Proceedings of Machine Learning Research 230:1–13, 2024 Conformal and Probabilistic Prediction with Applications

Entropy Reweighted Conformal Classification

Rui Luo RUILUO@CITYU.EDU.HK
City University of Hong Kong, Hong Kong, China
Nicolo Colombo NICOLO.COLOMBO@RHUL.AC.UK
Royal Holloway, University of London, Egham, Surrey, UK

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

Conformal Prediction (CP) is a powerful framework for constructing prediction sets with guaranteed coverage. However, recent studies have shown that integrating confidence calibration with CP can lead to a degradation in efficiency. In this paper, We propose an adaptive approach that considers the classifier's uncertainty and employs entropy-based reweighting to enhance the efficiency of prediction sets for conformal classification. Our experimental results demonstrate that this method significantly improves efficiency.

Keywords: Conformal prediction, entropy reweighting, confidence calibration, temperature scaling, neural networks.

1. Introduction

Conformal Prediction (CP) is a well-established framework for constructing prediction sets with guaranteed coverage, regardless of the underlying distribution of the data. However, the efficiency of CP prediction sets can be affected by the accuracy of the underlying classifier's uncertainty quantification.

Recent studies (Dabah and Tirer, 2024; Xi et al., 2024) have investigated the integration of confidence calibration and CP to improve the quality of the prediction sets. Confidence calibration aims to ensure that the predicted probabilities of a classifier reflect its true accuracy (Xi et al., 2024). By calibrating the classifier's probabilities, one can obtain more reliable uncertainty estimates, which can potentially lead to more efficient prediction sets. However, the results presented in Dabah and Tirer (2024); Xi et al. (2024) show that the efficiency of prediction sets degrades when confidence calibration is combined with CP. This raises the question of how to effectively integrate accurate uncertainty quantification with guaranteed coverage to obtain efficient prediction sets.

We propose a novel approach that applies entropy-based reweighting to conformal classification in order to improve the efficiency of prediction sets. This method leverages the uncertainty of the classifier to dynamically adjust the weights, leading to more efficient prediction sets.

We conduct extensive experiments on various datasets, including AG News, CARER, MNIST, and Fashion MNIST, to evaluate the effectiveness of our approach. By comparing our method with existing techniques, we demonstrate its superior performance in terms of prediction efficiency and accuracy. Our experimental results highlight the robustness and applicability of the proposed method across different types of data and classification tasks.

2. Related Work

Previous related works can be considered under two categories:

(1) Conformal Prediction for Calibrated Models

Methods like Platt's scaling (Platt et al., 1999), isotonic regression (Niculescu-Mizil and Caruana, 2005), spline-based probability calibration (Lucena, 2018), and temperature scaling (Guo et al., 2017) are used for post ad hoc calibration to adjust classifier confidence levels. Local Temperature Scaling (Ding et al., 2021) calibrates probabilities in multi-label segmentation by assigning a unique temperature to each pixel for location-specific calibration. Temperature-conditional Generative Flow Networks (Kim et al., 2023) incorporate a distinct pathway in their structure to adjust the policy's logits with the inverse temperature β, minimizing disruption to the network's parameters. They lack the coverage guarantee offered by CP. Moreover, temperature scaling, which is the most favored technique in this context, requires a considerable amount of validation data that must be representative of the training data. On the other hand, CP merely requires that the validation data (i.e., the calibration data) and the test data be exchangeable. However, integrating confidence calibration with CP fails to deliver the anticipated advantages. Dabah and Tirer (2024); Xi et al. (2024) empirically show that using post ad hoc calibration before applying CP increase the size of prediction size. The observation that overconfident models (e.g. low temperature in temperature scaling) produce small prediction sets but an exceedingly low temperature fails to achieve the desired coverage Xi et al. (2024) motivates a novel loss function which penalizes the under coverage of the prediction set. Cha et al. (2023) studied the impact of temperature scaling in Bayesian graph neural networks on the prediction set produced by CP for node classification task. As noted in Stutz et al. (2023), in complex tasks like the dermatology problem, expert disagreements result in a high-entropy conditional probability $P(Y|X)$ that deviates from a one-hot distribution, leading to a coverage gap when modeled using a majority-voting scheme.

