Conformalized Credal Set Predictors

Alireza Javanmardi
LMU Munich, MCML
Munich, Germany
alireza.javanmardi@ifi.lmu.de

David Stutz
Max Planck Institute for Informatics
Saarbrücken, Germany
david.stutz@mpi-inf.mpg.de

Eyke Hüllermeier
LMU Munich, MCML
Munich, Germany
eyke@ifi.lmu.de

Abstract

Credal sets are sets of probability distributions that are considered as candidates for an imprecisely known ground-truth distribution. In machine learning, they have recently attracted attention as an appealing formalism for uncertainty representation, in particular, due to their ability to represent both the aleatoric and epistemic uncertainty in a prediction. However, the design of methods for learning credal set predictors remains a challenging problem. In this paper, we make use of conformal prediction for this purpose. More specifically, we propose a method for predicting credal sets in the classification task, given training data labeled by probability distributions. Since our method inherits the coverage guarantees of conformal prediction, our conformal credal sets are guaranteed to be valid with high probability (without any assumptions on model or distribution). We demonstrate the applicability of our method on ambiguous classification tasks for uncertainty quantification.

1 Introduction

Representing and quantifying uncertainty is becoming increasingly important in machine learning (ML), particularly as ML models are employed in safety-critical application domains such as medicine or autonomous driving. In such domains, a distinction between so-called *aleatoric uncertainty* (AU) and *epistemic uncertainty* (EU) is often useful [21]. Broadly speaking, aleatoric uncertainty is due to the inherent randomness of the data-generating process, whereas epistemic uncertainty stems from the learner's lack of knowledge about the best predictive model. Thus, while the former is irreducible, the latter can, in principle, be reduced through additional information, e.g., by gathering additional data to learn from.

Representation of aleatoric and epistemic uncertainty requires formalism more expressive than standard probability distributions [22]. One such formalism which prevails in the recent ML literature is second-order probability distributions. Essentially, in a classification setting, these are distributions over distributions over classes. Models producing second-order distributions as predictions can be learned in a classical Bayesian way [16, 25] or using more recent approaches such as evidential deep learning [44]. Yet, approaches of that kind are not unproblematic and have been subject to criticism [8, 9]. Specifically, such approaches have been shown to misrepresent epistemic uncertainty. Another formalism suitable for representing both types of uncertainty is the concept of a *credal set*, which is well-established in the field of imprecise probability theory [58] and meanwhile also attracted attention in ML [23, 46]. Credal sets are (convex) sets of probability distributions that can be considered as candidates for an imprecisely known ground-truth distribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3714

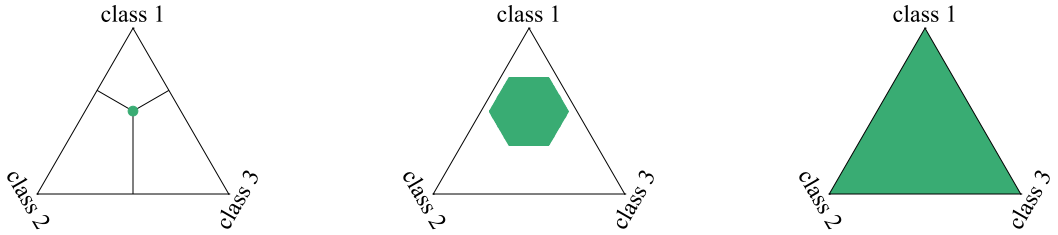


Figure 1: For the three-class classification setting, the space of probability distributions can be illustrated by a two-dimensional simplex: each point in the simplex corresponds to a probability distribution so that credal sets can be depicted as regions. The left case corresponds to the special case of a singleton (credal) set, i.e., a precise probability distribution, signifying aleatoric but no epistemic uncertainty. The case in the middle represents partial knowledge with a certain degree of (epistemic) uncertainty about the true distribution, and the right one corresponds to the case of complete ignorance, where nothing is known about the distribution.

Figure 1 shows examples of credal sets in a three-class scenario, where the space of distributions can be visualized by the two-dimensional probability simplex. Broadly speaking, the larger the credal set, the higher the epistemic uncertainty, and the more "in the middle" the set is located, i.e., the closer it is to the uniform distribution, the higher the aleatoric uncertainty.

Learning to predict second-order representations, such as credal sets or second-order distributions, from standard "zero-order" data — training instances together with observed class labels — has been shown to be difficult in that it typically provides biased estimates of epistemic uncertainty [8, 9]. To alleviate this problem, we assume "first-order" training data, i.e., instances associated with probability (frequency) distributions over the class labels. In other words, instances are labeled probabilistically instead of being assigned a deterministic class label. This type of data is becoming increasingly available in practice, for example, in the form of aggregations over multiple annotations per data instance [10, 31, 35, 63], and hence increasingly relevant in many applications [52, 53]. Moreover, such data facilitates second-order learning.

In this paper, we leverage conformal prediction (CP), a non-parametric approach for set-valued prediction rooted in classical frequentist statistics [57], to construct credal sets. With relatively mild assumptions similar to those in CP, this approach inherits the so-called marginal coverage from CP: Predicted sets are guaranteed to cover the true target with high probability. For us, this means that we can use any first-order or second-order predictive model to construct unbiased credal sets equipped with such a coverage guarantee. Specifically, we propose various nonconformity functions applicable to first- and second-order predictors to construct conformal credal sets. We also study the case where only a noisy version of first-order distributions is available, demonstrating that the coverage guarantee holds under a bounded noise assumption. On ChaosNLI [31], an ambiguous natural language inference task with multiple annotations per example, we show that our conformal credal sets are valid, i.e., covering the true ground-truth distribution with high probability while comparing the efficiency of different nonconformity functions. Together with experiments on CIFAR10-H [35]—a variant of the CIFAR10 test set containing class distributions from human annotations—we demonstrate how these credal sets enable practical quantification of aleatoric and epistemic uncertainty. We further complement this study with controlled experiments on synthetic data, specifically investigating the performance of credal set prediction in the presence of noisy data.

2 Background

2.1 Supervised Learning and Predictive Uncertainty

We consider the setting of (polychotomous) classification with label space $\mathcal{Y} = \{1, \dots, K\}$ and an instance space $\mathcal{X}$. As usual, we assume an underlying data-generating process in the form of a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, so that observations (X, Y) are i.i.d. samples from P. We denote by $\boldsymbol{\lambda}^{\boldsymbol{x}} = (\lambda_1^{\boldsymbol{x}}, \dots, \lambda_K^{\boldsymbol{x}})^\top \in \Delta^K$ the conditional probability distribution $P(\cdot \mid X = \boldsymbol{x})$, which we also consider as an element of the $(K-1)$-simplex Δ^K. Thus, the probability to observe $Y = k$ as an outcome for $\boldsymbol{x} \in \mathcal{X}$ is given by $\lambda_k^{\boldsymbol{x}}$.

https://doi.org/10.52202/079017-3714

Since the dependency between instances X and outcomes Y is non-deterministic, the prediction of Y given $X = \boldsymbol{x}$ is necessarily afflicted with uncertainty, even if the ground-truth distribution $\boldsymbol{\lambda}^{\boldsymbol{x}}$ is known. As already said, this uncertainty is commonly referred to as *aleatoric* [22]. Intuitively, the closer $\boldsymbol{\lambda}^{\boldsymbol{x}}$ to the uniform distribution $\boldsymbol{p}_{\text{uni}} = (1/K, \ldots, 1/K)^\top$, the higher the uncertainty, and the closer it is to a degenerate (Dirac) distribution assigning all probability mass to a single class (a corner point in Δ^K), the lower the uncertainty.

Instead of assuming $\boldsymbol{\lambda}^{\boldsymbol{x}}$ to be known, suppose now that only a prediction $\hat{\boldsymbol{\lambda}}^{\boldsymbol{x}}$ of this distribution is available. *Epistemic uncertainty* refers to the uncertainty about how well the latter approximates the former, and hence to the additional uncertainty in the prediction of outcome Y that is caused by the discrepancy between $\hat{\boldsymbol{\lambda}}^{\boldsymbol{x}}$ and the ground-truth $\boldsymbol{\lambda}^{\boldsymbol{x}}$. We seek to capture this discrepancy by means of credal sets $Q \in \mathcal{Q}_K \subset \Delta^K$, with the idea that $Q \ni \boldsymbol{\lambda}^{\boldsymbol{x}}$ holds with high probability. Typically, credal sets are assumed to be convex, and further restrictions might be imposed on $\mathcal{Q}_K$ for practical and computational reasons, for example, a restriction to convex polygons (with a finite number of extreme points).

2.2 Conformal Prediction

Conformal prediction provides a general framework for producing set-valued predictions with a certain guarantee of validity. In a supervised setting, consider data points of the form $Z = (X, U) \in \mathcal{X} \times \mathcal{U}$, and the task is to predict U given $X = \boldsymbol{x}$. We assume the space $\mathcal{Z} = \mathcal{X} \times \mathcal{U}$ to be equipped with a nonconformity measure $f : \mathcal{Z} \longrightarrow \mathbb{R}$ that quantifies the "strangeness" of $\boldsymbol{z}$, i.e., the higher $f(\boldsymbol{z})$, the less normal or expected the data point. Let $\mathcal{D}_{\text{calib}} \subset \mathcal{Z}$ be a (randomly generated) set of data points, called *calibration data*, and Z another data point that remains unobserved. Under the assumption of exchangeability, i.e., that the calibration data and the query point Z have been generated by an exchangeable process, we seek a so-called confidence set $C \subseteq \mathcal{U}$ that guarantees coverage:

$$\mathbb{P}(U \in C) \geq 1 - \alpha. \tag{1}$$

By a simple combinatorial argument [57], the confidence set C can be constructed as

$$C(\boldsymbol{x}) := \left\{ u \in \mathcal{U} \mid f(\boldsymbol{x}, u) \leq q(\mathcal{E}, \alpha') \right\}, \tag{2}$$

where $\mathcal{E} := \{f(\boldsymbol{z}) : \boldsymbol{z} \in \mathcal{D}_{\text{calib}}\}$ is the set of nonconformity scores, $\alpha' = |\mathcal{E}|^{-1}\lceil(1 + |\mathcal{E}|)(1 - \alpha)\rceil$, and $q(\mathcal{E}; \alpha')$ denotes the α'-quantile of $\mathcal{E}$. Importantly, the guarantee (1) holds regardless of the nonconformity function $f(\cdot)$, which, however, has an influence on the *efficiency* of the prediction: The more appropriate the function, the smaller the prediction set C tends to be. Normally, $f(\cdot)$ is not predefined but constructed in a data-driven way using training data $\mathcal{D}_{\text{train}}$. For example, a common approach is to train a predictor $\pi : \mathcal{X} \longrightarrow \mathcal{U}$ and then define $f(\boldsymbol{x}, u)$ in terms of $d(u, \pi(\boldsymbol{x}))$, where $d(\cdot, \cdot)$ is an appropriate distance function on $\mathcal{U}$. Replacing the point-prediction $\pi(\boldsymbol{x}) \in \mathcal{U}$ by the prediction set $C(\boldsymbol{x}) \subset \mathcal{U}$ can then be seen as "conformalizing" the predictor π.

