
QUANTIFYING ALEATORIC AND EPISTEMIC UNCERTAINTY WITH PROPER SCORING RULES

A PREPRINT

Paul Hofman

Institute of Informatics, LMU Munich
Munich Center for Machine Learning (MCML)
paul.hofman@ifi.lmu.de

Yusuf Sale

Institute of Informatics, LMU Munich
Munich Center for Machine Learning (MCML)
yusuf.sale@ifi.lmu.de

Eyke Hüllermeier

Institute of Informatics, LMU Munich
Munich Center for Machine Learning (MCML)
eyke@lmu.de

April 22, 2024

ABSTRACT

Uncertainty representation and quantification are paramount in machine learning and constitute an important prerequisite for safety-critical applications. In this paper, we propose novel measures for the quantification of aleatoric and epistemic uncertainty based on proper scoring rules, which are loss functions with the meaningful property that they incentivize the learner to predict ground-truth (conditional) probabilities. We assume two common representations of (epistemic) uncertainty, namely, in terms of a credal set, i.e. a set of probability distributions, or a second-order distribution, i.e., a distribution over probability distributions. Our framework establishes a natural bridge between these representations. We provide a formal justification of our approach and introduce new measures of epistemic and aleatoric uncertainty as concrete instantiations.

Keywords Uncertainty quantification · credal sets · second-order distributions · proper scoring rules

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 these and other domains, the deployment of ML comes with the urgent need for adequate methods of representing and quantifying uncertainty.

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 learner’s incomplete knowledge of this process. The latter can therefore be reduced by the acquisition of additional information, such as additional observations. Conversely, aleatoric uncertainty, being a characteristic of the data-generating process itself, is non-reducible (Hüllermeier and 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 and Gal, 2017), adversarial example detection (Smith and Gal, 2018), and data augmentation in Bayesian classification (Kapoor et al., 2022).

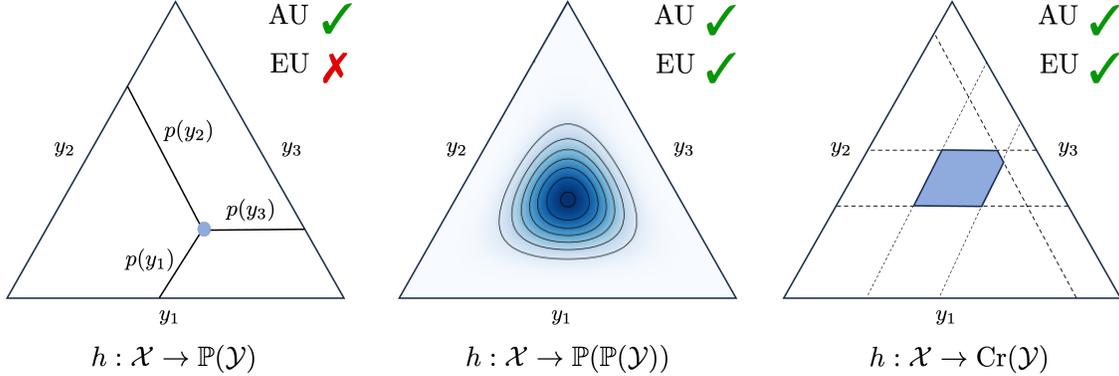


Figure 1: Uncertainty awareness in multi-class classification, illustrated on the probability simplex for $\mathcal{Y} = \{y_1, y_2, y_3\}$. From left to right: Probabilistic agent (AU, but *no* EU awareness), Bayesian agent (AU *and* EU awareness), and Levi agent (AU *and* EU awareness).

The Bayesian approach, in which the learner’s epistemic uncertainty is represented in terms of a probability distribution on the underlying model class, which in turn induces a second-order distribution on outcomes, prevails the ML literature so far. Uncertainty quantification for representations of that kind is commonly based on information-theoretic notions such as (Shannon) entropy and mutual information. Yet, this approach is not without criticism (Wimmer et al., 2023). In this paper, we address this criticism by proposing alternative, more general measures of uncertainty based on proper scoring rules. In addition, we extend this approach to the notion of *credal sets*, i.e., sets of probability distributions. Here, we assume a learner that produces (second-order) predictions in the form of sets of probability distributions on outcomes. For this representation, we also introduce new measures of epistemic and aleatoric uncertainty, thereby providing a flexible, unified framework to quantify uncertainty in different settings.