(2) Uncertainty Measures for Deep Classifiers

Various methods for quantifying uncertainty have been introduced for neural network-based classifiers. These uncertainty measures prove valuable in scenarios such as selecting samples based on uncertainty for active learning (Nguyen et al., 2022) and assigning uncertainty evaluations to individual data points (Chlaily et al., 2023; Hüllermeier and Waegeman, 2021). Additionally, uncertainty measures related to the concept of aleatoric and epistemic uncertainty have been discussed in Gruber et al. (2023). Rossellini et al. (2024) distinguishes between these two forms of uncertainty for constructing prediction intervals for conformalized quantile regressors. Zhu et al. (2008); Nguyen et al. (2022) have explored the balance between uncertainty and representativeness in the context of active learning. Defining $f_Y(X)$ as the classifier's predictive probability of categorizing the object with feature X as label Y, we can summarize some of existing uncertainty measures as follows:

1. *Entropy:* $-\sum_{Y=1}^{K} f_Y(X) \log f_Y(X)$.
2. *Smallest margin:* $\arg\max_{Y^* \in \{1,\ldots,K\}} f_{Y^*}(X) - \arg\max_{Y \in \{1,\ldots,K\} \setminus \{Y^*\}} f_Y(X)$.
3. *Gini impurity:* $\sum_{Y=1}^{K} (1 - f_Y(X)) f_Y(X)$.

4. *Threshold conformity score (Sadinle et al., 2019):* $1 - f_Y(X)$, which penalizes cases where f does not predict the observed label Y with high probability.

5. *Adaptive conformity score (Romano et al., 2020):* This is computed by summing up the sorted softmax values in a descending sequence

$$a(\mathbf{f}(X), Y) = \sum_{i=1}^{r(Y,\mathbf{f}(X))-1} f_{(i)}(X) + U f_{(r(Y,\mathbf{f}(X)))}(X), \tag{1}$$

where $f_{(1)}(X) > f_{(2)}(X) > \cdots > f_{(K)}(X)$ represent the order statistics of $\mathbf{f}(X)$, $U \sim \text{Uniform}(0,1)$ is independent of everything else, and $r(Y, \mathbf{f}(X))$ is $\mathbf{f}(X)$'s ranking of the label Y.

6. *Regularized Adaptive conformity score (Angelopoulos et al., 2021):* To tackle the long-tailed distribution issue inherent in softmax probabilities, Regularized Adaptive Prediction Sets (RAPS) eliminate classes that are less likely by imposing a penalty on classes that surpass a predetermined threshold:

$$a(\mathbf{f}(X), Y) = \sum_{i=1}^{r(Y,\mathbf{f}(X))-1} \big(f_{(i)}(X) + \lambda \mathbb{1}(i > k_{\text{reg}})\big) + U f_{(r(Y,\mathbf{f}(X)))}(X), \tag{2}$$

where $\lambda > 0$ discourages sets larger than k_{reg}.

7. *Rank-based conformity score (Luo and Zhou, 2024a):* The score function of RANK is defined as:

$$a(\mathbf{f}(X), Y) = \frac{r(Y, \mathbf{f}(X))}{K},$$

which assigns a score based on the rank of the estimated probability $\hat{f}_Y(X)$ among all the estimated probabilities for feature X. The rank is divided by K so that the range of the score is from 0 to 1. The prediction set gives higher priority to labels with larger ranks.

3. Entropy Reweighted Conformal Classification

We start by defining key terms related to classification with CP. Let the feature space be $\mathbb{R}^d$ and the label space be $\mathcal{Y} = \{1, \ldots, K\}$. Denote f an K-class classification model and $D = \{(X_n, Y_n) \in \mathbb{R}^d \times \mathcal{Y}\}_{n \in \mathcal{I}}$, $\mathcal{Y}$ a collection of i.i.d. random variables. Assume X_{N+1} is a test object with its label Y_{N+1} masked. The output of the classification model, denoted as $f(X_{N+1}) \in [0,1]^K$, represents the predicted probabilities of the test object belonging to each of the K classes. A conformal Prediction Set (PS) at X_{N+1}, is a subset of the label space, $C(X_{N+1}) \subseteq \mathcal{Y}$ that obeys

