Quantifying Aleatoric and Epistemic Uncertainty: A Credal Approach

Paul Hofman [1 2] **Yusuf Sale** [1 2] **Eyke Hüllermeier** [1 2]

Abstract

Uncertainty representation and quantification are paramount in machine learning, especially in safety-critical applications. In this paper, we propose a novel framework for the quantification of aleatoric and epistemic uncertainty based on the notion of credal sets, i.e., sets of probability distributions. Thus, we assume a learner that produces (second-order) predictions in the form of sets of probability distributions on outcomes. Practically, such an approach can be realized by means of ensemble learning: Given an ensemble of learners, credal sets are generated by including sufficiently plausible predictors, where plausibility is measured in terms of (relative) likelihood. We provide a formal justification for the framework and introduce new measures of epistemic and aleatoric uncertainty as concrete instantiations. We evaluate these measures both theoretically, by analysing desirable axiomatic properties, and empirically, by comparing them in terms of performance and effectiveness to existing measures of uncertainty in an experimental study.

1. Introduction

Understanding and handling uncertainty is a fundamental challenge in machine learning (ML) and artificial intelligence (AI) research. Due to the intrinsic complexities and variability of real-world data, coupled with the probabilistic nature of many ML algorithms, the latter are often subject to various forms of uncertainty. Unless properly addressed, this uncertainty can pose substantial limitations to the reliability of ML systems, which is especially problematic in applications with stringent safety considerations, such as in medicine and the healthcare sector (Lambrou et al., 2010; Senge et al., 2014; Yang et al., 2009).

In the broader scope of the literature, a distinction between *aleatoric* and *epistemic* uncertainty is usually made (Hora, 1996). Aleatoric uncertainty originates from the inherent stochastic nature of the data-generating process, while epistemic uncertainty is due to the learners incomplete knowledge of this process. The latter can therefore be reduced by acquiring additional information, such as new observations. Conversely, aleatoric uncertainty, being a characteristic of the data-generating process itself, is non-reducible (Hüllermeier & Waegeman, 2021). Representing and quantifying these types of uncertainties have become pivotal in recent ML research, including Bayesian deep learning (Depeweg et al., 2018; Kendall & Gal, 2017), adversarial example detection (Smith & Gal, 2018), and data augmentation in Bayesian classification (Kapoor et al., 2022).

The Bayesian approach, in which the learner's epistemic uncertainty is represented in terms of a (second-order) probability distribution on the underlying model class, prevails the ML literature so far. Yet, this approach is not without criticism (Wimmer et al., 2023). In this paper, we propose an alternative framework for the quantification of aleatoric and epistemic uncertainty based on the notion of *credal sets*, i.e., sets of probability distributions. Thus, we assume a learner that produces predictions in the form of sets of probability distributions on outcomes. For representations of that kind, we introduce new measures of aleatoric, and epistemic uncertainty. We evaluate these measures both theoretically, by analysing desirable axiomatic properties, and empirically, by comparing them in terms of performance and effectiveness to existing measures of uncertainty in an experimental study.

Our point of departure is a taxonomy of different types of uncertainty-aware learning algorithms, based on Shaker & Hüllermeier (2021), in the next section. In Section 3, we address the problem of uncertainty quantification and recall basic uncertainty measures for the probabilistic and Bayesian case. The conceptual basis of our credal approach is developed in Section 4, and the problem of learning credal predictors is addressed in Section 5. Our experimental studies are presented in Section 6, prior to concluding in Section 8.

[1] Institute of Informatics, LMU Munich, Germany [2] Munich Center for Machine Learning (MCML), Germany. Correspondence to: Paul Hofman <paul.hofman@ifi.lmu.de>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria. Copyright 2024 by the author(s).

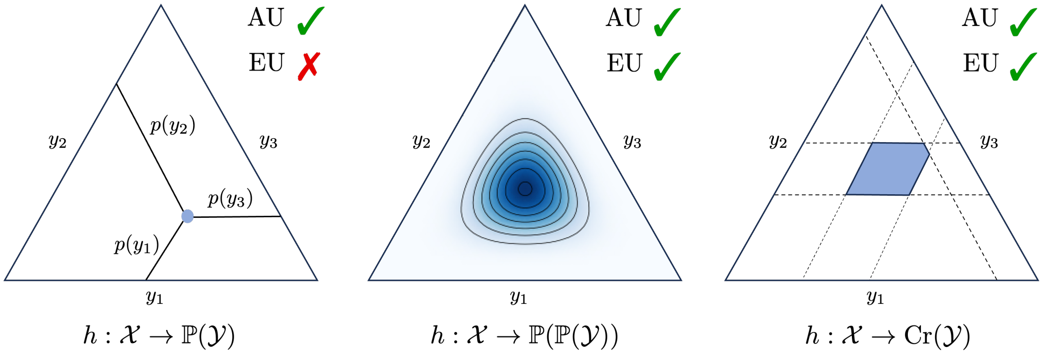
AU ✓
EU ✗
$p(y_2)$
$p(y_3)$
$p(y_1)$
y_1
y_2
y_3
$h : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$
AU ✓
EU ✓
y_1
y_2
y_3
$h : \mathcal{X} \to \mathbb{P}(\mathbb{P}(\mathcal{Y}))$
AU ✓
EU ✓
y_1
y_2
y_3
$h : \mathcal{X} \to \mathrm{Cr}(\mathcal{Y})$

Figure 1: Uncertainty awareness in multi-class classification, illustrated on the probability simplex for $\mathcal{Y} = \{y_1, y_2, y_3\}$. From *left* to *right*: Probabilistic agent (AU, but *no* EU awareness), Bayesian agent (AU *and* EU awareness), and Levi agent (AU *and* EU awareness).

2. Representation of Uncertainty

In this paper we consider the setting of classification. Let $\mathcal{X}$ be a (measurable) instance space and $\mathcal{Y}$ a label space, where we assume without loss of generality $\mathcal{Y} = \{y_1, \ldots, y_K\}$ for some $K \in \mathbb{N}$. Further, $\mathcal{D} = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ is a set of training data. The pairs $(\boldsymbol{x}_i, y_i)$ are realizations of random variables (X_i, Y_i), which are assumed to be independent and identically distributed (i.i.d.) according to some probability measure P on $\mathcal{X} \times \mathcal{Y}$. We also assume a hypothesis space $\mathcal{H}$ to be given, where each hypothesis $h \in \mathcal{H}$ is a probabilistic predictor $\mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$ that map instances $\boldsymbol{x}$ to a probability measure on $\mathcal{Y}$.

