction sets than CCP) Suppose the following inequality holds for any $y \in \mathcal{Y}$:*

$$\sum_{y\in\mathcal{Y}} \sigma_y \cdot \mathbb{P}_{Z_{n+1}}\Big[V(X_{n+1},y) \leq \widehat{Q}^{\textit{class}}_{1-\alpha}(y)\Big] \leq \sum_{y\in\mathcal{Y}} \mathbb{P}_{Z_{n+1}}\Big[V(X_{n+1},y) \leq \widehat{Q}^{\textit{class}}_{1-\alpha}(y)\Big].$$

Then k-CCP produces smaller expected prediction sets than CCP, i.e.,

$$\mathbb{E}_{X_{n+1}}[|\widehat{\mathcal{C}}^{\textit{k-CCP}}_{1-\tilde{\alpha}}(X_{n+1})|] \leq \mathbb{E}_{X_{n+1}}[|\widehat{\mathcal{C}}^{\textit{CCP}}_{1-\alpha}(X_{n+1})|].$$

Proof. (of Theorem 2)

The proof idea is to reduce the the cardinality of the prediction set made by k-CCP to that made by CCP in expectation.

$$\begin{aligned}
&\mathbb{E}_{Z_{n+1}}[|\widehat{\mathcal{C}}^{k\text{-CCP}}_{1-\tilde{\alpha}}(X_{n+1})|] = \mathbb{E}_{Z_{n+1}}\left[\sum_{y\in\mathcal{Y}}\mathbb{1}\Big[V(X_{n+1},y) \leq \widehat{Q}^{\text{class}}_{1-\tilde{\alpha}}(y),\ r_f(X_{n+1},y) \leq \widehat{k}(y)\Big]\right] \\
=& \sum_{y\in\mathcal{Y}} \mathbb{E}_{Z_{n+1}}\Big[\mathbb{1}[V(X_{n+1},y) \leq \widehat{Q}^{\text{class}}_{1-\tilde{\alpha}}(y),\ r_f(X_{n+1},y) \leq \widehat{k}(y)]\Big] \\
=& \sum_{y\in\mathcal{Y}} \mathbb{P}_{Z_{n+1}}\Big[V(X_{n+1},y) \leq \widehat{Q}^{\text{class}}_{1-\tilde{\alpha}}(y),\ r_f(X_{n+1},y) \leq \widehat{k}(y)\Big] \\
\overset{(a)}{=}& \sum_{y\in\mathcal{Y}} \sigma_y \cdot \mathbb{P}_{Z_{n+1}}\Big[V(X_{n+1},y) \leq \widehat{Q}^{\text{class}}_{1-\alpha}(y)\Big] \overset{(b)}{\leq} \sum_{y\in\mathcal{Y}} \mathbb{E}_{Z_{n+1}}\Big[\mathbb{1}[V(X_{n+1},y) \leq \widehat{Q}^{\text{class}}_{1-\alpha}(y)]\Big] \\
=& \mathbb{E}_{Z_{n+1}}\left[\sum_{y\in\mathcal{Y}}\mathbb{1}[V(X_{n+1},y) \leq \widehat{Q}^{\text{class}}_{1-\alpha}(y)]\right] = \mathbb{E}_{Z_{n+1}}[|\widehat{\mathcal{C}}^{\text{CCP}}_{1-\alpha}(X_{n+1})|],
\end{aligned} \tag{21}$$

where the equality (a) is due to the definitions of σ_y, and inequality (b) is due to the assumption

$$\sum_{y\in\mathcal{Y}} \sigma_y \cdot \mathbb{P}_{Z_{n+1}}\Big[V(X_{n+1},y) \le \widehat{Q}^{\text{class}}_{1-\alpha}(y)\Big] \le \sum_{y\in\mathcal{Y}} \mathbb{P}_{Z_{n+1}}\Big[V(X_{n+1},y) \le \widehat{Q}^{\text{class}}_{1-\alpha}(y)\Big].$$

This implies that k-CCP requires smaller prediction sets to guarantee the class-conditional coverage compared to CCP. □

C COMPLETE EXPERIMENTAL RESULTS

C.1 TRAINING DETAILS

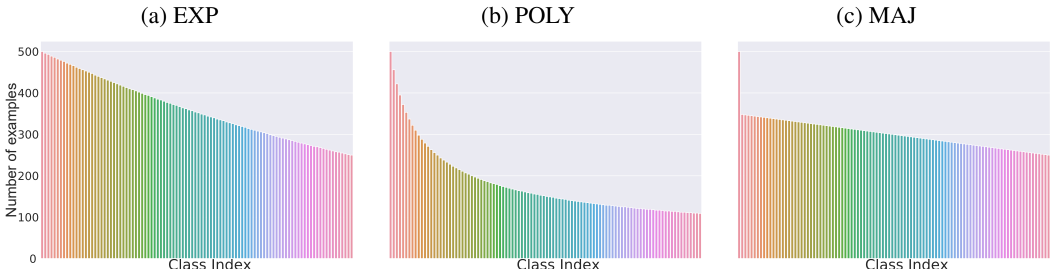
For CIFAR-10 and CIFAR-100, we train ResNet20 using LDAM loss function given in Cao *et al.* (2019) with standard mini-batch stochastic gradient descent (SGD) using learning rate 0.1, momentum 0.9, and weight decay $2e-4$ for 200 epochs. The batch size is 128. For experiments on mini-ImageNet, we use the same setting. For Food-101, the batch size is 256 and other parameters are kept the same.

C.2 CALIBRATION DETAILS

As mentioned in Section 5.1, we balanced split the validation set of CIFAR-10 and CIFAR-100, the number of calibration data is 5000. For mini-ImageNet, the number of calibration data is 15000. For Food-101, the total number is 12625. To compute the mean and standard deviation for the overall performance, we repeat calibration experiments for 10 times. Moreover, we select the g from the interval $[0.1, 1]$ with range 0.05 to find the minimal g that k-CCP and CCP achieves the target class-conditional coverage. We re-emphasize that the we have discussed the assumption in Theorem 2 and Remark 3.

The regularization parameter for RAPS scoring function is from the set $k_{reg} \in \{3, 5, 7\}$ and $\lambda \in \{0.001, 0.01, 0.1\}$ based on the empirical setting in `cluster-CP`. We select the combination of k_{reg} and λ for each experiments with same imbalanced type and imbalanced ratio on the same dataset, where the most of $APSS$ values of all methods are minimum. The hyper-parameter g is selected from the interval $[0.1, 1]$ with range 0.05 to find the minimal g that `CCP`, `cluster-CP`, and k-`CCP` achieve the target class-conditional coverage.

C.3 ILLUSTRATION OF IMBALANCED DATA


(a) EXP

(b) POLY

(c) MAJ

Figure 3: Illustrative examples of the different imbalanced distributions of the number of training examples per class index c on CIFAR-100

C.4 COMPLETE EXPERIMENT RESULTS

In this subsection, we report complete experimental results over four datasets, three decaying types, five imbalance ratios. Specifically, Table 3, 4, 5 report results on CIFAR-10 with three decaying types. Table 6, 7, 8 report results on CIFAR-100 with three decaying types. Table 9, 10, 11 report results on mini-ImageNet with three decaying types. Table 12, 13, 14 report results on Food-101 with three decaying types. Due to the limited time for rebuttal, we just report the results of `cluster-CP` with two imbalanced ratio $\rho = 0.1$ and $\rho = 0.5$. We will add all results in the final paper.

Figure 4 shows overall results on CIFAR-10 and CIFAR-100, including distribution of class-wise quantiles v.s. marginal quantile, histograms of class-conditional coverage and prediction set size achieved by `MCP`, `CCP`, `cluster-CP`, and k-`CCP`, and the histogram of condition numbers σ_y in Theorem 2, which are all corresponding to Figure 1 in main text.

Figure 5 shows the sensitivity of `CCP`, `cluster-CP`, and k-`CCP` for g on CIFAR-10 and CIFAR-100 with APS scoring function, which are all corresponding to Figure 2 in main text.

Figure 6, Figure 7, Figure 8, Figure 9, Figure 10 and Figure 11 show the class-conditional coverage and the corresponding prediction set sizes on EXP $\rho = 0.1$, EXP $\rho = 0.5$, POLY $\rho = 0.1$, POLY $\rho = 0.5$, MAJ $\rho = 0.1$, MAJ $\rho = 0.5$, respectively. This result on EXP $\rho = 0.1$ is in Figure 1 and Figure 4. Because of the same reason, we lack of visualization results of `cluster-CP` methods and we will add the complete visualization results in the final paper.