$$\text{Prob}(Y_{N+1} \in C(X_{N+1})) \geq 1 - \alpha, \tag{3}$$

where $\alpha \in (0,1)$ is a predefined confidence level and the probability is over D. We partition the set of indices $\mathcal{I}$ into $\mathcal{I}_1$ and $\mathcal{I}_2$, and then we build the PS using the split conformal method. The classification model f is trained using the training samples $\{(X_n, Y_n)\}_{n\in\mathcal{I}_1}$. The PS depends on the calibration samples $\{(X_n, Y_n)\}_{n\in\mathcal{I}_2}$, the corresponding predicted probabilities, $\mathbf{f}(\mathbf{X_n}) \in [0,1]^K$, and an arbitrary conformity function, $a(\mathbf{f}(\mathbf{X_n}), Y_n)$ which evaluates the goodness of a model prediction compared with the corresponding label. Marginal validity (3) holds if the test data and the calibration set, i.e., (X_{N+1}, Y_{N+1}) and $\{(X_n, Y_n)\}_{n\in\mathcal{I}_2}$ are exchangeable.

For classification problems, one can define a stronger property known as conditional coverage:

$$\text{Prob}(Y_{N+1} \in C | Y_{N+1} = y) \geq 1 - \alpha, \tag{4}$$

which is alternatively termed as label-conditional coverage (Ding et al., 2024; Löfström et al., 2015) to differentiate it from the feature-conditional coverage (Einbinder et al., 2022):

$$\text{Prob}(Y_{N+1} \in C | X_{N+1} = x) \geq 1 - \alpha, \tag{5}$$

3.1. Conformal Classification

By definition, the PS based on conformity function $a(\mathbf{f}(\mathbf{X_n}), Y_n)$ is:

$$C_A = \{y_{N+1} \in \mathcal{Y}, \sum_{n\in\mathcal{I}_2} \mathbf{1}\left(A_n \leq A_{N+1}\right) \leq n_\alpha\} \ = \{y_{N+1} \in \mathcal{Y}, A_{N+1} \leq Q_A\} \tag{6}$$

where $A_n = a(\mathbf{f}(\mathbf{X_n}), Y_n)$, $A_{N+1} = a(\mathbf{f}(X_{N+1}), y_{N+1})$, obeys (3) if Q_A is the $(1-\alpha)$-th sample quantile of $A_1, \ldots, A_N$ and $n_\alpha = \lceil (1-\alpha)(|\mathcal{I}_2| + 1) \rceil$. The meaning of the obtained PS depends on the definition of a. In regression tasks, it is natural to let $a = a(\hat{Y}, Y)$ be the distance between predicted and observed labels. In the classification setup, the model output is a discrete probability distribution. Finding a conformity function that produces useful PS in the classification setup is less straightforward. A popular choice, APS score (1), is

$$A_n = a(\mathbf{f}(\mathbf{X_n}), Y_n) = \sum_{i=1}^{r(Y_n, \mathbf{f}(X_n))-1} f_{(i)}(X_n) + U f_{(r(Y_n, \mathbf{f}(X_n)))}(X_n), \tag{7}$$

where $f_{(i)}(X_n)$ denotes the i-th largest element of the probability vector $\mathbf{f}(X_n)$, $r(Y_n, \mathbf{f}(X_n))$ is the rank of the true label Y_n in the probability vector $\mathbf{f}(X_n)$.

Although we use the APS score throughout our derivation and experiments, the entropy reweighting method is applicable and orthogonal to other score functions such as THR (Sadinle et al., 2019), RAPS (Angelopoulos et al., 2021), and SAPS (Huang et al., 2024).

3.2. Reweighted Conformity Scores

As C_A depends on all data points unregarding their attribute or label, the validity of C, i.e. the coverage guarantees in (3), is marginal over the attribute and label spaces. This means the uncertainty of the model on predicting the class probabilities is assumed to be constant. As for the regression case, data heteroskedasticity may make such marginal PS highly inefficient. In the Error Reweighted (ER) Conformal Prediction algorithm of