Our point of departure is a taxonomy of different types of uncertainty-aware learning algorithms introduced in the next section. In Section 3, we address the problem of uncertainty quantification and recall basic uncertainty measures for the probabilistic, Bayesian and Levi agent. We propose a novel construction of uncertainty measures for Bayesian and Levi agents based on (strictly) proper scoring rules in Section 4 prior to concluding the paper in Section 5.

2 Uncertainty Representation

Throughout this paper, we consider a formal setting of supervised learning for categorical prediction (classification) as follows: Let \mathcal{X} be a (measurable) instance space and \mathcal{Y} a label space, where $\mathcal{Y} = \{1, \dots, K\}$ for some $K \in \mathbb{N}$. Further, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ is a set of training data. The pairs (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 $\mathcal{X} \rightarrow \mathbb{P}(\mathcal{Y})$ that map instances x to a probability measure on \mathcal{Y} .

Given \mathcal{D} and \mathcal{H} , the learner induces a hypothesis h , the predictions $h(x) = \theta_x$ of which are considered as estimations of the ground-truth conditional distribution on \mathcal{Y} given $X = x$, denoted θ_x^* . For simplicity, we will often omit the (query) instance x as a subscript. As $\mathbb{P}(\mathcal{Y})$, the set of all 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 $\theta = (\theta_1, \dots, \theta_K)^\top$ and $\theta^* = (\theta_1^*, \dots, \theta_K^*)^\top$, respectively, where θ_k^* is the true probability $P(Y = k | X = x)$ of the k^{th} class (given x) and θ_k the estimate of this probability.

Given a query $x \in \mathcal{X}$ 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 $\theta \in \Delta_K$, is represented, we propose to distinguish four types of possible learners. Referring to the literature on decision under uncertainty, we call these learners deterministic agents, probabilistic agents, Bayesian agents, and Levi agents, respectively. An overview (of the last three) can be found in Figure 1.

2.1 Deterministic Agents

Very simple machine learning methods produce predictors in the form of mappings $\mathcal{X} \rightarrow \mathcal{Y}$, which assign to each query instance a predicted class label in a deterministic way. Obviously, a learner of that kind does not exhibit any sort of uncertainty-awareness, and does neither reflect aleatoric nor epistemic uncertainty.

2.2 Probabilistic Agents

A more common practice in contemporary machine learning is to consider probabilistic learners, which induce a hypothesis $h \in \mathcal{H}$ and use this hypothesis to make predictions. Given $\mathbf{x} \in \mathcal{X}$ as input, such a learner will predict a single probability distribution $h(\mathbf{x}) = \boldsymbol{\theta}$, which is considered as an estimation of the ground-truth conditional probability $\boldsymbol{\theta}^*$. We call a learner of that kind a *probabilistic agent*. As it fully commits to a single $h \in \mathcal{H}$, 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.

2.3 Bayesian Agents

Adhering to the principle of (strict) Bayesianism as advocated by statisticians such as De Finetti (De Finetti, 1980), a *Bayesian agent* will represent its belief in terms of a probability distribution on \mathcal{H} . Thus, instead of committing to a single hypothesis, the agent will assign a probability (density) $q(h)$ to each candidate $h \in \mathcal{H}$. Moreover, belief revision in the light of observed data \mathcal{D} is accomplished by replacing this distribution with the posterior $q(h | \mathcal{D})$.