Given $\mathcal{D}$ and $\mathcal{H}$, the learner induces a hypothesis h, the predictions $h(\boldsymbol{x}) = \boldsymbol{\theta}_{\boldsymbol{x}}$ of which are considered as estimations of the ground-truth conditional distribution on $\mathcal{Y}$ given $X = \boldsymbol{x}$, denoted $\boldsymbol{\theta}^*_{\boldsymbol{x}}$. For simplicity, we will often omit the (query) instance $\boldsymbol{x}$ as a subscript. As $\mathbb{P}(\mathcal{Y})$, the class of probability measures on $\mathcal{Y}$, can be identified with the $(K-1)$-simplex Δ_K, both the estimate and the ground truth can be considered as elements of this simplex, i.e., as vectors $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top$ and $\boldsymbol{\theta}^* = (\theta^*_1, \ldots, \theta^*_K)^\top$, respectively, where θ^*_k is the true probability $P(Y = k \,|\, X = \boldsymbol{x})$ of the k^{th} class (given $\boldsymbol{x}$) and θ_k the estimate of this probability.

Given a query $\boldsymbol{x} \in \mathcal{X}$ for which a prediction is sought, different learning methods proceed on the basis of different types of information. Depending on how the predictive uncertainty, i.e., the uncertainty about a predicted probability distribution $\boldsymbol{\theta} \in \Delta_K$, is represented, we propose to distinguish four types of possible learners. Referring to the literature on decision under uncertainty, we call these learners probabilistic, Bayesian, and Levi agent, respectively.

2.1. Probabilistic Agents

A common practice in machine learning is to consider learners that fully commit to a single hypothesis $h \in \mathcal{H}$ and use this hypothesis to make predictions. Given $\boldsymbol{x} \in \mathcal{X}$ as input, such a learner will predict a single probability distribution $h(\boldsymbol{x}) = \boldsymbol{\theta}$, which is considered as an estimation of the ground-truth conditional probability $\boldsymbol{\theta}^*$. We call a learner of that kind a *probabilistic agent*. Such an agent's uncertainty about the outcome y is purely aleatoric. At the level of the hypothesis space, the agent pretends full certainty, and hence the absence of any epistemic uncertainty.

2.2. Bayesian Agents

Following the principle of (strict) Bayesianism as promoted by statisticians like Bruno de Finetti (De Finetti, 1980), a *Bayesian agent* expresses its belief as a probability distribution over the hypothesis space $\mathcal{H}$. Rather than committing to one specific hypothesis, the agent assigns a probability (density) $q(h)$ to each potential hypothesis $h \in \mathcal{H}$. Furthermore, when new data $\mathcal{D}$ is observed, the belief is updated by substituting this distribution with the posterior $q(h \,|\, \mathcal{D})$.

Since every $h \in \mathcal{H}$ gives rise to a different probabilistic prediction $h(\boldsymbol{x})$, a Bayesian agent's belief about the outcome $y \in \mathcal{Y}$ is represented by a probability distribution of probability distributions (*viz.* a second-order distribution). Formally, this distribution is the image of q under the mapping $\mathcal{H} \longrightarrow \Delta_K, h \mapsto h(\boldsymbol{x})$:

$$p(\boldsymbol{\theta}) = \int_{\mathcal{H}} [\![h(\boldsymbol{x}) = \boldsymbol{\theta}]\!] \, d\, q(h \,|\, \mathcal{D}). \tag{1}$$

Hence, $p(\boldsymbol{\theta})$ is the probability (density) of the probabilistic prediction $\boldsymbol{\theta}$.

In addition to the second-order distribution p, known as the posterior predictive distribution, a Bayesian agent generally also induces a representative (first-order) distribution through Bayesian model averaging:

$$\bar{\boldsymbol{\theta}} = \int_{\mathcal{H}} h(\boldsymbol{x}) \, d\, q(h \,|\, \mathcal{D}). \tag{2}$$

2.3. Levi Agents

Instead of representing model uncertainty in terms of a distribution p on $\mathcal{H}$, this uncertainty could also be characterised in terms of an arguably even simpler model, namely, a subset $Q \subseteq \mathcal{H}$ of hypotheses. According to this model, each $h \in Q$ is deemed a possible candidate predictor, whereas all $h \notin Q$ are excluded as being implausible. Under the mapping $h \mapsto h(\boldsymbol{x})$, the set Q directly translates into a set $C \subseteq \Delta_K$ of possible class distributions:

$$C = \{\boldsymbol{\theta} = h(\boldsymbol{x}) \,|\, h \in Q\}. \tag{3}$$

In the literature, such a set of probability distributions is also referred to as a *credal set* (Walley, 1991). The reasonableness of taking decisions on the basis of sets of probability distributions (and thus deviating from strict Bayesianism) has been advocated by decision theorists like Isaac Levi (Levi, 1974; 1980). Correspondingly, we call a learner of this kind a *Levi agent*.

In the realm of machine learning, an approach of this kind is somewhat in line with the model of version space learning (Mitchell, 1977), i.e., the subset Q can be seen as a kind of version space. From an uncertainty representation point of view, a set $Q \subseteq \mathcal{H}$ appears to provide weaker information compared to a distribution $q \in \mathbb{P}(\mathcal{H})$. While this is true in a sense, a set-based representation does also have advantages. In particular, many have argued that probability distributions are less suitable for representing *ignorance* in the sense of a lack of knowledge (Dubois et al., 1996). For example, if the uniform distribution is taken as a model of *complete ignorance*, as commonly done in probability theory, it is no longer possible to distinguish between a complete lack of knowledge and precise knowledge about the equal probability of all outcomes. Apart from that, one has to admit that the specification of a meaningful (prior) distribution is a difficult task in a machine learning setting, where $\mathcal{H}$ is a very complex space.

Another problem of a (second-order) probabilistic model is caused by the measure-theoretic grounding and additive nature of probability, which implies that the uniform distribution is not invariant under nonlinear transformation. As a consequence, even when starting with a uniform distribution on $\mathcal{H}$, suggesting complete lack of knowledge about the right predictor, the image (1) will normally not be uniform on Δ_K. In other words, even if the learner is supposedly ignorant about the right predictor, it will pretend a certain degree of informedness about the prediction in a point $\boldsymbol{x} \in \mathcal{X}$. Even worse, different predictive distributions (and degrees of uncertainty) will be obtained for different instances $\boldsymbol{x} \in \mathcal{X}$.

3. Uncertainty Quantification

According to our discussion so far, different types of learners represent their information or "belief" about the outcome y for an instance $\boldsymbol{x}$ in different ways, e.g., in terms of a probability distribution, a second-order distribution, or a credal set. What we are interested in is a quantification of the epistemic and aleatoric (and maybe total) uncertainty associated with such representations. More formally, we are seeking a measure of epistemic uncertainty, EU, and a measure of aleatoric uncertainty, AU. In the following, we discuss this problem for probabilistic and Bayesian agents, which prevail in the machine learning literature so far.