Figure 12, Figure 13, and Figure 14 show the distribution of class-wise quantiles with EXP $\rho = 0.5$, POLY $\rho = 0.5$, and MAJ $\rho = 0.5$, respectively.

Figure 15, Figure 16, Figure 17, Figure 18, Figure 19 and Figure 20 verify the condition numbers σ_y on EXP $\rho = 0.1$, EXP $\rho = 0.5$, POLY $\rho = 0.1$, POLY $\rho = 0.5$, MAJ $\rho = 0.1$, MAJ $\rho = 0.5$, respectively.

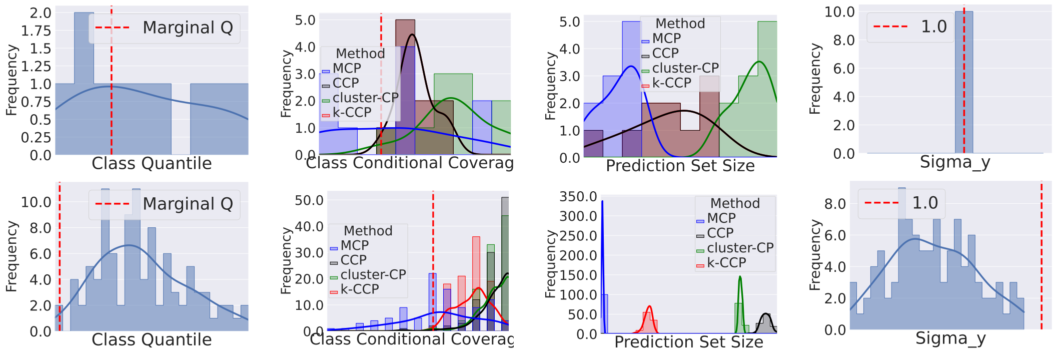
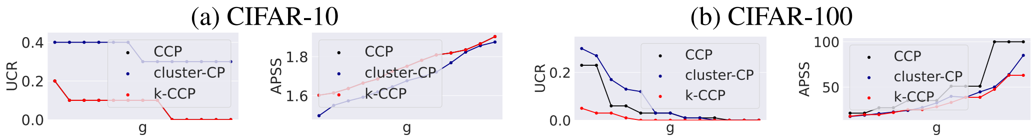


Figure 4: Justification experiments: CIFAR-10 in first row and CIFAR-100 in second row with ResNet20 model. First column: distribution of class-wise quantiles v.s. marginal quantile with imbalance type EXP and imbalance ratio $\rho = 0.5$. Second and third columns: histograms of class-conditional coverage and prediction set size achieved by `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with imbalance type EXP and imbalance ratio $\rho = 0.1$. The final column: the histogram of condition numbers σ_y in Theorem 2 with imbalance type EXP and imbalance ratio $\rho = 0.1$.



(a) CIFAR-10

(b) CIFAR-100

Figure 5: Results for under coverage ratio and average prediction set size achieved by `CCP`, `cluster-CP`, and $k-$`CCP` methods as a function of g using APS scoring function with imbalance type EXP for imbalance ratio $\rho = 0.1$. $k-$`CCP` degenerates to `CCP` in CIFAR-10, so overlopping with CCP (the black line overlaps with the red one).

Measure	Methods	EXP				
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.460 ± 0.025	0.390 ± 0.03	0.4 ± 0.032	0.4 ± 0.014	0.49 ± 0.026
	CCP	**0.110 ± 0.030**	**0.14 ± 0.029**	**0.140 ± 0.032**	**0.08 ± 0.031**	**0.040 ± 0.020**
	cluster-CP	0.160 ± 0.025	—	—	—	0.080 ± 0.012
	k-CCP	**0.110 ± 0.030**	**0.14 ± 0.029**	**0.140 ± 0.032**	**0.008 ± 0.031**	**0.040 ± 0.020**
APSS	MCP	1.132 ± 0.033	1.13 ± 0.03	1.169 ± 0.034	1.227 ± 0.039	1.406 ± 0.035
	CCP	1.481 ± 0.082	**1.508 ± 0.090**	**1.625 ± 0.083**	**1.711 ± 0.101**	**2.032 ± 0.096**
	cluster-CP	**1.445 ± 0.017**	—	—	—	2.323 ± 0.0115
	k-CCP	**1.481 ± 0.082**	**1.508 ± 0.090**	**1.625 ± 0.083**	**1.711 ± 0.101**	**2.032 ± 0.096**

Table 3: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type EXP and APS scoring function on dataset CIFAR-10. The dash symbol (—) means missing results that will be finished in the final version. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

Measure	Methods	POLY				
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.51 ± 0.033	0.54 ± 0.021	0.41 ± 0.03	0.47 ± 0.028	0.32 ± 0.031
	CCP	0.14 ± 0.032	**0.17 ± 0.028**	**0.12 ± 0.031**	**0.07 ± 0.025**	**0.01 ± 0.009**
	cluster-CP	**0.14 ± 0.021**	—	—	—	**0.01 ± 0.009**
	k-CCP	0.14 ± 0.032	**0.17 ± 0.028**	**0.12 ± 0.031**	**0.07 ± 0.025**	**0.01 ± 0.009**
APSS	MCP	1.17 ± 0.028	1.107 ± 0.028	1.138 ± 0.032	1.57 ± 0.033	1.214 ± 0.038
	CCP	**1.487 ± 0.09**	**1.465 ± 0.090**	**1.571 ± 0.086**	**1.652 ± 0.084**	**1.945 ± 0.087**
	cluster-CP	**1.612 ± 0.013**	—	—	—	**2.102 ± 0.015**
	k-CCP	**1.487 ± 0.09**	**1.465 ± 0.090**	**1.571 ± 0.086**	**1.652 ± 0.084**	**1.945 ± 0.087**

Table 4: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type POLY and APS scoring function on dataset CIFAR-10. The dash symbol (—) means missing results that will be finished in the final version. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

Measure	Methods	MAJ				
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.38 ± 0.019	0.33 ± 0.014	0.45 ± 0.029	0.51 ± 0.03	0.5 ± 0.014
	CCP	**0.12 ± 0.024**	**0.13 ± 0.025**	**0.11 ± 0.03**	**0.06 ± 0.021**	0.01 ± 0.009
	cluster-CP	0.15 ± 0.025	—	—	—	**0.009 ± 0.013**
	k-CCP	**0.12 ± 0.024**	**0.13 ± 0.025**	**0.11 ± 0.03**	**0.06 ± 0.021**	0.01 ± 0.009
APSS	MCP	1.17 ± 0.028	1.107 ± 0.028	1.138 ± 0.032	1.57 ± 0.033	1.406 ± 0.035
	CCP	**1.487 ± 0.09**	**1.465 ± 0.090**	**1.571 ± 0.086**	**1.652 ± 0.084**	**2.032 ± 0.096**
	cluster-CP	1.787 ± 0.019	—	—	—	2.969 ± 0.025
	k-CCP	**1.481 ± 0.082**	**1.508 ± 0.090**	**1.625 ± 0.083**	**1.711 ± 0.101**	**2.032 ± 0.096**

Table 5: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type MAJ and APS scoring function on dataset CIFAR-10. The dash symbol (—) means missing results that will be finished in the final version. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

Measure	Methods			EXP		
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.386±(0.009)	0.394±(0.010)	0.361±(0.012)	0.382±(0.011)	0.384±(0.018)
	CCP	0.009±(0.003)	0.015±(0.004)	0.011±(0.004)	0.016±(0.003)	0.011±(0.002)
	cluster-CP	0.004±(0.002)	—	—	—	0.004±(0.002)
	k-**CCP**	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**	**0.001±(0.001)**
APSS	MCP	10.303±(0.111)	10.848±(0.104)	12.480±(0.113)	12.909±(0.115)	14.544±(0.119)
	CCP	44.194±(0.514)	44.447±(0.566)	47.688±(0.569)	46.955±(0.500)	50.963±(0.482)
	cluster-CP	30.922±(0.454)	—	—	—	43.883±(1.070)
	k-**CCP**	**20.355±(0.357)**	**20.540±(0.356)**	**22.550±(0.306)**	**23.163±(0.265)**	**25.185±(0.279)**

Table 6: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type EXP and APS scoring function on dataset CIFAR-100. The dash symbol (—) means missing results that will be finished in the final version. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