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

$$p(\boldsymbol{\theta}) = \int_{\mathcal{H}} \llbracket h(\mathbf{x}) = \boldsymbol{\theta} \rrbracket dq(h | \mathcal{D}). \quad (1)$$

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

In addition to the second-order distribution p , called posterior predictive distribution, a Bayesian agent typically also induces a representative (first-order) distribution via Bayesian model averaging:

$$\bar{\boldsymbol{\theta}} := \text{bma}(p) = \int_{\mathcal{H}} h(\mathbf{x}) dq(h | \mathcal{D}).$$

2.4 Levi Agents

Instead of representing model uncertainty in terms of a distribution 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 $h \mapsto h(\mathbf{x})$, the set Q directly translates into a set $C \subseteq \Delta_K$ of possible class distributions:

$$C = \{\boldsymbol{\theta} = h(\mathbf{x}) \mid h \in Q\}.$$

In the literature, such a set of probability distributions is also referred to as a *credal set* (Levi, 1980; 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

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

$$S(\boldsymbol{\theta}) := - \sum_{k=1}^K \theta_k \cdot \log_2 \theta_k, \quad (2)$$

where $0 \log 0 = 0$ by definition. This measure is the most obvious candidate to quantify the aleatoric uncertainty of a probabilistic agent, i.e., $AU = S(\boldsymbol{\theta})$. As such an agent pretends to have precise knowledge about the predictive distribution, the epistemic uncertainty is 0.

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

3.2 Bayesian Agents

The Bayesian perspective, according to which the epistemic state of the learner is represented by the posterior on the hypothesis space, is quite common 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} . Moreover, epistemic uncertainty is considered as a property of the posterior $q(\cdot | \mathcal{D})$ resp. the induced distribution p on Δ_K : The less concentrated this distribution is, the higher the (epistemic) uncertainty of the learner.

A principled approach to measuring and separating aleatoric and epistemic uncertainty, which is now widely applied in machine learning (Houlsby et al., 2011; Depeweg et al., 2018; Mobiny et al., 2017), is based on a classical information-theoretic result, according to which entropy (of a random variable U) can be decomposed into conditional entropy (of U given another random variable V) and mutual information (of U and V). In our case, considering the distribution on Δ_K as a random variable Θ (with distribution p) and the outcome as a random variable Y , we obtain the following:

$$\begin{aligned} \text{TU} &= S(Y) = S(\bar{\boldsymbol{\theta}}) = - \sum_{k=1}^K \bar{\theta}_k \cdot \log_2(\bar{\theta}_k), \\ \text{AU} &= S(Y | \Theta) = - \int p(\boldsymbol{\theta}) \cdot S(\boldsymbol{\theta}) d\boldsymbol{\theta}, \\ \text{EU} &= S(Y) - S(Y | \Theta) = I(Y, \Theta) \end{aligned}$$

where I is mutual information.

3.3 Levi Agents

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 2.4. 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 criticisms 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 highly flexible uncertainty measures for credal sets.

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)

$$H(A) = \log(|A|), \quad (3)$$

Just like the Shannon entropy¹, this measure can be justified axiomatically¹.

Given the insight that conflict and non-specificity are two different, complementary sources of uncertainty, and (2) and (3) 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

$$\text{TU}(C) = \text{AU}(C) + \text{EU}(C), \quad (4)$$

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

As for the non-specificity part in (4), the following generalization of the Hartley measure to the case of graded possibilities has been proposed by various authors (Abellan and Moral, 2000):

$$\text{GH}(C) := \sum_{A \subseteq \mathcal{Y}} m_C(A) \log(|A|), \quad (5)$$

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

$$\nu_C(A) := \inf_{\theta \in C} q(A)$$

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

$$m_C(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \nu_C(B).$$

The measure (5) enjoys several desirable axiomatic properties, and its uniqueness was shown by Klir and 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:

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

Based on these measures, the following disaggregations of total uncertainty (4) have been proposed (Abellan et al., 2006):

$$S^*(C) = (S^*(C) - \text{GH}(C)) + \text{GH}(C) \quad (6)$$

$$S_*(C) = S_*(C) + (S^*(C) - S_*(C)) \quad (7)$$

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.

¹For example, see Chapter IX, pages 540–616, in the book by Rényi (1970).

Nevertheless, a fully satisfactory representation of aggregate uncertainty in the form (4), 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 (6–7) that are derived in terms of difference violate most of the desirable properties.

