Is the Volume of a Credal Set a Good Measure for Epistemic Uncertainty?

Yusuf Sale[1,3] **Michele Caprio**[2] **Eyke Hüllermeier**[1,3]

[1]Institute of Informatics, University of Munich (LMU), Germany
[2]PRECISE Center, Department of Computer and Information Science, University of Pennsylvania, USA
[3]Munich Center for Machine Learning, Germany

Abstract

Adequate uncertainty representation and quantification have become imperative in various scientific disciplines, especially in machine learning and artificial intelligence. As an alternative to representing uncertainty via one single probability measure, we consider credal sets (convex sets of probability measures). The geometric representation of credal sets as d-dimensional polytopes implies a geometric intuition about (epistemic) uncertainty. In this paper, we show that the volume of the geometric representation of a credal set is a meaningful measure of epistemic uncertainty in the case of binary classification, but less so for multi-class classification. Our theoretical findings highlight the crucial role of specifying and employing uncertainty measures in machine learning in an appropriate way, and for being aware of possible pitfalls.

1 INTRODUCTION

The notion of uncertainty has recently drawn increasing attention in machine learning (ML) and artificial intelligence (AI) due to the fields' burgeoning relevance for practical applications, many of which have safety requirements, such as in medical domains [Lambrou et al., 2010, Senge et al., 2014, Yang et al., 2009] or socio-technical systems [Varshney, 2016, Varshney and Alemzadeh, 2017]. These applications to safety-critical contexts show that a suitable representation and quantification of uncertainty for modern, reliable machine learning systems is imperative.

In general, the literature makes a distinction between *aleatoric* and *epistemic* uncertainties (AU and EU, respectively) [Hora, 1996]. While the former is caused by the inherent randomness of the data-generating process, EU results from the learner's lack of knowledge regarding the true underlying model; it also includes approximation uncertainty. Since EU can be reduced per se with further information (e.g., via data augmentation using semantic preserving transformations), it is also referred to as reducible uncertainty. In contrast, aleatoric uncertainty, as a property of the data-generating process, is irreducible [Hüllermeier and Waegeman, 2021]. The importance of distinguishing between different types of uncertainty is reflected in several areas of recent machine learning research, e.g. in Bayesian deep learning [Depeweg et al., 2018, Kendall and Gal, 2017], in adversarial example detection [Smith and Gal, 2018], or data augmentation in Bayesian classification [Kapoor et al., 2022]. A qualitative representation of total uncertainty, AU, and EU, and of their asymptotic behavior as the number of data points available to the learning agent increases, is given in Figure 1.

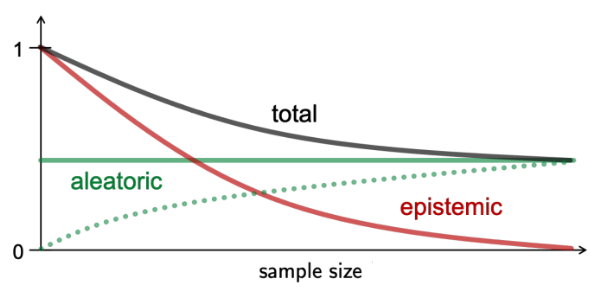

Figure 1: Qualitative behavior of total, aleatoric, and epistemic uncertainties depending on the sample size. The dotted line is the difference between total and epistemic uncertainties. This figure replicates [Hüllermeier, 2022, Figure 3].

Typically, uncertainty in machine learning, artificial intelligence, and related fields is expressed solely in terms of probability theory. That is, given a measurable space $(\Omega, \mathcal{A})$, uncertainty is entirely represented by defining one single probability measure P on $(\Omega, \mathcal{A})$. However, representing uncertainty in machine learning is not restricted to classical probability theory; various aspects of uncertainty representation and quantification in ML are discussed by Hüllermeier and Waegeman [2021]. *Credal sets*, i.e., (convex) sets of probability measures, are considered to be very popular

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

models of uncertainty representation, especially in the field of *imprecise probabilities* (IP) [Augustin et al., 2014, Walley, 1991]. Credal sets are also very appealing from an ML perspective for representing uncertainty, as they can represent both aleatoric and epistemic uncertainty (as opposed to a single probability measure). Numerous scholars emphasized the utility of representing uncertainty in ML via credal sets, e.g., *credal classification* [Zaffalon, 2002, Corani and Zaffalon, 2008] based on the Imprecise Dirichlet Model (IDM) [Walley, 1996], generalizing *Bayesian networks* to *credal classifiers* [Corani et al., 2012], or building *credal decision-trees* [Abellán and Moral, 2003].

Uncertainty representation via credal sets also requires a corresponding *quantification* of the underlying uncertainty, referred to as *credal uncertainty quantification* (CUQ). The task of (credal) uncertainty quantification translates to finding a suitable measure that can accurately reflect the uncertainty inherent to a credal set. In many ML applications, such as active learning [Settles, 2009] or classification with abstention, there is a need to quantify (predictive) uncertainty in a scalar way. Appropriate measures of uncertainty are often axiomatically justified [Bronevich and Klir, 2008, 2010].

Contributions. In this work, we consider the volume of the geometric representation of a credal set on the label space as a quite obvious and intuitively plausible measure of EU. We argue that this measure is indeed meaningful if we are in a binary classification setting. However, in a multi-class setting, the volume exhibits shortcomings that make it unsuitable for quantifying EU associated with a credal set.

Structure of the paper. The paper is divided as follows. Section 2 formally introduces the framework we work in, and Section 3 discusses the related literature. Section 4 presents our main findings, which are further discussed in Section 5. Proofs of our theoretical results are given in Appendix **??**, and (a version of) Carl-Pajor's theorem, intimately related to Theorem 1, is stated in Appendix **??**.