3 Conformal Credal Set Prediction

Our goal is to learn a credal set predictor $h : \mathcal{X} \longrightarrow \mathcal{Q}_K$, that is, a model that makes predictions in the form of credal sets, thereby representing both aleatoric and epistemic uncertainty. To this end, we assume access to first-order data, i.e., probabilistic training data of the form

$$\mathcal{D} = \left\{ \left(\boldsymbol{x}_1, \boldsymbol{\lambda}^{\boldsymbol{x}_1}\right), \ldots, \left(\boldsymbol{x}_N, \boldsymbol{\lambda}^{\boldsymbol{x}_N}\right) \right\} \subset \mathcal{X} \times \Delta^K. \tag{3}$$

The model h should be able to predict the (probabilistic) outcomes for new query instances in a reliable way. More specifically, suppose that $\boldsymbol{x}_{\text{new}}$ is a new query instance (following the same distribution as the training data) for which a prediction is sought. To correctly represent epistemic and aleatoric uncertainty, we want the credal prediction $Q = h(\boldsymbol{x}_{\text{new}})$ to be valid, meaning $Q \ni \boldsymbol{\lambda}^{\boldsymbol{x}_{\text{new}}}$ with high probability, while at the same time being informative such that the (epistemic) uncertainty reflected by Q is as small as possible.

We aim to construct the credal set predictor h by means of conformal prediction. Following the conformalization recipe outlined in Section 2.2, we partition $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{calib}}$, using the former for model training and the latter for calibration.

https://doi.org/10.52202/079017-3714

Algorithm 1 Conformal Credal Set Prediction

Input:

Data $\mathcal{D}$; error rate α; query instance $\boldsymbol{x}_{\text{new}}$.

Process:

Partition $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{calib}}$.
Train a first-order ($i = 1$) or a second-order ($i = 2$) predictor using $\mathcal{D}_{\text{train}}$.
Choose a nonconformity function f_i as in (5) or (8) that suits the trained predictor to obtain the set of scores $\mathcal{E}_i$.
Set $\alpha' = |\mathcal{E}_i|^{-1}\lceil(1 + |\mathcal{E}_i|)(1 - \alpha)\rceil$.

Output:

$h_i(\boldsymbol{x}_{\text{new}}) = \left\{ \boldsymbol{\lambda} \in \Delta^K \mid f_i(\boldsymbol{x}_{\text{new}}, \boldsymbol{\lambda}) \leq q(\mathcal{E}_i, \alpha') \right\}.$

Regarding the training step, we explore two learning strategies connected with two ways of defining a nonconformity function, which is pivotal in the calibration step. The first approach is based on training a standard (*first-order*) probability predictor, i.e., a probabilistic classifier $g : \mathcal{X} \longrightarrow \Delta^K$ that maps instances to the (first-order) probability distribution on $\mathcal{Y}$. This can be achieved, for example, by minimizing the cross-entropy loss between the ground truth and the predicted distributions, i.e.,

$$g = \underset{\bar{g} \in \mathcal{H}}{\operatorname{argmin}} \sum_{(\boldsymbol{x}_i, \boldsymbol{\lambda}^{\boldsymbol{x}_i}) \in \mathcal{D}_{\text{train}}} -\sum_{k=1}^{K} \lambda_k^{\boldsymbol{x}_i} \log(\bar{g}(\boldsymbol{x}_i)_k), \tag{4}$$

where $\mathcal{H}$ is a hypothesis space. Given a predictor $g(\cdot)$ of this kind, nonconformity is naturally defined in terms of a distance:

$$f_1(\boldsymbol{x}, \boldsymbol{\lambda}^{\boldsymbol{x}}) := d(\boldsymbol{\lambda}^{\boldsymbol{x}}, g(\boldsymbol{x})), \tag{5}$$

where $d(\cdot, \cdot)$ is a suitable distance function on Δ^K, such as total variation, Wasserstein distance, etc.

An alternative approach is motivated by recent work on (epistemic) uncertainty representation via *second-order* probability distributions [16, 25, 44]. A second-order learner $G : \mathcal{X} \longrightarrow \mathbb{P}(\Delta^K)$ maps each input $\boldsymbol{x}$ to a distribution over Δ^K. Given the training data of the form (3), meaningful learning in this context can be accomplished, for instance, by parameterizing the second-order distributions using Dirichlet distributions. Specifically, one can assume that each $\boldsymbol{x}$ is associated with a Dirichlet distribution characterized by the parameter vector $\boldsymbol{\theta}^{\boldsymbol{x}} \in \mathbb{R}^K_{\geq 1}$ with the probability density function

$$P(\boldsymbol{\lambda} \mid \boldsymbol{\theta}^{\boldsymbol{x}}) = \frac{1}{B(\boldsymbol{\theta}^{\boldsymbol{x}})} \prod_{k=1}^{K} \lambda_k^{\theta_k^{\boldsymbol{x}} - 1}, \tag{6}$$

where $B(\cdot)$ is the multivariate beta function. This way, $\boldsymbol{\lambda}^{\boldsymbol{x}}$ can be thought of as a sample from that distribution, i.e., $\boldsymbol{\lambda}^{\boldsymbol{x}} \sim \text{Dir}(\boldsymbol{\theta}^{\boldsymbol{x}})$. The model aims to find parameter vectors $\boldsymbol{\theta}^{\boldsymbol{x}}$s that minimize the negative log-likelihood loss

$$\sum_{(\boldsymbol{x}_i, \boldsymbol{\lambda}^{\boldsymbol{x}_i}) \in \mathcal{D}_{\text{train}}} \left(\log(B(\boldsymbol{\theta}^{\boldsymbol{x}_i})) - \sum_{k=1}^{K} (\theta_k^{\boldsymbol{x}_i} - 1) \log(\lambda_k^{\boldsymbol{x}_i}) \right). \tag{7}$$

Given a second-order predictor $\boldsymbol{\theta}^{\boldsymbol{x}}$, nonconformity can be defined as a decreasing function of likelihood, e.g., as 1 minus relative likelihood:

$$f_2(\boldsymbol{x}, \boldsymbol{\lambda}^{\boldsymbol{x}}) = 1 - \frac{P(\boldsymbol{\lambda}^{\boldsymbol{x}} \mid \boldsymbol{\theta}^{\boldsymbol{x}})}{\max_{\boldsymbol{\lambda} \in \Delta^K} P(\boldsymbol{\lambda} \mid \boldsymbol{\theta}^{\boldsymbol{x}})}. \tag{8}$$

Using the nonconformity function $f_i(\cdot)$ ($i \in \{1, 2\}$), we obtain the set of nonconformity scores by

$$\mathcal{E}_i := \left\{ f_i(\boldsymbol{x}_j, \boldsymbol{\lambda}^{\boldsymbol{x}_j}) \mid (\boldsymbol{x}_j, \boldsymbol{\lambda}^{\boldsymbol{x}_j}) \in \mathcal{D}_{\text{calib}} \right\}. \tag{9}$$

Accordingly, the credal set can be defined as

$$h_i(\boldsymbol{x}_{\text{new}}) := \left\{ \boldsymbol{\lambda} \in \Delta^K \mid f_i(\boldsymbol{x}_{\text{new}}, \boldsymbol{\lambda}) \leq q(\mathcal{E}_i, \alpha') \right\}. \tag{10}$$

Algorithm 1 outlines a summary of the proposed methods. In the following theorem, we state the validity of the predicted set, that is, the restatement of the conformal coverage guarantee [57] adjusted to our setting.

Theorem 3.1 (Validty Gaurantee)**.** *Let $\mathcal{P}$ denote the joint probability distribution on $(X, \Lambda) \in \mathcal{X} \times \Delta^K$. If data points in $\mathcal{D}_{calib}$ and $(\boldsymbol{x}_{new}, \boldsymbol{\lambda}^{\boldsymbol{x}_{new}})$ are drawn exchangeably from $\mathcal{P}$, then the conformal credal sets in (10) are valid, i.e.,*

$$\mathbb{P}\big(\boldsymbol{\lambda}^{\boldsymbol{x}_{new}} \in h_i(\boldsymbol{x}_{new})\big) \geq 1 - \alpha\,, \;\; \textit{for } i \in \{1, 2\}.$$

It is worth mentioning that both predictors can also be trained using zero-order data. A first-order predictor can be achieved through standard training with a cross-entropy loss function, while a second-order predictor can be achieved, for instance, through the means of evidential learning [44]. Hence, it is sufficient to have probabilistic calibration data in order to have a reasonable judgment of the predictors' performances and be able to conformalize them.

Moreover, while these two approaches can be compared in various ways, one immediate observation is that training a second-order predictor poses greater challenges. However, credal sets constructed using the second-order predictor exhibit a natural and superior adaptivity compared to those constructed by the first-order predictor. This is because, with the first-order predictor, once the quantile is determined during calibration, all distributions within a certain distance are included in the set for any given point at prediction time. On the other hand, with the second-order predictor, the resulting set depends on the calculated quantile as well as the skewness of the predicted distribution.

3.1 Generalization to Imprecise First-Order Data

So far, we (implicitly) assumed that ground-truth probability distributions $\boldsymbol{\lambda}^{\boldsymbol{x}_i}$ will be provided as calibration (and training) data. Needless to say, this assumption will rarely hold true in practice. Instead, observations will rather be noisy versions $\tilde{\boldsymbol{\lambda}}^{\boldsymbol{x}_i}$ of the true probabilities. Notably, such datasets emerge in scenarios where each data instance $\boldsymbol{x}$ is annotated by multiple human experts, which recently have attracted a lot of attention in the context of machine learning [10, 31, 35, 43, 63] and also conformal prediction [24, 52]. In this context, $\tilde{\boldsymbol{\lambda}}^{\boldsymbol{x}}$ denotes the distribution derived from aggregating annotator disagreements concerning the label of instance $\boldsymbol{x}$. Of course, conformal prediction can still be applied to noisy data of that kind, but the coverage guarantee will then only hold w.r.t. noisy labeling, i.e., $\mathbb{P}(\tilde{\boldsymbol{\lambda}}^{\boldsymbol{x}_{\text{new}}} \in h(\boldsymbol{x}_{\text{new}})) \geq 1 - \alpha$.

