

The Entropic Overhead of Conformal Correction

Senrong Xu
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
srxu@smail.nju.edu.cn

Tianyu Wang
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
monica.tianyu@gmail.com

Zenan Li
ETH Zurich
Zurich, Switzerland
zenan.li@inf.ethz.ch

Yuan Yao
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
targenardy@gmail.com

Taolue Chen
Birkbeck, University of London
London, United Kingdom
taolue.chen@gmail.com

Feng Xu
Xiaoxing Ma
State Key Laboratory for Novel
Software Technology, Nanjing
University
Nanjing, China
{xf,xxm}@nju.edu.cn

Abstract

Conformal prediction (CP) provides a comprehensive framework to produce statistically rigorous uncertainty sets for black-box machine learning models. To further improve the efficiency, conformal correction extends standard CP by wrapping the base model with an extra module. In this work, we empirically and theoretically identify a trade-off in conformal prediction: its efficiency is at odds with the entropy of model prediction. We further observe that existing conformal correction methods are essentially trading entropy for efficiency. Based on the above observations, we then propose an entropy-constrained conformal correction method, better exploring the Pareto frontier between efficiency and entropy. Extensive experimental results on both computer vision and graph datasets demonstrate the efficacy of the proposed method. For instance, it can significantly improve the efficiency of state-of-the-art CP methods by up to 34.4%, given an entropy threshold. An implementation of our conformal correction approach can be accessed at the following link: <https://github.com/Xusr1123/Conformal-Correction>.

CCS Concepts

• Computing methodologies → Machine learning.

Keywords

Conformal prediction, conformal correction, efficiency-entropy trade-off, pareto optimality

ACM Reference Format:

Senrong Xu, Tianyu Wang, Zenan Li, Yuan Yao, Taolue Chen, Feng Xu, and Xiaoxing Ma. 2026. The Entropic Overhead of Conformal Correction. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26)*, August 09–13, 2026, Jeju Island, Republic of Korea. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3770855.3817714>



This work is licensed under a Creative Commons Attribution 4.0 International License. *KDD '26, Jeju Island, Republic of Korea*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2259-2/2026/08
<https://doi.org/10.1145/3770855.3817714>

1 Introduction

For a decision-making process driven by machine learning (e.g., loan approval [38], fraud detection [32]), it is essential for the predictions to be accompanied by a level of confidence to quantify uncertainty [40, 48]. Conformal prediction (CP) is a promising uncertainty quantification method, providing statistically rigorous uncertainty sets for black-box machine learning models [3, 9, 41, 42]. In standard classification, for any test input x , the posterior distribution $\tilde{\pi}_y(x) = P(Y = y | X = x)$ on classes $[K] := \{1, \dots, K\}$ is calculated. Conformal prediction leverages an additional calibration step to guarantee a user-specified (marginal) coverage: by producing a prediction set $C(x) \subseteq [K]$, it guarantees the true class of x is included in $C(x)$ with a user-chosen probability, when the calibration samples are exchangeable with the test samples.

Currently, the main focus of CP is to improve the efficiency, i.e., reduce the size of the prediction set. This set essentially measures how many choices we need to consider (the first-level uncertainty). A recent progress in CP introduces extra training on the base model. For example, Bellotti [5] proposes the notion of conformal training and Stutz et al. [43] simulate the conformal prediction process during training; this approach is further extended to graph-structure data [20] by introducing an adapter, which performs an additional conformal-aware training step based on the fixed base model. We refer to this emerging class of approaches as *conformal correction*.¹ To be more concrete, given a base classifier $\tilde{M}$, we can obtain a conformal adapter $\hat{M}$, which takes $\tilde{\pi}(x)$ from $\tilde{M}$ as input and outputs $\hat{\pi}(x)$, together with $C(x)$ typically of a smaller size.

In this work, we argue that, in addition to efficiency, the probability distribution within the prediction set also matters. For example, the entropy of the prediction set measures how certain we are about $C(x)$ or the top-choices in $C(x)$ (the second-level uncertainty). Consider the following example where a patient is presented with a fever symptom, and two prediction sets are obtained from two machine learning models: $A: \{\text{Leukemia}, \text{Rheumatism}, \text{Influenza}\}$

¹In this paper, we adopt the latter setting as the conformal adapter only needs the output distribution of the base model (as input), which is more akin to traditional CP (in the sense that it is decoupled from the base model), and thus has broader applications in practice.

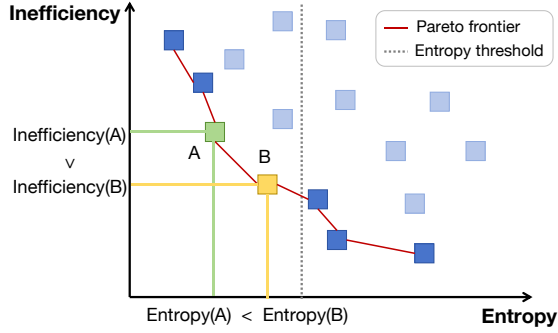


Figure 1: Pareto frontier between inefficiency and entropy. For both of them, the lower, the better. Given an entropy threshold, situation B is better than situation A.

with predictive probabilities 0.3, 0.3, and 0.3, and B : {*Leukemia*, *Rheumatism*, *Influenza*} with predictive probabilities 0.1, 0.1, and 0.7. These two sets share the same size (efficiency) but convey very different second-level uncertainty. For set A , doctors may need to investigate all three diseases due to the equal confidence, while set B naturally prioritizes *Influenza* for initial examination.

The above example demonstrates that, beyond efficiency, the prediction set should be as informative as possible. However, we observe that existing conformal correction methods typically trade the entropy of the prediction $\hat{\pi}(x)$ for better efficiency. That is, while the average size of $C(x)$ may be smaller, the corresponding prediction entropy increases, indicating that the prediction becomes more *uncertain*.

To be more specific, our experiments on CIFAR100 show that: (1) when the accuracy converges (e.g., after the 25-th epoch in Fig. 2(a)), the average size of prediction sets increases as the prediction entropy decreases; (2) a conformal correction method based on temperature scaling indeed trades entropy for better efficiency (cf. Fig. 3(a)). We further confirm the finding by showing that, for APS [36], the expected size of CP sets can be upper-bounded by the negative entropy (plus some positive constant; cf. Theorem 3 for a precise account). This gives theoretical evidence that the efficiency of CP sets produced by APS is at odds with the prediction entropy.

The trade-off between efficiency and entropy entails a Pareto perspective on CP, where different Pareto optima form a Pareto frontier as shown in Fig. 1. Conformal correction can thus be viewed as a traversal of the Pareto frontier. Technically, one can reduce the inefficiency significantly, but at the cost of an increased entropy, rendering such a reduction less meaningful. Instead, we argue that seeking for a better Pareto frontier is more crucial for conformal correction than simply adapting the trade-off. To this end, we propose a new method, i.e., **entropy-constrained conformal correction** (EC^3) to ameliorate the trade-off by controlling the entropy of conformal adapters via focal loss [31] and temperature scaling [14].

We conduct extensive experiments on computer vision (CV) and graph datasets to evaluate the effectiveness of EC^3 . The results show that our method can outperform the competitors by up to 34.4% in terms of efficiency given an entropy threshold; the qualitative analysis also indicates that our method can locate better Pareto

optimality with strong control over the model entropy. Furthermore, when EC^3 is adapted to provide (stronger) conditional coverage, it can significantly improve, for instance, the class coverage from 0.77 to 0.83 (for the CV dataset) and from 0.74 to 0.85 (for the graph dataset), respectively.

To summarize, the main contributions of the paper are: (1) we identify a trade-off between efficiency and prediction entropy in CP, which has not been fully investigated before, both empirically and theoretically; (2) we propose a new conformal correction method based on entropy control to improve the efficiency of CP, the effectiveness of which is confirmed by extensive experiments.

2 Preliminary

Notations. We focus on multiclass classification (with K classes). Assume $D = \{(X_i, Y_i)\}_{i=1}^{n+1}$ of i.i.d. (or simply exchangeable) observations sampled from an (unknown) testing distribution P_{XY} . We denote the (oracle) conditional distribution $P_{Y|X}$ by $\pi_y(x) = P(Y = y | X = x)$. Furthermore, a black-box base classifier undergoes adapting to prescribe prediction $\hat{\pi}_y(x)$. The prediction entropy is defined as $H(\hat{\pi}(x)) := -\sum_{k=1}^K \hat{\pi}_k(x) \log \hat{\pi}_k(x)$.

CP Framework. Given $D_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ as the calibration set and a user-defined miscoverage rate $\alpha \in (0, 1)$, CP typically proceeds in the following three steps:

(1) *Non-conformity score definition.* CP first heuristically defines a non-conformity score function $V(x, y)$, which indicates how the class y conforms to the predictive result $\hat{\pi}(x) = [\hat{\pi}_1(x), \dots, \hat{\pi}_K(x)]$. For example, the non-conformity score $V(x, y)$ can be defined as the sum of the probabilities of all K classes in $\hat{\pi}(x)$ except class y .