Algorithm 1 Entropy Reweighted Conformal Prediction with Calibrated Temperature

```
Input: Labeled data $\mathcal{D} = \{(X_n, Y_n)\}_{n \in [N]}$, unlabeled data $\mathcal{U} = \{X_{n'}\}_{n' \in [N']}$, coverage
       probability $1 - \alpha$, and a range of predefined temperature values $\{T_j\}_{j \in [M]}$.
Output: Prediction set $C_\alpha$ for unlabeld data $\mathcal{U}$.
Randomly split $[N] = \{1, 2, \dots, N\}$ into three (disjoint) parts $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$. Set $\mathcal{D}_1 =$
  $\{(X_n, Y_n) : i \in \mathcal{I}_1\}$, $\mathcal{D}_2 = \{(X_n, Y_n) : n \in \mathcal{I}_2\}$, and $\mathcal{D}_3 = \{(X_n, Y_n) : n \in \mathcal{I}_3\}$.
Train a classification model $f$ on $\mathcal{D}_1$ and use $f$ to obtain the logits on $\mathcal{D}_2$ and $\mathcal{D}_3$:

    $\mathbf{z}(X_n) = \mathbf{f}(X_n), n \in \mathcal{I}_2 \cup \mathcal{I}_3.$    (8)

for each temperature $T_j$ do
    Compute the reweighted probability vectors $\tilde{\mathbf{f}}(X_n)$ for each data point in $\mathcal{D}_2$ according
      to (23):

        $\tilde{f}_k(X_n) = \frac{\exp(\frac{z_k(X_n)}{H(X_n) \cdot T_j})}{\sum_{i=1}^{K} \exp(\frac{z_i(X_n)}{H(X_n) \cdot T_j})}, n \in \mathcal{I}_2.$    (9)

    Compute the reweighted conformity scores $\tilde{A}_n$ for each data point in $\mathcal{D}_2$ according to
      (24):

        $\tilde{A}_n = \sum_{i=1}^{r(Y_n, \tilde{\mathbf{f}}(X_n)) - 1} \tilde{f}_{(i)}(X_n) + U \tilde{f}_{(r(Y_n, \tilde{\mathbf{f}}(X_n)))}(X_n), n \in \mathcal{I}_2.$    (10)

    Construct the prediction sets $C_\alpha^{T_j}(X_n)$ for each data point in $\mathcal{D}_3$ as:

        $C_\alpha^{T_j}(X_n) = \left\{ y \in \mathcal{Y} : \tilde{A}_n(y) \geq Q_{\tilde{A}}^{T_j} \right\}, n \in \mathcal{I}_3,$    (11)

      where $Q_{\tilde{A}}^{T_j}$ is the $(1 - \alpha)$-th sample quantile of $\{\tilde{A}_n : (X_n, Y_n) \in \mathcal{D}_2\}$ and $\tilde{A}_n(y) =$
      $\sum_{i=1}^{r(y, \tilde{\mathbf{f}}(X_n)) - 1} \tilde{f}_{(i)}(X_n) + U \tilde{f}_{(r(y, \tilde{\mathbf{f}}(X_n)))}(X_n)$.
    Calculate the average size of the resulting prediction sets:

        $\mathrm{AvgSize}(T_j) = \frac{1}{|\mathcal{D}_3|} \sum_{n \in \mathcal{I}_3} |C_\alpha^{T_j}(X_n)|.$    (12)

end
Select the optimal temperature $T^* = \arg\min_{T_j} \mathrm{AvgSize}(T_j)$.
Compute the reweighted conformity scores $\tilde{A}_n$ for each $(X_n, Y_n) \in \mathcal{D}_2 \cup \mathcal{D}_3$ using $T^*$.
Compute the final prediction sets $C_\alpha^{T^*}(X)$ for each $X \in \mathcal{U}$ as:

    $C_\alpha^{T^*}(X) = \left\{ y \in \mathcal{Y} : \tilde{A}(y) \geq Q_{\tilde{A}}^{T^*} \right\},$    (13)

  where $\tilde{A}(y)$ is the reweighted conformity score for $X$ and label $y$, and $Q_{\tilde{A}}^{T^*}$ is the $(1 - \alpha)$-th
  sample quantile of $\{\tilde{A}_n(Y_i) : n \in \mathcal{I}_2 \cup \mathcal{I}_3\}$.
return the prediction sets $C_\alpha^{T^*}(X)$ for each $X \in \mathcal{U}$.
```