Practically, one may expect that the guarantees will hold for the ground truth as well, simply because calibration on noisy instead of clean data will tend to make prediction regions larger and hence more conservative. Moreover, since nonconformity is derived from a predictive model $g(\cdot)$ that seeks to recover ground-truth probabilities, the latter should conform at least as well as noisy distributions. Of course, this intuition is not a formal guarantee. In order to provide such a guarantee for the ground-truth probabilities, one obviously needs to make some assumptions. Concretely, let us make the following *bounded noise* assumption for the labeling process: The labeling noise is (stochastically) bounded in the sense that, given the nonconformity function f and a (small) probability $\delta > 0$, there exists a tolerance $\epsilon > 0$ such that the following holds all $\boldsymbol{x} \in \mathcal{X}$:

$$\mathbb{P}\left(|f(\boldsymbol{x}, \boldsymbol{\lambda}^{\boldsymbol{x}}) - f(\boldsymbol{x}, \tilde{\boldsymbol{\lambda}}^{\boldsymbol{x}})| < \epsilon\right) \geq 1 - \delta. \tag{11}$$

Theorem 3.2. *Let $\alpha > 0$ be any miscoverage rate, and suppose the bounded noise assumption holds. Let $q = q(\mathcal{E}, \tilde{\alpha})$ be the critical threshold on the noisy calibration data $\mathcal{D}_{calib}$ for miscoverage rate $\tilde{\alpha} = (\alpha - \delta)/(1 - \delta)$. Then, for any new query $\boldsymbol{x}_{new} \in \mathcal{X}$,*

$$\mathbb{P}\big(f(\boldsymbol{x}_{new}, \boldsymbol{\lambda}^{\boldsymbol{x}_{new}}) < q + \epsilon\big) \geq 1 - \alpha\,.$$

The proof is deferred to Appendix B. As a consequence of this result, a conformal predictor learned on the noisy data with modified miscoverage rate $\tilde{\alpha}$ can be turned into a valid predictor (with miscoverage rate α) for the ground-truth data by increasing the learned rejection threshold by ϵ, provided the bounded noise property (11) can be ascertained. Thus, if we denote the corresponding credal set predictor by h_ϵ, we can guarantee that $\mathbb{P}\big(\boldsymbol{\lambda}^{\boldsymbol{x}_{\text{new}}} \in h_\epsilon(\boldsymbol{x}_{\text{new}})\big) \geq 1 - \alpha$. For a comprehensive study of handling noisy data in conformal prediction, we refer to [19].

4 Experiments

For the sake of comparison, we examine different nonconformity functions in our experiments. When utilizing a first-order predictor, besides total variation (**TV**) and the First Wasserstein (**WS**)

https://doi.org/10.52202/079017-3714

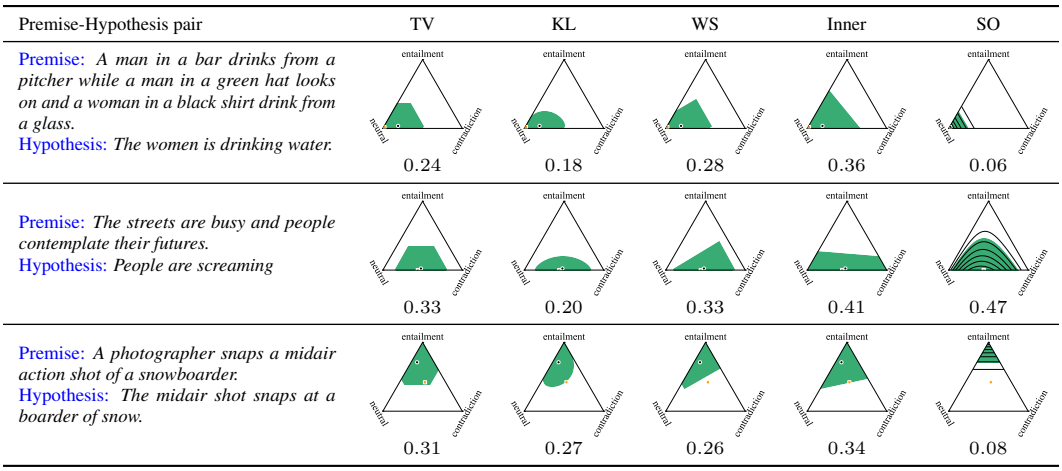


Figure 2: Various credal sets obtained for three instances from ChaosNLI dataset. The ground truth distributions are denoted by orange squares. Black circles indicate model predictions in cases employing a first-order learner (first four columns). For the last column, utilizing a second-order learner, the predicted second-order distributions are represented through contour plots. The miscoverage rate is $\alpha = 0.2$, and the efficiency of each credal set is depicted below it.

distance, we also investigate the Kullback–Leibler (**KL**) divergence and 1 minus the inner product (**Inner**) as nonconformity functions. For the second-order predictor, we consider 1 minus the relative likelihood (**SO**) as defined in (8). Furthermore, we demonstrate how our credal sets allow uncertainty quantification and the disentanglement of total uncertainty into epistemic and aleatoric components. We focus on a measure proposed by Abellán et al. [2], which regards the upper Shannon entropy H[1] of a given credal set Q as total uncertainty, the lower Shannon entropy as aleatoric, and the difference between upper and lower entropy as epistemic:

$$\underbrace{H^*(Q)}_{\text{total}} = \underbrace{H_*(Q)}_{\text{aleatoric}} + \underbrace{(H^*(Q) - H_*(Q))}_{\text{epistemic}} \quad \text{with } H^*(Q) := \max_{\boldsymbol{\lambda} \in Q} H(\boldsymbol{\lambda}),\ H_*(Q) := \min_{\boldsymbol{\lambda} \in Q} H(\boldsymbol{\lambda}). \quad (12)$$

We refer to [23] for other measures, along with a discussion of their corresponding strengths and weaknesses. Interestingly enough, the interval $[H_*(Q), H^*(Q)] \subseteq [0, 1]$ can be seen as an alternative characterization of the credal set Q, which helps address the challenge of visualization for $K > 3$.

In this section, we focus on experiments on two real-world datasets. Further information on these datasets and details about the learning models can be found in Appendix C. Additional experiments on synthetic data, including an illustrative example showing how the resulting credal sets change as epistemic uncertainty decreases, and experiments on the impact of imprecise first-order data, are provided in Appendix E. All implementations and experiments can be found on our GitHub repository.[2]

ChaosNLI Dataset. We start our experiments with a highly ambiguous dataset, ChaosNLI [31], where the task is to classify the textual entailment of a premise-hypothesis pair into three classes: entailment, contradiction, and neutral. We train both first-order and second-order predictors using this data and construct credal sets with all five nonconformity functions to facilitate a comprehensive comparison of these methods. In Figure 2, we compare the resulting credal sets of different nonconformity functions for three specific instances. As expected, even though CP should work regardless of the choice of nonconformity score function, this choice affects the size and geometry of the prediction set. To compare the prediction set size, aka *efficiency*, across different nonconformity functions, we discretize the simplex using a fine grid. The efficiency is gauged by considering the fraction of all distributions that lie within the predicted credal sets. We perform training, calibration, and testing using 10 different random seeds and depict the average coverage and efficiency results of each credal set predictor under different miscoverage rates (α) in Figure 3. Notably, the mean of the

[1] $H(\boldsymbol{\lambda}) := -\sum_{k=1}^{K} \lambda_k \log_K(\lambda_k)$ with $0 \log 0 = 0$ by definition.

[2] The link to the code: https://github.com/alireza-javanmardi/conformal-credal-sets

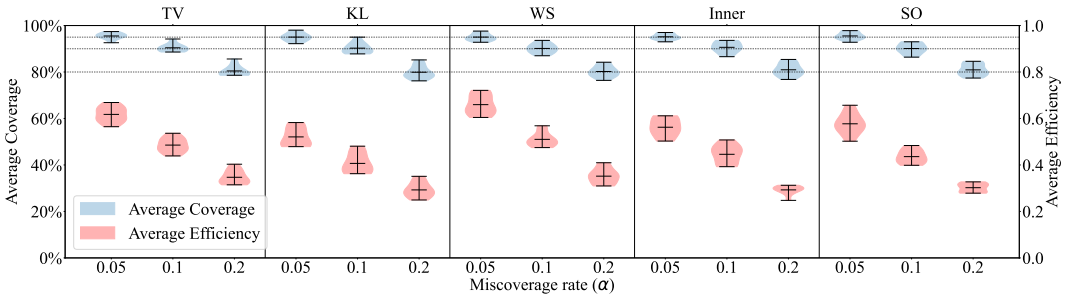
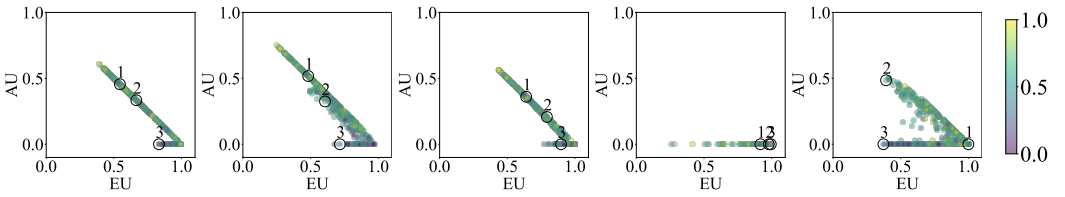


Figure 3: Coverage and efficiency results of different nonconformity functions applied on the ChaosNLI dataset. The horizontal dashed lines indicate the nominal coverage levels.



(a) **TV** (b) **KL** (c) **WS** (d) **Inner** (e) **SO**

Figure 4: Scatter plots of aleatoric versus epistemic uncertainty using various credal set predictors for the ChaosNLI dataset with $\alpha = 0.2$. The colors indicate the entropy of the ground truth distribution. The three cases numbered 1 to 3, correspond to the first to third rows of instances shown in Figure 5.

average coverage over the test data across various random seeds aligns with or exceeds the nominal value, consistent with the conformal prediction guarantee.

We calculate the uncertainty intervals as in (12) for all (test) instances. Figure 4 provides scatter plots of quantified AU vs. EU for different credal set predictors, with colors indicating the entropy of the ground-truth distribution. These plots serve as an evaluation of uncertainty quantification performance. Generally, given access to the first-order distribution, we expect the following: if the EU is low, the AU should align closely with the entropy of the ground-truth first-order distribution. However, when the EU is high, the quantified AU may vary, appearing either close to or far from this entropy. In contrast, with only standard zero-order data available, such direct evaluation isn't feasible, which is why indirect evaluation methods like OOD detection or accuracy-rejection curves are preferred.