(2) *Uncertainty calibration.* CP then evaluates the non-conformity score for each data point $(X_i, Y_i) \in D_{\text{cal}}$, resulting in the non-conformity score set $\{V(X_i, Y_i)\}_{i=1}^n$. Subsequently, it sets a threshold $\hat{\eta}$ as its $(1 - \alpha)(1 + 1/n)$ -quantile.

(3) *Prediction set construction.* For a new sample X_{n+1} , conformal prediction computes the corresponding prediction set by $C(X_{n+1}) = \{y \in [K] \mid V(X_{n+1}, y) \leq \hat{\eta}\}$.

Traditional CP is model-agnostic, as it only requires prediction from the base model. Moreover, various non-conformity scores can be used to instantiate the framework [2, 36]. For instance, the Adaptive Prediction Set (APS) [36] first sorts the predicted results in descending order, i.e., $\hat{\pi}_{(1)}(x) \geq \hat{\pi}_{(2)}(x) \geq \dots \geq \hat{\pi}_{(K)}(x)$. The non-conformity score is then defined by the cumulative probabilities from the most likely class to the observed class y in the calibration step, i.e., $V(x, y) = \sum_{i=1}^y \hat{\pi}_{(i)}(x)$.

CP Evaluation. Traditional methods evaluate the quality of prediction sets in terms of two dimensions, i.e., efficiency and coverage. (In)efficiency captures the average size of the prediction sets, i.e., $\mathbb{E}(|C(X_{n+1})|)$; for inefficiency, the smaller, the better. For coverage, CP ensures the marginal coverage, viz., the true class Y_{n+1} is in $C(X_{n+1})$ with a probability of at least $1 - \alpha$, i.e.,

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

In certain cases, we also expect the conditional coverage to exceed $1 - \alpha$ for each x , i.e., $P(Y_{n+1} \in C(x) \mid X_{n+1} = x) \geq 1 - \alpha$. In this work, we focus on adaptive conformal prediction approaches [2, 19, 36], which uniquely account for conditional coverage in classification [40].

3 Efficiency and Entropy Trade-off

In this section, we explore the trade-off between efficiency and entropy in conformal prediction, both empirically and theoretically.

Existing work [20, 43] has demonstrated that conformal correction can significantly reduce the sizes of CP sets. To this end, various sorting-smooth techniques [6, 34] are employed to encode the size of CP sets in the loss function. The general optimization objective for conformal correction can be formulated as

$$\min \mathcal{L}_f := \mathcal{L}_{\text{class}} + \beta \cdot \mathcal{L}_{\text{ineff}}, \quad (1)$$

where $\mathcal{L}_{\text{class}}$ is the standard cross-entropy loss function for classification, $\mathcal{L}_{\text{ineff}}$ is the inefficiency loss function aiming to reduce the prediction set size, and β is a hyperparameter.

Empirical Observations. Eq. (1) integrates inefficiency and classification losses by a weighted sum. However, a critical issue is whether the two losses, $\mathcal{L}_{\text{class}}$ and $\mathcal{L}_{\text{ineff}}$, can achieve their minima simultaneously. To explore this problem, we train a classifier only using $\mathcal{L}_{\text{class}}$ and plot the test curves of accuracy, efficiency, and entropy in Fig. 2(a) and 2(b). An interesting observation is that the accuracy and efficiency increase together while the entropy decreases during the initial training stage. When the accuracy converges, as the entropy of the models drops (meaning the model becomes more certain), the efficiency decreases (i.e., the average set size gets bigger). This phenomenon suggests a trade-off between efficiency and entropy may exist when the model is fully trained.

Theoretical Explanations. We confirm the empirical observation by analyzing the impact of prediction entropy on CP efficiency in the context of APS, which is the most classic adaptive conformal prediction method. We first define the average non-conformity score

$$\bar{V}(\hat{\pi}(x)) = \frac{1}{K} \sum_{i=1}^K V(\hat{\pi}(x), i),$$

where $V(\hat{\pi}(x), i)$ refers to the APS non-conformity score of the i -th class. The average non-conformity score indicates the overall performance of the predictive result $\hat{\pi}(x)$ conforming to the class set $[K]$. Then, we establish a relationship between the average non-conformity score $\bar{V}(\hat{\pi}(x))$ and the prediction entropy $H(\hat{\pi}(x))$.

PROPOSITION 1. *For a given sample point x and the corresponding predictive distribution $\hat{\pi}(x)$, the average non-conformity score is upper-bounded by the prediction entropy. Namely,*

$$\bar{V}(\hat{\pi}(x)) \leq \min(C_K + 1 - H(\hat{\pi}(x)), 1 + H(\hat{\pi}(x))),$$

with constant $C_K := \log(\sum_{k=1}^K \exp(-\frac{k-1}{K}))$.

The proof is given in Appendix A.1. Additionally, we use binary classification ($K = 2$) to illustrate this proposition by plotting the curves of $\bar{V}(\hat{\pi}(x))$ and its two upper bounds in Fig. 2(c). In this case, the entropy $H(\hat{\pi}(x))$ is determined by $\hat{\pi}_{(1)}(x)$ as $\hat{\pi}_{(2)}(x) = 1 - \hat{\pi}_{(1)}(x)$ and $\hat{\pi}_{(1)}(x) \geq \hat{\pi}_{(2)}(x)$. It can be observed that, with the decrease of entropy $H(\hat{\pi}(x))$, the tighter bound shifts from $C_K + 1 - H(\hat{\pi}(x))$ (the orange curve) to $1 + H(\hat{\pi}(x))$ (the blue curve).

Proposition 1 illustrates the relationship between the average non-conformity score and the entropy of model prediction. Moreover, the derived analysis is consistent with our empirical observation: when the entropy $H(\hat{\pi}(x))$ is sufficiently small, the average non-conformity is bounded by $1 + H(\hat{\pi}(x))$, allowing them to increase simultaneously. However, as $H(\hat{\pi}(x))$ becomes larger, $C_K + 1 - H(\hat{\pi}(x))$ becomes the tighter bound, which would prohibit the average non-conformity score from growing, leading to a trade-off in between.

Next, we extend the upper bound to the $(1 - \alpha)$ -quantile $\hat{\eta}$.

PROPOSITION 2. *Given a sample subset $C_{\hat{\eta}} := \{(X, Y) \mid V(X, Y) \geq \hat{\eta}\}$ in D , the $(1 - \alpha)$ -quantile $\hat{\eta}$ is upper bounded*

$$\hat{\eta} \leq \mathbb{E}[\bar{V}(X) \mid C_{\hat{\eta}}] + C_{(\pi, K)} + \tau,$$

with the probability at least $1 - \exp(-\frac{2\alpha\tau^2n}{(1-\hat{\eta})^2})$, where n is the size of the calibration set, τ is a positive constant, and the constant $C_{(\pi, K)} := \mathbb{E}[\sqrt{2(\log(K) - H(\pi(X)))} \mid C_{\hat{\eta}}]$.

The proof is given in Appendix A.2. Note that as n increases, $1 - \exp(-\frac{2\alpha\tau^2n}{(1-\hat{\eta})^2})$ tends to 1. Moreover, as the second term and third term of the upper bound are both constants, Proposition 2 effectively bounds $\hat{\eta}$ by the average non-conformity score $\bar{V}(X)$ from a sample subset $C_{\hat{\eta}}$.

By combining Proposition 1 and Proposition 2, we can finally establish the trade-off between the expected size of conformal prediction sets $\mathbb{E}[|C(X)|]$ and the entropy of model prediction.

THEOREM 3. *Let $\mu = \mathbb{P}(H(\hat{\pi}(X)) \geq \frac{1}{2}C_K \mid C_{\hat{\eta}})$. We have that*

$$\mathbb{E}[|C(X)|] \leq \underbrace{K(1 - \alpha)(1 - 2\mu)\mathbb{E}[H(\hat{\pi}(x)) \mid C_{\hat{\eta}}]}_{(\star)} + O(K).$$

The proof is provided in Appendix A.3.

REMARK 4. By Theorem 3, we can see that, when $\mu \geq \frac{1}{2}$ (i.e., the entropy of a majority of x in the subset $C_{\hat{\eta}}$ is greater than $\frac{1}{2}C_K$), $1 - 2\mu < 0$ holds and so does term $(\star)$, which entails that $\mathbb{E}[|C(X)|]$ is at odds with the expected entropy $\mathbb{E}[H(\hat{\pi}(x)) \mid C_{\hat{\eta}}]$. Otherwise, the term $(\star)$ is positive, allowing a potential synergy between efficiency and entropy.

Intuitively, for APS, the trade-off between efficiency and entropy will be present when the entropy is sufficiently large (roughly, greater than $\frac{1}{2}C_K$). It is not hard to see that $\frac{1}{2}C_K$ is a monotonically increasing function in K . This suggests that, when K is relatively small, the interval $[0, \frac{1}{2}C_K]$ is narrow, and thus the trade-off will be largely dominating.

Note that although our theoretical analysis is conducted using APS as a case study, it can be extended to other adaptive conformal prediction methods due to similar non-conformity score designs (e.g., RAPS [2] is a regularized version of APS). This theoretical expectation is further supported by the empirical results presented in Appendix B.5.