4 Proper Scoring Rule-based Measures

In this section, we propose a novel construction of uncertainty measures for Bayesian and Levi agents based on (strictly) proper scoring rules. This idea has already been mentioned by Sale et al. (2023), who proposed it as a generalization of the variance-based measures introduced in that paper, and has (independently) also been elaborated on in the very recent paper by Kotelevskii and Panov (2024), albeit only for the Bayesian case. Before explaining it in more detail, let us first introduce some notation and revisit definitions and characterizations related to proper scoring rules.

In the domain of probabilistic forecasting and decision-making, proper scoring rules are instrumental for the rigorous assessment and comparison of predictive models. These scoring rules, originating from the seminal works of Savage (1971) and further developed by Gneiting and Raftery (2005), provide a mechanism to assign numerical scores to probability forecasts, rewarding accuracy and honesty in predictions. Proper scoring rules, such as the Brier score and the log score, are uniquely characterized by their *properness*—a property ensuring that the forecaster’s expected score is optimized only when announcing probabilities that correspond to their true beliefs. This property is essential for eliciting truthful and well-calibrated probabilistic forecasts, a critical aspect in various fields. Employing these scoring rules leads to more reliable forecasting models, which in turn improve the processes of decision-making in scenarios that rely on probabilistic forecasting.

Recall that \mathcal{Y} denotes a finite label space in our supervised learning framework. Further, we identify the class of all probability measures on the label space with the $(K - 1)$ -simplex Δ_K . In the following we will denote the extended real line by $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$. A function defined on \mathcal{Y} taking values in $\bar{\mathbb{R}}$ is Δ_K -quasi-integrable if it is measurable with respect to $\sigma(\mathcal{Y})$, a suitable² σ -algebra on \mathcal{Y} , and is quasi-integrable with respect to all $\theta \in \Delta_K$. We assume scoring rules to be negatively oriented, thus taking a classical ML perspective where we wish to minimize the corresponding loss.

Definition 1 A scoring rule is a function $\ell : \Delta_K \times \mathcal{Y} \rightarrow \bar{\mathbb{R}}$ such that $\ell(\hat{\theta}, \cdot)$ is Δ_K -quasi-integrable for all $\hat{\theta} \in \Delta_K$. We further write

$$L_\ell(\hat{\theta}, \theta^*) = \mathbb{E}_{Y \sim \theta^*}[\ell(\hat{\theta}, Y)] \quad (8)$$

to denote the expected loss. A scoring rule is called proper if

$$L_\ell(\theta^*, \theta^*) \leq L_\ell(\hat{\theta}, \theta^*) \quad \text{for all } \hat{\theta}, \theta^* \in \Delta_K. \quad (9)$$

The scoring rule is strictly proper, if (9) holds with equality if and only if $\hat{\theta} = \theta^*$.

It is well known that scoring rules and their corresponding expected losses can be decomposed into a *divergence* term and an *entropy* term Gneiting and Raftery (2005); Kull and Flach (2015):

$$D_\ell(\hat{\theta}, \theta^*) = L_\ell(\hat{\theta}, \theta^*) - L_\ell(\theta^*, \theta^*)$$

$$H_\ell(\theta^*) = L_\ell(\theta^*, \theta^*)$$

The former represents the expected loss, or risk, of predicting $\hat{\theta}$ when the ground truth is θ^* , while the latter captures the expected loss that materializes even when the ground truth θ^* is predicted. This highlights the inherent connection to the quantification of epistemic and aleatoric uncertainty: $H_\ell(\theta^*)$ is the irreducible part of the risk, and hence relates to aleatoric uncertainty, whereas $D_\ell(\hat{\theta}, \theta^*)$ is the “excess risk” that is purely due to the learner’s imperfect knowledge and could be reduced by improving that knowledge.

In the following, we will further define the relation between (strictly) proper scoring rules and uncertainty quantification in Bayesian agents and Levi agents, showing how the currently used measures are related and how we propose to improve them.

²In the following, we assume without loss of generality $\sigma(\mathcal{Y}) = 2^{\mathcal{Y}}$.