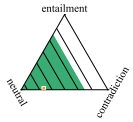
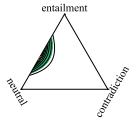
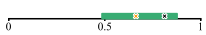
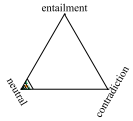
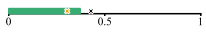
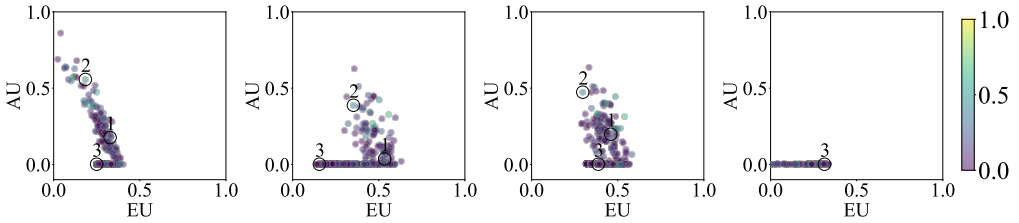
	Premise	Hypothesis	Q	$[H_*(Q), H^*(Q)]$
High EU	*The purpose of the Diwan-i-Khas is hotly disputed; it is not necessarily the hall of private audience that its name implies.*	*The hall is not know many people.*		
Low EU, High AU	*For example, Bruce Barton's The Man Nobody Knows , a best seller in 1925-26, portrays Jesus as the ultimate businessman.*	*Bruce Barton's, "The Man Nobody Knows", a best seller in 1925-26, is known as the best example of Jesus as the ultimate businessman.*		
Low EU, Low AU	*A woman in a long-sleeved shirt checks her phone as a man in a leather jacket passes behind her.*	*The man is passing by on his way to the store.*		

Figure 5: Different uncertainty situations given the predicted credal sets generated by **SO** method for the ChaosNLI dataset with $\alpha = 0.2$.

https://doi.org/10.52202/079017-3714



(a) **TV** (b) **KL** (c) **WS** (d) **Inner**

Figure 6: Scatter plots of aleatoric versus epistemic uncertainty using various credal set predictors for the CIFAR10-H dataset with $\alpha = 0.2$. The colors indicate the entropy of the ground truth distribution. The three cases numbered 1 to 3, correspond to the first to third rows of instances shown in Figure 7.

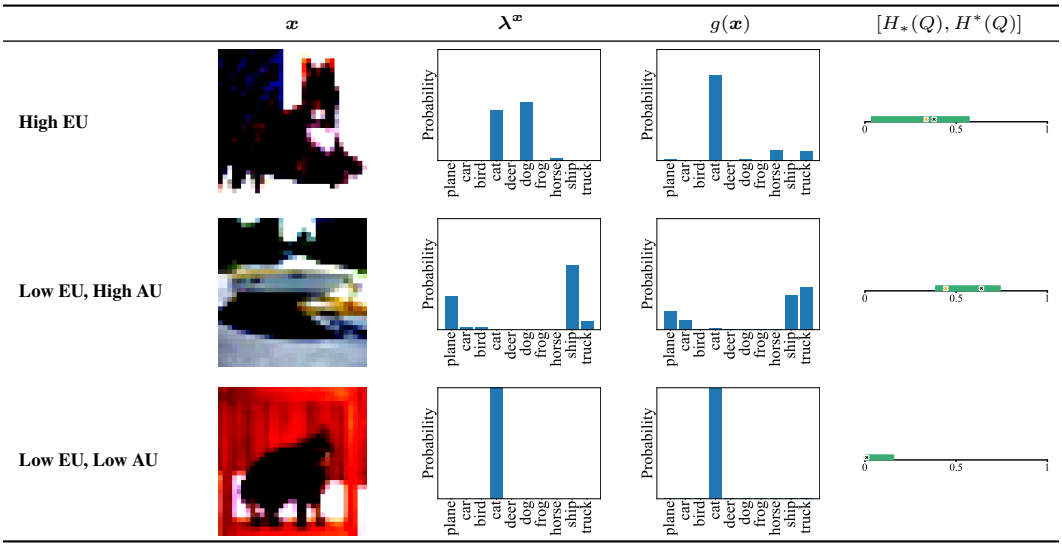


Figure 7: Different uncertainty situations given the predicted credal sets generated by **KL** method for the CIFAR10-H dataset with $\alpha = 0.2$.

For the **SO** method, three uncertainty cases are chosen from the results: high EU, low EU with high AU, and low EU with low AU as shown in Figure 5. When visualizing the uncertainty intervals, we display the entropy of the ground truth distribution in orange and the predicted entropy for the first-order predictor or the mean of the Dirichlet distribution for the second-order predictor in black. However, note that covering the entropy of a distribution with the uncertainty interval does not equate to covering the distribution itself with the credal set.

CIFAR-10H Dataset. We also apply our methods to the CIFAR-10H dataset [35], which contains the distributions on the classes for the CIFAR-10 test set derived from human annotations. This time, we simply use a first-order predictor pre-trained on the CIFAR-10 training set (i.e., standard zero-order training data) and only use CIFAR-10H for calibration and testing. Since the visualization of credal sets is no longer possible for this dataset, we limit our attention to uncertainty quantification analysis. Figure 6 illustrates scatter plots of quantified AU versus EU for different credal set predictors. Three uncertainty cases are chosen from the results and shown in Figure 7. For instance, in the first row, the model's predicted distribution ($g(\boldsymbol{x})$) heavily favors "cat" over "dog", while human-derived distributions assign nearly equal probabilities to both classes. This relatively high epistemic uncertainty is effectively captured by the uncertainty interval.

Further Discussions. Given a model and a data point, different methods may exhibit different behaviors in terms of uncertainty representation, stemming from their distinct approaches to generating the sets. For instance, it has been observed that **KL** tends to limit the set's expansion towards high entropy regions compared to other methods. The **Inner** method, on the other hand, tends to construct

credal sets by incorporating corners (degenerate distributions), thereby setting the lower entropy to zero. Consequently, it may fail to effectively represent different uncertainty components in a reliable manner. Moreover, this method has the potential to yield empty credal sets, meaning that even the prediction of the first-order predictor may be excluded from the credal set.

From an uncertainty quantification perspective and considering the measure in (12), it becomes apparent that set size may not always accurately reflect epistemic uncertainty. In fact, a credal set of a certain size positioned in the middle of simplex may contain less epistemic uncertainty than the same set located around a corner.

The observed dependency between quantified aleatoric and epistemic uncertainty in Figures 4 and 6 is partly due to the chosen measure in (12): Generally, higher epistemic uncertainty implies a larger uncertainty interval, resulting in a lower value for aleatoric uncertainty. Thus, this dependency is inevitable in the high EU region. Another unintended behavior observed in the **TV**, **KL**, and **WS** approaches is that low epistemic uncertainty coincides *solely* with high aleatoric uncertainty. This issue arises mainly from the lack of adaptivity in the credal sets constructed by first-order predictors. In Appendix D, we propose a method to enhance the adaptivity of these approaches using the concept of normalized nonconformity functions [33].

5 Limitations

The methods we propose are promising but still subject to certain limitations. For instance, for our methods to perform effectively, we require first-order data, at least for calibration. While such data is becoming increasingly available in practice, it is not accessible for all datasets and domains. Besides, for our generalization to the case of imprecise first-order data, practical implementation depends on a meaningful choice of the hyperparameters ϵ and δ to ensure inference that is both valid and efficient. When labels are based on relative frequencies (as in the case of multiple annotators), classical statistical methods might apply. However, determining an appropriate choice of ϵ and δ for broader practical problems remains an open issue. Another challenge lies in representing credal sets as subsets of the probability simplex. Although the credal sets can always be precisely described by equation (10), for the nonconformity functions used in this work, there are no closed-form equations for the resulting credal sets. Instead, the sets are represented implicitly through the nonconformity threshold. Numerical approximation is feasible but generally limited to scenarios with a small number of classes. The representation issue is also connected to computing uncertainty measures for quantifying epistemic and aleatoric uncertainty in a credal set that involves the computation of specific set characteristics [26, 41].

6 Conclusion and Future Work

Conformal credal set prediction connects machine learning with imprecise probability theory and offers a novel data-driven approach to constructing predictions that effectively capture both aleatoric and epistemic uncertainty. Thereby, it provides the basis of a new approach to reliable, uncertainty-aware machine learning. Leveraging the inherent validity of the conformal prediction framework, our conformalized credal sets are assured to cover the ground truth distributions with high probability. We have explored different nonconformity functions within this novel setting and evaluated their performance through numerical experiments.

There are several promising directions for future work. One avenue is to extend our method to standard (zero-order) data. While it has been shown that learning a second-order predictor from such data is challenging [8, 9], whether similar problems apply to credal predictors constructed from such data is not yet fully clear. Additionally, exploring other nonconformity functions that lead to closed-form solutions for credal sets or enhance efficiency and reduce uncertainty is worth considering. Finally, constructing a set of labels from our proposed credal sets presents an intriguing opportunity for further research, especially as such sets can provide more information compared to standard conformal prediction sets.

The broader impact of our work is the advancement of Machine Learning models towards better uncertainty-aware predictions, and we do not foresee any negative societal impacts.

 https://doi.org/10.52202/079017-3714

Acknowledgment

Alireza Javanmardi was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): Project number 451737409.

References

[1] J. Abellán and S. Moral. Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems*, 2003.

[2] J. Abellán, G. J. Klir, and S. Moral. Disaggregated total uncertainty measure for credal sets. *International Journal of General Systems*, 35, 2006.

[3] G. Abercrombie, V. Rieser, and D. Hovy. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. *arXiv preprint arXiv:2301.10684*, 2023.

[4] A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

[5] L. Aroyo and C. Welty. The three sides of crowdtruth. *Human Computation*, 1, 2014.

[6] L. Aroyo and C. Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36, 2015.

[7] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49, 2021.

[8] V. Bengs, E. Hüllermeier, and W. Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[9] V. Bengs, E. Hüllermeier, and W. Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[10] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[11] C. Bhagavatula, R. Le Bras, C. Malaviya, K. Sakaguchi, A. Holtzman, H. Rashkin, D. Downey, W.-t. Yih, and Y. Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations*, 2019.

[12] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015.

[13] L. Cella and R. Martin. Validity, consonant plausibility measures, and conformal prediction. *International Journal of Approximate Reasoning*, 141, 2022.

[14] G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 2008.

[15] G. Corani, A. Antonucci, and M. Zaffalon. Bayesian networks with imprecise probabilities: Theory and application to classification. *Data Mining: Foundations and Intelligent Paradigms: Volume 1: Clustering, Association and Classification*, 2012.

[16] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*. PMLR, 2018.

[17] T. G. Dietterich and J. Hostetler. Conformal prediction intervals for markov decision process trajectories. *arXiv preprint arXiv:2206.04860*, 2022.

[18] A. Dumitrache, F. Mediagroep, L. Aroyo, and C. Welty. A crowdsourced frame disambiguation corpus with ambiguity. In *Proceedings of NAACL-HLT*, 2019.

[19] S. Feldman, B.-S. Einbinder, S. Bates, A. N. Angelopoulos, A. Gendler, and Y. Romano. Conformal prediction is robust to dispersive label noise. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204. PMLR, 2023.

[20] A. Fisch, T. Schuster, T. Jaakkola, and R. Barzilay. Conformal prediction sets with limited false positives. In *International Conference on Machine Learning*. PMLR, 2022.

[21] S. C. Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54, 1996.