3.1. Probabilistic Agents: Entropy

Recall that a probabilistic agent represents predictive uncertainty in terms of a distribution $\boldsymbol{\theta}$ on $\mathcal{Y}$. The most well-known measure of uncertainty of a single probability distribution is the (Shannon) entropy, which, in our case of a discrete $\mathcal{Y}$, is given as

$$S(\boldsymbol{\theta}) := -\sum_{k=1}^{k} \theta_k \cdot \log_2 \theta_k, \tag{4}$$

with the convention that $0 \log 0 = 0$. This measure is the most straightforward candidate to quantify the aleatoric uncertainty of a probabilistic agent, i.e., $\mathrm{AU} = S(\boldsymbol{\theta})$. Since such an agent assumes it has exact knowledge of the predictive distribution, the epistemic uncertainty is 0.

The Shannon entropy can be justified axiomatically, and various axiomatic systems have been proposed in the literature (Csiszár, 2008).

3.2. Bayesian Agents: Conditional Entropy and Mutual Information

The Bayesian paradigm, which represents the learner's epistemic state by the posterior distribution over the hypothesis space, is widely accepted in the machine learning community. Most recently, the problem of predictive uncertainty estimation has attracted specific attention in the field of deep neural networks. Corresponding methods typically seek to quantify (total) uncertainty on the basis of the predictive posterior distribution on $\mathcal{Y}$. Additionally, epistemic uncertainty is viewed as a characteristic of the posterior $q(\cdot \,|\, \mathcal{D})$ or the resultant distribution $p(\cdot)$ on Δ_K: The less concentrated this distribution is, the higher the (epistemic) uncertainty of the learner.

A well-established method for quantifying and distinguishing between aleatoric and epistemic uncertainty, now prevalent in machine learning (Houlsby et al., 2011; Depeweg et al., 2018; Mobiny et al., 2017), relies on a classical information-theoretic result. This result states that the entropy of a random variable U can be decomposed into the conditional entropy of U given another random variable V and the mutual information between U and V. In our context, by treating the distribution on Δ_K as a random variable Θ (with distribution p) and the outcome as a random variable Y, we derive the following:

$$\mathrm{TU}(p) = S(Y) = S(\bar{\boldsymbol{\theta}}) = -\sum_{k=1}^{K} \bar{\theta}_k \cdot \log_2(\bar{\theta}_k), \quad (5)$$

$$\mathrm{AU}(p) = S(Y \mid \Theta) = -\int p(\boldsymbol{\theta}) \cdot S(\boldsymbol{\theta}) \, d\boldsymbol{\theta}, \quad (6)$$

$$\mathrm{EU}(p) = S(Y) - S(Y \mid \Theta) = \mathrm{I}(Y, \Theta), \quad (7)$$

where I is mutual information.

4. A Credal Approach

The Bayesian approach is clearly meaningful and has produced promising results in practice. Yet, it has recently been criticised for various reasons. Some conceptual problems of second-order distributions for representing the learner's belief have already been mentioned in Section 2.3. Moreover, Wimmer et al. (2023) provide a critical discussion of the quantification of uncertainty in terms of mutual information and conditional entropy (Section 3.2), and demonstrate that these measures may show counter-intuitive behavior. Much of the problems revealed have to do with with "averaging" of (first-order) uncertainty over the learner's belief, i.e., over the second-order distribution.

We take these criticism as motivation to elaborate on the Levi agent as an alternative to the Bayesian approach, which essentially means replacing the second-order distribution by a credal set for representing the learner's epistemic state. More specifically, we propose a novel class of uncertainty measures for credal sets, and show that these measures are highly flexible and enjoy desirable theoretical properties. An empirical analysis will then follow in the next section.

4.1. Existing Credal Uncertainty Measures

In the uncertainty literature, there is quite some work on defining uncertainty measures for credal sets and related representations. Here, aleatoric and epistemic uncertainty are also referred to as *conflict* (randomness, discord) and *non-specificity*, respectively (Yager, 1983). The standard uncertainty measure in classical possibility theory (where uncertain information is simply represented in the form of subsets $A \subseteq \mathcal{Y}$ of possible alternatives) is the Hartley measure (Hartley, 1928)

$$H(A) = \log(|A|), \quad (8)$$

Just like the Shannon entropy, this measure can be justified axiomatically (Rényi, 1970).

Given the insight that conflict and non-specificity are two different, complementary sources of uncertainty, and (4) and (8) as well-established measures of these two types of uncertainty, a natural question in the context of credal sets is to ask for a generalized representation

$$\mathrm{TU}(C) = \mathrm{AU}(C) + \mathrm{EU}(C), \quad (9)$$

where AU is a generalization of the Shannon entropy, and EU a generalization of the Hartley measure.

As for the non-specificity part in (9), the following generalization of the Hartley measure to the case of graded possibilities has been proposed by various authors (Abellan & Moral, 2000):

$$\mathrm{GH}(C) := \sum_{A \subseteq \mathcal{Y}} \mathrm{m}_Q(A) \, \log(|A|) \, , \quad (10)$$

where $\mathrm{m}_C \, : \, 2^{\mathcal{Y}} \longrightarrow [0,1]$ is the Möbius inverse of the capacity function $\nu : 2^{\mathcal{Y}} \longrightarrow [0,1]$ defined by

$$\nu_Q(A) := \inf_{q \in Q} q(A) \quad (11)$$

for all $A \subseteq \mathcal{Y}$, that is,

$$\mathrm{m}_C(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \nu_C(B). \quad (12)$$

The measure (10) enjoys several desirable axiomatic properties, and its uniqueness was shown by Klir & Mariano (1987).

The generalization of the Shannon entropy as a measure of conflict turned out to be more difficult. The upper and lower Shannon entropy play an important role in this regard:

$$S^*(C) := \max_{\boldsymbol{\theta} \in C} S(\boldsymbol{\theta}), \quad S_*(C) := \min_{\boldsymbol{\theta} \in C} S(\boldsymbol{\theta}). \quad (13)$$

Based on these measures, the following disaggregations of total uncertainty (9) have been proposed (Abellan et al., 2006):

$$S^*(C) = \big(S^*(C) - \mathrm{GH}(C)\big) + \mathrm{GH}(C), \quad (14)$$

$$S^*(C) = S_*(C) + \big(S^*(C) - S_*(C)\big). \quad (15)$$

In both cases, upper entropy serves as a measure of total uncertainty, which is again justified on an axiomatic basis. In the first case, the generalized Hartley measure is used for quantifying epistemic uncertainty, and aleatoric uncertainty is obtained as the difference between total and epistemic uncertainty. In the second case, epistemic uncertainty is specified in terms of the difference between upper and lower entropy.

Nevertheless, a fully satisfactory representation of aggregate uncertainty in the form (9), with all three measures having nice theoretical properties, has not yet been found for the case of credal sets. Both S^* and GH appear to be well justified and enjoy strong axiomatic properties. To a slightly lesser extent, this is also true for S_* (this measure violates the property of monotonicity). However, those measures in (14–15) that are derived in terms of difference violate most of the desirable properties.