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

Loss	Aleatoric	Epistemic
log	$\mathbb{E}_{\theta \sim p}[S(\theta)]$	$\mathbb{E}_{\hat{\theta} \sim p}[\mathbb{E}_{\theta \sim p}[D_{KL}(\theta \hat{\theta})]]$
brier	$\mathbb{E}_{\theta \sim p}[1 - \sum_{k=1}^K \theta_k^2]$	$\mathbb{E}_{\hat{\theta} \sim p}[\mathbb{E}_{\theta \sim p}[\sum_{k=1}^K (\hat{\theta}_k - \theta_k)^2]]$
spherical	$\mathbb{E}_{\theta \sim p}[1 - \ \theta\ _2]$	$\mathbb{E}_{\hat{\theta} \sim p}[\mathbb{E}_{\theta \sim p}[\ \theta\ _2 - \sum_{k=1}^K \hat{\theta}_k \theta_k / \ \theta\ _2]]$
zero-one	$\mathbb{E}_{\theta \sim p}[1 - \max \theta_k]$	$\mathbb{E}_{\hat{\theta} \sim p}[\mathbb{E}_{\theta \sim p}[\max \theta_k - \theta_{k=\text{argmax } \hat{\theta}_k}]]$

4.1 Bayesian Agents

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:

$$\text{EU}(p) = \mathbb{E}_{\theta \sim p}[D_\ell(\bar{\theta}, \theta)] \quad (10)$$

$$= \mathbb{E}_{\theta \sim p}[L_\ell(\bar{\theta}, \theta) - L_\ell(\theta, \theta)]$$

$$= \underbrace{\mathbb{E}_{\theta \sim p}[L_\ell(\bar{\theta}, \theta)]}_{\text{TU}(p)} - \underbrace{\mathbb{E}_{\theta \sim p}[L_\ell(\theta, \theta)]}_{\text{AU}(p)} \quad (11)$$

where L_ℓ is instantiated with ℓ as 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 θ . Intuitively, this is plausible: The more uncertain the learner is about the true θ (i.e., the more dispersed p), the more it can gain by getting to know this distribution.

The connection to proper scoring rules is also quite obvious:

- Total uncertainty in (11) is the expected (log-)loss of the learner when predicting optimally ($\bar{\theta}$) on the basis of its uncertain belief p . It corresponds to the expectation (with regard to p) of the expected loss (8) with ℓ the log-loss. Broadly speaking, we average the score of the prediction $\bar{\theta}$ over the potential ground-truths $\theta \sim p$.
- Aleatoric uncertainty is the expected loss that remains, even when the learner is perfectly informed about the ground-truth θ before predicting. Again, we average over the potential ground-truths $\theta \sim p$.
- Epistemic uncertainty is the difference between the two, i.e., the expected loss reduction due to information about θ .

This definition is clearly meaningful, but as mentioned before, may not have desirable behaviour when instantiated with the log-loss. Therefore, we modify and extend the Bayesian approach as follows: First, we allow for loss functions other than log-loss. Second, we no longer assume that the agent fully commits to the expectation $\bar{\theta}$ as a prediction. Instead, it predicts at random according to p . Hence, our measures of uncertainty become:

$$\text{EU}(p) = \mathbb{E}_{\hat{\theta} \sim p}[\mathbb{E}_{\theta \sim p}[D_\ell(\hat{\theta}, \theta)]] \quad (12)$$

$$\text{AU}(p) = \mathbb{E}_{\theta \sim p}[H_\ell(\theta, \theta)]$$

$$\text{TU}(p) = \text{AU}(p) + \text{EU}(p) = \mathbb{E}_{\hat{\theta} \sim p}[\mathbb{E}_{\theta \sim p}[L_\ell(\hat{\theta}, \theta)]]$$

Compared to (10), the measure (12) can be seen as a more faithful representation of the agent’s belief and its (epistemic) uncertainty. According to the latter, uncertainty is defined in terms of the expected loss when the agent predicts θ according to its true belief Q . As opposed to this, (10) defines uncertainty in terms of the expected loss when the agent chooses the *optimal* (loss-minimizing) prediction $\hat{\theta}$ (which is given by $\bar{\theta}$). As a result, the (first-order) prediction behavior of the agent may deviate from its (second-order) belief, and in essence, this is exactly what the theoretical results by Bengs et al. (2022) have shown.