Measure	Methods			POLY		
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.395±(0.010)	0.382±(0.014)	0.409±(0.013)	0.383±(0.015)	0.410±(0.010)
	CCP	0.011±(0.003)	0.008±(0.002)	0.016±(0.003)	0.011±(0.004)	0.015±(0.003)
	cluster-CP	0.001±(0.001)	0.008±(0.002)	0.016±(0.003)	0.011±(0.004)	0.015±(0.003)
	k-**CCP**	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**
APSS	MCP	15.730±(0.126)	16.738±(0.170)	17.670±(0.165)	20.422±(0.211)	25.888±(0.197)
	CCP	49.896±(0.490)	54.011±(0.572)	56.018±(0.529)	59.893±(0.438)	64.366±(0.390)
	cluster-CP	56.696±(0.393)	—	—	—	63.208±(0.364)
	k-**CCP**	**25.843±(0.300)**	**26.851±(0.222)**	**29.655±(0.286)**	**31.933±(0.218)**	**37.035±(0.245)**

Table 7: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type POLY and APS scoring function on dataset CIFAR-100 that will be finished in the final version. The dash symbol (—) means missing results. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

Measure	Methods			MAJ		
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.352±(0.010)	0.412±(0.014)	0.361±(0.009)	0.374±(0.010)	0.401±(0.008)
	CCP	0.017±(0.003)	0.008±(0.002)	0.018±(0.004)	0.011±(0.004)	0.008±(0.004)
	cluster-CP	0.005±(0.002)	—	—	—	0.019±(0.005)
	k-**CCP**	**0.002±(0.001)**	**0.001±(0.001)**	**0.000±(0.000)**	**0.000±(0.000)**	**0.001±(0.001)**
APSS	MCP	11.680±(0.117)	14.034±(0.128)	15.171±(0.121)	18.516±(0.152)	23.796±(0.159)
	CCP	48.323±(0.548)	49.193±(0.403)	53.688±(0.537)	55.024±(0.402)	64.640±(0.621)
	cluster-CP	33.623±(0.395)	—	—	—	50.382±(0.711)
	k-**CCP**	**21.196±(0.320)**	**24.058±(0.282)**	**25.501±(0.249)**	**29.344±(0.315)**	**35.630±(0.232)**

Table 8: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type MAJ and APS scoring function on dataset CIFAR-100 that will be finished in the final version. The dash symbol (—) means missing results. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

Measure	Methods			EXP		
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.408 ± 0.008	0.424 ± 0.008	0.406 ± 0.009	0.405 ± 0.01	0.414 ± 0.012
	CCP	0.007 ± 0.003	0.001 ± 0.001	0.0 ± 0.0	0.002 ± 0.002	0.001 ± 0.001
	cluster-CP	0.009 ± 0.003	—	—	—	0.002 ± 0.001
	k-**CCP**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**
APSS	MCP	9.705 ± 0.101	9.498 ± 0.102	9.438 ± 0.098	9.361 ± 0.092	8.931 ± 0.093
	CCP	26.666 ± 0.415	30.437 ± 0.352	29.777 ± 0.409	30.007 ± 0.33	34.867 ± 0.445
	cluster-CP	27.786 ± 0.307	—	—	—	33.114 ± 0.418
	k-**CCP**	**18.129 ± 0.454**	**17.546 ± 0.453**	**18.944 ± 0.381**	**18.81 ± 0.368**	**17.769 ± 0.463**

Table 9: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type EXP and APS scoring function on dataset mini-ImageNet. The dash symbol (—) means missing results that will be finished in the final version. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

Measure	Methods			POLY		
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.412 ± 0.013	0.4 ± 0.009	0.427 ± 0.011	0.407 ± 0.01	0.41 ± 0.009
	CCP	0.005 ± 0.002	0.002 ± 0.001	0.003 ± 0.001	0.001 ± 0.001	0.001 ± 0.001
	cluster-CP	0.028 ± 0.005	—	—	—	0.015 ± 0.003
	k-**CCP**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**
APSS	MCP	9.81 ± 0.102	9.838 ± 0.091	9.801 ± 0.107	9.528 ± 0.099	9.665 ± 0.101
	CCP	26.620 ± 0.369	30.236 ± 0.331	30.912 ± 0.401	31.639 ± 0.422	29.852 ± 0.36
	cluster-CP	21.273 ± 0.369	—	—	—	25.550 ± 0.279
	k-**CCP**	**17.784 ± 0.438**	**17.751 ± 0.466**	**19.388 ± 0.441**	**19.342 ± 0.443**	**19.153 ± 0.412**

Table 10: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type POLY and APS scoring function on dataset mini-ImageNet. The dash symbol (—) means missing results that will be finished in the final version. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

Measure	Methods			MAJ		
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.408 ± 0.009	0.405 ± 0.012	0.424 ± 0.01	0.43 ± 0.01	0.411 ± 0.007
	CCP	0.011 ± 0.004	0.005 ± 0.002	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	cluster-CP	0.015 ± 0.005	—	—	—	0.011 ± 0.003
	k-**CCP**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**
APSS	MCP	9.84 ± 0.091	9.929 ± 0.112	9.817 ± 0.092	9.499 ± 0.094	9.123 ± 0.086
	CCP	27.306 ± 0.377	31.114 ± 0.456	30.741 ± 0.345	30.608 ± 0.433	34.186 ± 0.32
	cluster-CP	25.288 ± 0.226	—	—	—	25.229 ± 0.352
	k-**CCP**	**18.11 ± 0.414**	**17.874 ± 0.511**	**19.711 ± 0.439**	**19.592 ± 0.376**	**18.594 ± 0.439**

Table 11: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type MAJ and APS scoring function on dataset mini-ImageNet. The dash symbol (—) means missing results that will be finished in the final version. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

Measure	Methods			EXP		
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.444 ± 0.007	0.452 ± 0.007	0.439 ± 0.005	0.427 ± 0.006	0.364 ± 0.01
	CCP	0.001 ± 0.001	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	cluster-CP	0.007 ± 0.003	—	—	—	0.003 ± 0.002
	k-CCP	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**	**0.0 ± 0.0**
APSS	MCP	9.57 ± 0.076	9.687 ± 0.075	10.437 ± 0.064	11.404 ± 0.076	13.998 ± 0.089
	CCP	40.408 ± 0.378	41.156 ± 0.405	40.881 ± 0.398	42.207 ± 0.356	60.762 ± 0.531
	cluster-CP	28.828 ± 0.294	—	—	—	44.885 ± 0.589
	k-CCP	**17.281 ± 0.225**	**17.294 ± 0.202**	**17.928 ± 0.21**	**18.63 ± 0.241**	**20.61 ± 0.222**

Table 12: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type EXP and APS scoring function on dataset Food-101. The dash symbol (—) means missing results that will be finished in the final version. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

Measure	Methods			POLY		
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.466 ± 0.009	0.446 ± 0.011	0.456 ± 0.006	0.465 ± 0.007	0.451 ± 0.008
	CCP	**0.001 ± 0.001**	**0.001 ± 0.001**	0.002 ± 0.001	**0.0 ± 0.0**	0.001 ± 0.001
	cluster-CP	0.007 ± 0.002	—	—	—	0.010 ± 0.003
	k-CCP	**0.001 ± 0.001**	**0.001 ± 0.001**	**0.001 ± 0.001**	0.001 ± 0.001	**0.001 ± 0.001**
APSS	MCP	12.267 ± 0.079	12.349 ± 0.085	13.533 ± 0.09	14.357 ± 0.08	16.468 ± 0.095
	CCP	45.148 ± 0.342	45.572 ± 0.355	46.134 ± 0.347	47.788 ± 0.407	65.672 ± 0.515
	cluster-CP	32.873 ± 0.307	—	—	—	38.326 ± 0.248
	k-CCP	**20.452 ± 0.209**	**20.503 ± 0.192**	**21.606 ± 0.206**	**22.62 ± 0.187**	**24.771 ± 0.192**

Table 13: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type POLY and APS scoring function on dataset Food-101. The dash symbol (—) means missing results that will be finished in the final version. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