4.2. Novel Credal Uncertainty Measures

Recall that, in the Bayesian case, the learner holds belief in the form of a probability distribution p on Δ_K, and epistemic uncertainty is defined in terms of mutual information. The latter can also be written as follows:

$$\mathrm{EU}(p) = \mathbb{E}_{Y\sim\bar{\boldsymbol{\theta}}}\ \ell(\bar{\boldsymbol{\theta}}, Y) - \mathbb{E}_{\boldsymbol{\theta}\sim p}\mathbb{E}_{Y\sim\boldsymbol{\theta}}\ \ell(\boldsymbol{\theta}, Y), \quad (16)$$

where $\ell(\cdot,\cdot)$ is the log-loss. That is, EU is the *gain* — in terms of loss reduction — the learner can expect when predicting, not on the basis of the uncertain knowledge p, but only after being revealed the true $\boldsymbol{\theta}$. Intuitively, this is plausible: The more uncertain the learner is about the true $\boldsymbol{\theta}$ (i.e., the more dispersed p), the more it can gain by getting to know this distribution. We can also summarize this as follows:

- Total uncertainty is the expected (log-)loss of the learner when predicting optimally on the basis of its uncertain belief p.
- Aleatoric uncertainty is the expected loss that remains, even when the learner is perfectly informed about the ground-truth $\boldsymbol{\theta}$ before predicting.
- Epistemic uncertainty is the difference between the two, i.e., the expected loss reduction due to information about $\boldsymbol{\theta}$.

We modify and extend this approach as follows: First, as motivated above, we present belief in terms of a credal set instead of a distribution. Second, we allow for loss functions other than log-loss. Thus, "maxing" over a (credal) set instead of averaging over a distribution, we define epistemic uncertainty in terms of the maximal gain:

$$\mathrm{EU}(C) := \max_{\boldsymbol{\theta},\boldsymbol{\theta}'\in C} D_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}'), \quad (17)$$

with the ℓ-divergence

$$D_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}') := \mathbb{E}_{Y\sim\boldsymbol{\theta}}\ \left\{\ell(\boldsymbol{\theta}', Y) - \ell(\boldsymbol{\theta}, Y)\right\}. \quad (18)$$

The latter is the expected regret (excess loss) when predicting $\boldsymbol{\theta}'$ although the ground-truth is $\boldsymbol{\theta}$. For the aleatoric uncertainty, we obtain lower and upper bounds as follows:

$$\underline{\mathrm{AU}}(C) := \inf_{\boldsymbol{\theta}\in C} H_\ell(\boldsymbol{\theta}), \quad (19)$$

$$\overline{\mathrm{AU}}(C) := \sup_{\boldsymbol{\theta}\in C} H_\ell(\boldsymbol{\theta}), \quad (20)$$

with H_ℓ the ℓ-entropy of $\boldsymbol{\theta}$ given by

$$H_\ell(\boldsymbol{\theta}) := \mathbb{E}_{Y\sim\boldsymbol{\theta}}\,\ell(\boldsymbol{\theta}, Y). \quad (21)$$

Our approach, as detailed above, relies on the choice of a suitable loss function $\ell(\cdot,\cdot)$. Natural candidates for this loss are (strictly) proper scoring rules (Gneiting & Raftery, 2005). In this work, we consider four commonly-used proper scoring rules. Table 1 shows the losses and their respective aleatoric and epistemic component. Quite recently, the efficacy of (strictly) proper scoring rules has been further explored in the context of second-order uncertainty quantification (Sale et al., 2023a;c).

Further, if we accept the idea of an additive aggregation of aleatoric and epistemic uncertainty into a measure of total uncertainty, we obtain the following lower and upper bound for the latter:

$$\underline{\mathrm{TU}}(C) = \underline{\mathrm{AU}} + \mathrm{EU}, \quad (22)$$

$$\overline{\mathrm{TU}}(C) = \overline{\mathrm{AU}} + \mathrm{EU}\,. \quad (23)$$

Of course, suitable measures of aleatoric, epistemic, and total uncertainty associated with a credal set should meet certain theoretical properties. Such properties have been proposed by Abellán & Klir (2005); Jiroušek & Shenoy (2018), and more recently adapted by Hüllermeier et al. (2022) and Sale et al. (2023b) in a machine learning context. As shown by the following theorem, our proposed measures fulfills several desirable properties. In the following we will denote the set of all credal sets on $\mathbb{P}(\mathcal{Y})$ by $\mathrm{Cr}(\mathcal{Y})$.

Theorem 4.1. *If the loss-function $\ell : \Delta_K \times \mathcal{Y} \longrightarrow \mathbb{R}$ is continuous in $\boldsymbol{\theta} \in \Delta_K$, it fulfills the following properties:*

(i) Continuity: Lower and upper bounds for TU *and* AU *as well as* EU *are continuous functionals.*

(ii) Monotonicity: for all $C, P \in Cr(\mathcal{Y})$ such that $C \subseteq P$, we have $\mathrm{EU}(C) \leq \mathrm{EU}(P)$; the same holds for $\overline{\mathrm{AU}}$ and $\overline{\mathrm{TU}}$, respectively.

(iii) Precise probabilities: for all $C \in Cr(\mathcal{Y})$ such that $C = \{\boldsymbol{\theta}\}$, we have $\mathrm{EU}(C) = 0$. Then, the lower and upper bounds for TU *and* AU*, respectively, coincide.*

Additionally, if $\ell(\cdot,\cdot)$ is a proper scoring rule, lower and upper bounds for TU, AU, *and* EU *are non-negative.*

Let us finally consider some exemplary instantiations of our family of measures, i.e., concrete measures that are obtained by fixing a loss function $\ell(\cdot,\cdot)$. First, for the special case of the log-loss, we recover lower and upper entropy for AU. Moreover, as D_ℓ is the KL divergence, EU is the maximal KL divergence between any pair of distributions $\boldsymbol{\theta}, \boldsymbol{\theta}' \in C$.

Table 1: Proper scoring rules and their decomposition into aleatoric and epistemic uncertainty for the Levi agent.