Table 1 has an overview of common proper scoring rules and the proposed decomposition into aleatoric and epistemic uncertainty for Bayesian agents. Using log-loss as an instantiation results in expected conditional entropy as a measure of aleatoric uncertainty and expected pairwise KL-divergence for epistemic uncertainty, which coincides with a recently proposed measure of uncertainty by Schweighofer et al. (2023).

Table 2: Proper scoring rules and their decomposition into aleatoric and epistemic uncertainty for the Levi agent.

Loss	Aleatoric (upper\lower)	Epistemic
log	$\sup_{\theta \in C} \inf_{\hat{\theta} \in C} S(\theta)$	$\max_{\hat{\theta}, \theta \in C} D_{KL}(\theta \hat{\theta})$
brier	$\sup_{\theta \in C} \inf_{\hat{\theta} \in C} 1 - \sum_{k=1}^K \theta_k^2$	$\max_{\hat{\theta}, \theta \in C} \sum_{k=1}^K (\hat{\theta}_k - \theta_k)^2$
spherical	$\sup_{\theta \in C} \inf_{\hat{\theta} \in C} 1 - \ \theta\ _2$	$\max_{\hat{\theta}, \theta \in C} \ \theta\ _2 - \sum_{k=1}^K \hat{\theta}_k \theta_k / \ \theta\ _2$
zero-one	$\sup_{\theta \in C} \inf_{\hat{\theta} \in C} 1 - \max \theta_k$	$\max_{\hat{\theta}, \theta \in C} \max \theta_k - \theta_{k=\text{argmax } \hat{\theta}_k}$

4.2 Levi Agents

For Levi agents, we have no way of computing an expectation, thus we “max” over a (credal) set C instead of averaging over a distribution and define epistemic uncertainty in terms of the maximal gain:

$$\text{EU}(C) := \max_{\hat{\theta}, \theta \in C} D_\ell(\hat{\theta}, \theta),$$

with the ℓ -divergence

$$D_\ell(\hat{\theta}, \theta) := L_\ell(\hat{\theta}, \theta) - L_\ell(\theta, \theta).$$

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

$$\underline{\text{AU}}(C) := \inf_{\theta \in C} H_\ell(\theta),$$

$$\overline{\text{AU}}(C) := \sup_{\theta \in C} H_\ell(\theta),$$

with H_ℓ the ℓ -entropy of θ given by

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

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:

$$\underline{\text{TU}}(C) = \underline{\text{AU}}(C) + \text{EU}(C),$$

$$\overline{\text{TU}}(C) = \overline{\text{AU}}(C) + \text{EU}(C).$$

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 $\hat{\theta}, \theta \in C$.

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 $C = \{\theta = (\theta_{pos}, \theta_{neg}) \mid 1/2 < \theta_{pos} \leq 1\}$. 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 zero-one-loss:

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

Table 2 gives an overview of common proper scoring rules and the proposed decomposition into aleatoric and epistemic uncertainty for Levi agents.

5 Concluding Remarks

In the light of recent literature on uncertainty quantification criticising uncertainty measures defined in the Bayesian settings and the lack of a unified approach to quantify uncertainty for second-order distributions (Bayesian agents) and credal sets (Levi agents) as representations, we propose novel uncertainty measures for the quantification of aleatoric and epistemic uncertainty based on proper scoring rules.

We showed how the commonly used uncertainty measures in the Bayesian setting can be seen as an instantiation of proper scoring rule decomposition and extended this idea to overcome the aforementioned criticisms. Additionally, we extend this idea to the setting of Levi agents, which still lacks commonly accepted measures for the quantification of uncertainty.

In future work, we will extend the formal analysis of the proposed measures and elaborate on their properties. We will also conduct empirical studies to see how the measures behave in different practical scenarios.