2 UNCERTAINTY IN ML AND AI

Uncertainty is a crucial concept in many academic and applied disciplines. However, since its definition depends on the specific context a scholar works in, we now introduce the formal framework of supervised learning within which we will examine it.

Let $(\mathcal{X}, \sigma(\mathcal{X}))$, and $(\mathcal{Y}, \sigma(\mathcal{Y}))$ be two measurable spaces, where $\sigma(\mathcal{X})$, and $\sigma(\mathcal{Y})$ are suitable σ-algebras. We will refer to $\mathcal{X}$ as *instance space* (or equivalently, input space) and to $\mathcal{Y}$ as *label space*. Further, the sequence $\{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, is called *training data*. The pairs (x_i, y_i) are realizations of random variables (X_i, Y_i), which are assumed independent and identically distributed (i.i.d.) according to some probability measure P on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{X} \times \mathcal{Y}))$.

Definition 1 (Credal set)**.** *Let $(\Omega, \mathcal{A})$ be a generic measurable space and denote by $\mathcal{M}(\Omega, \mathcal{A})$ the set of all (countably additive) probability measures on $(\Omega, \mathcal{A})$. A convex subset $\mathcal{P} \subseteq \mathcal{M}(\Omega, \mathcal{A})$ is called a credal set.*

Note that in Definition 1, the assumption of convexity is quite natural and considered to be rational (see, e.g., Levi [1980]). It is also mathematically appealing, since, as shown by Walley [1991, Section 3.3.3], the "lower boundary" $\underline{P}$ of $\mathcal{P}$, defined as $\underline{P}(A) := \inf_{P \in \mathcal{P}} P(A)$, for all $A \in \mathcal{A}$ and called the *lower probability* associated with $\mathcal{P}$, is coherent [Walley, 1991, Section 2.5].

Further, in a supervised learning setting, we assume a *hypothesis space* $\mathcal{H}$, where each hypothesis $h \in \mathcal{H}$ maps a query instance $\mathbf{x_q}$ to a probability measure P on $(\mathcal{Y}, \sigma(\mathcal{Y}))$. We distinguish between different "degrees" of uncertainty-aware predictions, which are depicted in Table 1.

Predictor	AU aware?	EU aware?
Hard label prediction: $h : \mathcal{X} \longrightarrow \mathcal{Y}$	✖	✖
Probabilistic prediction: $h : \mathcal{X} \longrightarrow \mathcal{M}(\mathcal{Y}, \sigma(\mathcal{Y}))$	✔	✖
Credal prediction: $h : \mathcal{X} \longrightarrow \mathrm{Cr}(\mathcal{Y})$	✔	✔

Table 1: Aleatoric uncertainty (AU) and epistemic uncertainty (EU) awareness of different predictors.

We denote by $\mathrm{Cr}(\mathcal{Y})$ the set of all credal sets on $(\mathcal{Y}, \sigma(\mathcal{Y}))$. While probabilistic predictions $h(\mathbf{x_q}) = \hat{y}$ fail to capture the epistemic part of the (predictive) uncertainty, predictions in the form of credal sets $h(\mathbf{x_q}) = \mathcal{P} \subseteq \mathcal{M}(\mathcal{Y}, \sigma(\mathcal{Y}))$ account for both types of uncertainty. It should also be remarked that representing uncertainty is not restricted to the credal set formalism. Another possible framework to represent AU and EU is that of second-order distributions; they are commonly applied in Bayesian learning and have been recently inspected in the context of uncertainty quantification by Bengs et al. [2022].

In this paper, we restrict our attention to the credal set representation. Given a credal prediction set, it remains to properly quantify the uncertainty encapsulated in it using a suitable measure. Credal set representations are often illustrated in low dimensions (usually $d = 2$ or $d = 3$). Examples of such geometrical illustrations can be found in the context of machine learning in [Hüllermeier and Waegeman, 2021] and in imprecise probability theory in [Walley, 1991, Chapter 4]. This suggests that a credal set and its geometric representation are strictly intertwined. We will show in the following sections that this intuitive view can have disastrous consequences in higher dimensions and that one should exercise caution in this respect. Furthermore, it remains to be discussed whether a geometric viewpoint on (predictive) uncertainty quantification is in fact sensible.

3 MEASURES OF CREDAL UNCERTAINTY

In this section we examine some axiomatically defined properties of (credal) uncertainty measures. For a more detailed discussion of various (credal) uncertainty measures in machine learning and a critical analysis thereof, we refer to Hüllermeier et al. [2022].

Let S denote the Shannon entropy [Shannon, 1948], whose discrete version is defined as

$$S : \mathcal{M}(\mathcal{Y}, \sigma(\mathcal{Y})) \to \mathbb{R},$$
$$P \mapsto S(P) := -\sum_{y \in \mathcal{Y}} P(\{y\}) \log_2 P(\{y\}).$$

A suitable measure of credal uncertainty $U : \text{Cr}(\mathcal{Y}) \to \mathbb{R}$ should satisfy the following axioms proposed by Abellán and Klir [2005], Jiroušek and Shenoy [2018]:

- A1 *Non-negativity* and *boundedness*:
  - (i) $U(\mathcal{P}) \geq 0$, for all $\mathcal{P} \in \text{Cr}(\mathcal{Y})$;
  - (ii) there exists $u \in \mathbb{R}$ such that $U(\mathcal{P}) \leq u$, for all $\mathcal{P} \in \text{Cr}(\mathcal{Y})$.
- A2 *Continuity*: U is a continuous functional.
- A3 *Monotonicity*: for all $\mathcal{Q}, \mathcal{P} \in \text{Cr}(\mathcal{Y})$ such that $\mathcal{Q} \subset \mathcal{P}$, we have $U(\mathcal{Q}) \leq U(\mathcal{P})$.
- A4 *Probability consistency*: for all $\mathcal{P} \in \text{Cr}(\mathcal{Y})$ such that $\mathcal{P} = \{P\}$, we have $U(\mathcal{P}) = S(P)$.
- A5 *Sub-additivity*: Suppose $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2$, and let $\mathcal{P}$ be a joint credal set on $\mathcal{Y}$ such that $\mathcal{P}'$ is the marginal credal set on $\mathcal{Y}_1$ and $\mathcal{P}''$ is the marginal credal set on $\mathcal{Y}_2$, respectively. Then, we have
$$U(\mathcal{P}) \leq U(\mathcal{P}') + U(\mathcal{P}''). \tag{1}$$
- A6 *Additivity*: If $\mathcal{P}'$ and $\mathcal{P}''$ are independent, (1) holds with equality.

In axiom A6, independence refers to a suitable notion for independence of credal sets, see e.g. Couso et al. [1999]. An axiomatic definition of properties for uncertainty measures is a common approach in the literature [Pal et al., 1992, 1993]. Examples of credal uncertainty measures that satisfy some of the axioms A1–A6 are the maximal entropy [Abellan and Moral, 2003] and the generalized Hartley measure [Abellán and Moral, 2000].

Recall that the lower probability $\underline{P}$ of $\mathcal{P}$ is defined as $\underline{P}(A) := \inf_{P \in \mathcal{P}} P(A)$, for all $A \in \sigma(\mathcal{Y})$, and call *upper probability* its conjugate $\overline{P}(A) := 1 - \underline{P}(A^c) = \sup_{P \in \mathcal{P}} P(A)$, for all $A \in \sigma(\mathcal{Y})$. Since we are concerned with the fundamental question of whether the volume functional is a suitable measure for *epistemic* uncertainty, we replace A4 with the following axiom that better suits our purposes.

- A4' *Probability consistency*: $U(\mathcal{P})$ reduces to 0 as the distance between $\overline{P}(A)$ and $\underline{P}(A)$ goes to 0, for all $A \in \sigma(\mathcal{Y})$.

While A4' addresses solely the epistemic component of uncertainty assoicated with the credal set $\mathcal{P}$, A4 incorporates the aleatoric uncertainty. Finally, we introduce a seventh axiom that subsumes a desirable property of U proposed by Hüllermeier et al. [2022, Theorem 1.A3-A5].

- A7 *Invariance*: U is invariant to rotation and translation.

Call d the cardinality of the label space $\mathcal{Y}$. In the next section, we will note that many of these axioms are satisfied by the volume operator in the case $d = 2$ but can no longer be guaranteed for $d > 2$.

4 GEOMETRY OF EPISTEMIC UNCERTAINTY

As pointed out in Section 3, there is no unambiguous measure of (credal) uncertainty for machine learning purposes. In this section, we present a measure for EU rooted in the geometric concept of volume and show how it is well-suited for a binary classification setting, while it loses its appeal when moving to a multi-class setting.

Since we are considering a classification setting, we assume that $\mathcal{Y}$ is a finite Polish space so that $|\mathcal{Y}| = d$, for some natural number $d \geq 2$. We also let $\sigma(\mathcal{Y}) = 2^{\mathcal{Y}}$ to work with the finest possible σ-algebra of $\mathcal{Y}$; the results we provide still hold for any coarser σ-algebra.[1] Because $\mathcal{Y}$ is Polish, $\mathcal{M}(\mathcal{Y}, \sigma(\mathcal{Y}))$ is Polish as well. In particular, the topology endowed to $\mathcal{M}(\mathcal{Y}, \sigma(\mathcal{Y}))$ is the weak topology, which – because we assumed $\mathcal{Y}$ to be finite – coincides with the topology $\tau_{\|\cdot\|_2}$ induced by the Euclidean norm. Consider a credal set $\mathcal{P} \subset \mathcal{M}(\mathcal{Y}, \sigma(\mathcal{Y}))$, which can be seen as the outcome of a procedure involving an imprecise Bayesian neural network (IBNN) [Caprio et al., 2023a], or an imprecise neural network (INN) [Caprio et al., 2023b]; an ensemble-based approach is proposed by Shaker and Hüllermeier [2020].

Since $\mathcal{Y} = \{y_1, \dots, y_d\}$, each element $P \in \mathcal{P}$ can be seen as a d-dimensional probability vector, $P = (p_1, \dots, p_d)^\top$, where $p_j = P(\{y_j\})$, $j \in \{1, \dots, d\}$, $p_j \geq 0$, for all $j \in \{1, \dots, d\}$, and $\sum_{j=1}^d p_j = 1$. This entails that if we denote by Δ^{d-1} the unit simplex in $\mathbb{R}^d$, we have $\mathcal{P} \subset \Delta^{d-1}$, which means that $\mathcal{P}$ is a convex body inscribed in Δ^{d-1}.[2]

Intuitively, the "larger" $\mathcal{P}$ is, the higher the credal uncertainty. A natural way of capturing the size of $\mathcal{P}$, then, ap-

[1] Call τ the topology on $\mathcal{Y}$. The ideas expressed in this paper can be easily extended to the case where $\mathcal{Y}$ is not Polish. We require it to convey our results without dealing with topological subtleties.