Loss	Aleatoric (upper\lower)	Epistemic
log	$\sup_{\boldsymbol{\theta}\in C}\backslash\inf_{\boldsymbol{\theta}\in C}\ S(\boldsymbol{\theta})$	$\max_{\boldsymbol{\theta},\boldsymbol{\theta}'\in C}\ \mathrm{D}_{\mathrm{KL}}(\boldsymbol{\theta}\,\Vert\,\boldsymbol{\theta}')$
Brier	$\sup_{\boldsymbol{\theta}\in C}\backslash\inf_{\boldsymbol{\theta}\in C}\ 1-\sum_{k=1}^{K}\theta_k^2$	$\max_{\boldsymbol{\theta},\boldsymbol{\theta}'\in C}\ \sum_{k=1}^{K}(\theta'_k-\theta_k)^2$
spherical	$\sup_{\boldsymbol{\theta}\in C}\backslash\inf_{\boldsymbol{\theta}\in C}\ 1-\Vert\boldsymbol{\theta}\Vert_2$	$\max_{\boldsymbol{\theta},\boldsymbol{\theta}'\in C}\ \Vert\boldsymbol{\theta}\Vert_2-\sum_{k=1}^{K}\theta'_k\theta_k/\Vert\boldsymbol{\theta}\Vert_2$
zero-one	$\sup_{\boldsymbol{\theta}\in C}\backslash\inf_{\boldsymbol{\theta}\in C}\ 1-\max\theta_k$	$\max_{\boldsymbol{\theta},\boldsymbol{\theta}'\in C}\ \max\theta_k-\theta_{k=\operatorname{argmax}\theta'_k}$

Allowing for other losses increases flexibility and allows one to capture uncertainty of different kind. For example, log-loss essentially captures uncertainty regarding the outcome y eventually observed. From the perspective of the learner, however, this uncertainty might not be the most relevant one. Instead, the learner might be more interested in the uncertainty about the best *decision* to make, i.e., the best prediction. These uncertainties are clearly not the same. For example, consider the case of binary classification and suppose that $C = \{\boldsymbol{\theta} = (\theta_{pos}, \theta_{neg}) \mid 1/2 < \theta_{pos} \leq 1\}$. In this case, it is clear that the learner should predict positive. In other words, there is no (epistemic) uncertainty about the right decision, although the uncertainty about the outcome is still rather high. An exemplary instantiation accounting for decision uncertainty is the 0/1 loss, which serves as another interesting instantiation:

$$\ell(\boldsymbol{\theta}, y) = \begin{cases} 0, & \text{if } y = \operatorname{argmax}_k \theta_k \\ 1, & \text{otherwise.} \end{cases} \tag{24}$$

5. Learning Credal Predictors

In the previous sections, we motivated the notion of a Levi agent and provided its conceptual basis in terms of uncertainty representation and quantification. However, we did not yet address the question of how to realise a Levi agent algorithmically, i.e., how to learn a classifier that produces predictions in the form of credal sets.

Several approaches have been proposed within the imprecise probabilities literature (Zaffalon, 2002; Corani & Zaffalon, 2008), where credal predictors are often based on the imprecise Dirichlet model (IDM) (Walley, 1996). Furthermore, recent advancements have seen methods from the imprecise probability literature adapted for application in deep learning (Wang et al., 2024; Caprio et al., 2023).

In this paper, we adopt another way of learning credal sets, which is based on the statistical notion of (relative) likelihood function (Cattaneo, 2007; Antonucci et al., 2011). In the spirit of a confidence region, this approach considers all hypothesis that are sufficiently plausible, namely those whose relative likelihood is above a certain threshold level $\alpha \in [0, 1]$. In the imprecise-probabilities literature, the resulting credal set is often referred to as α-cut (credal set).

Given training data $\mathcal{D}$, the relative likelihood of a hypothesis $h \in \mathcal{H}$ is defined as

$$\mathcal{L}_{\mathcal{H}}(h) := \frac{L(h)}{L(\hat{h}_{ML})} = \frac{L(h)}{\sup_{h'\in\mathcal{H}} L(h')}, \tag{25}$$

where $L(h) = \prod_{i=1}^{N} p(y_i \mid h, \boldsymbol{x})$ denotes the likelihood of the hypothesis $h \in \mathcal{H}$, and $\hat{h}_{ML}$ the maximum likelihood predictor. Then, given $\alpha \in [0, 1]$ and $\boldsymbol{x} \in \mathcal{X}$, we obtain the following credal set:

$$C_\alpha := \{h(\boldsymbol{x}) \in Q \mid \mathcal{L}_{\mathcal{H}}(h) \geq \alpha\}. \tag{26}$$

The parameter α determines the "cautiousness" of the learner: The smaller α, the less hypotheses will be rejected as implausible candidates. The choice of α is facilitated through its clear semantics. For example, setting $\alpha = 0.2$ means that an h will only be excluded if there is another hypothesis whose likelihood is at least 5 times higher. The likelihood-based approach is also attractive from an ensemble learning perspective: Rather than including the predictions from all ensemble members in the credal set, those with comparatively low likelihood can be excluded. Note that α-cuts are not restrictive, since we can recover the initial credal set by C_α with $\alpha = 0$. In practice, we find that the hypotheses of an ensemble will have high relative likelihoods after training and as such we don't explicitly remove hypotheses.

6. Experiments

Since ground-truth uncertainty degrees are normally not available, the evaluation of uncertainty measures is a non-trivial problem. Here, we perform experiments to study the effectiveness of the proposed measures in different "downstream" tasks: prediction with abstention, out-of-distribution (OoD) detection, and active learning. The code for the experiments can be found in a Github repository[1].

[1] https://github.com/pwhofman/credal-uncertainty

6.1. Accuracy-Rejection Curves

Accuracy-rejection curves (ARCs) can be generated by allowing the predictor to abstain from making decision on part of the data. Naturally, a predictor equipped with a good uncertainty quantification method can abstain from making decisions on instances where the uncertainty is high. As such, the accuracy should increase as the percentage of abstained instances increases.

We train an ensemble of 5 probabilistic neural networks on the FashionMNIST (Xiao et al., 2017) and the Food101 (Bossard et al., 2014) datasets. A detailed description of the model architecture and training setup is provided in Appendix B.

The predictor should abstain from instances that are associated with a high aleatoric uncertainty. However, in the presence of epistemic uncertainty a single estimate for the aleatoric uncertainty may not be accurate. Our proposal provides both lower and upper bounds for aleatoric uncertainty guaranteeing that the true uncertainty is captured within these bounds. We use the upper bound for aleatoric uncertainty as a basis for rejection, ensuring that the predictor avoids instances that might exhibit high aleatoric uncertainty.

We compare our measures of upper AU with to the baseline of random rejection. It's worth noting that our measure for upper aleatoric uncertainty, when instantiated with log-loss, aligns with the upper entropy. All accuracies are reported as the mean over 5 runs, the standard deviation depicted by the shaded area. Figure 2 shows the accuracy-rejection curves for the FashionMNIST and Food101 dataset.

The accuracy-rejection curves for all measures exhibit similar behaviors: The curves increase monotonically until reaching 100% accuracy. The measure based on the spherical loss performs best for both datasets. As anticipated, random rejection leads to a flat ARC.

6.2. Out-of-Distribution Detection

We use out-of-distribution (OoD) detection to assess epistemic uncertainty. A model is trained on a dataset, which is referred to as the in-distribution (iD) data, and we compute the epistemic uncertainty on instances of the iD test set. Then, the model receives instances from another dataset, referred to as the out-of-distribution dataset. For these test instances, we also compute epistemic uncertainty. Naturally, the model, which has not seen the OoD data before, should have higher epistemic uncertainty for these instances. To evaluate how well the epistemic uncertainties of the iD and OoD instances are separated, we compute the AUROC.