[22] E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110, 2021.

[23] E. Hüllermeier, S. Destercke, and M. H. Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Uncertainty in Artificial Intelligence*. PMLR, 2022.

[24] A. Javanmardi, Y. Sale, P. Hofman, and E. Hüllermeier. Conformal prediction with partially labeled data. In *Conformal and Probabilistic Prediction with Applications*. PMLR, 2023.

[25] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[26] G. Klir and M. Wierman. *Uncertainty-based information: elements of generalized information theory*, volume 15. Springer Science & Business Media, 1999.

[27] A. Kovashka, O. Russakovsky, L. Fei-Fei, K. Grauman, et al. Crowdsourcing in computer vision. *Foundations and Trends® in computer graphics and Vision*, 10, 2016.

[28] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[29] J. Lienen, C. Demir, and E. Hüllermeier. Conformal credal self-supervised learning. In *Conformal and Probabilistic Prediction with Applications*. PMLR, 2023.

[30] H. Linusson, U. Johansson, and H. Boström. Efficient conformal predictor ensembles. *Neurocomputing*, 397, 2020.

[31] Y. Nie, X. Zhou, and M. Bansal. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020.

[32] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning*. Springer, 2002.

[33] H. Papadopoulos, A. Gammerman, and V. Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, 2008.

[34] E. Pavlick and T. Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7, 2019.

[35] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626, 2019.

[36] D. Reidsma and R. op den Akker. Exploiting 'subjective'annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, 2008.

[37] Y. Romano, E. Patterson, and E. Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[38] Y. Romano, M. Sesia, and E. Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 2020.

[39] P. Röttger, B. Vidgen, D. Hovy, and J. Pierrehumbert. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

[40] M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 2019.

[41] Y. Sale, M. Caprio, and E. Hüllermeier. Is the volume of a credal set a good measure for epistemic uncertainty? In *Uncertainty in Artificial Intelligence*. PMLR, 2023.

 https://doi.org/10.52202/079017-3714

[42] M. Schaekermann, E. Law, A. C. Williams, and W. Callaghan. Resolvable vs. irresolvable ambiguity: A new hybrid framework for dealing with uncertain ground truth. In *1st Workshop on Human-Centered Machine Learning at SIGCHI*, volume 2016, 2016.

[43] L. Schmarje, V. Grossmann, C. Zelenka, S. Dippel, R. Kiko, M. Oszust, M. Pastell, J. Stracke, A. Valros, N. Volkmann, et al. Is one annotation enough?-a data-centric image classification benchmark for noisy and ambiguous label estimation. *Advances in Neural Information Processing Systems*, 35, 2022.

[44] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[45] M. Sesia and Y. Romano. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34, 2021.

[46] M. H. Shaker and E. Hüllermeier. Aleatoric and epistemic uncertainty with random forests. In *International Symposium on Intelligent Data Analysis*. Springer, 2020.

[47] R. Snow, B. O'connor, D. Jurafsky, and A. Y. Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 2008.

[48] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*. IEEE, 2008.

[49] L. Steinberger and H. Leeb. Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801*, 2016.

[50] D. Stutz, K. D. Dvijotham, A. T. Cemgil, and A. Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2021.

[51] D. Stutz, A. T. Cemgil, A. G. Roy, T. Matejovicova, M. Barsbey, P. Strachan, M. Schaekermann, J. Freyberg, R. Rikhye, B. Freeman, et al. Evaluating ai systems under uncertain ground truth: a case study in dermatology. *arXiv preprint arXiv:2307.02191*, 2023.

[52] D. Stutz, A. G. Roy, T. Matejovicova, P. Strachan, A. T. Cemgil, and A. Doucet. Conformal prediction under ambiguous ground truth. *Transactions on Machine Learning Research*, 2023.

[53] A. Uma, D. Almanea, and M. Poesio. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 2022.

[54] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72, 2021.

[55] V. Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74, 2015.

[56] V. Vovk, I. Nouretdinov, V. Fedorova, I. Petej, and A. Gammerman. Criteria of efficiency for set-valued classification. *Annals of Mathematics and Artificial Intelligence*, 81, 2017.

[57] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer Nature, 2022.

[58] P. Walley. *Statistical reasoning with imprecise probabilities*, volume 42. Springer, 1991.

[59] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1996.

[60] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018.

[61] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[62] M. Zaffalon. The naive credal classifier. *Journal of statistical planning and inference*, 2002.

[63] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Haberle, Y. Hua, R. Huang, et al. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8, 2020.

Appendix A: Extended Related Work

Credal sets are widely used as models for representing uncertainty, notably within the domain of imprecise probabilities [58]. As already mentioned, they can represent both types of uncertainty, aleatoric and epistemic. In the context of data analysis and statistical inference, credal sets are often used as robust models of prior information, namely for modeling imprecise information about the prior in Bayesian inference [59].

In machine learning, credal sets have been used for generalizing some of the standard methods, including naive Bayes [14, 62], Bayesian networks [15], and decision trees [1]. Typically, these approaches generalize simple frequentist inference to robust Bayesian inference, making use of an imprecise version of the Dirichlet model (a conjugate prior for the multinomial distribution). Compared to our approach, these methods are learning on standard (zero-order) training data. Moreover, despite representing uncertainty in predictions, they do not provide any formal guarantees.

Conformal prediction [57], briefly introduced in Section 2.2, has recently gained attention for various applications in machine learning, especially for classification tasks [4, 20, 38, 40, 50]. These methods mostly focus on split conformal prediction using a held-out calibration set [32], overcoming computational limitations of earlier transductive or bagging approaches [7, 30, 49, 55, 57]. While tackling classification tasks, our method for constructing conformal credal sets has more similarity with conformal regression [37, 45], particularly in multivariate settings [17], as we essentially conformalize the simplex space of categorical distributions. Our nonconformity scores differ, however, in that they are specific for distributions rather than considering general multivariate spaces. This work also relates to work on appropriate measures of inefficiency [56] as measuring the inefficiency of our conformal credal sets is non-trivial. Most closely related to our work is the recent work by Stutz et al. [52], who consider conformal prediction in settings with high aleatoric uncertainty. However, we explicitly target the construction of conformal credal sets, while Stutz et al. [52] mainly focus on constructing confidence sets of classes. The connection between CP and credal sets has also been explored by Cella and Martin [13] and Lienen et al. [29]. However, their emphasis lies on standard (zero-order) data, which fails to represent uncertainty truthfully. Furthermore, the sets generated by their methods are consistently confined to the simplex corners (around degenerate distributions) and display notably conservative behavior.

First-order data. In settings with high aleatoric uncertainty, labeling each example with a single, unique class is clearly insufficient. In practice, this is typically captured by high disagreement among annotators – a problem particularly common in natural language tasks [3, 5, 6, 18, 34, 36, 39, 42] or even in computer vision [10, 31, 35, 43, 63]. Handling this disagreement has received considerable attention lately [54] as it offers to go beyond this zero-order information. For example, recent work on evaluation with disagreeing annotators [51] argues the use of these annotations to get approximate first-order information for evaluation. This approach is becoming more and more viable with crowdsourcing tools [27, 47, 48] being an integral component of the benchmark, making multiple annotations per data instance more accessible. We follow a similar approach in our construction of conformal credal sets.

Appendix B: Proof of Theorem 3.2

Proof. Let A denote the event $f(\boldsymbol{x}_{\text{new}}, \boldsymbol{\lambda}^{\boldsymbol{x}_{\text{new}}}) < q + \epsilon$ and $\tilde{A}$ the event $f(\boldsymbol{x}_{\text{new}}, \tilde{\boldsymbol{\lambda}}^{\boldsymbol{x}_{\text{new}}}) < q$. We have

$$\begin{aligned} P(A) &\geq P(A \wedge \tilde{A}) \\ &= P(\tilde{A}) \cdot P(A \,|\, \tilde{A}) \\ &= P(\tilde{A}) \cdot (1 - P(\neg A \,|\, \tilde{A})) \end{aligned}$$

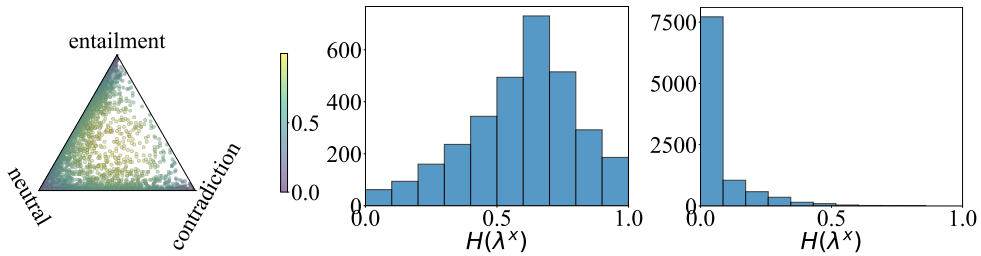
Since $\neg A$ means that $f(\boldsymbol{x}_{\text{new}}, \boldsymbol{\lambda}^{\boldsymbol{x}_{\text{new}}}) \geq q + \epsilon$, the conditional event $\neg A \,|\, B$ implies a violation of the closeness condition in (11), wherefore the probability $P(\neg A \,|\, B)$ is upper-bounded by δ according to (11). Therefore, noting that $P(\tilde{A}) \geq 1 - \tilde{\alpha}$ is the standard guarantee by CP,

$$\begin{aligned} P(A) &\geq P(\tilde{A}) \cdot (1 - P(\neg A \,|\, \tilde{A})) \\ &\geq (1 - \tilde{\alpha}) \cdot (1 - \delta) \\ &= 1 - \alpha \,. \end{aligned}$$

□

https://doi.org/10.52202/079017-3714

Appendix C: Real-World Datasets and Models Overview



(a) Scatter plot of the ChaosNLI distributions.

(b) Histogram of the entropies of ChaosNLI distributions.

(c) Histogram of the entropies of CIFAR10-H distributions.

Figure 8: General overview of real-world datasets.

ChaosNLI [31] (License: CC BY-NC 4.0 DEED) is an English Natural Language Inference (NLI) dataset that captures the inherent variability in human judgments of textual entailment. Here, the classes are *entailment*, *neutral*, and *contradiction* for each premise-hypothesis pair. Instances in this dataset are selected from the development sets of SNLI [12], MNLI [60], and AbductiveNLI [11], for which the majority vote was less than three among the five human annotators. These instances were then given to 100 independent humans for annotation, given strict annotation guidelines. We combine the chaos-SNLI and chaos-MNLI subsets, resulting in a dataset of 3113 datapoints. For model training, we leverage a language model from the Hugging Face `transformers` library [61], initially trained on SNLI and MultiNLI datasets for classification tasks[3]. We utilize the last hidden layer output of this model to embed the premise-hypothesis pairs from all 3113 instances into 768-dimensional vectors, serving as inputs for the learning model. To split the data, we randomly select 500 instances for calibration, 500 for testing, and the remaining for training.