Measure	Methods			MAJ		
		$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
UCR	MCP	0.462 ± 0.008	0.469 ± 0.009	0.47 ± 0.01	0.459 ± 0.007	0.467 ± 0.006
	CCP	**0.0 ± 0.0**	**0.001 ± 0.001**	0.003 ± 0.001	**0.0 ± 0.0**	**0.0 ± 0.0**
	cluster-CP	0.007 ± 0.003	—	—	—	0.005 ± 0.003
	k-CCP	**0.0 ± 0.0**	**0.001 ± 0.001**	**0.002 ± 0.001**	**0.0 ± 0.0**	0.002 ± 0.001
APSS	MCP	9.964 ± 0.078	10.742 ± 0.075	11.57 ± 0.088	12.654 ± 0.085	16.256 ± 0.088
	CCP	41.453 ± 0.335	43.252 ± 0.385	44.884 ± 0.348	45.492 ± 0.288	66.633 ± 0.622
	cluster-CP	33.258 ± 0.450	—	—	—	46.430 ± 0.337
	k-CCP	**19.398 ± 0.223**	**18.97 ± 0.217**	**20.375 ± 0.218**	**21.172 ± 0.242**	**25.164 ± 0.197**

Table 14: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type MAJ and APS scoring function on dataset Food-101. The dash symbol (—) means missing results that will be finished in the final version. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

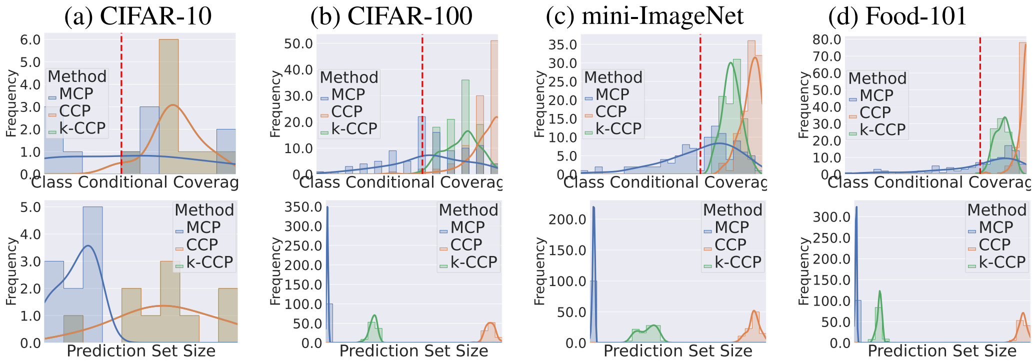

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 6: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by `MCP`, `CCP`, and k−`CCP` methods using ResNet20 model on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type EXP for imbalance ratio $\rho = 0.1$. It is clear that k-CCP has more densely distributed class-conditional coverage above 0.9 (the target $1-\alpha$ class-conditional coverage) than CCP with significantly smaller prediction sets on CIFAR-100, mini-ImageNet and Food-101.

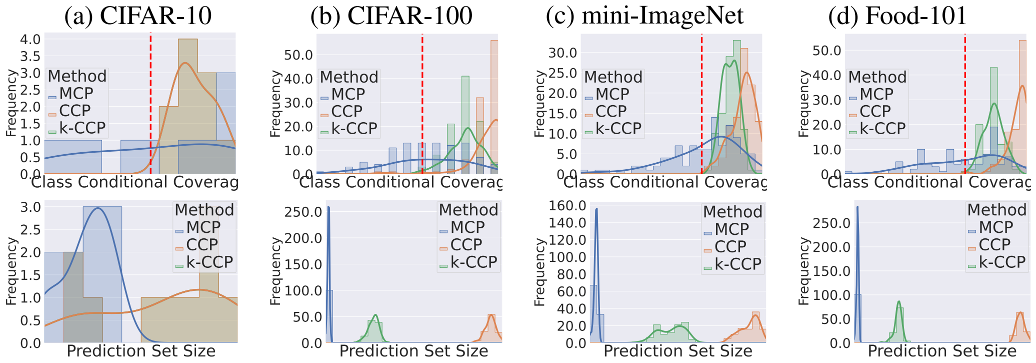

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 7: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by `MCP`, `CCP`, and k−`CCP` methods using ResNet20 model on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type EXP for imbalance ratio $\rho = 0.5$. It is clear that k-CCP has more densely distributed class-conditional coverage above 0.9 (the target $1-\alpha$ class-conditional coverage) than CCP with significantly smaller prediction sets on CIFAR-100, mini-ImageNet and Food-101.

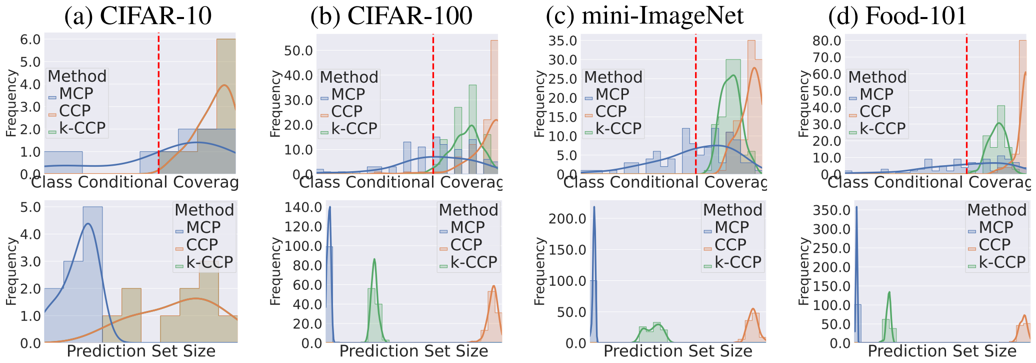
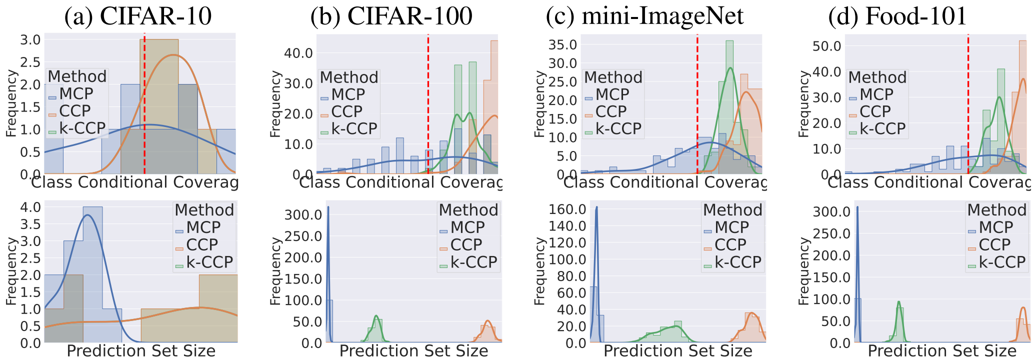
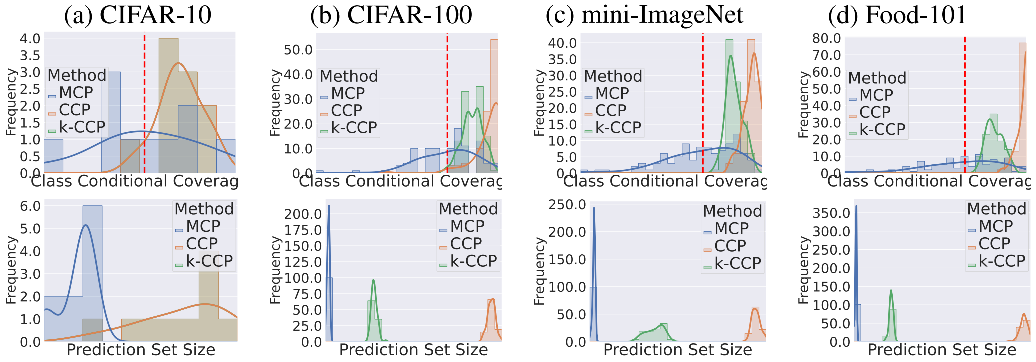

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 8: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by `MCP`, `CCP`, and k−`CCP` methods using ResNet20 model on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type POLY for imbalance ratio $\rho = 0.1$. It is clear that k-CCP has more densely distributed class-conditional coverage above 0.9 (the target $1 - \alpha$ class-conditional coverage) than CCP with significantly smaller prediction sets on CIFAR-100, mini-ImageNet and Food-101.