We train an ensemble of 5 neural networks on FashionMNIST (iD) and use MNIST (LeCun et al., 1998) and KMNIST (Clanuwat et al., 2018) as the OoD datasets. We compute the mean and standard deviation of the AUROC over 5 runs. The same is done for Food101, where the OoD datasets are SVHN (Netzer et al., 2011) and CIFAR100 (Krizhevsky et al., 2009). We refer to the supplementary material for more details on model architecture and training setup.

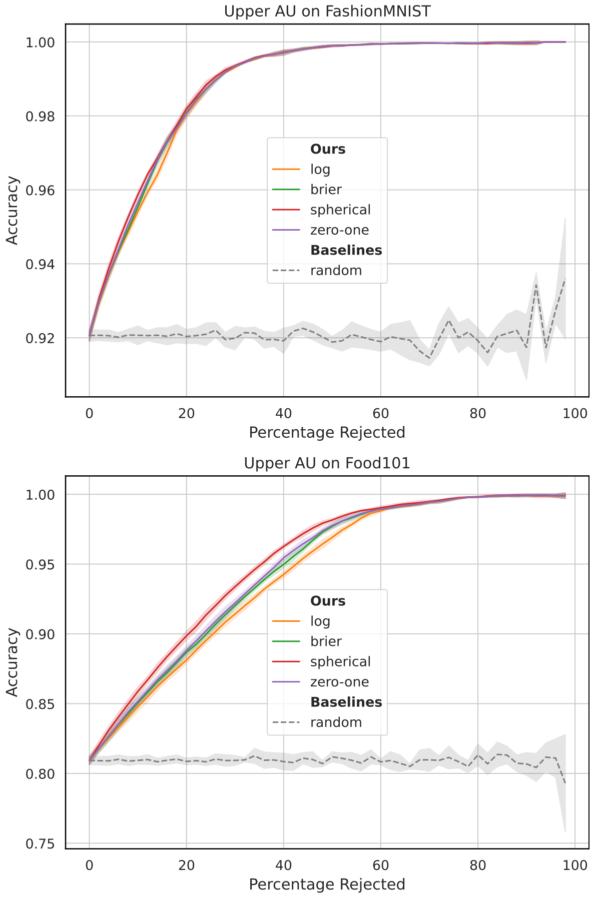

Figure 2: Accuracy-Rejection curve on FashionMNIST *(top)* and Food101 *(bottom)*. Upper aleatoric uncertainty is used as the rejection criterion. The shaded areas represent the standard deviations over five runs.

Table 2 shows the results for the networks trained on FashionMNIST or Food101. Instantiated with the log-loss, our measure performs across the four datasets.

6.3. Active Learning

We conduct active learning experiments to highlight the flexibility of our proposed measure. As our measure allows for different instantiations, it also allows us to model different types of uncertainty, as discussed in Section 4.2. In active learning, querying instances with highest uncertainty should give the most information (Nguyen et al., 2022). However,

Table 2: Out-of-Distribution Detection using epistemic uncertainty. The mean and standard deviation over five runs are reported. Best performance is in **bold**.

iD	OoD	log	Brier	spherical	zero-one	Hartley	entropy
FMNIST	MNIST	**0.871**±0.017	0.812±0.019	0.818±0.019	0.742±0.015	0.815±0.019	0.826±0.019
	KMNIST	**0.973**±0.002	0.94±0.004	0.946±0.003	0.863±0.006	0.944±0.004	0.942±0.003
Food101	SVHN	**0.700**±0.04	0.572±0.027	0.588±0.038	0.669±0.007	0.479±0.023	0.681±0.07
	CIFAR-100	**0.805**±0.015	0.66±0.016	0.681±0.016	0.697±0.008	0.576±0.022	0.775±0.021

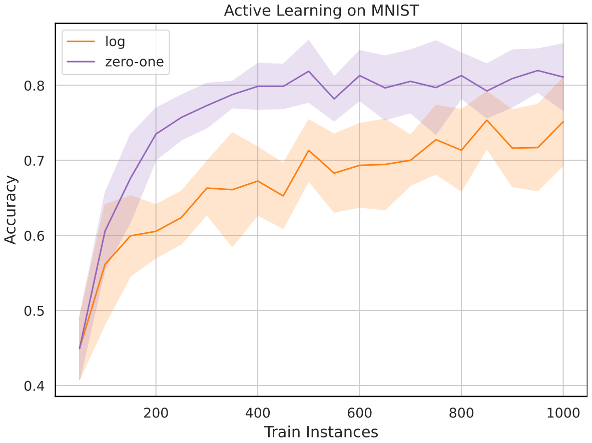

Figure 3: Active learning using MNIST.

the type of uncertainty can be relevant here. As the ultimate goal is high accuracy, thus making the correct decision, the learner should query points for which it is maximally uncertain in terms of the decision it should make.

We train an ensemble of 10 neural networks on a small amount of data and iteratively allow it to query a set of new instances from a pool of available data. We use the instantiations of our epistemic uncertainty measure with log-loss and zero-one-loss.

The test accuracy is recorded over 5 runs and the mean and standard deviation are reported. The model architecture, training details and other details can be found in Appendix B. Figure 3 shows the test accuracy of the model as it iteratively receives more data to train on.

In pool-based active learning, the performance of different methods tends to converge in the end, because the overlap of the training sets necessarily increases when the complement (namely the pool) decreases. What is more important, therefore, is the performance in the beginning. As can be seen in Figure 3, our zero-one-loss-based uncertainty sampling has a clear advantage here, and reaches a higher accuracy much faster. This shows the importance of quantifying decision uncertainty in a setting where a model is strictly evaluated based on the *decisions* it makes.

7. Limitations and Future Work

In this work, we used an ensemble of learners to generate a credal set. Although this works well empirically, we *do not offer any guarantees* on the inclusion of the ground truth conditional distribution in the set. An interesting next step would be to study how a credal set with statistical guarantees can be learned. Recent work by Javanmardi et al. (2024) has focused on this question. Furthermore, we *do not assume a convex set*, while credal sets are usually assumed to be convex sets of probability measures. Future work could explore whether this assumption is meaningful in machine learning and how it affects theoretical and empirical results.

8. Conclusion

In light of recent literature on uncertainty quantification and criticism of uncertainty measures defined in the Bayesian setting, we advocate the use of credal sets as an alternative representation of epistemic uncertainty in machine learning. In our approach, such credal sets are constructed using α-cuts of the relative likelihood function, which offers a natural way to establish an ordering of predictive models and harmonises quite well with ensemble-based learning.

We also proposed a new family of measures for quantifying aleatoric and epistemic uncertainty, which is inspired by the measures commonly used for second-order distributions in the Bayesian setting (conditional entropy and mutual information), but adapts them to the case of credal sets and generalises them to loss-functions other than log-loss. We show that this family of measures exhibits appealing theoretical properties and propose concrete instantiations for capturing different kinds of uncertainty. Empirically, we demonstrate the versatility and flexibility of our measures in different scenarios where the learner's uncertainty-awareness is key to strong performance (learning with abstention, distribution shift, active sampling).