As for the learner, we employ a deep neural network consisting of three hidden layers with 256, 64, and 16 units, utilizing ReLU as the activation function. Prior to the output layer, a dropout layer with a rate of 0.3 is incorporated. The same model architecture serves both first- and second-order predictors, differing only in the activation functions of the output layers. For the first-order predictor, softmax is used, while for the second-order predictor, ReLU is employed. Learning is facilitated using the Adam optimizer with a learning rate of 10^{-4}, utilizing cross-entropy as the loss function for the first-order predictor and negative log-likelihood for the second-order predictor.

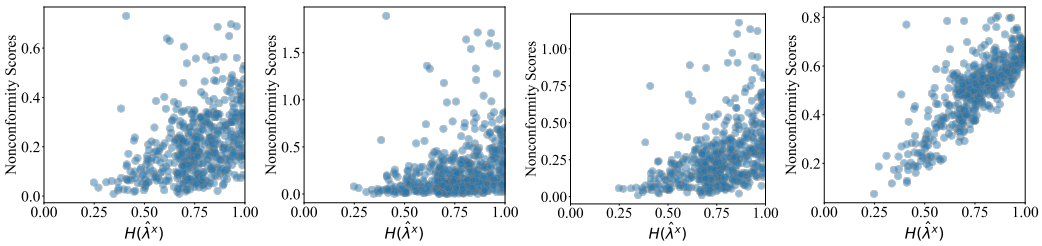
CIFAR10-H [35] (License: CC BY-NC-SA 4.0 DEED) is a dataset of soft labels that capture human perceptual uncertainty for the 10000 images of CIFAR-10 test set [28]. A total of $511,400$ human classifications were gathered via Amazon Mechanical Turk, excluding participants who performed poorly on obvious images. On average, each image received 51 judgments, with the number of judgments per image ranging from 47 to 63. As shown in Figure 8, the histogram of the entropies of CIFAR-10H distributions shows that this dataset is less ambiguous compared to ChaosNLI. For this dataset, we don't perform any model training and instead use a model pre-trained on the CIFAR-10 training images[4]. We utilize the model's predicted distributions along with the CIFAR-10H distributions to construct credal sets, focusing solely on first-order approaches. We randomly select 1000 instances for testing and the rest for calibration.

Appendix D: Enhancing Adaptivity in First-Order Predictor Credal Sets

As mentioned earlier, with methods based on the first-order predictor, once the quantile is determined during calibration, all distributions within a certain distance are included in the set for any given point at prediction time. This indicates that these methods lack adaptivity and do not account for the local heterogeneity of the data. Given the uncertainty measure in (12), for a given quantile or

[3]The model can be found at `https://huggingface.co/cross-encoder/nli-deberta-base` with Apache License 2.0

[4]The pre-trained model can be found at `https://github.com/huyvnphan/PyTorch_CIFAR10` with MIT license.



(a) **TV** (b) **KL** (c) **WS** (d) **Inner**

Figure 9: Scatter plots of calibration scores versus the entropy of the first-order predictor's output for different nonconformity scores for the ChaosNLI dataset.

set size, when the model's predicted distribution is near the simplex corner, the difference between the highest and lowest entropies is much greater than when the predicted distribution is located in the middle. That is why, in Figure 4, small epistemic uncertainty is observed only for high aleatoric cases for methods based on the first-order predictor. To handle this issue, one approach involves utilizing the normalized nonconformity score by dividing the score of each point by (a measure of) its dispersion at that point [33]. As depicted in Figure 9, we've noticed a direct correlation between the score's dispersion and the entropy of the predicted distributions for the ChaosNLI dataset; specifically, higher entropy corresponds to greater variability in the scores. Therefore, we propose the following normalized score:

$$\tilde{f}_1(\boldsymbol{x}, \boldsymbol{\lambda}^{\boldsymbol{x}}) := \frac{f_1(\boldsymbol{x}, \boldsymbol{\lambda}^{\boldsymbol{x}})}{H(\hat{\boldsymbol{\lambda}}^{\boldsymbol{x}}) + \tau}\,, \tag{13}$$

where $H(\hat{\boldsymbol{\lambda}}^{\boldsymbol{x}})$ denotes the entropy of the predicted distribution at point $\boldsymbol{x}$ and τ is a hyperparameter that prevents division by zero and controls the influence of $H(\hat{\boldsymbol{\lambda}}^{\boldsymbol{x}})$ on the scores (the higher the τ, the lower the influence). The credal set for $\boldsymbol{x}_{\text{new}}$ will be constructed as

$$\left\{ \boldsymbol{\lambda} \in \Delta^K \mid f_1(\boldsymbol{x}_{\text{new}}, \boldsymbol{\lambda}) \leq q(\tilde{\mathcal{E}}_1, \alpha')(H(\hat{\boldsymbol{\lambda}}^{\boldsymbol{x}_{\text{new}}}) + \tau) \right\},$$

where $\tilde{\mathcal{E}}_1$ is the set of normalized nonconformity scores. Figure 10 compares the scatter plots of AU versus EU for all first-predictor-based models with their adaptive counterparts with $\tau = 0.1$. The adaptive approach decreases the average epistemic uncertainty. For example, the positions of the three instances of Figure 5 moved closer to their corresponding values of the **SO** method.

https://doi.org/10.52202/079017-3714

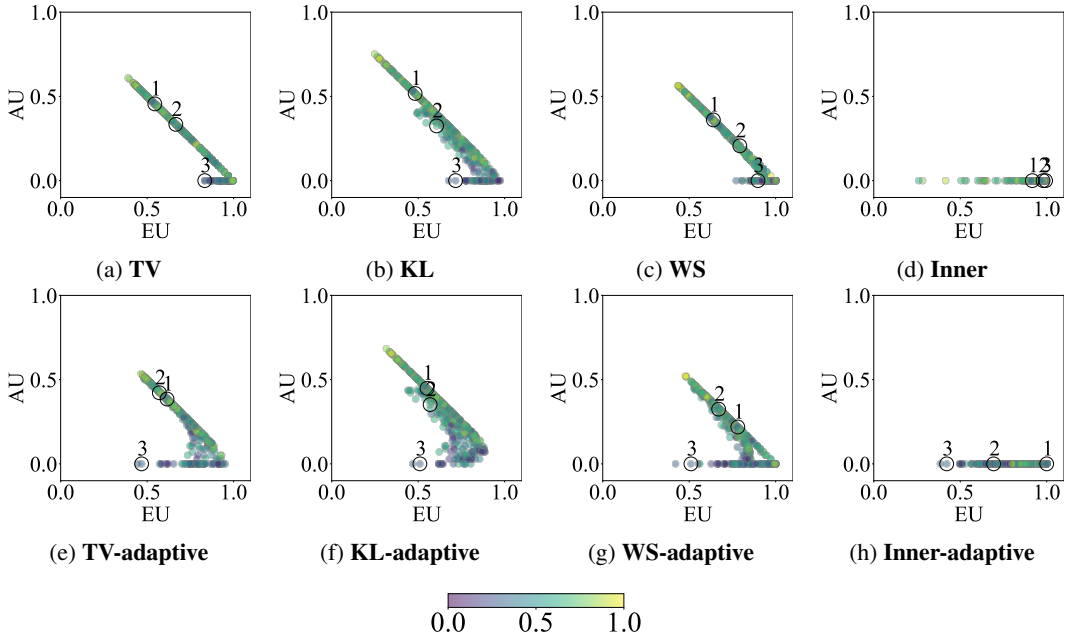


Figure 10: Scatter plots of aleatoric versus epistemic uncertainty using various credal set predictors for the ChaosNLI dataset with $\alpha = 0.2$. The colors indicate the entropy of the ground truth distribution. The three cases numbered 1 to 3, correspond to the first to third rows of instances shown in Figure 5.

Appendix E: Synthetic Data

Throughout this paper, we generate synthetic data for a K-class classification task as follows: we consider d-dimensional features $X \in \mathbb{R}^d$, where $X_1, \ldots, X_d$ are independent standard normal random variables. Subsequently, we generate a random matrix $\beta \in \mathbb{R}^{d \times K}$, with its elements drawn independently from the standard normal distribution. To define the ground truth probability over the classes for object X, we use the following formulation:

$$\lambda_k^{\boldsymbol{x}} := \mathbb{P}(Y = k|\boldsymbol{x}) = \frac{Z_k(\boldsymbol{x})}{\sum_j Z_j(\boldsymbol{x})}, \tag{14}$$

where $Z(\boldsymbol{x}) := \exp(\boldsymbol{x}^\top \beta)$. As for the learner (both first-order and second-order predictors), in the subsequent experiments, we utilize the same deep neural network described in Appendix C for the ChaosNLI dataset, with the input layer adopting the d-dimensional feature vector.

E.1 Illustrative Example

To illustrate the primary concept of credal sets and uncertainty decomposition, we present a straightforward yet intuitive example. Given the data-generating process in (14), we generate 3000 samples with $d = 10$ and $K = 3$ solely to facilitate the visualization of the credal sets, splitting them equally between training, calibration, and testing datasets. Initially, we train a model with only 10 points, gradually increasing to 50, 100, 500 and 1000. At each stage, we calibrate the model using all the calibration data points and evaluate its performance on the test points. From the test data, we select three examples characterized by high entropy (almost uniform distribution), relatively high entropy (uniform across two classes), and low entropy (almost a Dirac distribution). We then plot their credal sets alongside lower and upper entropies for varying numbers of training data points.

Figure 11 provides intriguing insights! As evident in the first column, where the model was trained with only 10 data points, there exists a significant epistemic uncertainty across all three cases. This indicates that none of the predictions are reliable, and it's challenging to predict a single label for any of the given examples. Moving to the third column, we observe a significant reduction in epistemic uncertainties. In the second example (second row), it indicates that the total uncertainty is primarily due to epistemic uncertainty, whereas for the other two cases, it's a result of high aleatoric uncertainty with some epistemic uncertainty. This suggests that gathering more data (and reducing the total uncertainty) may not significantly enhance the capability to predict a single label for the latter cases. Figure 12 illustrates the evolution of scatter plots of AU versus EU for all first-predictor methods as the number of training data increases.

E.2 Experiments with Imprecise First-Order Data

We considered another set of experiments with synthetic data to illustrate the impact of imprecise first-order data, particularly to showcase the behavior of the proposed credal sets when we only have access to an approximation of the ground truth distributions. Our experiment revolves around a K-class classification task with $K \in \{3, 4, 6, 8, 10\}$. For each K, we generate $N = 1500$ samples with $d = 10$ using (14) to construct the datasets $\mathcal{D}^K = \{(\boldsymbol{x}_i, \boldsymbol{\lambda}^{\boldsymbol{x}_i})\}_{i=1}^N$. To obtain imprecise versions of $\mathcal{D}^K$, we employ a sampling approach. Specifically, we independently sample each distribution $\boldsymbol{\lambda}^{\boldsymbol{x}_i}$ m times and utilize relative frequencies to create its noisy counterpart $\tilde{\boldsymbol{\lambda}}_m^{\boldsymbol{x}_i}$. We represent the resulting dataset as $\mathcal{D}_m^K = \{(\boldsymbol{x}_i, \tilde{\boldsymbol{\lambda}}_m^{\boldsymbol{x}_i})\}_{i=1}^N$. We repeat this process four times with $m \in \{1, 5, 10, 100\}$.