4 Conformal Correction Methods

In this section, we present a new method, EC³, for conformal correction. In general, EC³ is based on Section 3, searching for better Pareto optima via controlling the entropy of model predictions. Then, we directly utilize temperature scaling to explore the Pareto

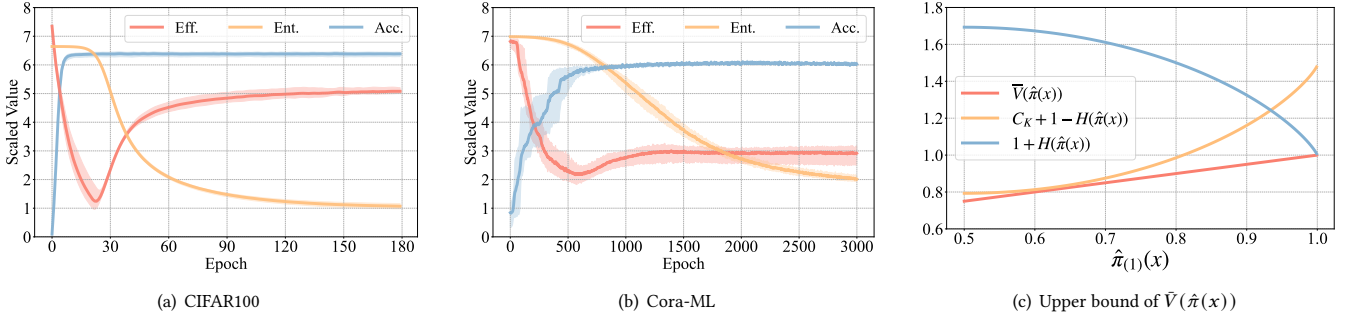


Figure 2: Fig. (a) and (b) are the results of training only with $\mathcal{L}_{\text{class}}$ on CIFAR100 and Cora-ML, respectively; Fig.(c) is an illustration of Proposition 1 when $K = 2$. (a) and (b) depict the efficiency and entropy on the test set, and there is a trade-off between them when the accuracy reaches the top. (c) shows that the tight upper-bound of $\bar{V}(\hat{\pi}(x))$ consists of two pieces.

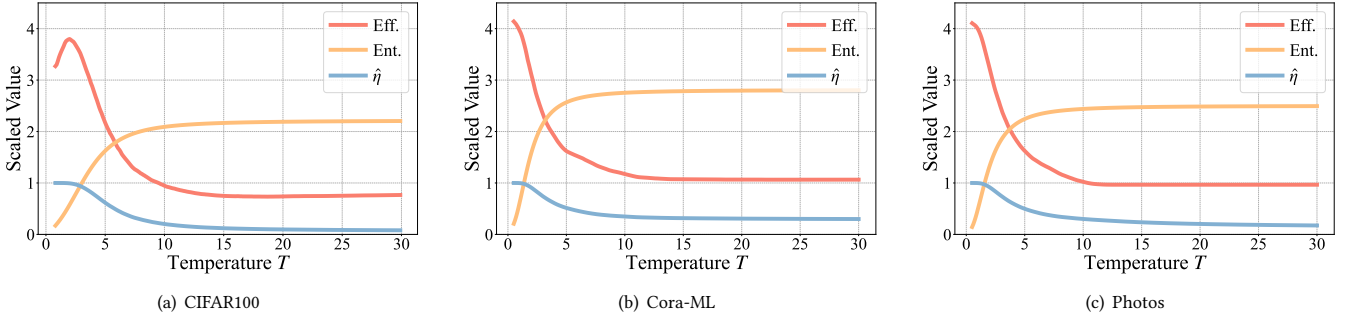


Figure 3: Efficiency and entropy curves of APS (cf. Section 2) on the test set after temperature scaling w.r.t. T when $\alpha = 0.1$. The efficiency of APS is at odds with the entropy of model prediction in most cases.

frontier, and extend EC³ to improve the user-specified conditional coverage.

4.1 Entropy-Constrained Conformal Correction

As discussed in Section 3, there is a fundamental trade-off between CP efficiency and prediction entropy. A natural way to search for the Pareto frontier is to introduce a positive entropy term into Eq. (1), and balance it with $\mathcal{L}_{\text{ineff}}$. However, in our case, the cross-entropy loss $\mathcal{L}_{\text{class}}$ already implicitly enforces entropy reduction.² Furthermore, we observe that directly optimizing the original training loss often results in a rapid decline in efficiency, leading to low-entropy solutions with poor efficiency (e.g., after the 30-th epoch in Fig. 2(a)). Therefore, we add a negative entropy term into Eq. (1) to counter the rapid decline in efficiency, which enables a more fine-grained control of entropy during conformal correction, viz.,

$$\min \mathcal{L}_f = \mathcal{L}_{\text{class}} + \beta \cdot \mathcal{L}_{\text{ineff}} - \gamma \cdot H(\hat{\pi}(x)), \quad (2)$$

where $\gamma \geq 0$ is a hyperparameter controlling the weight of the entropy term.

²Minimizing cross-entropy loss essentially encourages the predicted distribution to approximate a sharp distribution (e.g., the one-hot label vector), rendering a trained model with low entropy, which is also known as the over-confident problem [14].

There are three competing optimization objectives in Eq. (2), making the optimization challenging. Fortunately, the following inequality for $\mathcal{L}_{\text{class}}$ and $H(\hat{\pi}(x))$ holds [31]:

$$\mathcal{L}_{\text{focal}} \geq \text{KL}(\pi(x) \parallel \hat{\pi}(x)) - \gamma \cdot H(\hat{\pi}(x)),$$

where $\mathcal{L}_{\text{focal}} = -\sum_{k=1}^K (1 - \hat{\pi}_k(x))^\gamma \pi_k(x) \log \hat{\pi}_k(x)$ is the form of focal loss [31] and $\text{KL}(\pi(x) \parallel \hat{\pi}(x))$ is the KL-divergence between the ground-truth distribution $\pi(x)$ and the model prediction $\hat{\pi}(x)$. Since $\text{KL}(\pi(x) \parallel \hat{\pi}(x))$ can be reduced to the cross-entropy, this inequality allows us to directly optimize the upper bound for $\mathcal{L}_{\text{class}} - \gamma \cdot H(\hat{\pi}(x))$, i.e., $\mathcal{L}_{\text{focal}}$. Thus, we rewrite the objective in Eq. (2) as

$$\min \mathcal{L}_f = \mathcal{L}_{\text{focal}} + \beta \cdot \mathcal{L}_{\text{ineff}}. \quad (3)$$

Compared with Eq. (1), the above objective enjoys two advantages. From the view of multi-objective optimization, the proposed minimization objective can flexibly adjust the trade-off between the CP efficiency and the entropy of prediction by controlling γ . Specifically, if we prefer CP efficiency over entropy, γ should be augmented to increase the efficiency at the cost of entropy. Otherwise, we should lower the value of γ . When $\gamma = 0$, Eq. (3) degrades into Eq. (1).

Additionally, in contrast to directly penalizing the entropy, focal loss presents a more flexible approach to balance classification loss and entropy regularization through the coefficient $(1 - \hat{\pi}_k(x))^\gamma$.

Specifically, the focal loss can effectively control the strength of the classification loss based on the sharpness of the model prediction $\hat{\pi}(x)$. This adaptation facilitates locating better solutions in the trade-off between entropy and efficiency.

Pareto Frontier Exploration via Temperature Scaling. When a better Pareto optimum is achieved by Eq. (3), we can next traverse the Pareto frontier from this Pareto optimum via temperature scaling—a common trick used in model calibration [14]—to flexibly regulate the entropy. Specifically, it rephrases the softmax function as follows:

$$\hat{\pi}_i(x) = \frac{\exp(\hat{\pi}_i(x)/T)}{\sum_{j=1}^K \exp(\hat{\pi}_j(x)/T)},$$

where $T > 0$ is the temperature controlling the uncertainty of models. With T increasing, the prediction entropy $H(\hat{\pi}(x))$ becomes larger. Practically, temperature scaling is based on the grid search, which is simple and convenient to use.

Note that one can directly use temperature scaling, which can be viewed as a straightforward conformal correction method with a single parameter T , to adapt the trade-off between efficiency and entropy. Fig. 3 plots the results of APS over T on CIFAR100, Cora-ML, and Photos datasets.

Fig. 3(a) depicts the result of the CIFAR100 dataset. We observe an initial concurrent increase in both efficiency and entropy as T rises. However, as T continues to increase, a trade-off emerges between these two metrics: efficiency decreases while entropy increases. This observation aligns with Theorem 3. That is, when entropy is relatively low (corresponding to lower values of T), efficiency is upper-bounded by the entropy; conversely, as entropy increases (with higher values of T), efficiency is upper-bounded by the *negative* entropy (plus some positive constant).

Fig. 3(b) and 3(c) depict the result of the Cora-ML and Photos datasets. In contrast, only the trade-off between the efficiency and entropy can be observed. This is because the number of classes therein is significantly smaller than that in the CIFAR100 dataset, which is consistent with Remark 4. In addition to the results on CIFAR100, Cora-ML, and Photos datasets, we relegate the rest to Appendix B.4.

Nevertheless, directly using temperature scaling without the objective in Eq. (3) will lead to a suboptimal Pareto frontier (cf. Fig. 8 in Appendix B.7).