[2] In the remaining part of the paper, we denote by $\mathcal{P}$ both the credal set and its geometric representation, as no confusion arises.

pears to be its volume $\text{Vol}(\mathcal{P})$. Notice that $\text{Vol}(\mathcal{P})$ is a bounded quantity: its value is bounded from below by 0 and from above by $\sqrt{d}/[(d-1)!]$, the volume of the whole unit simplex Δ^{d-1}. The latter corresponds to the case where $\mathcal{P} = \Delta^{d-1}$, that is, to the case of completely vacuous beliefs: the agent is only able to say that the probability of A is in $[0,1]$, for all $A \in \mathcal{F}$. In this sense, the volume is a measure of the size of set $\mathcal{P}$ that increases the more uncertain the agent is about the elements of $\mathcal{F}$. This argument shows that $\text{Vol}(\mathcal{P})$ is well suited to capture credal uncertainty. But why is it appropriate to describe EU?[3] Think of the extreme case where EU does not exist, so that the agent faces AU only. In that case, they would be able to specify a unique probability measure $P \in \mathcal{M}(\mathcal{Y}, \sigma(\mathcal{Y}))$ (or equivalently, $P \in \Delta^{d-1}$), and $\text{Vol}(\{P\}) = 0$. Hence, if $\text{Vol}(\mathcal{P}) > 0$, then this means that the agent faces EU. In addition, let $(\mathcal{P}_n)_{n\in\mathbb{N}}$ be a sequence of credal sets on $(\mathcal{Y}, \sigma(\mathcal{Y}))$ representing successive refinements of $\mathcal{P}$ computed as new data becomes available to the agent.[4] If, after observing enough evidence, the EU is resolved, that is, if $\lim_{n\to\infty}[\overline{P}_n(A) - \underline{P}_n(A)] = 0$ for all $A \in \mathcal{F}$, we see that the following holds. Sequence $(\mathcal{P}_n)_{n\in\mathbb{N}}$ converges – say in the Hausdorff metric – as $n \to \infty$ to $\mathcal{P}^\star \subset \mathcal{M}(\mathcal{Y}, \sigma(\mathcal{Y}))$ such that $|\mathcal{P}^\star| = |\mathcal{P}_n|$, for all n, and all the elements of $\mathcal{P}^\star$ are equal to $P^\star$, the (unique) probability measure that encapsulates the AU.[5] Through the learning process, we refine our estimates for the "true" underlying aleatoric uncertainty (pertaining to $P^\star$), which is left after all the EU is resolved. Then, the geometric representation of $\mathcal{P}^\star$ is a point whose volume is 0. Hence, we have that the volume of $\mathcal{P}_n$ converges from above to 0 (that is, it possesses the continuity property), which is exactly the behavior we would expect as EU resolves.

As we shall see, while this intuitive explanation holds if $d = 2$, for $d > 2$, continuity fails, thus making the volume not suited to capture EU in a multi-class classification setting. We also show in Theorem 1 that the volume lacks robustness in higher dimensions. Small perturbations to the boundary of a credal set make its volume vary significantly. This may seriously hamper the results of a study, leading to potentially catastrophic consequences in downstream tasks.

4.1 Vol($\mathcal{P}$): A GOOD MEASURE FOR EU, BUT ONLY IF $d = 2$

Let $d = 2$ so that $\mathcal{P}$ is a subset of Δ^1, the segment linking the points $(1,0)$ and $(0,1)$ in a 2-dimensional Cartesian plane. Notice that in this case, the volume $\text{Vol}(\mathcal{P})$ corresponds to the length of the segment. In this context, $\text{Vol}(\mathcal{P})$ is an appealing measure to describe the EU associated with the credal set $\mathcal{P}$.

Proposition 1. *Vol($\cdot$) satisfies axioms A1–A3, A4', A5 and A7 of Section 3.*

Let us now discuss additivity (axiom A6 of Section 3). Suppose the label space $\mathcal{Y} = \{(y_1, y_2), (y_3, y_4)\}$ can be written as $\mathcal{Y}_1 \times \mathcal{Y}_2$, where $\mathcal{Y}_1 = \{y_1, y_3\}$ and $\mathcal{Y}_2 = \{y_2, y_4\}$. Let $\mathcal{P}$ be a joint credal set on $\mathcal{Y}$ such that $\mathcal{P}'$ is the marginal credal set on $\mathcal{Y}_1$ and $\mathcal{P}''$ is the marginal credal set on $\mathcal{Y}_2$. In the proof of Proposition 1, we show that if $y_1 \neq y_3$ and $y_2 \neq y_4$,[6] then the volume is sub-additive. Suppose instead now that $y_1 = y_3 = y_\star$, so that $|\mathcal{Y}| = |\mathcal{Y}_2| = 2$ and $|\mathcal{Y}_1| = 1$.[7] Then, the marginal $\text{marg}_{\mathcal{Y}_1}(P) = P'$ of any $P \in \mathcal{P}$ on $\mathcal{Y}_1$ will give probability 1 to $y_\star$; in formulas, $P'(y_\star) = 1$. This entails that $\mathcal{P}' = \{P'\}$ is a singleton and that its geometric representation is a point.[8] Then, for all $P \in \mathcal{P}$, $P((y_1, y_2)) = P''(y_2)$ and $P((y_3, y_4)) = P''(y_4)$, where $\text{marg}_{\mathcal{Y}_2}(P) = P''$ is the marginal of any $P \in \mathcal{P}$ on $\mathcal{Y}_2$.

In turn, this line of reasoning implies that $\text{Vol}(\mathcal{P}') + \text{Vol}(\mathcal{P}'') = 0 + \text{Vol}(\mathcal{P}) = \text{Vol}(\mathcal{P})$, which shows that the volume is additive in this case.