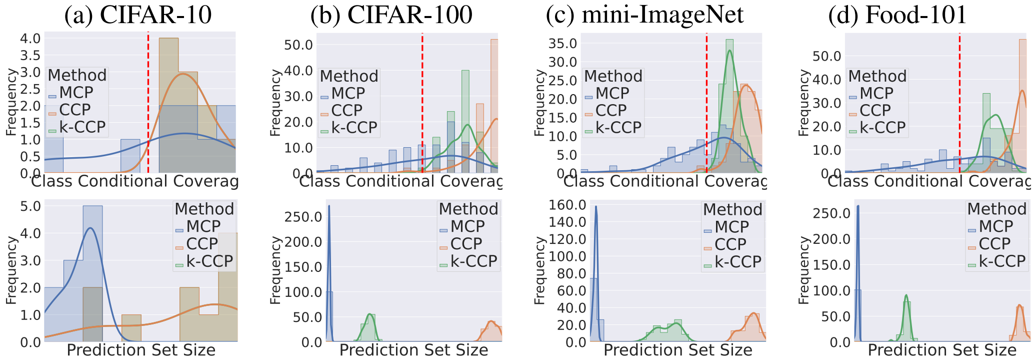
(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 9: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by `MCP`, `CCP`, and $k-$`CCP` methods using ResNet20 model on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type POLY for imbalance ratio $\rho = 0.5$. It is clear that k-CCP has more densely distributed class-conditional coverage above 0.9 (the target $1 - \alpha$ class-conditional coverage) than CCP with significantly smaller prediction sets on CIFAR-100, mini-ImageNet and Food-101.



(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 10: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by `MCP`, `CCP`, and $k-$`CCP` methods using ResNet20 model on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type MAJ for imbalance ratio $\rho = 0.1$. It is clear that k-CCP has more densely distributed class-conditional coverage above 0.9 (the target $1 - \alpha$ class-conditional coverage) than CCP with significantly smaller prediction sets on CIFAR-100, mini-ImageNet and Food-101.



(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 11: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by `MCP`, `CCP`, and $k-$`CCP` methods using ResNet20 model on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type MAJ for imbalance ratio $\rho = 0.5$. It is clear that k-CCP has more densely distributed class-conditional coverage above 0.9 (the target $1 - \alpha$ class-conditional coverage) than CCP with significantly smaller prediction sets on CIFAR-100, mini-ImageNet and Food-101.

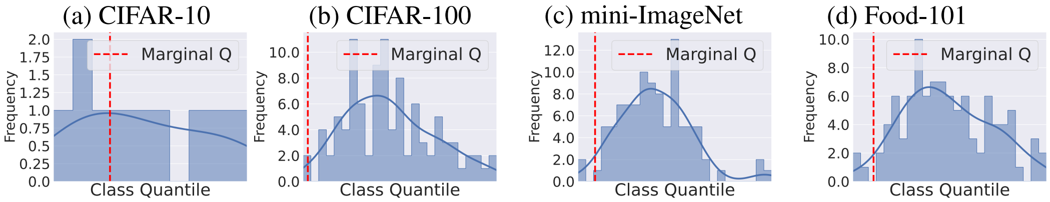

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 12: Distribution of class-wise quantiles with ResNet20 model on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type EXP and imbalance ratio $\rho = 0.5$. This result verifies that the deviation of class-wise quantiles from the marginal quantile can easily happen, i.e., the assumption of class-conditional under-coverage for MCP in Proposition 1.

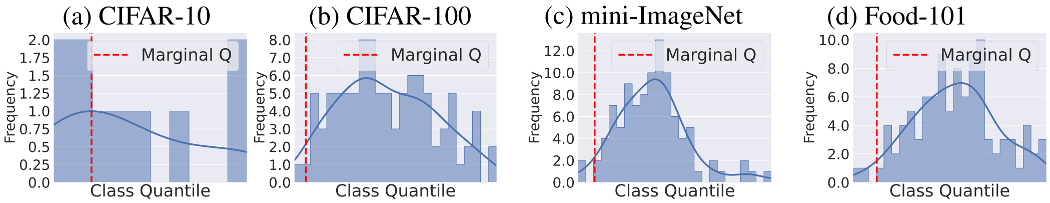

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 13: Distribution of class-wise quantiles with ResNet20 model on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type POLY and imbalance ratio $\rho = 0.5$. This result verifies that the deviation of class-wise quantiles from the marginal quantile can easily happen, i.e., the assumption of class-conditional under-coverage for MCP in Proposition 1.

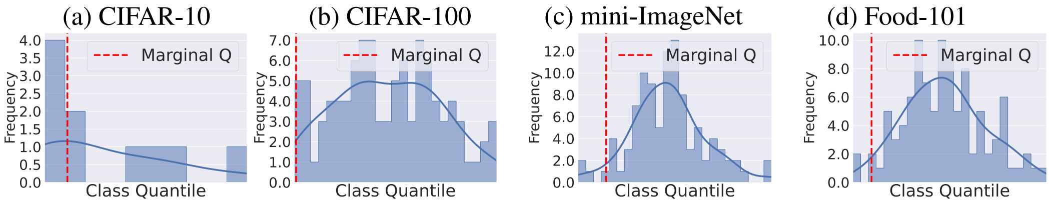

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 14: Distribution of class-wise quantiles with ResNet20 model on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type MAJ and imbalance ratio $\rho = 0.5$. This result verifies that the deviation of class-wise quantiles from the marginal quantile can easily happen, i.e., the assumption of class-conditional under-coverage for MCP in Proposition 1.

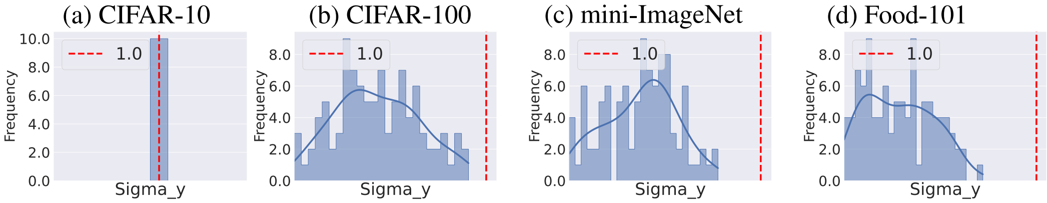

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 15: Verification of condition numbers $\{\sigma_y\}_{y=1}^{C}$ in Theorem 2 with $\rho = 0.1$ EXP. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Theorem 2, and thus confirms that k-CCP produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

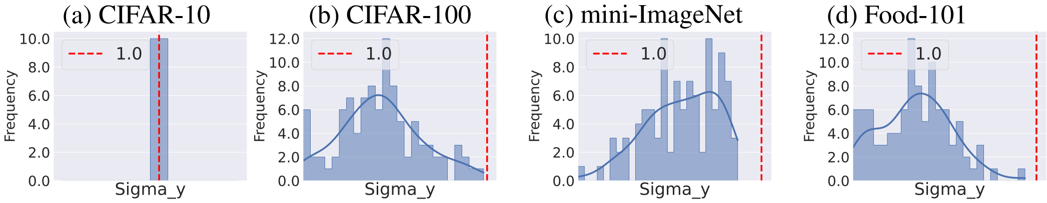

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 16: Verification of condition numbers $\{\sigma_y\}_{y=1}^{C}$ in Theorem 2 with $\rho = 0.5$ EXP. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Theorem 2, and thus confirms that k-CCP produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

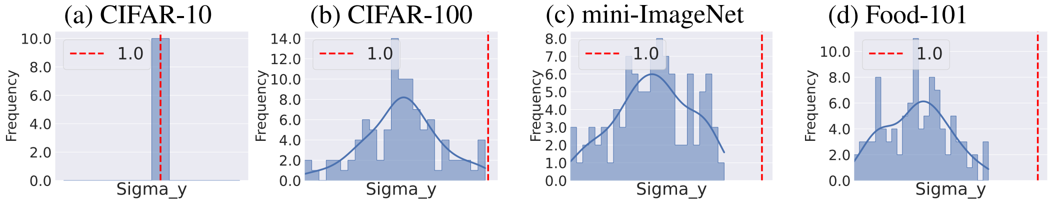

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 17: Verification of condition numbers $\{\sigma_y\}_{y=1}^{C}$ in Theorem 2 with $\rho = 0.1$ POLY. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Theorem 2, and thus confirms that k-CCP produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

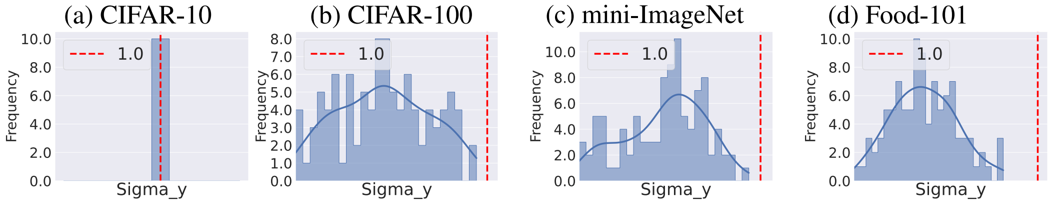

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 18: Verification of condition numbers $\{\sigma_y\}_{y=1}^{C}$ in Theorem 2 with $\rho = 0.5$ POLY. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Theorem 2, and thus confirms that k-CCP produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