4.2 Extensions to Conditional Coverage

Recall from Section 2 that conditional coverage is stronger than marginal coverage. The flexibility of the EC³ approach allows us to adjust user-specified conditional coverage adaptively. Take the class conditional coverage (i.e., the coverage of the sample subsets with the same true class [53]) as an example. We define the following class conditional coverage loss function for each class k ,

$$\mathcal{L}_k := -\frac{1}{|D_{\text{cal}}^k|} \sum_{(x,y) \in D_{\text{cal}}^k} \mathbb{I}(y \in C(x)),$$

where $D_{\text{cal}}^k := \{(x,y) \in D_{\text{cal}} \mid y = k\}$ and $\mathbb{I}(\cdot)$ is the indicator function. We then obtain a new minimization objective to improve

the class conditional coverage during conformal correction,

$$\min \mathcal{L}_f = \mathcal{L}_{\text{focal}} + \beta \cdot \mathcal{L}_{\text{ineff}} - \frac{1}{K} \sum_{k=1}^K \mathcal{L}_k.$$

We refer to this method as “EC³ (Cond)” which will be evaluated in Section 5.2.3.

5 Experiments

In this section, we conduct the extensive experiments to showcase the effectiveness of EC³.

5.1 Experimental Setup

Datasets. We conduct main experiments on five datasets, including CIFAR10, CIFAR100 [24], Cora-ML [30], CS [39], and Photos [29]. The former two datasets are from the computer vision (CV) domain and the latter three are graph-structure datasets. For CV datasets, CIFAR10 and CIFAR100 consist of 60,000 32×32 colour images in 10 and 100 classes, respectively. For graph datasets, there are 2,995/18,333/7,650 nodes with 2,879/6,805/745 features and 16,346/163,788/238,162 edges in Cora-ML, CS, and Photos. The number of classes in these graph datasets is 7, 15 and 8, respectively. Following Huang et al. [20], we randomly split each dataset into the training set D_{train} , validation set D_{valid} , calibration set D_{cal} , and testing set D_{test} with the ratio 2:1:4:3. We perform 100 random splits of calibration/testing sets, and report the average results and standard deviations to suppress randomness.

Base and Adapter Models. For CV tasks, we apply ResNet [15], PreResNet [16], and DenseNet [18] as the base models, and use MLP as the conformal adapter model. For graph tasks, we use GCN [22], GAT [46], and SGC [50] as the base models and use GAT as the conformal adapter model following Huang et al. [20]. For a fair comparison, we train each of the base models five times and report the average results to avoid fluctuations from randomness.

Baselines. We select the state-of-the-art methods, i.e., ConfTr [43] and CF-GNN [20] from CV and graph domains, respectively. Note that while CF-GNN strictly follows the conformal correction framework, ConfTr was initially proposed to retrain the base model. In our experiments, we adapt ConfTr to the conformal correction framework (i.e., applying its optimization objective as an adapter after the base model is trained); further performance improvements are observed.

Evaluation Metrics. For efficiency, we use the standard *average size* of CP sets as the metric. For marginal coverage, we can directly compute its value. For conditional coverage, we consider WSC and SSCV metrics [2, 36]. All these metrics are widely adopted by existing work [20].

Implementations. To construct the CP sets, we consider both APS [36] and RAPS [2]. We report the APS results in the main body of the paper, and the results on RAPS are included in Appendix B.5. We set the threshold of the prediction entropy to be $(1 - \epsilon) \exp(\log K)$, and use $\epsilon = 1/4$ by default. We also set the mis-coverage rate $\alpha = 0.1$, hyperparameter $\beta = 0.1$ and $\gamma = 4$ by default. All the experiments are carried out on NVIDIA GeForce RTX 3090. More implementation details, such as the hyperparameters of base models and conformal adapters, are presented in Appendix B.1.

Table 1: Efficiency results of conformal correction methods on CV and graph datasets when $\alpha = 0.1$. The results are the average of five runs of pre-trained model, each with 100 runs of conformal splits on datasets. Baselines are ConfTr and CF-GNN on CV and graph datasets, respectively. The best results are in shadow. The proposed EC³ achieves the best efficiency performance and maintains the marginal coverage.

DATASET	MODEL	CP		BASELINE		EC ³	
		COVERAGE	EFFICIENCY	COVERAGE	EFFICIENCY	COVERAGE	EFFICIENCY
CIFAR10	RESNET56	0.90 $\pm$.00	5.41 $\pm$.11	0.90 $\pm$.00	1.31 $\pm$.15	0.90 $\pm$.00	1.23 $\pm$.06
	PRERESNET110	0.90 $\pm$.00	5.54 $\pm$.07	0.90 $\pm$.00	1.25 $\pm$.16	0.90 $\pm$.00	1.18 $\pm$.04
	DENSENET100	0.90 $\pm$.00	5.52 $\pm$.03	0.90 $\pm$.00	1.30 $\pm$.15	0.90 $\pm$.00	1.09 $\pm$.04
CIFAR100	RESNET56	0.90 $\pm$.00	23.06 $\pm$.66	0.90 $\pm$.00	19.83 $\pm$ 1.94	0.90 $\pm$.00	18.05 $\pm$ 2.71
	PRERESNET110	0.90 $\pm$.00	25.93 $\pm$.37	0.90 $\pm$.00	17.62 $\pm$ 1.98	0.90 $\pm$.00	15.27 $\pm$ 1.27
	DENSENET100	0.90 $\pm$.00	30.00 $\pm$ 2.44	0.90 $\pm$.00	13.29 $\pm$ 1.61	0.90 $\pm$.00	10.87 $\pm$ 1.42
CORA-ML	GCN	0.90 $\pm$.00	4.00 $\pm$.19	0.90 $\pm$.00	1.85 $\pm$.26	0.90 $\pm$.00	1.50 $\pm$.13
	GAT	0.90 $\pm$.00	3.92 $\pm$.13	0.90 $\pm$.00	1.94 $\pm$.39	0.90 $\pm$.00	1.68 $\pm$.13
	SGC	0.90 $\pm$.00	4.01 $\pm$.13	0.90 $\pm$.00	1.81 $\pm$.26	0.90 $\pm$.00	1.58 $\pm$.09
CS	GCN	0.90 $\pm$.00	8.37 $\pm$.22	0.90 $\pm$.00	4.45 $\pm$.38	0.90 $\pm$.00	3.13 $\pm$.21
	GAT	0.90 $\pm$.00	6.92 $\pm$.36	0.90 $\pm$.00	4.47 $\pm$.50	0.90 $\pm$.00	3.03 $\pm$.36
	SGC	0.90 $\pm$.00	8.37 $\pm$.26	0.90 $\pm$.00	4.36 $\pm$.33	0.90 $\pm$.00	3.13 $\pm$.25
PHOTOS	GCN	0.90 $\pm$.00	4.00 $\pm$.14	0.90 $\pm$.00	2.07 $\pm$.38	0.90 $\pm$.00	1.54 $\pm$.14
	GAT	0.90 $\pm$.00	2.36 $\pm$.24	0.90 $\pm$.00	2.69 $\pm$.30	0.90 $\pm$.00	2.13 $\pm$.08
	SGC	0.90 $\pm$.00	4.00 $\pm$.06	0.90 $\pm$.00	2.17 $\pm$.28	0.90 $\pm$.00	1.53 $\pm$.06

Table 2: Efficiency results of conformal correction methods on CV and graph datasets when $\alpha = 0.2$. The results are the average of five runs of pre-trained model, each with 100 runs of conformal splits on datasets. Baselines are ConfTr and CF-GNN on CV and graph datasets, respectively. The best results are in shadow. The proposed EC³ still generally outperforms the competitors.

DATASET	MODEL	CP		BASELINE		EC ³	
		COVERAGE	EFFICIENCY	COVERAGE	EFFICIENCY	COVERAGE	EFFICIENCY
CIFAR10	RESNET56	0.80 $\pm$.00	4.01 $\pm$.04	0.80 $\pm$.00	1.06 $\pm$.01	0.80 $\pm$.00	1.04 $\pm$.01
	PRERESNET110	0.80 $\pm$.00	4.22 $\pm$.09	0.80 $\pm$.00	1.01 $\pm$.02	0.80 $\pm$.00	1.02 $\pm$.01
	DENSENET100	0.80 $\pm$.00	4.20 $\pm$.02	0.80 $\pm$.00	0.99 $\pm$.00	0.80 $\pm$.00	0.98 $\pm$.00
CIFAR100	RESNET56	0.80 $\pm$.00	13.38 $\pm$.36	0.80 $\pm$.00	6.17 $\pm$.75	0.80 $\pm$.00	2.93 $\pm$.09
	PRERESNET110	0.80 $\pm$.00	14.98 $\pm$.24	0.80 $\pm$.00	4.72 $\pm$.10	0.80 $\pm$.00	4.25 $\pm$.17
	DENSENET100	0.80 $\pm$.00	18.13 $\pm$ 1.28	0.80 $\pm$.00	3.20 $\pm$.09	0.80 $\pm$.00	2.76 $\pm$.16
CORA-ML	GCN	0.80 $\pm$.00	3.10 $\pm$.23	0.80 $\pm$.00	1.41 $\pm$.10	0.80 $\pm$.00	1.03 $\pm$.03
	GAT	0.80 $\pm$.00	3.07 $\pm$.11	0.80 $\pm$.00	1.47 $\pm$.15	0.80 $\pm$.00	1.10 $\pm$.06
	SGC	0.80 $\pm$.00	3.11 $\pm$.15	0.80 $\pm$.00	1.42 $\pm$.12	0.80 $\pm$.00	1.03 $\pm$.04
CS	GCN	0.80 $\pm$.00	6.12 $\pm$.23	0.80 $\pm$.00	2.66 $\pm$.12	0.80 $\pm$.00	1.73 $\pm$.14
	GAT	0.80 $\pm$.00	5.18 $\pm$.19	0.80 $\pm$.00	2.53 $\pm$.23	0.80 $\pm$.00	1.74 $\pm$.17
	SGC	0.80 $\pm$.00	6.15 $\pm$.25	0.80 $\pm$.00	2.41 $\pm$.29	0.80 $\pm$.00	1.84 $\pm$.15
PHOTOS	GCN	0.80 $\pm$.00	3.11 $\pm$.14	0.80 $\pm$.00	1.58 $\pm$.31	0.80 $\pm$.00	1.24 $\pm$.07
	GAT	0.80 $\pm$.00	1.83 $\pm$.19	0.80 $\pm$.00	1.74 $\pm$.47	0.80 $\pm$.00	1.61 $\pm$.12
	SGC	0.80 $\pm$.00	3.13 $\pm$.05	0.80 $\pm$.00	1.44 $\pm$.28	0.80 $\pm$.00	1.35 $\pm$.09