Acknowledgements

Yusuf Sale is supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

References

Abellan, J. and Moral, S. A non-specificity measure for convex sets of probability distributions. *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8: 357–367, 2000.

Abellan, J., Klir, J., and Moral, S. Disaggregated total uncertainty measure for credal sets. *Int. Journal of General Systems*, 35(1), 2006.

Abellán, J. and Klir, G. J. Additivity of uncertainty measures on credal sets. *International Journal of General Systems*, 34(6):691–713, 2005.

Antonucci, A., Cattaneo, M., and Corani, G. Likelihood-based naive credal classifier. In *ISIPTA*, volume 11, pp. 21–30. Citeseer, 2011.

Bossard, L., Guillaumin, M., and Gool, L. V. Food-101 - mining discriminative components with random forests. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, pp. 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29. URL https://doi.org/10.1007/978-3-319-10599-4_29.

Caprio, M., Dutta, S., Jang, K. J., Lin, V., Ivanov, R., Sokolsky, O., and Lee, I. Imprecise bayesian neural networks. *arXiv preprint arXiv:2302.09656*, 2023.

Cattaneo, M. E. *Statistical decisions based directly on the likelihood function*. PhD thesis, ETH Zurich, 2007.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature, 2018.

Corani, G. and Zaffalon, M. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 9(4), 2008.

Csiszár, I. Axiomatic characterizations of information measures. *Entropy*, 10:261–273, 2008.

De Finetti, B. Foresight: It's logical laws, it's subjective sources. In Kyburg, H. and Smokler, H. (eds.), *Studies in Subjective Probability*. R.E. Krieger, New York, 1980.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., and Udluft, S. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pp. 1184–1193. PMLR, 2018.

Dubois, D., Prade, H., and Smets, P. Representing partial ignorance. *IEEE Transactions on Systems, Man and Cybernetics, Series A*, 26(3):361–377, 1996.

Gneiting, T. and Raftery, A. Strictly proper scoring rules, prediction, and estimation. Technical Report 463R, Department of Statistics, University of Washington, 2005.

Hartley, R. Transmission of information. *Bell Syst. Tech. Journal*, 7(3):535–563, 1928.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Hora, S. C. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2-3):217–223, 1996.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Hüllermeier, E. and Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Hüllermeier, E., Destercke, S., and Shaker, M. H. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Uncertainty in Artificial Intelligence*, pp. 548–557. PMLR, 2022.

Javanmardi, A., Stutz, D., and Hüllermeier, E. Conformalized credal set predictors. *CoRR*, abs/2402.10723, 2024. doi: 10.48550/ARXIV.2402.10723. URL https://doi.org/10.48550/arXiv.2402.10723.

Jiroušek, R. and Shenoy, P. P. A new definition of entropy of belief functions in the Dempster–Shafer theory. *International Journal of Approximate Reasoning*, 92:49–65, 2018.

Kapoor, S., Maddox, W. J., Izmailov, P., and Wilson, A. G. On uncertainty, tempering, and data augmentation in bayesian classification. *arXiv preprint arXiv:2203.16481*, 2022.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Klir, G. and Mariano, M. On the uniqueness of possibilistic measure of uncertainty and information. *Fuzzy Sets and Systems*, 24(2):197–219, 1987.

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

Lambrou, A., Papadopoulos, H., and Gammerman, A. Reliable confidence measures for medical diagnosis with evolutionary algorithms. *IEEE Transactions on Information Technology in Biomedicine*, 15(1):93–99, 2010.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Levi, I. On indeterminate probabilities. *Journal of Philosophy*, 71:391–418, 1974.

Levi, I. *The Enterprise of Knowledge*. MIT Press, Cambridge, 1980.

Mitchell, T. M. Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1*, pp. 305–310, 1977.

Mobiny, A., Nguyen, H., Moulik, S., Garg, N., and Wu, C. DropConnect is effective in modeling uncertainty of Bayesian networks. *CoRR*, abs/1906.04569, 2017.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Nguyen, V., Shaker, M. H., and Hüllermeier, E. How to measure uncertainty in uncertainty sampling for active learning. *Mach. Learn.*, 111(1):89–122, 2022. doi: 10.1007/s10994-021-06003-9. URL https://doi.org/10.1007/s10994-021-06003-9.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E. Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8024–8035, 2019.

Rényi, A. *Probability Theory*. North-Holland, Amsterdam, 1970.

Sale, Y., Bengs, V., Caprio, M., and Hüllermeier, E. Second-order uncertainty quantification: A distance-based approach. *arXiv preprint arXiv:2312.00995*, 2023a.

Sale, Y., Caprio, M., and Hüllermeier, E. Is the volume of a credal set a good measure for epistemic uncertainty? In *Uncertainty in Artificial Intelligence*, pp. 1795–1804. PMLR, 2023b.

Sale, Y., Hofman, P., Wimmer, L., Hüllermeier, E., and Nagler, T. Second-order uncertainty quantification: Variance-based measures. *arXiv preprint arXiv:2401.00276*, 2023c.

Senge, R., Bösner, S., Dembczyński, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., and Hüllermeier, E. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, 2014.

Shaker, M. H. and Hüllermeier, E. Ensemble-based uncertainty quantification: Bayesian versus credal inference. *CoRR*, abs/2107.10384, 2021. URL https://arxiv.org/abs/2107.10384.

Smith, L. and Gal, Y. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

Tornede, T., Tornede, A., Fehring, L., Gehring, L., Graf, H., Hanselle, J., Mohr, F., and Wever, M. PyExperimenter: Easily distribute experiments and track results. *Journal of Open Source Software*, 8(84):5149, 2023. doi: 10.21105/joss.05149. URL https://doi.org/10.21105/joss.05149.

Walley, P. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.

Walley, P. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):3–34, 1996.

Wang, K., Shariatmadar, K., Manchingal, S. K., Cuzzolin, F., Moens, D., and Hallez, H. Creinns: Credal-set interval neural networks for uncertainty estimation in classification tasks. *arXiv preprint arXiv:2401.05043*, 2024.

Wimmer, L., Sale, Y., Hofman, P., Bischl, B., and Hüllermeier, E. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?, 2023.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yager, R. Entropy and specificity in a mathematical theory of evidence. *International Journal of General Systems*, 9: 249–260, 1983.

Yang, F., Wang, H.-z., Mi, H., Lin, C.-d., and Cai, W.-w. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC bioinformatics*, 10(1):1–14, 2009.

Zaffalon, M. The naive credal classifier. *Journal of statistical planning and inference*, 105(1):5–21, 2002.

A. Proofs

Proof of Theorem 4.1

Let $\ell : \Delta_K \times \mathcal{Y} \longrightarrow \mathbb{R}$ be continuous in $\boldsymbol{\theta} \in \Delta_K$.