This situation corresponds to an instance of strong independence (SI) [Couso et al., 1999, Section 3.5]. We have SI if and only if

$$\mathcal{P} = \text{Conv}(\{P \in \mathcal{M}(\mathcal{Y}, \sigma(\mathcal{Y}))\colon \text{marg}_{\mathcal{Y}_1}(P) \in \mathcal{P}' \text{ and } \text{marg}_{\mathcal{Y}_2}(P) \in \mathcal{P}''\}). \tag{2}$$

In other words, there is complete lack of interaction between the probability measure on $\mathcal{Y}_1$ and those on $\mathcal{Y}_2$. To see that this is the case, recall that $\mathcal{P}$ is a credal set, and so is convex; recall also that $\mathcal{P}' = \{P'\}$ is a singleton. Then, pick any $P^{ex} \in \text{ex}(\mathcal{P})$, where $\text{ex}(\mathcal{P})$ denotes the set of extreme elements of $\mathcal{P}$. We have that $\text{marg}_{\mathcal{Y}_1}(P^{ex}) = P'$, and so $\text{marg}_{\mathcal{Y}_2}(P^{ex}) \in \text{ex}(\mathcal{P}'')$. With a slight abuse of notation, we can write $\text{ex}(\mathcal{P}) = \{P'\} \times \text{ex}(\mathcal{P}'')$. This immediately implies that the equality in (2) holds. As pointed out in [Couso et al., 1999, Section 3.5], SI implies independence of the marginal sets, epistemic independence of the marginal experiments, and independence in the selection [Couso et al., 1999, Sections 3.1, 3.4, and 3.5, respectively]. It is, therefore, a rather strong notion of independence.

The volume is also trivially additive if $(y_1, y_2) = (y_3, y_4)$, but in that case $\mathcal{Y}$ would be a multiset.

The argument put forward so far can be summarized in the following proposition.

[3] The concept of volume has been explored in the imprecise probabilities literature, see e.g., Bloch [1996], [Cuzzolin, 2021, Chapter 17], and Seidenfeld et al. [2012], but, to the best of our knowledge, has never been tied to the notion of epistemic uncertainty. More in general, the geometry of imprecise probabilities has been studied, e.g., by Anel [2021], Cuzzolin [2021].

[4] Clearly $|\mathcal{P}_n| = |\mathcal{P}|$, for all n.

[5] Technically $\mathcal{P}^\star$ is a multiset, that is, a set where multiple instances for each of its elements are allowed.

[6] This implies that $|\mathcal{Y}| = |\mathcal{Y}_1| = |\mathcal{Y}_2| = 2$.

[7] A similar argument will hold if we assume $y_2 = y_4 = y^\star$, so that $|\mathcal{Y}| = |\mathcal{Y}_1| = 2$ and $|\mathcal{Y}_2| = 1$.

[8] Or, alternatively, $\mathcal{P}'$ is a multiset whose elements are all equal.

Proposition 2. *Let $\mathcal{Y} = \{(y_1, y_2), (y_3, y_4)\}$. Vol$(\cdot)$ satisfies axiom A6 if we assume the instance of SI given by either of the following*

- $y_1 = y_3$,
- $y_2 = y_4$,
- $y_1 = y_3$ *and* $y_2 = y_4$.

If $d > 2$, the volume ceases to be an appealing measure for EU. This is because quantifying the uncertainty associated with a credal set becomes challenging due to the dependency of the volume on the dimension. So far, we have written Vol in place of Vol_{d-1} to ease notation, but for $d > 2$ the dimension with respect to which the volume is taken becomes crucial. Let us give a simple example to illustrate this.

Example 1. *Let $d = 3$, so that the unit simplex is $\Delta^{3-1} = \Delta^2$, the triangle whose extreme points are $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ in a 3-dimensional Cartesian plane (the purple triangle in Figure 2). Consider a sequence $(\mathcal{P}_n)$ of credal sets whose geometric representations are triangles, and suppose their height reduces to 0 as $n \to \infty$, so that the (geometric representation of) $\mathcal{P}_\infty$ – the limit of $(\mathcal{P}_n)$ in the Hausdorff metric – is a segment. The limiting set $\mathcal{P}_\infty$, then, is not of full dimensionality that is, its geometric representation is a proper subset of Δ^1, while the geometric representation of $\mathcal{P}_n$ is a proper subset of Δ^2, for all n. This implies that Vol$_2(\mathcal{P}_\infty) = 0$, but – unless $\mathcal{P}_\infty$ is a degenerate segment, i.e. a point – Vol$_1(\mathcal{P}_\infty) > 0$. As we can see, the EU has not resolved, yet $\mathcal{P}_\infty$ has a zero 2-dimensional volume; this is clearly undesirable. It is easy to see how this problem exacerbates in higher dimensions.*

There are two possible ways one could try to circumvent the issue in Example 1; alas, both exhibit shortcomings, that is, at least one of the axioms A1–A3, A4', A5–A7 in Section 3 is not satisfied. The first one is to consider the volume operator $\text{Vol}(\mathcal{P})$ as the volume taken with respect to the space in which set $\mathcal{P}$ is of full dimensionality. In this case, we immediately see how A2 fails. Considering again the sequence in Example 1, we would have a sequence $(\mathcal{P}_n)$ whose volume $\text{Vol}_2(\mathcal{P}_n)$ is going to zero. However, in the limit, its volume $\text{Vol}_1(\mathcal{P}_\infty)$ would be positive. Axiom A3 fails as well: consider a credal set $\mathcal{P}$ whose representation is a triangle having base b and height h and suppose $h < 2$. Consider then a credal set $\mathcal{Q} \subsetneq \mathcal{P}$ whose representation is a segment having length $\ell = b$. Then, $\text{Vol}_2(\mathcal{P}) = b \cdot h/2 < b$, while $\text{Vol}_1(\mathcal{Q}) = \ell = b$.