Given each dataset $\mathcal{D}_m^K$, we randomly partition data points into 1300 training, 100 calibration, and 100 test instances and perform the proposed methodologies accordingly. Again, we repeat this process ten times with different random seeds for each dataset $\mathcal{D}_m^K$. Due to the computational complexity in calculating efficiency for cases with $K > 3$, we utilize the quantile of the calibration nonconformity scores as an efficiency metric. In Figure 13, we represent the overall result under different K and m values. It can be observed that the coverage is fulfilled across almost all scenarios, including $m = 1$ with degenerate distributions. This observed behavior is somewhat intuitive. The model endeavors to learn the underlying probabilistic relationship between X and Y, even given the noisy data [5]. Consequently, during calibration with noisy instances, the nonconformity scores of noise-free

[5] Of course, this holds under some reasonable assumptions on noise.

https://doi.org/10.52202/079017-3714

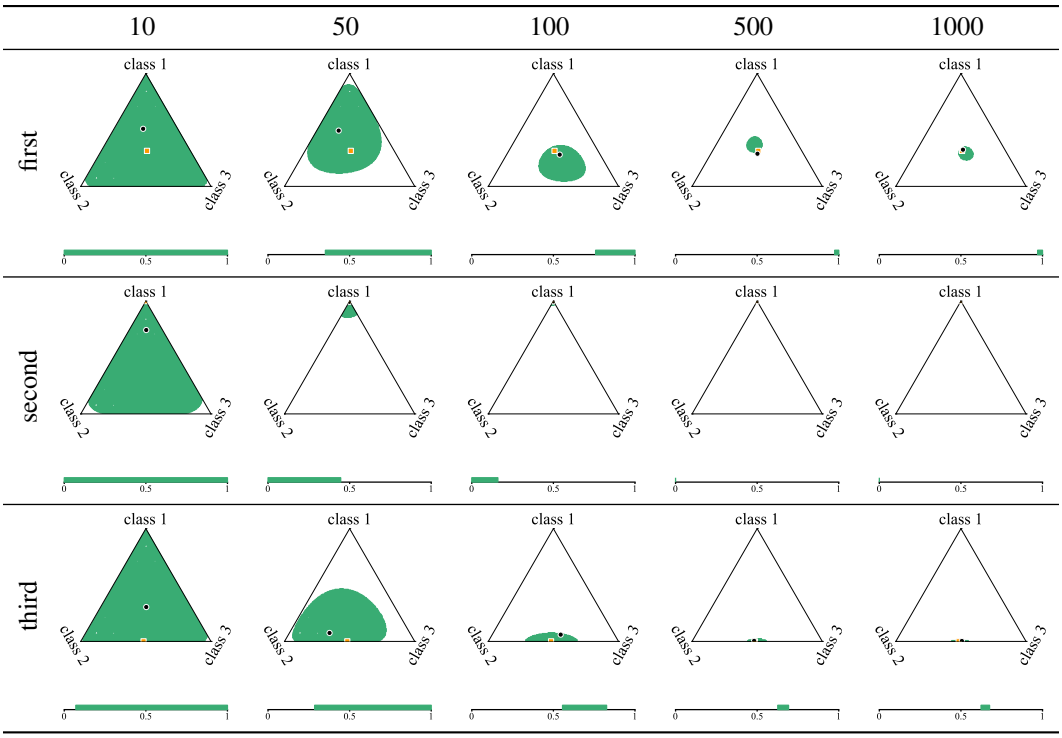


Figure 11: Credal sets generated by the **KL** method for three examples of illustrative synthetic data with $\alpha = 0.1$. Each row represents one of three examples, with columns illustrating the evolution of the predicted credal set based on varying numbers of training data points. The lower and upper entropy of each credal set is displayed below it.

instances are mostly upper-bounded by the scores of their noisy counterparts, resulting in more conservative sets that effectively cover the ground-truth distributions. It can also be seen that the quantile of the nonconformity scores shrinks as m increases. In Figure 14, we illustrate the evolution of the credal sets as m changes from 1 to 100 for different nonconformity functions when $K = 3$. For this case, the full comparison of efficiency and coverage across various nonconformity functions is provided in Figure 15.

Appendix F: Experiments Compute Resources

For all experiments, we used an Intel(R) Core(TM) i7-11800H CPU with 16.0 GB of RAM. Model training and calibration steps are quite fast and take only a few minutes to run. In constructing the credal sets, we use high-resolution simplex discretization and determine whether each distribution from the discretized simplex belongs to the credal set via exhaustive search. For the **WS** method, this process can take up to a few seconds per data instance, while for all other methods, it takes less than a second. The uncertainty quantification was also completed in less than a second.

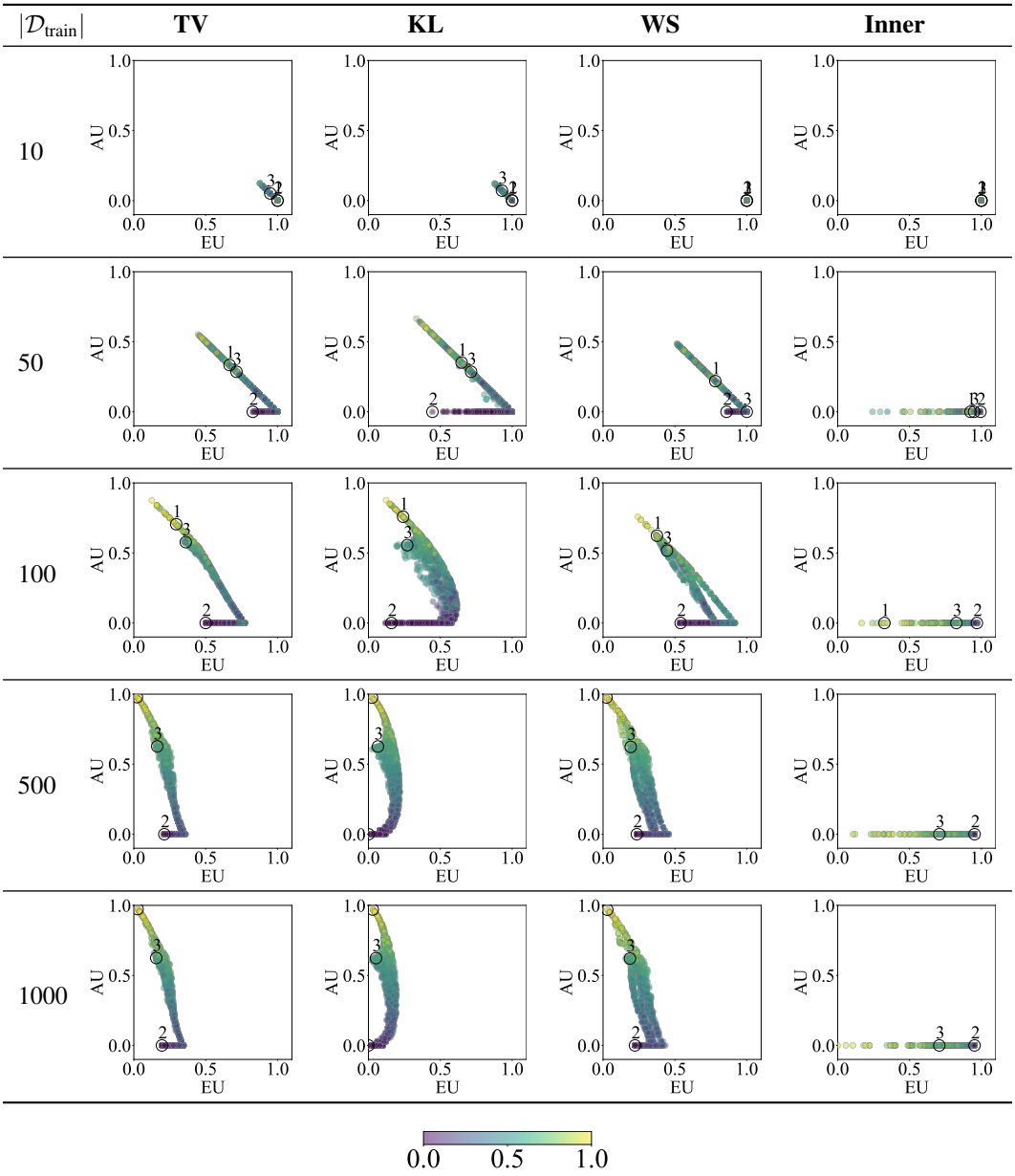
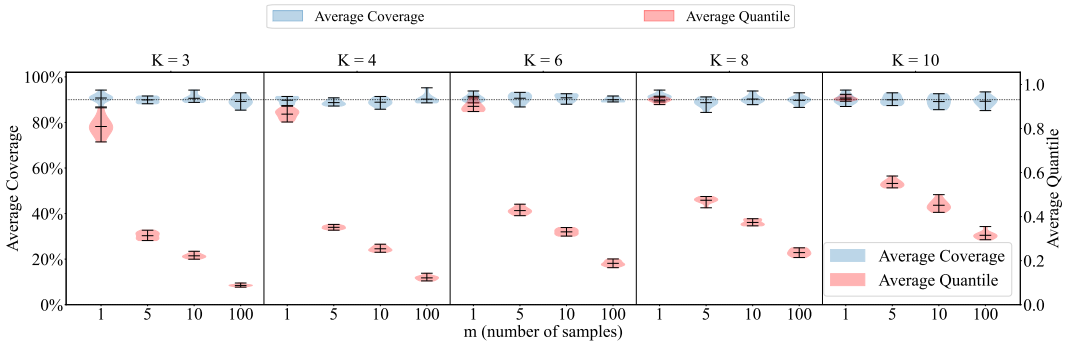
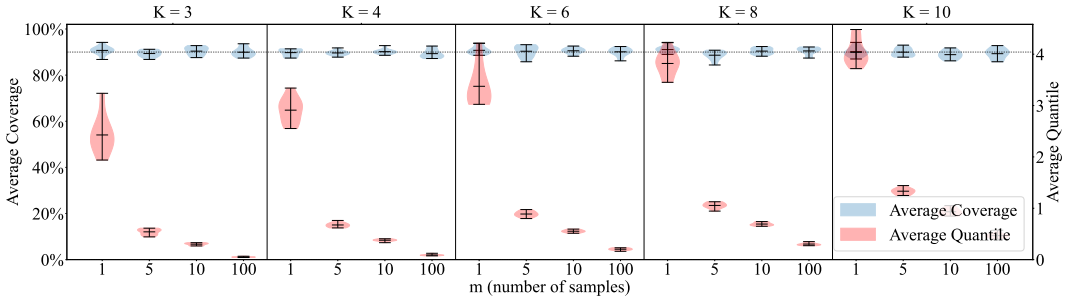


Figure 12: Scatter plots of aleatoric versus epistemic uncertainty using various credal set predictors given different numbers of training data for illustrative synthetic data with $\alpha = 0.1$. The colors indicate the entropy of the ground truth distribution. The three cases, numbered 1 to 3, correspond to the first to third rows of instances shown in Figure 11.