(1) Since $\ell : \Delta_K \times \mathcal{Y} \longrightarrow \mathbb{R}$ is continuous in $\boldsymbol{\theta} \in \Delta_K$ it follows from the linearity of the expectation that the lower and upper bounds of AU and EU are continuous. The lower and upper bounds of TU, as a sum of two continuous functions, are also continuous.

(2) Let $C, P \in \mathrm{Cr}(\mathcal{Y})$ such that $C \subseteq P$. We have the following:

$$\begin{aligned}
\mathrm{EU}(C) &= \max_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in C} D_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq \max_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in P} D_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathrm{EU}(P) \\
\overline{\mathrm{AU}}(C) &= \sup_{\boldsymbol{\theta} \in C} H_\ell(\boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta} \in P} H_\ell(\boldsymbol{\theta}) = \overline{\mathrm{AU}}(P) \\
\overline{\mathrm{TU}}(C) &= \overline{\mathrm{AU}}(C) + \mathrm{EU}(C) \leq \overline{\mathrm{AU}}(P) + \mathrm{EU}(P) = \overline{\mathrm{TU}}(P)
\end{aligned}$$

This proves monotonicity of the corresponding measures.

(3) Let $C \in \mathrm{Cr}(\mathcal{Y})$ such that $C = \{\boldsymbol{\theta}\}$. Then, we have immediately

$$\mathrm{EU}(C) = \max_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in C} D_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}') = D_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$$

Similarly, the lower and upper bounds TU, and AU, respectively, coincide.

Now, let $\ell(\cdot, \cdot)$ be a proper scoring rule, i.e., $\mathbb{E}_{y \sim \boldsymbol{\theta}} \ell(\boldsymbol{\theta}, y) \leq \mathbb{E}_{y \sim \boldsymbol{\theta}} \ell(\boldsymbol{\theta}', y)$, where $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{P}(\mathcal{Y})$ and $y \in \mathcal{Y}$. Thus, $0 \leq \mathbb{E}_{y \sim \boldsymbol{\theta}} \ell(\boldsymbol{\theta}', y) - \mathbb{E}_{y \sim \boldsymbol{\theta}} \ell(\boldsymbol{\theta}, y)$, this yields as desired

$$\mathrm{EU}(Q) = \max_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in Q} \{\mathbb{E}_{y \sim \boldsymbol{\theta}} \ell(\boldsymbol{\theta}', y) - \mathbb{E}_{y \sim \boldsymbol{\theta}} \ell(\boldsymbol{\theta}, y)\} \geq 0.$$

Since we assume that $\ell(\cdot, \cdot)$ is non-negative, it is easy to see that the lower and upper bounds for TU and AU, respectively, are non-negative. This concludes the proof. □

Derivation of measures with 0/1 loss

We show the following with respect to the 0/1-loss instantiation of our measures:

$$\begin{aligned}
D_\ell(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \mathbb{E}_{y \sim \boldsymbol{\theta}} \left\{\ell(\boldsymbol{\theta}', y) - \ell(\boldsymbol{\theta}, y)\right\} \\
&= \sum_{y \in \mathcal{Y}} \boldsymbol{\theta}(y) \ell(\boldsymbol{\theta}', y) - \sum_{y \in \mathcal{Y}} \boldsymbol{\theta}(y) \ell(\boldsymbol{\theta}, y) \\
&= \sum_{y \neq \mathrm{argmax}_{y' \in \mathcal{Y}} \boldsymbol{\theta}'(y')} \boldsymbol{\theta}(y) - \sum_{y \neq \mathrm{argmax}_{y' \in \mathcal{Y}} \boldsymbol{\theta}(y')} \boldsymbol{\theta}(y) \\
&= (1 - \boldsymbol{\theta}(\operatorname*{argmax}_{y' \in \mathcal{Y}} \boldsymbol{\theta}'(y'))) - (1 - \boldsymbol{\theta}(\operatorname*{argmax}_{y' \in \mathcal{Y}} \boldsymbol{\theta}(y'))) \\
&= \max_{y \in \mathcal{Y}} \boldsymbol{\theta}(y) - \boldsymbol{\theta}\left(\operatorname*{argmax}_{y' \in \mathcal{Y}} \boldsymbol{\theta}'(y')\right)
\end{aligned}$$

The lower and upper bound of AU simplify, for $Q \in \mathrm{Cr}(\mathcal{Y})$, to:

$$\underline{\mathrm{AU}}(Q) = \inf_{\boldsymbol{\theta} \in Q} \left(1 - \max_{y \in \mathcal{Y}} \boldsymbol{\theta}(y)\right), \qquad \overline{\mathrm{AU}}(Q) = \sup_{\boldsymbol{\theta} \in Q} \left(1 - \max_{y \in \mathcal{Y}} \boldsymbol{\theta}(y)\right)$$

B. Experimental Details

In the following we provide the necessary details regarding the experimental setting. The code is written in Python 3.10.12 with PyTorch (Paszke et al., 2019) and all experiments are done using with the help of PyExperimenter (Tornede et al., 2023).

B.1. Models

We use the following base models.

The **Multi-Layer Perceptron** (MLP) has 784 - 200 - 200 - 10 neurons and ReLU activation functions at every layer, except for the last which uses a softmax function to transform the logits into probabilities.

The **Convolutional Neural Network** (CNN), based on (LeCun et al., 1998), has two convolutional layers with 20 and 50 5 by 5 filers, respectively. Both convolutional layers are followed by a max-pooling operation with a 2 by 2 filter. The convolutional layers are followed by two fully-connected layers with dimensions 800 by 500 and 500 by 10, respectively. All layers use a ReLU activation function, except for the last which uses a softmax to transform the logits into probabilities.

The PyTorch implementation of the **ResNet** (He et al., 2016), pretrained on ImageNet (Deng et al., 2009), with a final fully-connected layer of 512 by 10 and a softmax function to give probabilistic output for 10 classes.

We generate ensemble outputs by taking the mean of the base model outputs.

B.2. Downstream Tasks

All downstream tasks are run five times and for each run a new ensemble is trained.

Accuracy-Rejection Curves We train the CNN on FMNIST for 20 epochs using the Adam optimizer (Kingma & Ba, 2015) with the PyTorch default parameters. We train the ResNet that has been pre-trained on ImageNet for 5 epochs on Food101 using Adam with the default parameters. We randomly sample 10000 instances from the dedicated test split of the datasets and use this to generate the accuracy-rejection curves.

Out-of-Distribution Detection For Out-of-Distribution Detection we use the models that were trained for the Accuracy-Rejection task. We also sample 10000 instances from the test sets of the in-Distribution and Out-of-Distribution dataset.

Active Learning We use an ensemble of 10 MLPs trained using the Adam optimizer and the default parameters. The 10 learners are initially trained using 50 and can acquire 50 new instances in every one of 20 rounds. The models are trained for 5 epochs every round. The accuracy of the models is computed on the test set.