5.2 Experimental Results

5.2.1 Efficiency Comparison. Give an entropy threshold as mentioned in Section 5.1, we first compare the best efficiency and the corresponding marginal coverage of training-based conformal correction methods when $\alpha = 0.1$, the results of which are given in

Table 1. We also report the entropy results (Appendix B.2) and the accuracy results (Appendix B.3).

On the CV datasets, the proposed EC³ performs better than the baseline ConfTr in terms of efficiency over all three pre-trained models on both datasets, while it keeps the marginal coverage at the

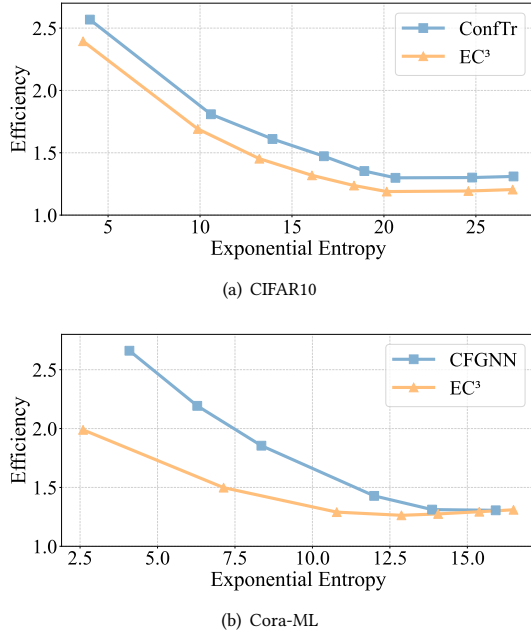


Figure 4: Pareto optima of different conformal correction methods. Compared with baselines, the proposed EC^3 obtains the best Pareto frontier via achieving a better balance between efficiency and entropy on both CIFAR10 and Cora-ML.

same time. For example, EC^3 is 18.2% more efficient than ConfTr for the pre-trained model DenseNet100 on CIFAR100. Similarly, EC^3 achieves significant efficiency improvements from 12.7% to 34.4% on the graph datasets.

To study the influence of α on conformal correction, we report the results of efficiency and marginal coverage when $\alpha = 0.2$ on CV and graph datasets in Table 2. Generally speaking, the proposed EC^3 still outperforms other baselines on both CV and graph datasets. Specifically, EC^3 significantly improves the efficiency of APS by up to 52.5% on all datasets except CIFAR10. On CIFAR10, since the accuracy of base models is relatively high (e.g., 0.89 in DenseNet100), conformal correction methods easily ameliorate the efficiency of both our method EC^3 and baseline ConfTr to achieve good efficiency.

The significant efficiency improvement can be attributed to the theoretical analysis of the efficiency-entropy tradeoff which EC^3 is built on. In particular, the explicit modeling of entropy enables EC^3 to achieve a better balance between efficiency and entropy within an acceptable entropy range.

5.2.2 Pareto Frontiers. We explore the Pareto frontier of efficiency and entropy for all conformal correction methods with temperature scaling. We can control entropy by adjusting the temperature T and select sufficient values of T to cover the entropy range. The results of the Pareto frontier are shown in Fig. 4.

It can be observed that EC^3 (the orange curve) achieves a better Pareto frontier (i.e., lower efficiency given the same entropy) than

Table 3: Conditional coverage results on CIFAR10 and Cora-ML before/after conformal correction. WSC closer to 0.9 and smaller SSCV indicate better performance. Conformal correction does not worsen the conditional coverage.

DATASET	METHOD	WSC	SSCV
CIFAR10	CP	0.88 $\pm$.01	0.45 $\pm$.01
	CONFTr	0.89 $\pm$.00	0.22 $\pm$.18
	TS	0.88 $\pm$.01	0.19 $\pm$.02
	EC^3	0.89 $\pm$.00	0.15 $\pm$.05
CORA-ML	CP	0.90 $\pm$.00	0.10 $\pm$.00
	CF-GNN	0.90 $\pm$.00	0.09 $\pm$.05
	TS	0.90 $\pm$.00	0.09 $\pm$.01
	EC^3	0.90 $\pm$.00	0.09 $\pm$.05

Table 4: Distances between the class coverage below 0.9 and the target coverage 0.9 with L1-norm and L2-norm in Fig. 5. EC^3 (Cond) improves the class conditional coverage successfully by up to 73%.

MODEL	CIFAR10		CORA-ML	
	L1	L2	L1	L2
EC^3	0.25	0.15	0.16	0.16
EC^3 (COND)	0.21	0.11	0.04	0.04
IMP.	18%	26%	73%	73%

other conformal correction methods (the blue curve) in both Fig. 4(a) and Fig. 4(b). The contrast of Pareto frontiers further confirms the positive impact of the entropy control on the efficiency-entropy trade-off, in terms of seeking better Pareto optima.

5.2.3 Conditional Coverage. One advantage of our method is its flexibility in accommodating conditional coverage. Here, we present the conditional coverage results before and after conformal correction by the metrics WSC and SSCV in Table 3. It is clear that conformal correction can tacitly improve conditional coverage to some extent.

Next, we evaluate the effectiveness of the conditional conformal correction presented in Section 4.2. In Fig. 5, we plot the histogram of the coverage of different classes, without and with the conditional conformal correction, i.e., EC^3 and EC^3 (Cond). As mentioned in Section 2, class conditional coverage requires that the coverage of each class is greater than $1 - \alpha$. Hence, we only need to examine the classes whose coverage is below $1 - \alpha$. In Fig. 5, we can observe that the conditional conformal correction improves most of the class coverages below 0.9. In particular, the lowest class coverages are increased from 0.77 to 0.83 and from 0.74 to 0.85, respectively. Moreover, we compute the distance between the class coverage below 0.9 and the target coverage of 0.9 with L1-norm and L2-norm in Table 4. The results show that EC^3 (Cond) significantly reduces such distances (up to 73%).

5.2.4 Extensions to LLM. The above experiments are mainly for classification problems. Here, we evaluate our approach on the

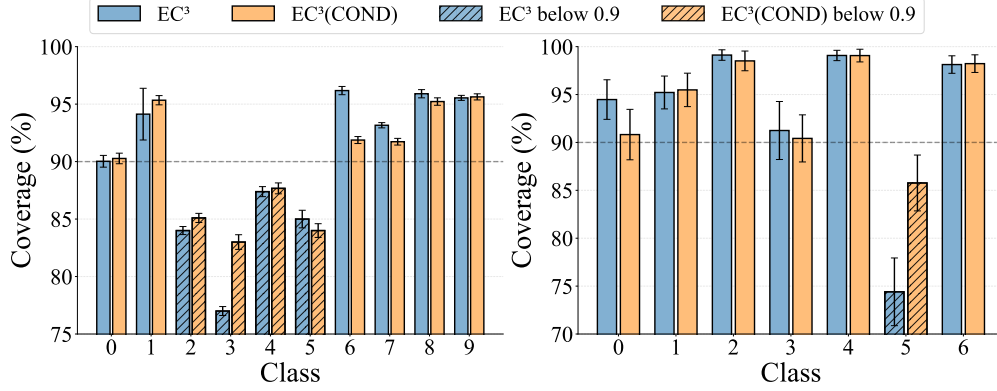


Figure 5: Class conditional coverage results on CIFAR10 (left) and Cora-ML (right). The class coverage below 0.9 is in shadow. Our method EC^3 (Cond) increases most of the class coverages below 0.9.

Table 5: Coverage, entropy and efficiency results of conformal correction methods on LLMs for the question answering task.