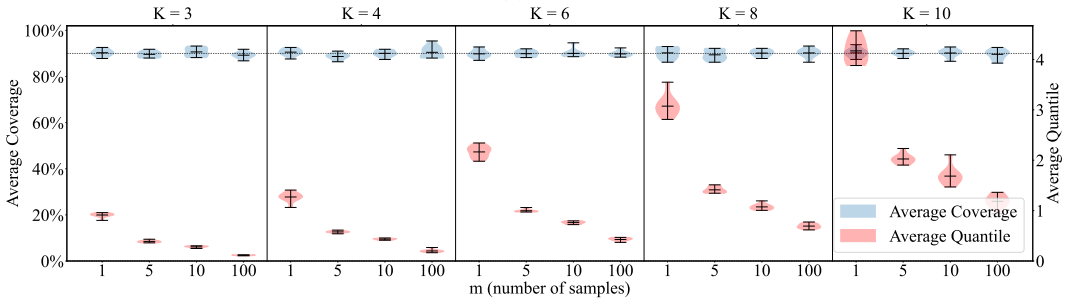
https://doi.org/10.52202/079017-3714



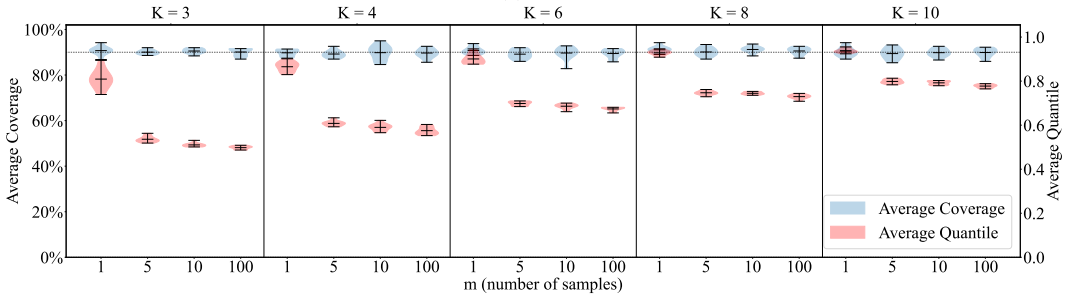
(a) **TV**



(b) **KL**



(c) **WS**



(d) **Inner**

Figure 13: Coverage and quantile results for synthetic data with imprecise first-order distributions, where the ground truth distributions are approximated by observing m samples from them. The horizontal dashed lines indicate the nominal coverage level $1 - \alpha = 0.9$.

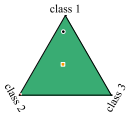
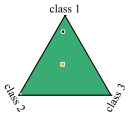
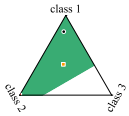
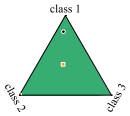
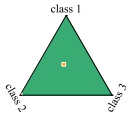
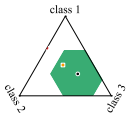
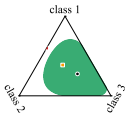
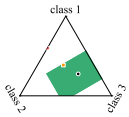
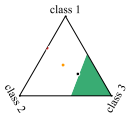
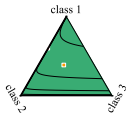
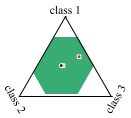
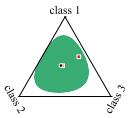
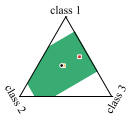
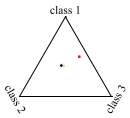
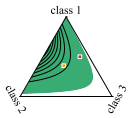
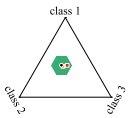
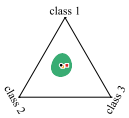
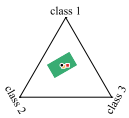
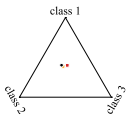
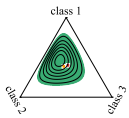
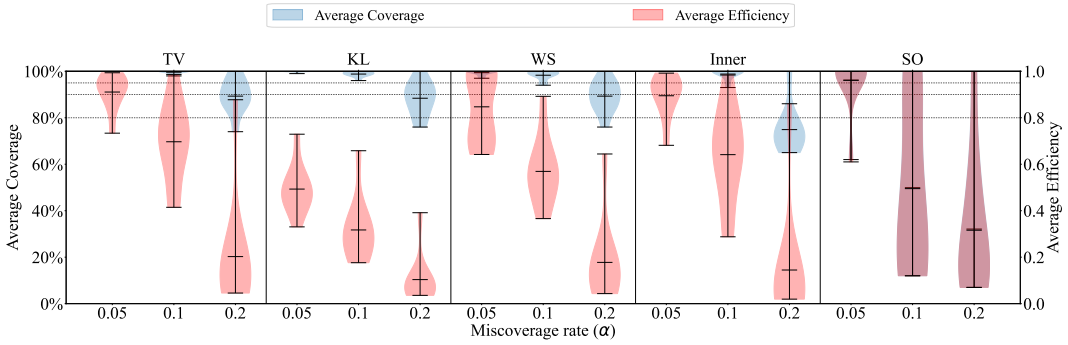
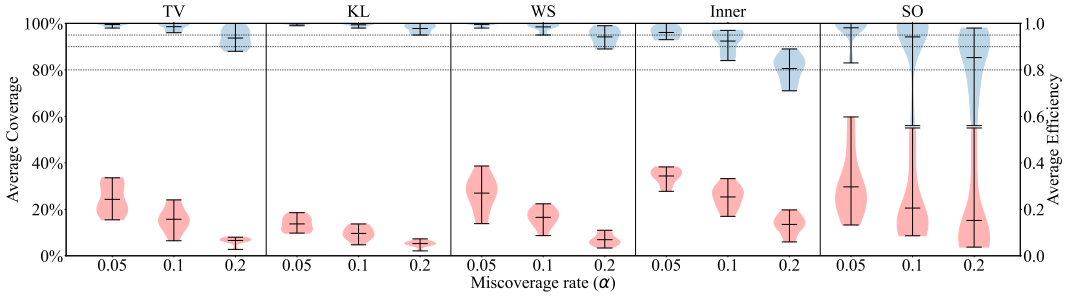
m	TV	KL	WS	Inner	SO
1	 1.00	 1.00	 0.70	 1.00	 1.00
5	 0.46	 0.67	 0.39	 0.22	 1.00
10	 0.64	 0.58	 0.68	 0.00	 0.74
100	 0.06	 0.09	 0.09	 0.00	 0.43

Figure 14: Credal sets derived using various credal set predictors for an example from synthetic data with imprecise first-order distributions. Rows correspond to the number of samples utilized for distribution estimation. The ground truth distribution is marked by an orange square, and its noisy versions are denoted by red squares. In cases employing a first-order learner (first four columns), model predictions are denoted by black circles. The predicted second-order distributions are illustrated via contour plots in the last column, where a second-order learner is employed. The miscoverage rate is $\alpha = 0.05$, and the efficiency of each credal set is indicated below it.

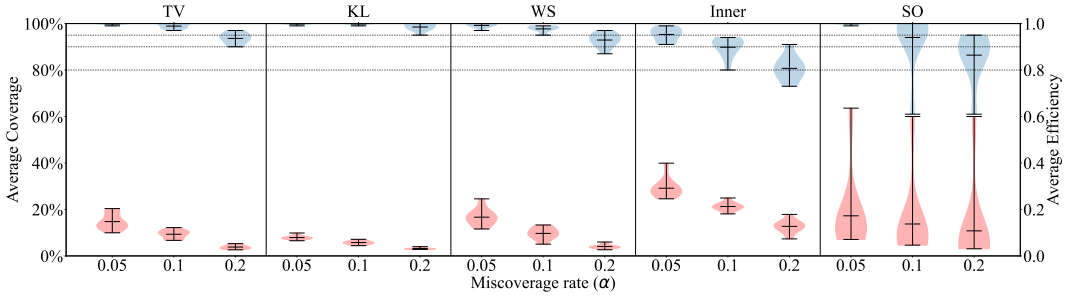
https://doi.org/10.52202/079017-3714



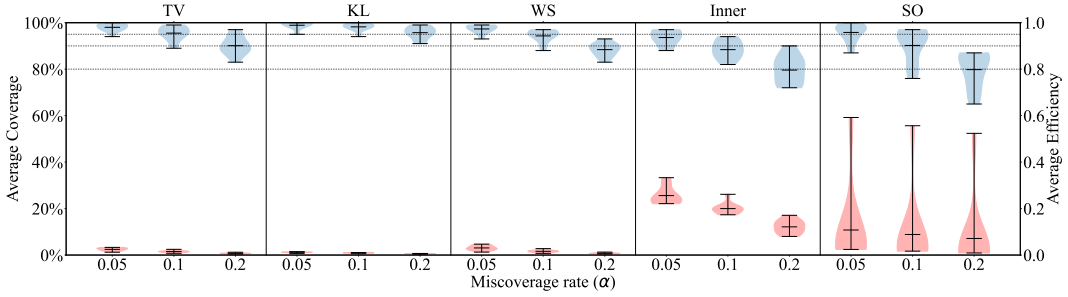
(a) $m = 1$.



(b) $m = 5$.



(c) $m = 10$.



(d) $m = 100$.

Figure 15: Coverage and efficiency results of different nonconformity functions applied on the synthetic data with imprecise first-order distributions ($K = 3$). The horizontal dashed lines indicate the nominal coverage levels.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contribution: Providing a data-driven method to construct credal sets using conformal prediction, assuming access to probabilistically labeled data.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-3714

Answer: [Yes]

Justification: The proof for Theorem 3.1 is not provided, as it is essentially a restatement of the conformal prediction validity theorem adjusted to our setting with the reference provided. As for Theorem 3.2, its proof can be found in Appendix B. All assumptions have been stated.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All experiment details, including models, data, etc., are provided in the paper or its Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Please see https://github.com/alireza-javanmardi/conformal-credal-sets.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details are provided in the paper or its Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Where applicable, we do so. For instance, see Figure 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

https://doi.org/10.52202/079017-3714

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide a broader impact statement in the conclusion of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our methods do not pose a risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provide licenses and citations to all used datasets, models, and libraries.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

https://doi.org/10.52202/079017-3714

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: We do not introduce any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.