The second one is to consider lift probability sets; let us discuss this idea in depth. Let $d, d' \in \mathbb{N}$, and let $d' < d$. Call

$$O(d', d) := \{V \in \mathbb{R}^{d' \times d} : VV^\top = I_{d'}\},$$

where $I_{d'}$ is the d'-dimensional identity matrix. That is, $O(d', d)$ is the *Stiefel manifold* of $d' \times d$ matrices with orthonormal rows [Cai and Lim, 2022]. Then, for any $V \in O(d', d)$ and any $b \in \mathbb{R}^{d'}$, define

$$\varphi_{V,b} : \mathbb{R}^d \to \mathbb{R}^{d'}, \quad x \mapsto \varphi_{V,b}(x) := Vx + b.$$

Suppose now that, for some n, (the geometric representation of) $\mathcal{P}_n$ is a proper subset of Δ^{d-1}, while (the geometric representation of) $\mathcal{P}_{n+1}$ is a proper subset of $\Delta^{d'-1}$. Pick any $V \in O(d', d)$ and any $b \in \mathbb{R}^{d'}$; an embedding of $\mathcal{P}_{n+1}$ in Δ^{d-1} is a set K such that for all $x \in K$, there exists a probability vector $p \in \mathcal{P}_{n+1}$ such that $\varphi_{V,b}(x) = p$. Call $\Phi^+(\mathcal{P}_{n+1}, d)$ the set of embeddings of $\mathcal{P}_{n+1}$ in Δ^{d-1}, and assume that it is nonempty.

Then, define

$$\breve{\mathcal{P}}_{n+1} := \underset{K \in \Phi^+(\mathcal{P}_{n+1}, d)}{\arg\min} \left|\text{Vol}_{d-1}(K) - \text{Vol}_{d'-1}(\mathcal{P}_{n+1})\right|;$$

we call it the *lift probability set* for the heuristic similarity with lift zonoids [Mosler, 2002]. We define it in this way because we want the d-dimensional set whose (full dimensionality) volume is the closest possible to the (d'-dimensional) volume of $\mathcal{P}_{n+1}$. A simple example is the following. Suppose the geometric representation of $\mathcal{P}_n$ is a proper subset of Δ^2, and that the geometric representation of $\mathcal{P}_{n+1}$ is a proper subset of Δ^1. So the former is a subset of $\mathbb{R}^2$, and the latter is a segment in $\mathbb{R}$. Then, a possible $\breve{\mathcal{P}}_{n+1}$ is any triangle in Δ^2 whose height h is 2 and whose base length b is equal to the length ℓ of the segment representing $\mathcal{P}_{n+1}$. This because the area of such $\breve{\mathcal{P}}_{n+1}$ is $b \cdot h/2$; if $h = 2$ and $b = \ell$, then $\text{Vol}_2(\breve{\mathcal{P}}_{n+1}) = \text{Vol}_1(\mathcal{P}_{n+1})$, which is what we wanted. A visual representation is given in Figure 2.

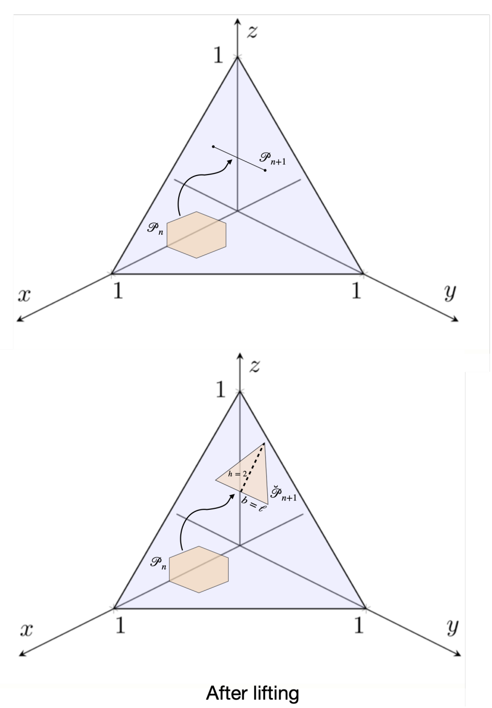

Figure 2: A visual representation of a lift probability set.

Notice that $\breve{\mathcal{P}}_{n+1}$ is well defined because $\Phi^+(\mathcal{P}_{n+1}, d) \subset 2^{\Delta^{d-1}}$, and Δ^{d-1} is compact.[9] We can then compare $\text{Vol}_{d-1}(\mathcal{P}_n)$ and of $\text{Vol}_{d-1}(\breve{\mathcal{P}}_{n+1})$, and also compute the relative quantity

$$\frac{\left|\text{Vol}_{d-1}(\mathcal{P}_n) - \text{Vol}_{d-1}(\breve{\mathcal{P}}_{n+1})\right|}{\text{Vol}_{d-1}(\mathcal{P}_n)}$$

that captures the variation in volume between $\mathcal{P}_n$ and $\breve{\mathcal{P}}_{n+1}$. Alas, in this case, too, it is easy to see how A2 fails. Consider the same sequence as in Example 1. We would have that $\text{Vol}_2(\mathcal{P}_n)$ goes to zero as $n \to \infty$, but $\text{Vol}_2(\breve{\mathcal{P}}_\infty) > 0$. Axiom A3 may fail as well since we could find credal sets $\mathcal{P} \subset \Delta^{d-1}$ and $\mathcal{Q} \subset \Delta^{d'-1}$ such that $\mathcal{Q} \subsetneq \mathcal{P}$, but $\breve{\mathcal{Q}} \not\subset \mathcal{P}$.