Metric	CP	CONFTr	EC^3
COVERAGE	$0.81 \pm .00$	$0.80 \pm .00$	$0.81 \pm .00$
EFFICIENCY	$2.92 \pm .01$	$2.48 \pm .08$	$2.32 \pm .15$
ENTROPY	$0.04 \pm .01$	$1.84 \pm .11$	$1.89 \pm .07$

question answering task using the TruthfulQA dataset [27]. The prompt we use is shown as follows.

This is a 4-choice question that you should answer: `\{ question \}`.
 Put the final results within `\textbackslash boxed \{ \{ \} \}`, e.g., `\textbackslash boxed \{ \{ A \} \}`.
 The correct answer to this question is: "."

For each question, we sample 100 Chain-of-Thought responses from Llama-2-7b-chat [45] and record the answer distribution (based on self-consistency) to perform conformal prediction. Table 5 shows the results of APS when $\alpha = 0.2$, considering the low accuracy 48.2%, and our method achieves a 6.5% improvement in efficiency over the ConfTr baseline while maintaining a comparable level of entropy. Moreover, the results after temperature scaling are provided (cf. Fig. 7(b) in Appendix B). The experimental results demonstrate that the efficiency-entropy trade-off of conformal prediction is also present in LLM generation tasks.

Additional experimental results such as parameter sensitivity analysis are reported in Appendix B.

6 Related Work

In this section, we briefly review the related work including conformal prediction and conformal correction.

Conformal Prediction. Uncertainty quantification [1] aims to provide calibrated uncertainty estimates for machine learning models, enabling reliable decision-making in critical applications. It has

been widely studied in classification and regression, where typical methods include Bayesian methods [12], confidence calibration [14], and model-agnostic frameworks that construct uncertainty intervals [35]. However, most of these methods fail to provide rigorous statistical guarantees regarding coverage, especially in non-i.i.d. settings.

CP, as a UQ method, distinguishes itself in providing guaranteed coverage, regardless of the underlying model or data distribution. It has been applied to diverse domains, including image classification [37], object detection [44], and large language models [25]. Most CP methods rely on splitting the dataset into a training set and a held-out calibration set to estimate non-conformity scores, as proposed in split conformal prediction [26]. Other extensions, such as jackknife methods [4] and cross-validation-based approaches [47], can further enhance CP's flexibility and applicability.

Some work has focused on improving the efficiency and adaptability of CP by refining non-conformity scores. Adaptive Prediction Sets (APS) [36] introduced a score function that accumulates sorted softmax probabilities; Regularized Adaptive Prediction Sets (RAPS) [2] extended APS by adding penalties to tail classes and Sorted Adaptive Prediction Sets (SAPS) [19] substituted probability values with sort orders, both resulting in more efficient prediction sets with minimal computational overhead. On a different dimension, considerable efforts have been devoted to proving (stronger) conditional coverage. These include SCP+conditional calibration [13], randomly localized conformal prediction [17], and posterior conformal prediction [54], to name a few.

CP has also been explored in graph-structured data, where node dependencies pose unique challenges. The first work [49] proposed a margin-based score for graphs. Afterwards, more methods, such as NAPS [7], DAPS [53], and a Jackknife+ variant for GCNs [21], have been developed to incorporate graph-specific properties such as homophily, further improving their performance.

Conformal Correction. These techniques aim to improve the performance of CP by modifying the model's output via extra training. Although computationally intensive, model correction represents a significant advancement in enhancing CP's performance. For instance, ConfTr [43] proposes non-conformity loss

functions designed to align scores with a uniform distribution. Similar approaches have also demonstrated effectiveness in graph-structured data [28]. CF-GNN [20], for example, introduces an additional correction model that utilizes the graph's topology to adjust the model's output.

Our approach also optimizes CP during training by modifying the model's output. Furthermore, our work investigates the relationship between the entropy and efficiency of CP. While the experimental results of ConfTS [51] and [10] corroborate part of this relationship, we provide a comprehensive theoretical framework, which is not limited to temperature scaling. By introducing a novel loss function, we achieve superior efficiency and prediction performance compared to existing correction-based optimization methods. Correia et al. [8] provide a lower bound of the expected size of the conformal prediction sets (i.e., inefficiency). In contrast, our work provides an upper bound, which is more important for improving efficiency. We mention that, in addition to efficiency, some work has also focused on conditional coverage of CP through correction [11, 23, 52] and discussed the stability of conformal training [33].

7 Conclusion

In this paper, we have demonstrated that a decrease in the inefficiency is often accompanied by an increase in the prediction entropy for conformal prediction. We have also provided a theoretical analysis explaining this phenomenon. Both lead to the conclusion that CP efficiency may be at odds with the prediction entropy in most cases. The trade-off between them hints at a Pareto optimality view of conformal correction, for which we have proposed a new method EC³. Experiments on both CV and graph datasets showcase that it outperforms the existing baselines. Currently, our theoretical analysis and methods mainly target adaptive conformal prediction, which are the mainstream conformal methods for classification [40]. In the future, we plan to explore this trade-off with other conformal prediction methods.

Acknowledgment

This work is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118), the Frontier Technologies R&D Program of Jiangsu (BF2024059), the National Natural Science Foundation of China (No. 62572229), and the "111 Center" (No. B26023).

References

- [1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion* 76 (2021), 243–297.
- [2] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. 2020. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193* (2020).
- [3] Varun Babbar, Umang Bhatt, and Adrian Weller. 2022. On the utility of prediction sets in human-ai teams. *arXiv preprint arXiv:2205.01411* (2022).
- [4] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. 2021. Predictive inference with the jackknife+. (2021).
- [5] Anthony Bellotti. 2020. Constructing normalized nonconformity measures based on maximizing predictive efficiency. In *Conformal and Probabilistic Prediction and Applications*. PMLR, 41–54.
- [6] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. 2020. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*. PMLR, 950–959.
- [7] Jase Clarkson. 2023. Distribution free prediction sets for node classification. In *International Conference on Machine Learning*. PMLR, 6268–6278.
- [8] Alvaro Correia, Fabio Valerio Massoli, Christos Louizos, and Arash Behboodi. 2024. An information theoretic perspective on conformal prediction. *Advances in Neural Information Processing Systems* 37 (2024), 101000–101041.
- [9] Jesse C Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. 2024. Conformal prediction sets improve human decision making. *arXiv preprint arXiv:2401.13744* (2024).
- [10] Lahav Dabah and Tom Tirer. 2024. On temperature scaling and conformal prediction of deep classifiers. *arXiv preprint arXiv:2402.05806* (2024).
- [11] Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. 2022. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in Neural Information Processing Systems* 35 (2022), 22380–22395.
- [12] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.
- [13] Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. 2024. Conformal Prediction With Conditional Guarantees. *arXiv:2305.12616 [stat.ME]* <https://arxiv.org/abs/2305.12616>
- [14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 630–645.
- [17] Rohan Hore and Rina Foygel Barber. 2024. Conformal prediction with local weights: randomization enables local guarantees. *arXiv:2310.07850 [stat.ME]* <https://arxiv.org/abs/2310.07850>
- [18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.
- [19] Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. 2024. Conformal prediction for deep classifier via label ranking. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, Article 816, 17 pages.
- [20] Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. 2024. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems* 36 (2024).
- [21] Jian Kang, Qinghai Zhou, and Hanghang Tong. 2022. JuryGCN: quantifying jackknife uncertainty on graph convolutional networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 742–752.
- [22] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [23] Shayan Kiyani, George J Pappas, and Hamed Hassani. 2024. Length optimization in conformal prediction. *Advances in Neural Information Processing Systems* 37 (2024), 99519–99563.
- [24] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
- [25] Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404* (2023).
- [26] Jing Lei, Alessandro Rinaldo, and Larry Wasserman. 2015. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence* 74 (2015), 29–43.
- [27] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958 [cs.CL]* <https://arxiv.org/abs/2109.07958>
- [28] Pranav Maneriker, Aditya T Vadlamani, Anutam Srinivasan, Yuntian He, Ali Payani, and Srinivasan Parthasarathy. 2024. Conformal Prediction: A Theoretical Note and Benchmarking Transductive Node Classification in Graphs. *arXiv preprint arXiv:2409.18332* (2024).
- [29] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.
- [30] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval* 3 (2000), 127–163.
- [31] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems* 33 (2020), 15288–15299.
- [32] Abed Mutemi and Fernando Bacao. 2024. E-commerce fraud detection based on machine learning techniques: Systematic literature review. *Big Data Mining and*