Papadopoulos et al. (2008)) and the localized conformal prediction of Guan (2023), the regression conformity scores are rescaled by a pre-trained function to boost the adaptivity of the obtained prediction intervals. More formally, ER computes the prediction intervals using a *transformed conformity score*, $B_n = \frac{A_n}{\gamma + g(X_n)}$, where g is a pre-trained model of the conditional residuals, i.e. $g(X_n) \approx \mathrm{E}_{Y_n|X_n}(|\mathbf{f}(\mathbf{X_n}) - Y_n|)$ and $\gamma > 0$ a regularization parameter. The corresponding PS

$$C_B = \{y_{N+1}, B_{N+1} \leq Q_B\}, \tag{14}$$

where Q_B is the $(1-\alpha)$-th sample quantile of $B_{n_{\text{train}}+1}, \ldots, B_N$. When $A_n = |\mathbf{f}(\mathbf{X_n}) - Y_n|$, this is equivalent to $B_n = \frac{A_n}{\gamma + \mathrm{E}_{A_n|X_n}(A_n)}$. This work is about transferring the ER idea to the classification domain. Colombo (2023) considers a set of parameterized transformations of the conformity scores,

$$B(X, Y) = b(A(X, Y), g(X, Y), \theta), \tag{15}$$

and propose to train them in a CP-aware sense. In this formalism, the ER approach by Papadopoulos et al. (2008) corresponds to setting $b(a, g) = \frac{a}{\gamma + g}$. The benefit of ER approach is that it improves feature-conditional coverage (5).

3.3. Entropy-based Reweighting

In the previous section, we emphasized the concept of reweighting the conformity scores based on the conditional residuals. However, since these residuals are estimated using a separate model g, there is a risk of model mis-specification if the model doesn't capture the underlying data distribution correctly. To address this, a logical approach would be to use information derived from the same model used for classification.

A natural extension of the ER approach to the classification setting involves setting $g(X) = H(X)$, where the entropy inherently adapts to and reflects the classification model's uncertainty about an individual data point.

Denote $\mathbf{z}(\mathbf{X}) = (z_1(X), \ldots, z_K(X))$ as the logit vector produced by the classification model, then the probability vector can be written as

$$\mathbf{f}(X) = \frac{\exp(\mathbf{z}(\mathbf{X}))}{\sum_{k=1}^K \exp z_k(X)}. \tag{16}$$

We consider reweighting the logit vector by the entropy of the corresponding probability vector. The entropy $H(X)$ of the probability vector $\mathbf{f}(X) = (f_1(X), \ldots, f_K(X))$ is defined as:

$$H(X) = -\sum_{k=1}^K f_k(X) \log f_k(X). \tag{17}$$

Substituting the softmax function into the entropy formula, we have:

$$
\begin{aligned}
H(X) &= -\sum_{k=1}^{K} \frac{\exp(z_k(X))}{\sum_{j=1}^{K}\exp(z_j(X))} \log\left(\frac{\exp(z_k(X))}{\sum_{j=1}^{K}\exp(z_j(X))}\right) && (18)\\
&= -\sum_{k=1}^{K} \frac{\exp(z_k(X))}{\sum_{j=1}^{K}\exp(z_j(X))} \left(z_k(X) - \log\left(\sum_{j=1}^{K}\exp(z_j(X))\right)\right) && (19)\\
&= -\sum_{k=1}^{K} f_k(X) z_k(X) + \log\left(\sum_{j=1}^{K}\exp(z_j(X))\right)\sum_{k=1}^{K} f_k(X) && (20)\\
&= -\sum_{k=1}^{K} f_k(X) z_k(X) + \log\left(\sum_{j=1}^{K}\exp(z_j(X))\right). && (21)
\end{aligned}
$$

Adding another tunable temperature parameter T, the reweighted logit vector becomes:

$$
\tilde{z}_k(X) = \frac{z_k(X)}{H(X)\cdot T},\ k = 1, \ldots, K. \tag{22}
$$

The resulting reweighted probability vector can be obtained by applying the softmax function to the reweighted logits:

$$
\tilde{f}_k(X) = \frac{\exp(\tilde{z}_k(X))}{\sum_{j=1}^{K}\exp(\tilde{z}_j(X))} = \frac{\exp(\frac{z_k(X)}{H(X)\cdot T})}{\sum_{j=1}^{K}\exp(\frac{z_j(X)}{H(X)\cdot T})}. \tag{23}
$$

The temperature parameter T controls the sharpness of the reweighted probability distribution. When $T \to 0$, the distribution becomes more concentrated on the class with the highest reweighted logit. Conversely, when $T \to \infty$, the distribution becomes more uniform. By adjusting T, we can control the influence of entropy-based reweighting on the resulting probability distribution. In our experiments, we follow the cross-validation procedure outlined in Yang and Kuchibhotla (2024) and the weighted aggregation idea in Luo and Zhou (2024b) to find the optimal temperature parameter using a separate validation set (see Algorithm 1).