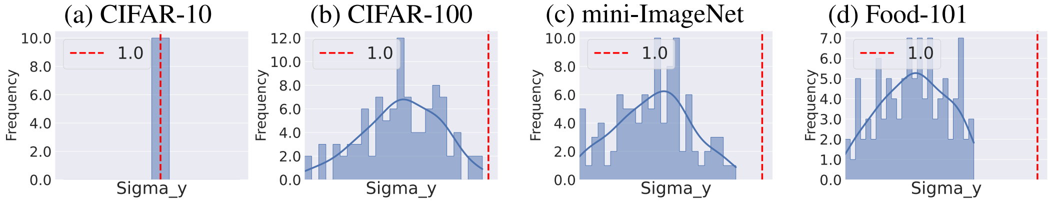

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 19: Verification of condition numbers $\{\sigma_y\}_{y=1}^{C}$ in Theorem 2 with $\rho = 0.1$ MAJ. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Theorem 2, and thus confirms that k-CCP produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

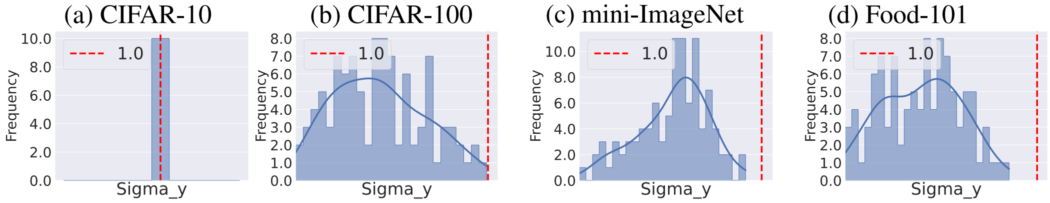

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 20: Verification of condition numbers $\{\sigma_y\}_{y=1}^{C}$ in Theorem 2 with $\rho = 0.5$ MAJ. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Theorem 2, and thus confirms that k-CCP produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

D EXPERIMENTS WITH CLUSTER-CP USING APS SCORE FUNCTION

With the same model, evaluation metrics, and APS score function Romano *et al.* (2020), we add the comparison experiments with cluster-CP Ding *et al.* (2023) [3] on four datasets and summarize the results in Table 15. We highlight that we also select the g from the interval $[0.1, 1]$ with increments of 0.05 to find the minimal g that k-CCP and cluster-CP achieves the target class-conditional coverage.

Based on the results in Table 15, we make the following observations: (i) cluster-CP and k-CCP can guarantee the class-conditional coverage; and (ii) k-CCP significantly outperforms cluster-CP on three datasets by producing smaller prediction sets.

Measure	Methods	EXP		POLY		MAJ	
		$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$
				CIFAR-10			
UCR	MCP	0.460 ± 0.025	0.490 ± 0.026	0.510 ± 0.033	0.320 ± 0.030	0.380 ± 0.020	0.500 ± 0.014
	CCP	**0.110 ± 0.030**	**0.040 ± 0.020**	0.140 ± 0.032	**0.010 ± 0.001**	**0.120 ± 0.024**	0.010 ± 0.001
	cluster-CP	0.160 ± 0.025	0.080 ± 0.012	**0.140 ± 0.021**	**0.010 ± 0.001**	0.150 ± 0.025	**0.09 ± 0.013**
	k-**CCP**	**0.110 ± 0.030**	**0.040 ± 0.020**	0.140 ± 0.032	**0.010 ± 0.001**	**0.120 ± 0.024**	0.010 ± 0.001
APSS	MCP	1.132 ± 0.033	1.406 ± 0.045	1.117 ± 0.028	1.214 ± 0.038	1.196 ± 0.032	2.039 ± 0.046
	CCP	1.481 ± 0.082	**2.032 ± 0.096**	**1.487 ± 0.090**	**1.945 ± 0.087**	**1.765 ± 0.093**	**2.964 ± 0.123**
	cluster-CP	**1.445 ± 0.017**	2.323 ± 0.015	1.612 ± 0.013	2.102 ± 0.015	1.787 ± 0.019	2.969 ± 0.025
	k-**CCP**	1.481 ± 0.082	**2.032 ± 0.096**	**1.487 ± 0.090**	**1.945 ± 0.087**	**1.765 ± 0.093**	**2.964 ± 0.123**
				CIFAR-100			
UCR	MCP	0.386 ± 0.009	0.384 ± 0.018	0.395 ± 0.010	0.411 ± 0.010	0.352 ± 0.010	0.401 ± 0.008
	CCP	0.009 ± 0.003	0.011 ± 0.002	0.011 ± 0.003	0.015 ± 0.003	0.017 ± 0.003	0.008 ± 0.004
	cluster-CP	0.004 ± 0.002	0.004 ± 0.002	0.001 ± 0.001	0.004 ± 0.002	0.005 ± 0.002	0.019 ± 0.005
	k-**CCP**	**0.000 ± 0.000**	**0.001 ± 0.001**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.002 ± 0.001**	**0.001 ± 0.001**
APSS	MCP	10.303 ± 0.111	14.544 ± 0.119	15.729 ± 0.126	25.888 ± 0.197	11.680 ± 0.117	23.796 ± 0.159
	CCP	44.194 ± 0.514	50.963 ± 0.481	49.895 ± 0.489	64.366 ± 0.389	48.323 ± 0.548	64.640 ± 0.621
	cluster-CP	30.922 ± 0.454	43.883 ± 1.070	56.696 ± 0.393	63.208 ± 0.364	33.623 ± 0.395	50.382 ± 0.711
	k-**CCP**	**20.355 ± 0.357**	**25.185 ± 0.278**	**25.843 ± 0.300**	**37.034 ± 0.244**	**21.196 ± 0.320**	**35.630 ± 0.232**
				mini-ImageNet			
UCR	MCP	0.408 ± 0.008	0.414 ± 0.012	0.412 ± 0.013	0.410 ± 0.0018	0.408 ± 0.010	0.411 ± 0.007
	CCP	0.007 ± 0.003	0.001 ± 0.001	0.005 ± 0.002	0.001 ± 0.001	0.011 ± 0.004	0.003 ± 0.001
	cluster-CP	0.009 ± 0.003	0.002 ± 0.001	0.028 ± 0.005	0.015 ± 0.003	0.015 ± 0.005	0.011 ± 0.003
	k-**CCP**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
APSS	MCP	9.705 ± 0.101	8.930 ± 0.093	9.810 ± 0.101	9.665 ± 0.101	9.840 ± 0.091	9.123 ± 0.086
	CCP	26.666 ± 0.415	34.867 ± 0.445	26.620 ± 0.369	29.852 ± 0.360	27.306 ± 0.377	29.200 ± 0.379
	cluster-CP	27.786 ± 0.307	33.114 ± 0.418	21.273 ± 0.369	25.550 ± 0.279	25.288 ± 0.226	25.229 ± 0.352
	k-**CCP**	**18.129 ± 0.453**	**17.769 ± 0.463**	**17.784 ± 0.438**	**19.153 ± 0.412**	**18.110 ± 0.414**	**18.594 ± 0.439**
				Food-101			
UCR	MCP	0.444 ± 0.007	0.364 ± 0.010	0.466 ± 0.009	0.451 ± 0.008	0.462 ± 0.008	0.467 ± 0.006
	CCP	0.001 ± 0.001	**0.000 ± 0.000**	**0.001 ± 0.001**	**0.001 ± 0.001**	**0.000 ± 0.000**	**0.000 ± 0.000**
	cluster-CP	0.007 ± 0.003	0.003 ± 0.002	0.007 ± 0.002	0.010 ± 0.003	0.007 ± 0.003	0.005 ± 0.003
	k-**CCP**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.001 ± 0.001**	**0.001 ± 0.001**	**0.000 ± 0.000**	**0.000 ± 0.000**
APSS	MCP	9.570 ± 0.076	13.998 ± 0.089	12.267 ± 0.079	16.468 ± 0.095	9.964 ± 0.078	23.796 ± 0.159
	CCP	40.408 ± 0.378	60.762 ± 0.531	45.148 ± 0.342	65.6723 ± 0.515	41.453 ± 0.335	66.633 ± 0.622
	cluster-CP	28.828 ± 0.294	44.885 ± 0.589	32.873 ± 0.307	38.326 ± 0.248	33.258 ± 0.450	46.430 ± 0.337
	k-**CCP**	**17.281 ± 0.225**	**20.610 ± 0.222**	**20.452 ± 0.209**	**24.771 ± 0.192**	**19.398 ± 0.223**	**26.584 ± 0.191**

Table 15: Results comparing k-`CCP` and `cluster-CP` with ResNet-20 model and APS score function under different imbalance ratios $\rho = 0.5$ and $\rho = 0.1$. We set UCR of k-`CCP` the same as or better than that of `cluster-CP` for a fair comparison of prediction set size. The APSS results show that k-`CCP` significantly outperforms `cluster-CP` in terms of the average prediction set size over all settings on CIFAR-100, mini-ImageNet, and Food-101.