- Analytics* 7, 2 (2024), 419–444.
- [33] Sima Noorani, Orlando Romero, Nicolo Dal Fabbro, Hamed Hassani, and George J. Pappas. 2025. Conformal Risk Minimization with Variance Reduction. *arXiv:2411.01696* [cs.LG]. <https://arxiv.org/abs/2411.01696>
- [34] Felix Petersen, Christian Borgelt, Hilde Kuehne, and Oliver Deussen. 2021. Differentiable sorting networks for scalable sorting and ranking supervision. In *International Conference on Machine Learning*. PMLR, 8546–8555.
- [35] Yaniv Romano, Evan Patterson, and Emmanuel Candes. 2019. Conformalized quantile regression. *Advances in neural information processing systems* 32 (2019).
- [36] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems* 33 (2020), 3581–3591.
- [37] Mauricio Sadinle, Jing Lei, and Larry Wasserman. 2019. Least ambiguous set-valued classifiers with bounded error levels. *J. Amer. Statist. Assoc.* 114, 525 (2019), 223–234.
- [38] Tiyas Sarkar, Manik Rakhra, Vikrant Sharma, and Amanpreet Singh. 2024. An Empirical Comparison of Machine Learning Techniques for Bank Loan Approval Prediction. In *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)*. IEEE, 137–143.
- [39] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. 2018. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868* (2018).
- [40] Ralph C Smith. 2024. *Uncertainty quantification: theory, implementation, and applications*. SIAM.
- [41] Eleni Straitouri and Manuel Gomez Rodriguez. 2023. Designing decision support systems using counterfactual prediction sets. *arXiv preprint arXiv:2306.03928* (2023).
- [42] Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. 2023. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*. PMLR, 32633–32653.
- [43] David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. 2022. Learning optimal conformal classifiers. In *International Conference on Learning Representations*.
- [44] Jiaye Teng, Chuan Wen, Dinghui Zhang, Yoshua Bengio, Yang Gao, and Yang Yuan. 2023. Predictive inference with feature conformal prediction. In *The Eleventh International Conference on Learning Representations*.
- [45] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288* [cs.CL]. <https://arxiv.org/abs/2307.09288>
- [46] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. 2017. Graph attention networks. *stat* 1050, 20 (2017), 10–48550.
- [47] Vladimir Vovk. 2015. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence* 74 (2015), 9–28.
- [48] Vladimir Vovk, Alexander Gammernan, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Vol. 29. Springer.
- [49] Pivithuru Wijegunawardana, Raluca Gera, and Sucheta Soundarajan. 2020. Node Classification with Bounded Error Rates. In *Complex Networks XI: Proceedings of the 11th Conference on Complex Networks CompleNet 2020*. Springer, 26–38.
- [50] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. 2019. Simplifying graph convolutional networks. In *International conference on machine learning*. PMLR, 6861–6871.
- [51] HuaJun Xi, Jianguo Huang, Kangdao Liu, Lei Feng, and Hongxin Wei. 2024. Delving into Temperature Scaling for Adaptive Conformal Prediction. (2024).
- [52] Senrong Xu, Yanke Zhou, Yuhao Tan, Zenan Li, Yuan Yao, Taolue Chen, Feng Xu, and Xiaoxing Ma. 2026. Fair Conformal Classification via Learning Representation-Based Groups. *arXiv preprint arXiv:2605.12195* (2026).
- [53] Soroush H Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. 2023. Conformal prediction sets for graph neural networks. In *International Conference on Machine Learning*. PMLR, 12292–12318.
- [54] Yao Zhang and Emmanuel J. Candès. 2024. Posterior Conformal Prediction. *arXiv:2409.19712* <https://arxiv.org/abs/2409.19712>

A Technical proofs for theoretical results

A.1 Proof of Propostion 1

PROOF. Recall that the non-conformity scores of APS are defined by $V(x, y) = \hat{\pi}_{(1)}(x) + \dots + \hat{\pi}_{(y)}(x)$, where $\hat{\pi}_{(1)}(x) \geq \dots \geq \hat{\pi}_{(K)}(x)$ are (ordered) probabilities of the model prediction. Hence, we have

$$\begin{aligned} \bar{V}(\hat{\pi}(x)) &= \frac{1}{K} \sum_{i=1}^K V(x, i) \\ &= \sum_{k=1}^K \frac{1}{K} (\hat{\pi}_{(1)}(x) + \dots + \hat{\pi}_{(k)}(x)) \\ &= \hat{\pi}_{(1)}(x) + \frac{K-1}{K} \hat{\pi}_{(2)}(x) + \dots + \frac{1}{K} \hat{\pi}_{(K)}(x). \end{aligned} \quad (4)$$

Consider the function $F_1(\hat{\pi}) := \bar{V}(\hat{\pi}(x)) - H(\hat{\pi}(x))$, we can find that $F_1(\hat{\pi})$ is a convex function w.r.t. $\hat{\pi}(x)$, since the negative entropy is convex and the average non-conformity score is linear of $\hat{\pi}(x)$. Therefore, the upper bound of $F_1(\hat{\pi})$ is achieved at the boundary of constraints $\{\hat{\pi}_{(1)}(x) \geq \dots \geq \hat{\pi}_{(K)}(x), \hat{\pi}_{(1)}(x) + \dots + \hat{\pi}_{(K)}(x) = 1\}$. Furthermore, it can be observed that $\bar{V}(\hat{\pi}(x))$ achieves the maximum value 1 and $-H(\hat{\pi}) = \sum_{k=1}^K \hat{\pi}_{(k)} \log \hat{\pi}_{(k)}$ reaches the maximum value 0 simultaneously when $\hat{\pi}_{(1)}(x) = 1$ and $\hat{\pi}_{(2)}(x) = \dots = \hat{\pi}_{(K)}(x) = 0$. Therefore, we have $F_1(\hat{\pi}) \leq 1$, which implies $\bar{V}(\hat{\pi}(x)) \leq 1 + H(\hat{\pi}(x))$.

We also define the function $F_2(\hat{\pi}) := \bar{V}(\hat{\pi}(x)) + H(\hat{\pi}(x))$. Similarly, we can find that $F_2(\hat{\pi}(x))$ is concave w.r.t. $\hat{\pi}(x)$. To analyze the upper bound of $F_2(\hat{\pi})$, we formulate the following optimization problem:

$$\begin{aligned} \max_{\pi} \quad & \bar{V}(\hat{\pi}(x)) + H(\hat{\pi}(x)) \\ \text{s.t.} \quad & \sum_{k=1}^K \hat{\pi}_{(k)}(x) = 1, \\ & \hat{\pi}_{(k)}(x) \geq \hat{\pi}_{(k+1)}(x) \quad k = 0, 1, \dots, K-1, \end{aligned}$$

We temporarily drop the inequality constraints and compute the Lagrangian as

$$\mathcal{L}_F(\hat{\pi}) = H(\hat{\pi}(x)) + \bar{V}(\hat{\pi}(x)) - \lambda \left(\sum_{k=1}^K \hat{\pi}_{(k)}(x) - 1 \right),$$

where λ is the Lagrangian multiplier. Vanishing the partial derivatives, we obtain

$$\frac{\partial \mathcal{L}_F(\hat{\pi})}{\partial \hat{\pi}_{(k)}(x)} = \log \hat{\pi}_{(k)}(x) + \frac{k-1}{K} - \lambda = 0.$$

Hence, we have

$$\hat{\pi}_{(k)}(x) = \exp\left(\lambda - \frac{k-1}{K}\right).$$

Note that $\hat{\pi}_{(k)}(x) \geq \hat{\pi}_{(k+1)}(x)$ also holds for $i = 0, 1, \dots, t-1$. Using the equation $\sum_{k=1}^K \hat{\pi}_{(k)}(x) = 1$, we can derive that

$$\exp(\lambda) = \frac{1}{\sum_{k=1}^K \exp\left(-\frac{k-1}{K}\right)}.$$

In other words, the optimal solution is the output of the Softmax function for logits $[0, -\frac{1}{K}, \dots, -\frac{K-1}{K}]$. By substituting the optimal

solution, we have the maximum value of $F_2(\hat{\pi}(x))$:

$$\begin{aligned}
F_2(\hat{\pi}(x)) &\leq -\sum_{k=1}^K \left(\lambda - \frac{k-1}{K}\right) \exp\left(\lambda - \frac{k-1}{K}\right) \\
&\quad + \left(1 - \frac{k-1}{K}\right) \exp\left(\lambda - \frac{k-1}{K}\right) \\
&= (-\lambda + 1) \sum_{k=1}^K \exp\left(\lambda - \frac{k-1}{K}\right) \\
&= -\lambda + 1 = \log\left(\sum_{k=1}^K \exp\left(-\frac{k-1}{K}\right)\right) + 1.
\end{aligned} \tag{5}$$

Therefore, we have

$$\bar{V}(\hat{\pi}(x)) \leq \log\left(\sum_{k=1}^K \exp\left(-\frac{k-1}{K}\right)\right) + 1 - H(\hat{\pi}(x)).$$

Putting the two upper bounds together, we have

$$\bar{V}(\hat{\pi}(x)) \leq \min(C_K + 1 - H(\hat{\pi}(x)), 1 + H(\hat{\pi}(x))),$$

where $C_K = \log\left(\sum_{k=1}^K \exp\left(-\frac{k-1}{K}\right)\right)$. Furthermore, we can see that the first upper bound strictly holds only when $H(\hat{\pi}(x)) \geq \frac{1}{2}C_K$. We complete the proof. $\square$

We plot the function image of $\frac{1}{2}C_K$ using Wolfram and find that $\frac{1}{2}C_K$ monotonically increases as K rising. Thus, $C_K + 1 - H(\hat{\pi}(x))$ is a better upper bound than $1 + H(\hat{\pi}(x))$ in most cases, when K is relatively small.