The APS score (7) for the entropy reweighted probability vector (23) thus becomes:

$$
\tilde{A}_n = a(\tilde{\mathbf{f}}(X_n), Y_n) = \sum_{i=1}^{r(Y_n, \tilde{\mathbf{f}}(X_n))-1} \tilde{f}_{(i)}(X_n) + U\tilde{f}_{(r(Y_n,\tilde{\mathbf{f}}(X_n)))}(X_n), \tag{24}
$$

where $\tilde{\mathbf{f}}(X_n) = (\tilde{f}_1(X_n), \ldots, \tilde{f}_K(X_n))$ is the reweighted probability vector for input X_n and $r(Y_n, \tilde{\mathbf{f}}(X_n))$ is the rank of the true label Y_n in the reweighted probability vector $\tilde{\mathbf{f}}(X_n)$. We present the full procedure for constructing prediction sets with the entropy reweighted method in Algorithm 1.

4. Experiments

In this section, we evaluate the performance of our proposed method. We conducted experiments on four datasets: AG News (Zhang et al., 2015), CelebA Attributes (CARER) (Liu et al., 2015), Fashion MNIST (Xiao et al., 2017), and MNIST (LeCun et al., 2010). For the AG News and CARER datasets, we use BERT as the classifier, while for the MNIST and Fashion MNIST datasets, we use a multi-layer neural network.

The experiment evaluates the calibration performance of classifiers trained on the four datasets. The calibration performance is assessed using various score functions, including the Entropy Reweighted (ER) score function proposed in this work, APS (Romano et al., 2020), RAPS (Angelopoulos et al., 2021), and SAPS (Huang et al., 2024), across different desired coverage levels $1 - \alpha$ ranging from 0.90 to 0.99 with a step size of 0.01. We employ the split conformal method for all score functions. To account for variability, the experiment is repeated 10 times using different random splits of the data into calibration and test sets. The evaluation metrics used are coverage and size. Coverage measures the proportion of test instances for which the true label falls within the predicted prediction set, while size represents the average number of labels included in the prediction sets.

The results are presented in Figure 4, which shows the coverage-size plots for each dataset separately. Each plot displays the trade-off between coverage and size for different score functions. The proposed ER score function is labeled as “ER (Ours)” in the plots. From the plots, we observe that the ER score function achieves competitive performance compared to other score functions across all datasets. It maintains good coverage while yielding relatively small confidence set sizes. The results demonstrate the effectiveness of the proposed entropy-based reweighting approach in improving the calibration performance of the classifiers.