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., Klir, J., and Moral, S. (2006). Disaggregated total uncertainty measure for credal sets. *Int. Journal of General Systems*, 35(1).
- Abellan, J. and Moral, S. (2000). A non-specificity measure for convex sets of probability distributions. *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8:357–367.
- Bengs, V., Hüllermeier, E., and Waegeman, W. (2022). Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in Neural Information Processing Systems*.
- Csiszár, I. (2008). Axiomatic characterizations of information measures. *Entropy*, 10:261–273.
- De Finetti, B. (1980). Foresight: It’s logical laws, it’s subjective sources. In Kyburg, H. and Smokler, H., editors, *Studies in Subjective Probability*. R.E. Krieger, New York.
- Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., and Udluft, S. (2018). Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR.
- Dubois, D., Prade, H., and Smets, P. (1996). Representing partial ignorance. *IEEE Transactions on Systems, Man and Cybernetics, Series A*, 26(3):361–377.
- Gneiting, T. and Raftery, A. (2005). Strictly proper scoring rules, prediction, and estimation. Technical Report 463R, Department of Statistics, University of Washington.
- Hartley, R. (1928). Transmission of information. *Bell Syst. Tech. Journal*, 7(3):535–563.
- Hora, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2-3):217–223.
- Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.
- Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506.
- Kapoor, S., Maddox, W. J., Izmailov, P., and Wilson, A. G. (2022). On uncertainty, tempering, and data augmentation in bayesian classification. *arXiv preprint arXiv:2203.16481*.
- Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30.
- Klir, G. and Mariano, M. (1987). On the uniqueness of possibilistic measure of uncertainty and information. *Fuzzy Sets and Systems*, 24(2):197–219.
- Kotelevskii, N. and Panov, M. (2024). Predictive uncertainty quantification via risk decompositions for strictly proper scoring rules. *CoRR*, abs/2402.10727.

- Kull, M. and Flach, P. A. (2015). Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In Appice, A., Rodrigues, P. P., Costa, V. S., Soares, C., Gama, J., and Jorge, A., editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I*, volume 9284 of *Lecture Notes in Computer Science*, pages 68–85. Springer.
- Lambrou, A., Papadopoulos, H., and Gammerman, A. (2010). Reliable confidence measures for medical diagnosis with evolutionary algorithms. *IEEE Transactions on Information Technology in Biomedicine*, 15(1):93–99.
- Levi, I. (1974). On indeterminate probabilities. *Journal of Philosophy*, 71:391–418.
- Levi, I. (1980). *The Enterprise of Knowledge*. MIT Press, Cambridge.
- Mitchell, T. M. (1977). Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1*, pages 305–310.
- Mobiny, A., Nguyen, H., Moulik, S., Garg, N., and Wu, C. (2017). DropConnect is effective in modeling uncertainty of Bayesian networks. *CoRR*, abs/1906.04569.
- Rényi, A. (1970). *Probability Theory*. North-Holland, Amsterdam.
- Sale, Y., Hofman, P., Wimmer, L., Hüllermeier, E., and Nagler, T. (2023). Second-order uncertainty quantification: Variance-based measures. *arXiv preprint arXiv:2401.00276*.
- Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.
- Schweighofer, K., Aichberger, L., Ielanskyi, M., and Hochreiter, S. (2023). Introducing an improved information-theoretic measure of predictive uncertainty. *CoRR*, abs/2311.08309.
- Senge, R., Bösner, S., Dembczyński, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., and Hüllermeier, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29.
- Smith, L. and Gal, Y. (2018). Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*.
- Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall.
- Wimmer, L., Sale, Y., Hofman, P., Bischl, B., and Hüllermeier, E. (2023). Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in Artificial Intelligence*, pages 2282–2292. PMLR.
- Yager, R. (1983). Entropy and specificity in a mathematical theory of evidence. *International Journal of General Systems*, 9:249–260.
- Yang, F., Wang, H.-z., Mi, H., Lin, C.-d., and Cai, W.-w. (2009). Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC bioinformatics*, 10(1):1–14.