A.2 Proof of Proposition 2

PROOF. For a given sample point x and its oracle distribution $\pi(x)$, we have

$$\mathbb{E}[V(X, Y) | X = x] = \sum_{k=1}^K \pi_{(k)}(x) \cdot \sum_{j=1}^k \hat{\pi}_{(j)}(x) \tag{6}$$

Next, we bridge $\bar{V}(x)$ and $\mathbb{E}[V(X, Y) | X = x]$. We can obtain that

$$\begin{aligned}
|\mathbb{E}[V(X, Y) | X = x] - \bar{V}(x)| &= \left| \sum_{k=1}^K \left(\pi_{(k)}(x) - \frac{1}{K}\right) \cdot \sum_{j=1}^k \hat{\pi}_{(j)}(x) \right| \\
&\leq \max\left\{ \sum_{j=1}^k \hat{\pi}_{(j)}(x), k = 1, \dots, K \right\} \\
&\quad \cdot \sum_{k=1}^K \left| \pi_{(k)}(x) - \frac{1}{K} \right| \\
&= 2\delta_{\text{TVD}}(\pi(x), \mathcal{U}(K)),
\end{aligned}$$

where the inequation is derived by the Hölder inequality, and $\delta_{\text{TVD}}(\cdot, \cdot)$ refers to the total variation distance.

Using the Pinsker inequality, we have

$$\delta_{\text{TVD}}(\pi(x), \mathcal{U}(K)) \leq \sqrt{\frac{1}{2} \delta_{\text{KL}}(\pi(x) \| \mathcal{U}(K))} = \sqrt{\frac{1}{2} (\log(K) - H(\pi(x)))}.$$

Putting together, we obtain

$$|\mathbb{E}[V(X, Y) | X = x] - \bar{V}(x)| \leq \sqrt{2(\log(K) - H(\pi(x)))}.$$

Next, we define $C_{\hat{\eta}} := \{(X, Y) | V(X, Y) \geq \hat{\eta}\}$, and seek for the upper bound of $\mathbb{E}[V(X, Y) | C_{\hat{\eta}}]$ from the above result:

$$\mathbb{E}[V(X, Y) | C_{\hat{\eta}}] \leq \mathbb{E}[\bar{V}(X) | C_{\hat{\eta}}] + \mathbb{E}[\sqrt{2(\log(K) - H(\pi(X)))} | C_{\hat{\eta}}].$$

By Hoeffding's inequality and $V(X, Y) \in [\hat{\eta}, 1]$, and $|C_{\hat{\eta}}| = \alpha n$, we have

$$\mathbb{P}\left(\frac{1}{n\alpha} \sum_{(X, Y) \in C_{\hat{\eta}}} V(X, Y) - \mathbb{E}[V(X, Y) | C_{\hat{\eta}}] \geq \tau\right) \leq \exp\left(-\frac{2\alpha\tau^2 n}{(1 - \hat{\eta})^2}\right),$$

where τ is an arbitrary positive constant. Since we have $\sum_{(X, Y) \in C_{\hat{\eta}}} V(X, Y) \geq n\alpha\hat{\eta}$, we can derive

$$\begin{aligned}
\exp\left(-\frac{2\alpha\tau^2 n}{(1 - \hat{\eta})^2}\right) &\geq \mathbb{P}\left(\frac{1}{n\alpha} \sum_{(X, Y) \in C_{\hat{\eta}}} V(X, Y) - \mathbb{E}[V(X, Y) | C_{\hat{\eta}}] \geq \tau\right) \\
&\geq \mathbb{P}(\hat{\eta} - \mathbb{E}[V(X, Y) | C_{\hat{\eta}}] > \tau)
\end{aligned}$$

Substituting the bound of $\mathbb{E}[V(X, Y) | C_{\hat{\eta}}]$ in, we have

$$\begin{aligned}
\exp\left(-\frac{2\alpha\tau^2 n}{(1 - \hat{\eta})^2}\right) &\geq \mathbb{P}(\hat{\eta} > \mathbb{E}[V(X, Y) | C_{\hat{\eta}}] + \tau) \\
&\geq \mathbb{P}(\hat{\eta} > \mathbb{E}[\bar{V}(X) | C_{\hat{\eta}}] \\
&\quad + \mathbb{E}[\sqrt{2(\log(K) - H(\pi(X)))} | C_{\hat{\eta}}] + \tau),
\end{aligned}$$

which completes the proof. $\square$

A.3 Proof of Theorem 3

PROOF. According to the definition of $C(x)$ in terms of APS, we first have

$$|C(x)| = S(\hat{\pi}(x), \hat{\eta}) = \sum_{k=1}^K u(\hat{\eta} - V(x, k)),$$

where $u(a)$ is a unit step function, i.e., if $a \geq 0$, $u(a) = 1$; otherwise, $u(a) = 0$. Since $u(a) \leq a + 1$ holds on $-1 \leq a \leq 1$, we have

$$|C(x)| \leq K\hat{\eta} + K - \sum_{k=1}^K V(x, k) = K(\hat{\eta} - \bar{V}(\hat{\pi}(x)) + 1).$$

Transform the above equation into the expectation form, we get

$$\mathbb{E}[|C(X)|] \leq K(\hat{\eta} - \mathbb{E}[\bar{V}(\hat{\pi}(X))]) + 1.$$

Furthermore, for $C_{\hat{\eta}} = \{(X, Y) | V(X, Y) \geq \hat{\eta}\}$ and its complementary set $\bar{C}_{\hat{\eta}}$, we have

$$\begin{aligned}
\mathbb{E}[\bar{V}(\hat{\pi}(X))] &= (1 - \alpha)\mathbb{E}[\bar{V}(\hat{\pi}(X)) | V(X, Y) \leq \hat{\eta}] \\
&\quad + \alpha\mathbb{E}[\bar{V}(\hat{\pi}(X)) | V(X, Y) > \hat{\eta}],
\end{aligned}$$

and put it and the bound of $\hat{\eta}$ into the above equation:

$$\begin{aligned}
\frac{1}{K}\mathbb{E}[|C(X)|] &\leq \mathbb{E}[\bar{V}(\hat{\pi}(X)) | C_{\hat{\eta}}] \\
&\quad - (1 - \alpha)\mathbb{E}[\bar{V}(\hat{\pi}(X)) | \bar{C}_{\hat{\eta}}] - \alpha\mathbb{E}[\bar{V}(\hat{\pi}(X)) | C_{\hat{\eta}}] + C \\
&= (1 - \alpha)(\mathbb{E}[\bar{V}(\hat{\pi}(X)) | C_{\hat{\eta}}] - \mathbb{E}[\bar{V}(\hat{\pi}(X)) | \bar{C}_{\hat{\eta}}]) + C,
\end{aligned}$$

where constant $C := \mathbb{E}[\sqrt{2(\log(K) - H(\pi(X)))} | C_{\hat{\eta}}] + \tau + 1$. Using the lower bound $\bar{V}(\hat{\pi}(x)) \geq \frac{K+1}{2K}$, we have

$$\frac{1}{K}\mathbb{E}[|C(X)|] \leq (1 - \alpha)(\mathbb{E}[\bar{V}(\hat{\pi}(X)) | C_{\hat{\eta}}] - \frac{K+1}{2K}) + C.$$

Given the assumption that $\mathbb{P}(H(\hat{\pi}(X)) \geq \frac{1}{2}C_K \mid C_{\hat{\eta}}) \geq \mu$ ($\mu \in [0, 1]$) and $\mathcal{D} := \{(X, Y) \mid H(\hat{\pi}(X)) \geq \frac{1}{2}C_K\}$, we obtain that

$$\begin{aligned}
\frac{1}{K}\mathbb{E}(|C(X)|) &\leq (1-\alpha)\mathbb{E}[\tilde{V}(\hat{\pi}(X)) \mid C_{\hat{\eta}}] - (1-\alpha)\frac{K+1}{2K} + C \\
&\leq (1-\alpha)(\mu(C_K - 1 - \mathbb{E}[H(\hat{\pi}(x)) \mid \mathcal{D} \cap C_{\hat{\eta}}]) \\
&\quad + (1-\mu)(1 + \mathbb{E}[H(\hat{\pi}(x)) \mid \bar{\mathcal{D}} \cap C_{\hat{\eta}}])) \\
&\quad - (1-\alpha)\frac{K+1}{2K} + C \\
&\leq (1-\alpha)(1-2\mu)\mathbb{E}[H(\hat{\pi}(x)) \mid C_{\hat{\eta}}] \\
&\quad + (1-\alpha)(\mu C_K - \frac{K+1}{2K}) + (1-\alpha)(1-2\mu) + C.
\end{aligned}$$

By combining with the inequality derived in Proposition 1, where the final step uses $\mathbb{E}[H(\hat{\pi}(x)) \mid \bar{\mathcal{D}} \cap C_{\hat{\eta}}] \leq \mathbb{E}[H(\hat{\pi}(x)) \mid C_{\hat{\eta}}] \leq \mathbb{E}[H(\hat{\pi}(x)) \mid \mathcal{D} \cap C_{\hat{\eta}}]$. Clearly, there is an $O(\log K)$ constant term, as $\mathbb{E}[\sqrt{2(\log(K) - H(\pi(X)))} \mid C_{\hat{\eta}}]$ and $C_K := \log(\sum_{k=1}^K \exp(-\frac{k-1}{K})) < \log K$ for $K \geq 2$. $\square$

B Further experiment details

Additional experimental details and results are provided in the Appendix B file at our repository (<https://github.com/Xusr1123/Conformal-Correction>).