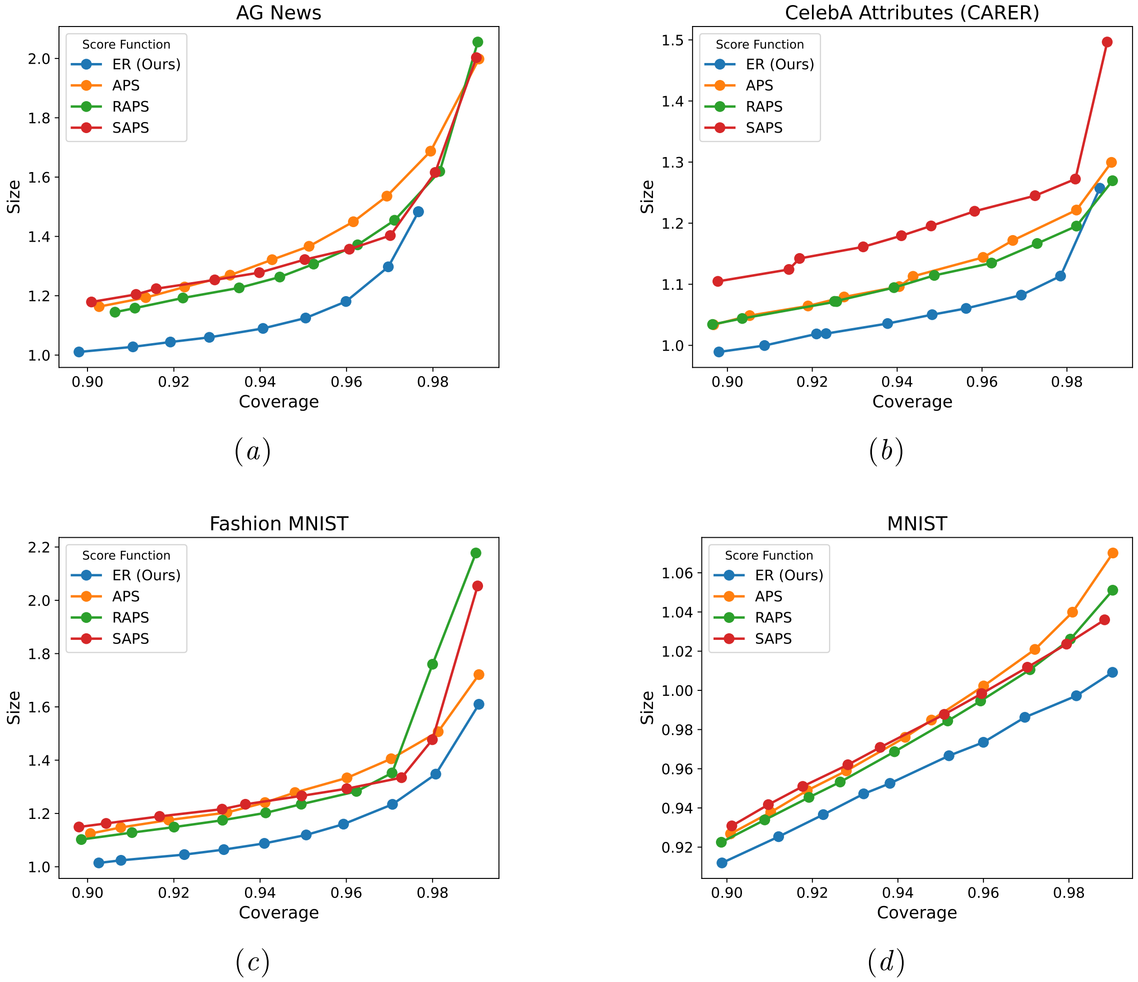
Table 1 indicates that the proposed reweighting method outperforms the baseline model in terms of size across different α values.

To explain the superior performance of the proposed ER method, we categorize the cases into four scenarios: (1) the model was correct ($Y_n = \arg\max_y f_y(X_n)$) and the entropy is low (indicating the model is certain about its prediction); (2) the model was correct and the entropy is high; (3) the model was incorrect and the entropy is low; and (4) the model was incorrect and the entropy is high. According to entropy reweighting, in the third scenario, the conformity score will adjust in preferable directions, meaning the score will increase and make the incorrect label harder to be included in the prediction set. Since this scenario is very common in over-confident deep network models, entropy reweighting will positively influence the scoring process and the construction of the prediction set.

5. Limitations and Future Work

Although the experimental results indicate enhancements in both the conditional coverage and the efficiency of the PS, the current analysis remains largely empirical. We aim to enhance the theoretical robustness of the entropy reweighting approach by pursuing the following future directions.

1. Temperature spline-based calibration: Instead of estimating a single temperature parameter T, we propose a temperature spline by estimating a feature-conditional



(a)

(b)

(c)

(d)

Figure 1: Size vs. Coverage plots for different datasets and score functions.

temperature function $T(X)$. This added flexibility could enable us to enhance the conditional coverage further.

2. Sample reweighting approach: Furthermore, we consider to reweight the samples according to localized CP approach (Guan, 2023), where we examine a local region around the test sample. Another option is by extending the calibration training idea in Colombo (2024) to classification settings.

3. Application in graph-structure datasets: We plan to apply entropy reweighted method to graph-structured datasets to explore how graph topology influences model uncertainty quantification. We will focus on key applications such as node classification (Cha et al., 2023), link prediction (Luo et al., 2023), and edge weight prediction (Luo and Colombo, 2024), which is crucial for weighted graphs in various domains.