E COMPARISON EXPERIMENTS USING RAPS SCORE FUNCTION

With the same model, evaluation metrics and RAPS score function Angelopoulos *et al.* (2020), we add the comparison experiments with MCP, CCP, and cluster-CP on four datasets with different imbalanced type and imbalance ratio $\rho = 0.5$ and $\rho = 0.1$. The regularization parameter for RAPS score function is from the set $k_{reg} \in \{3, 5, 7\}$ and $\lambda \in \{0.001, 0.01, 0.1\}$. We select the combination of k_{reg} and λ for each experiments with same imbalanced type and imbalanced ratio on same

---

[3] https://github.com/tiffanyding/class-conditional-conformal/tree/main

dataset, where the most of $APSS$ values of all methods are minimum. The overall performance is summarized in Table 16. We highlight that we also select the g from the interval $[0.1, 1]$ with increments of 0.05 to find the minimal g that CCP, cluster-CP, and k-CCP achieves the target class conditional coverage.

Based on the results in Table 16, we make the following observations: (i) CCP, cluster-CP, and k-CCP can guarantee the class-conditional coverage; and (ii) k-CCP significantly outperforms CCP and cluster-CP on three datasets by producing smaller prediction sets.

Measure	Methods	EXP		POLY		MAJ	
		$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$
				CIFAR-10			
UCR	MCP	0.460 ± 0.021	0.490 ± 0.026	0.500 ± 0.031	0.290 ± 0.030	0.380 ± 0.019	0.500 ± 0.014
	CCP	**0.010 ± 0.020**	**0.020 ± 0.013**	**0.080 ± 0.030**	0.050 ± 0.021	**0.090 ± 0.022**	**0.040 ± 0.015**
	cluster-CP	0.160 ± 0.025	0.080 ± 0.012	0.140 ± 0.021	**0.040 ± 0.015**	0.150 ± 0.025	0.120 ± 0.013
	k-CCP	**0.010 ± 0.020**	**0.020 ± 0.013**	**0.080 ± 0.030**	0.050 ± 0.021	**0.090 ± 0.022**	**0.040 ± 0.015**
APSS	MCP	1.143 ± 0.004	1.419 ± 0.013	1.118 ± 0.004	1.233 ± 0.006	1.196 ± 0.032	2.043 ± 0.016
	CCP	1.502 ± 0.007	**2.049 ± 0.013**	**1.558 ± 0.010**	**1.776 ± 0.012**	**1.786 ± 0.020**	**2.628 ± 0.012**
	cluster-CP	**1.493 ± 0.017**	2.323 ± 0.015	1.612 ± 0.013	1.981 ± 0.013	1.787 ± 0.019	2.968 ± 0.024
	k-CCP	1.502 ± 0.007	**2.049 ± 0.013**	**1.558 ± 0.010**	**1.776 ± 0.012**	**1.786 ± 0.020**	**2.628 ± 0.012**
				CIFAR-100			
UCR	MCP	0.388 ± 0.007	0.384 ± 0.018	0.394 ± 0.010	0.402 ± 0.010	0.352 ± 0.010	0.401 ± 0.007
	CCP	0.008 ± 0.002	0.011 ± 0.002	0.011 ± 0.003	0.015 ± 0.003	0.015 ± 0.003	0.008 ± 0.004
	cluster-CP	0.004 ± 0.002	0.004 ± 0.002	0.001 ± 0.001	0.003 ± 0.002	0.005 ± 0.002	0.018 ± 0.006
	k-CCP	**0.000 ± 0.000**	**0.001 ± 0.001**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.001 ± 0.001**	**0.000 ± 0.000**
APSS	MCP	10.300 ± 0.080	14.554 ± 0.107	15.755 ± 0.103	25.850 ± 0.150	11.684 ± 0.091	23.708 ± 0.137
	CCP	44.243 ± 0.340	50.969 ± 0.345	49.877 ± 0.354	64.247 ± 0.234	48.337 ± 0.355	64.580 ± 0.536
	cluster-CP	30.971 ± 0.454	43.883 ± 1.073	56.656 ± 0.354	63.113 ± 0.397	33.656 ± 0.388	50.365 ± 0.701
	k-CCP	**20.355 ± 0.005**	**25.185 ± 0.011**	**25.843 ± 0.006**	**37.035 ± 0.005**	**21.197 ± 0.005**	**35.631 ± 0.007**
				mini-ImageNet			
UCR	MCP	0.408 ± 0.008	0.410 ± 0.011	0.412 ± 0.014	0.410 ± 0.008	0.403 ± 0.010	0.412 ± 0.007
	CCP	**0.007 ± 0.003**	0.003 ± 0.001	0.018 ± 0.002	0.004 ± 0.002	**0.000 ± 0.000**	0.005 ± 0.002
	cluster-CP	0.008 ± 0.004	0.003 ± 0.001	0.031 ± 0.005	0.002 ± 0.001	0.004 ± 0.002	0.011 ± 0.004
	k-CCP	0.009 ± 0.003	**0.000 ± 0.000**	**0.004 ± 0.002**	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
APSS	MCP	9.703 ± 0.076	9.003 ± 0.067	9.806 ± 0.079	9.714 ± 0.075	9.865 ± 0.060	9.146 ± 0.063
	CCP	26.689 ± 0.177	29.750 ± 0.219	21.352 ± 0.196	26.266 ± 0.218	36.535 ± 0.196	25.641 ± 0.217
	cluster-CP	27.466 ± 0.268	32.991 ± 0.434	21.212 ± 0.298	36.061 ± 0.475	32.085 ± 0.424	25.269 ± 0.375
	k-CCP	**15.101 ± 0.003**	**18.418 ± 0.003**	**15.331 ± 0.003**	**17.465 ± 0.003**	**17.388 ± 0.003**	**17.167 ± 0.004**
				Food-101			
UCR	MCP	0.445 ± 0.007	0.363 ± 0.010	0.466 ± 0.008	0.450 ± 0.007	0.457 ± 0.008	0.465 ± 0.006
	CCP	0.002 ± 0.001	**0.000 ± 0.000**	0.002 ± 0.001	**0.001 ± 0.001**	0.001 ± 0.001	0.008 ± 0.002
	cluster-CP	0.003 ± 0.002	0.003 ± 0.002	0.007 ± 0.002	0.005 ± 0.002	0.007 ± 0.002	0.006 ± 0.003
	k-CCP	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.001 ± 0.001**	**0.001 ± 0.001**	**0.000 ± 0.000**	**0.002 ± 0.001**
APSS	MCP	9.580 ± 0.037	14.039 ± 0.055	12.327 ± 0.046	16.541 ± 0.060	10.040 ± 0.051	16.293 ± 0.047
	CCP	40.411 ± 0.285	60.790 ± 0.395	36.550 ± 0.141	41.755 ± 0.153	32.957 ± 0.224	36.797 ± 0.139
	cluster-CP	28.919 ± 0.287	44.583 ± 0.667	32.928 ± 0.358	41.785 ± 0.220	32.983 ± 0.518	46.078 ± 0.312
	k-CCP	**17.282 ± 0.004**	**20.610 ± 0.006**	**20.452 ± 0.002**	**24.771 ± 0.004**	**19.398 ± 0.006**	**25.163 ± 0.002**

Table 16: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` under different imbalance ratios $\rho = 0.5$ and $\rho = 0.1$ using the RAPS score function. The regularization parameter for RAPS score function is selected from the set $[3, 5, 7]$ and $[0.001, 0.01, 0.1]$. We select the best results as each element in the table. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size. The APSS results show that k-`CCP` significantly outperforms `CCP` and `cluster-CP` in terms of average prediction set size over all settings on CIFAR-100, mini-ImageNet, and Food-101.