4.2 LACK OF ROBUSTNESS IN HIGHER DIMENSIONS

In this section, we show how, if we measure the EU associated with a credal set on the label space using the volume, as the number of labels grows, "small" changes of the uncertainty representation may lead to catastrophic consequences in downstream tasks.

For a generic compact set $K \in \mathbb{R}^d$ and a positive real r, the *r-packing* of K, denoted by $\text{Pack}_r(K)$, is the collection of sets K' that satisfy the following properties

(i) $K' \subset K$,

(ii) $\cup_{x \in K'} B_r^d(x) \subset K$, where $B_r^d(x)$ denotes the ball of radius r in space $\mathbb{R}^d$ centered at x,

(iii) the elements of $\{B_r^d(x)\}_{x \in K'}$ are pairwise disjoint,

(iv) there does not exist $x' \in K$ such that (i)-(iii) are satisfied by $K' \cup \{x'\}$.

The *packing number* of K, denoted by $N_r^{\text{pack}}(K)$, is given by $\max_{K' \in \text{Pack}_r(K)} |K'|$. We also let $K_r^\star := \arg\max_{K' \in \text{Pack}_r(K)} |K'|$ and $\tilde{K}_r := \cup_{x \in K_r^\star} B_r^d(x)$. Notice that

$$\text{Vol}(\tilde{K}_r) = c(r, d, K)\text{Vol}(K), \tag{3}$$

where

$$\begin{aligned} &c(r, d, K) \in (0, 1], \text{ for all } r > 0, \\ \text{and} \quad &c(r, d, K) \leq c(r - \epsilon, d, \breve{K}), \text{ for all } \epsilon > 0, \end{aligned} \tag{4}$$

where $\breve{K}$ is any compact set in $\mathbb{R}^d$, possibly different than K. That is, we can always find a real number $c(r, d, K)$ depending on the dimension d of the Euclidean space, on the radius r of the balls, and on the set K of interest, that relates the volume of K and that of $\tilde{K}_r$. Being in $(0, 1]$, it takes into account the fact that since $\tilde{K}_r$ is a union of pairwise disjoint balls *within* K, its volume cannot exceed that of K. This is easy to see in Figure 3. The second condition in (4) states that irrespective of the compact set of interest, we retain more of the volume of the original set if we pack it using balls of a smaller radius.

To give a simple illustration, consider $r_1, r_2 > 0$ such that $r_1 \leq r_2$. Then, by (3) and (4), we have that $\text{Vol}(K) - \text{Vol}(\tilde{K}_{r_2}) = \text{Vol}(K)[1 - c(r_2, d, K)] \geq \text{Vol}(K)[1 - c(r_1, d, K)] = \text{Vol}(K) - \text{Vol}(\tilde{K}_{r_1})$. This means that the difference in volume between K and $\tilde{K}_{r_2}$ is larger than that between K and $\tilde{K}_{r_1}$.

Let $\mathcal{K}(\mathbb{R}^d)$ be the class of compact sets in $\mathbb{R}^d$, and call $c(r, d) := \max_{K \in \mathcal{K}(\mathbb{R}^d)} c(r, d, K)$. As r goes to 0, $c(r, d)$ increases to its optimal value that we denote as $c^\star(d)$. The values of $c^\star(d)$ have only been found for $d \in \{1, 2, 3, 8, 24\}$ [Cohn et al., 2017, Viazovska, 2017]. The fact that $c(r, d)$ increases as r decreases to 0 captures the idea that using balls of smaller radius leads to a better approximation of the volume of the compact set K in $\mathbb{R}^d$ that is being packed.

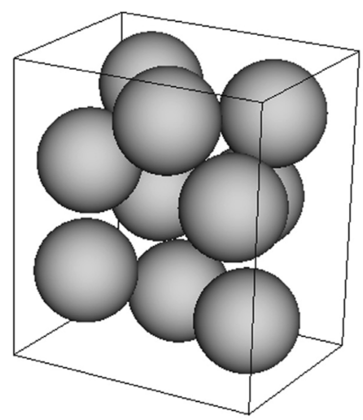

Figure 3: A representation of $\tilde{K}_r$, for some $r > 0$, where K is a parallelepiped in $\mathbb{R}^3$. This figure replicates [Hifi and Yousef, 2019, Figure 4].

Suppose our credal set $\mathcal{P}$ is compact, so to be able to use the concepts of r-packing and packing number. Consider then a set $\mathcal{Q} \subset \mathcal{M}(\Omega, \mathcal{F})$ that satisfies the following three properties:

(a) $\mathcal{Q} \subsetneq \mathcal{P}$, so that $\mathcal{Q}' := \mathcal{P} \setminus \mathcal{Q} \neq \emptyset$,

(b) $d_H(\mathcal{P}, \mathcal{Q}) = \epsilon$, for some $\epsilon > 0$,

(c) ϵ is such that we can find $r > 0$ for which $N_r^{\text{pack}}(\mathcal{P}) \geq N_{r-\epsilon}^{\text{pack}}(\mathcal{Q}')$.