Dataset	Score Function	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
		Coverage	Size	Coverage	Size	Coverage	Size
AG News	ER (Ours)	0.977	**1.484**	0.951	**1.125**	0.898	**1.011**
	APS	0.991	1.998	0.951	1.367	0.903	1.163
	RAPS	0.990	2.055	0.952	1.307	0.906	1.145
	SAPS	0.990	2.003	0.950	1.322	0.901	1.179
CARER	ER (Ours)	0.988	**1.257**	0.948	**1.050**	0.898	**0.989**
	APS	0.991	1.300	0.944	1.113	0.897	1.034
	RAPS	0.991	1.270	0.949	1.115	0.897	1.034
	SAPS	0.990	1.497	0.948	1.196	0.898	1.105
Fashion MNIST	ER (Ours)	0.991	**1.610**	0.951	**1.119**	0.903	**1.014**
	APS	0.991	1.721	0.948	1.279	0.901	1.125
	RAPS	0.990	2.178	0.950	1.235	0.899	1.102
	SAPS	0.990	2.054	0.950	1.266	0.898	1.149
MNIST	ER (Ours)	0.990	**1.009**	0.952	**0.967**	0.899	**0.912**
	APS	0.990	1.070	0.948	0.985	0.901	0.927
	RAPS	0.990	1.051	0.952	0.984	0.899	0.923
	SAPS	0.988	1.036	0.951	0.988	0.901	0.931

Table 1: Results for different α values.

References

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

Seohyeon Cha, Honggu Kang, and Joonhyuk Kang. On the temperature of bayesian graph neural networks for conformal prediction. In *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*, 2023. URL https://openreview.net/forum?id=Nrs8BA84br.

Saloua Chlaily, Debanshu Ratha, Pigi Lozou, and Andrea Marinoni. On measures of uncertainty in classification. *IEEE Transactions on Signal Processing*, 2023.

Nicolo Colombo. On training locally adaptive cp. In *Conformal and Probabilistic Prediction with Applications*, pages 384–398. PMLR, 2023.

Nicolo Colombo. Normalizing flows for conformal regression. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://openreview.net/forum?id=acgwLdoB3d.

Lahav Dabah and Tom Tirer. On calibration and conformal prediction of deep classifiers. *arXiv preprint arXiv:2402.05806*, 2024.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local temperature scaling for probability calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6889–6899, 2021.

Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *arXiv preprint arXiv:2205.05878*, 2022.

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint arXiv:2305.16703*, 2023.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Jianguo Huang, HuaJun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=b3pYoZfcoo.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

Minsu Kim, Joohwan Ko, Dinghuai Zhang, Ling Pan, Taeyoung Yun, Woo Chang Kim, Jinkyoo Park, and Yoshua Bengio. Learning to scale logits for temperature-conditional GFlownets. In *NeurIPS 2023 AI for Science Workshop*, 2023. URL https://openreview.net/forum?id=xRqAvRNUmq.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Tuve Löfström, Henrik Boström, Henrik Linusson, and Ulf Johansson. Bias reduction through conditional conformal prediction. *Intelligent Data Analysis*, 19(6):1355–1375, 2015.

Brian Lucena. Spline-based probability calibration. *arXiv preprint arXiv:1809.07751*, 2018.

Rui Luo and Nicolo Colombo. Conformal load prediction with transductive graph autoencoders. *arXiv preprint arXiv:2406.08281*, 2024.

Rui Luo and Zhixin Zhou. Trustworthy classification through rank-based conformal prediction sets. *arXiv preprint arXiv:2407.04407*, 2024a.

Rui Luo and Zhixin Zhou. Weighted aggregation of conformity scores for classification, 2024b. URL https://arxiv.org/abs/2407.10230.

Rui Luo, Buddhika Nettasinghe, and Vikram Krishnamurthy. Anomalous edge detection in edge exchangeable social network models. In *Conformal and Probabilistic Prediction with Applications*, pages 287–310. PMLR, 2023.

Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2022.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

Harris Papadopoulos, Alex Gammerman, and Volodya Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pages 64–69, 2008.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Raphael Rossellini, Rina Foygel Barber, and Rebecca Willett. Integrating uncertainty awareness into conformalized quantile regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1548. PMLR, 2024.

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.

David Stutz, Abhijit Guha Roy, Tatiana Matejovicova, Patricia Strachan, Ali Taylan Cemgil, and Arnaud Doucet. Conformal prediction under ambiguous ground truth. *arXiv preprint arXiv:2307.09302*, 2023.

Huajun Xi, Jianguo Huang, Lei Feng, and Hongxin Wei. Does confidence calibration help conformal prediction? *arXiv preprint arXiv:2402.04344*, 2024.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Yachong Yang and Arun Kumar Kuchibhotla. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, pages 1–13, 2024.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *22nd International Conference on Computational Linguistics, Coling 2008*, pages 1137–1144, 2008.