F COMPARISON EXPERIMENTS WITH CONFORMAL TRAINING MODEL

To verify the suitability of k-CCP, we add the experiments by performing calibration using MCP, CCP, cluster-CP, and k-CCP on the conformal training model Stutz *et al.* (2021) [4] in CIFAR-100 dataset. The imbalance type is EXP and the imbalance ratio is $\rho = 0.1$. The regularization parameters of conformal training model are $r \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$. We have selected the best results for each method and summarized results in Table 17. It is clear that k-CCP outperforms CCP and cluster-CP, and improvement is not caused by RAPS. Due to the limited time for rebuttal and

---

[4] https://github.com/google-deepmind/conformal_training/tree/main

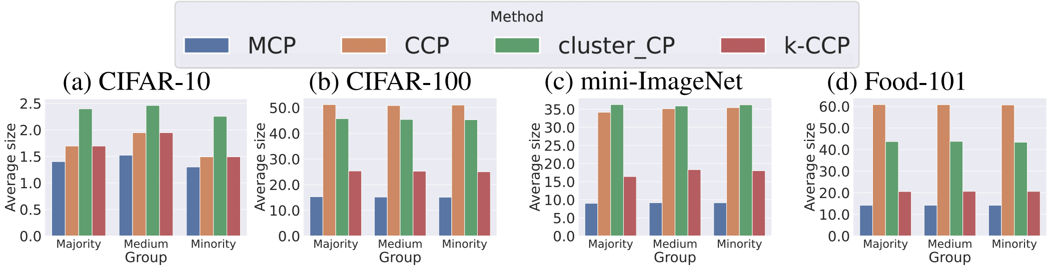
large computational cost of training conformal classifiers, we plan to add the corresponding results for mini-ImageNet and Food-101 in the final paper.

Measure	Methods	CIFAR-100	
		APS	RAPS
UCR	MCP	0.355 ± 0.014	0.377 ± 0.011
	CCP	0.041 ± 0.004	0.046 ± 0.004
	cluster-CP	0.027 ± 0.008	0.039 ± 0.007
	k-CCP	**0.000** ± **0.000**	**0.000** ± **0.000**
APSS	MCP	19.786 ± 0.141	15.175 ± 0.127
	CCP	38.505 ± 0.301	37.439 ± 0.593
	cluster-CP	36.055 ± 0.636	36.193 ± 0.687
	k-CCP	**31.281** ± **0.007**	**31.283** ± **0.006**

Table 17: Results comparing `MCP`, `CCP`, `cluster-CP`, and k-`CCP` with ResNet-18 model by conformal training loss proposed by Stutz *et al.* (2021) and imbalance ratio $\rho = 0.1$ EXP on dataset CIFAR-100. We set UCR of k-`CCP` the same as or better than that of `CCP` and `cluster-CP` for a fair comparison of prediction set size.

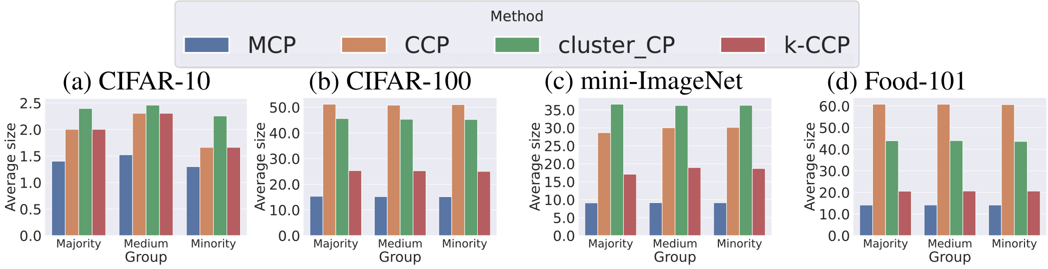
G ILLUSTRATION OF GROUP-WISE AVERAGE PREDICTION SIZE

To visualize the average prediction prediction set size gap between group class, we add the experiments with same model, imbalance type and imbalance ratio on four datasets. We select the top $1/4$ classes of largest number of data to the majority group. Similarly, we select the bottom $1/4$ classes of smallest number of data to the minority group, and the remaining $1/2$ classes to the medium group.



(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 21: Comparison of average prediction set size with $\rho = 0.1$ EXP and APS score function. These results show that there exists small gap between the average predictions set size on majority, medium, and minority groups.

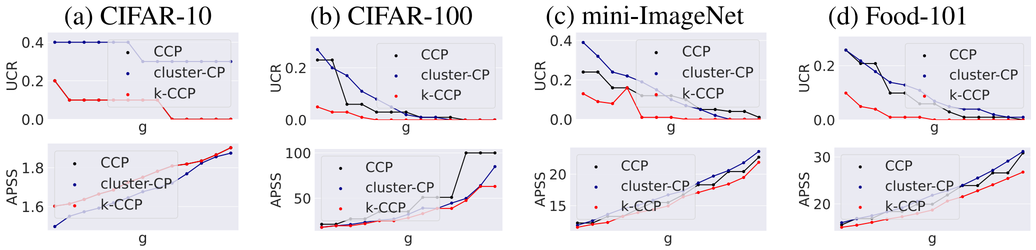


(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 22: Comparison of average prediction set size with $\rho = 0.1$ EXP and RAPS score function. These results show that there exists small gap between the average predictions set size on majority, medium, and minority groups.

H ABLATION STUDY FOR HYPER-PARAMETER g WITHIN k-CCP

We add the ablation study to verify how the hyper-parameter g affects the performance of CCP, Cluster-CP, and k-CCP. We use the under coverage ratio (UCR) and average prediction



(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

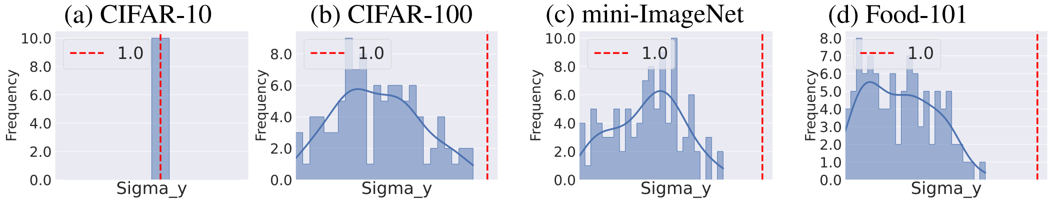
Figure 23: Under coverage ratio (Top row) and average prediction set size (Bottom row) achieved by `CCP`, `cluster-CP`, and k`-CCP` methods using ResNet20 model and RAPS score function on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type EXP for imbalance ratio $\rho = 0.1$. k`-CCP` degenerates to `CCP` in CIFAR-10, so there are only two lines (the black line overlaps with the red line).

set size (APSS) as metrics, which are introduced in Section 5.1. We set the range of g from $\{0.1, 0.15, \cdots, 0.7\}$ on four datasets. Based on the results in Figure 2, 5, and 23, the UCR and APSS of k-CCP are much smaller than CCP and Cluster-CP with the same g value.

I VERIFICATION OF σ_y ON CALIBRATION SET

To investigate the validity of σ_y on calibration datasets, we add experiments with imbalance ratio $\rho = 0.1$ and imbalance type EXP on four datasets.

Figure 24 verifies the validity of Theorem 2 and confirms that optimized trade-off between the coverage with inflated quantile and the constraint with calibrated rank leads to smaller prediction sets. Experiments even show a stronger condition ($\sigma_y \leq 1$ for all y) than the weighted aggregation condition in (12). It also confirms that the condition number $\{\sigma_y\}_{y=1}^C$ could be evaluated on calibration datasets without testing datasets and thus decreases the computation cost. We notice that k-CCP degenerates to CCP on CIFAR-10, so $\sigma_y = 1$ for all y and there is no trade-off. On the other three datasets, we observe significant conditions for the optimized trade-off in k-CCP.



(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101

Figure 24: Verification of condition numbers $\{\sigma_y\}_{y=1}^C$ in Theorem 2 with $\rho = 0.1$ EXP on calibration datasets. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Theorem 2, and thus confirms that the conditional number $\{\sigma_y\}_{y=1}^C$ could be evaluated on calibration datasets without testing datasets and thus decrease the computation cost.