Property (a) tells us that $\mathcal{Q}$ is a proper subset of $\mathcal{P}$. Let d_2 denote the metric induced by the Euclidean norm $\|\cdot\|_2$. Property (b) tells us that the Hausdorff distance

$$d_H(\mathcal{P}, \mathcal{Q}) = \max\left\{\max_{P \in \mathcal{P}} d_2(P, \mathcal{Q}), \max_{Q \in \mathcal{Q}} d_2(\mathcal{P}, Q)\right\} \tag{5}$$

between $\mathcal{P}$ and $\mathcal{Q}$ is equal to some $\epsilon > 0$. Property (c) ensures that ϵ is "not too large". To understand why, notice

[9] If the arg min is not a singleton, pick any of its elements.

that if ϵ is "large", that is, if it is close to r, then the packing number of $\mathcal{Q}' \subsetneq \mathcal{P}$ using balls of radius $r-\epsilon$ can be larger than the packing number of $\mathcal{P}$ using balls of radius r.[10] Requiring (c) ensures us that this does not happen, and therefore that ϵ is "small". A representation of $\mathcal{P}$ and $\mathcal{Q}$ satisfying (a)–(c) is given in Figure 4. A (possibly very small) change in uncertainty representation is captured by a situation in which the agent specifies credal set $\mathcal{Q}$ in place of $\mathcal{P}$. We are ready to state the main result of this section.

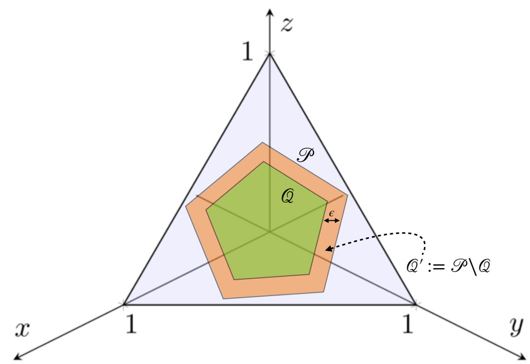

Figure 4: A representation of $\mathcal{P}$ (the orange pentagon) and $\mathcal{Q}$ (the green pentagon) satisfying (a)-(c) when the dimension of state space Ω is $d=3$. The unit simplex Δ^2 in $\mathbb{R}^3$ is given by the purple triangle whose vertices are the elements of the basis of $\mathbb{R}^3$, i.e., $e_1=(1,0,0)$, $e_2=(0,1,0)$, and $e_2=(0,0,1)$.

Theorem 1. *Let Ω be a finite Polish space so that $|\Omega|=d$, and let $\mathcal{F}=2^\Omega$. Pick any compact set $\mathcal{P}\subset\mathcal{M}(\Omega,\mathcal{F})$, and any set $\mathcal{Q}$ that satisfies (a)-(c). The following holds*

$$\frac{Vol(\mathcal{P})-Vol(\mathcal{Q}')}{Vol(\mathcal{P})} \geq 1-\left(1-\frac{\epsilon}{r}\right)^d. \tag{6}$$

Notice that we implicitly assumed that at least a $\mathcal{Q}$ satisfying (a)-(c) exists. We have that $[\text{Vol}(\mathcal{P})-\text{Vol}(\mathcal{Q}')]/\text{Vol}(\mathcal{P})\in[0,1]$; in light of this, since $1-(1-\epsilon/r)^d\to 1$ as $d\to\infty$, Theorem 1 states that as d grows, most of the volume of $\mathcal{P}$ concentrates near its boundary.

As a result, if we use the volume operator as a metric for the EU, this latter is very sensitive to perturbations of the boundary of the (geometric representation of the) credal set; this is problematic for credal sets in the context of ML. Suppose we are in a multi-classification setting such that the cardinality of $\mathcal{Y}$ is some large number d. Suppose that two different procedures produce two different credal sets on $\mathcal{Y}$; call one $\mathcal{P}$ and the other $\mathcal{Q}$, and suppose $\mathcal{Q}$ satisfies (a)-(c). This means that the uncertainty representations associated with the two procedures differ only by a "small amount". For instance, this could be the result of an agent specifying "slightly different" credal prior sets. This may well happen since defining the boundaries of credal sets is usually quite an arbitrary task to perform. Then, this would result in a (possibly massive) underestimation of the epistemic uncertainty in the results of the analysis, which would potentially translate in catastrophic consequence in downstream tasks. In Example 2, we describe a situation in which Theorem 1 is applied to credal prior sets.

Example 2. *Assume for simplicity that the parameter space Θ is finite and that its cardinality is $\mathfrak{c}$. Suppose an agent faces complete ignorance regarding the probabilities to assign to the elements of 2^Θ. Although tempting, there is a pitfall in choosing the whole simplex $\Delta^{\mathfrak{c}-1}$ as the credal prior set. As shown by Walley [1991, Chapter 5], completely vacuous beliefs – captured by choice of $\Delta^{\mathfrak{c}-1}$ as a credal prior set – cannot be Bayes-updated. This means that the posterior credal set will again be $\Delta^{\mathfrak{c}-1}$: no large amount of data is enough to swamp the prior. Instead, suppose that the agent considers a credal prior set $\Delta_\epsilon^{\mathfrak{c}-1}$ that satisfies (a)–(c). If $\mathfrak{c}$ is large enough, this would mean that Vol$(\Delta_\epsilon^{\mathfrak{c}-1})$ is much smaller than Vol$(\Delta^{\mathfrak{c}-1})$.*

Two remarks are in order. First, in the binary classification setting (that is, when $d=2$), the lack of robustness of the volume highlighted by Theorem 1 is not an issue since $1-(1-\epsilon/r)^d$ is approximately 1 only when the cardinality $|\mathcal{Y}|=d$ is large. Second, Theorem 1 is intimately related to Carl-Pajor's Theorem [Ball and Pajor, 1990, Theorem 1]; this implies that in the future, more techniques from high-dimensional geometry may become useful in the study of epistemic, and potentially also aleatoric, uncertainties.[11]

5 CONCLUSION

Credal sets provide a flexible and powerful formalism for representing uncertainty in various scientific disciplines. In particular, uncertainty representation via credal sets can capture different degrees of uncertainty and allow for a more nuanced representation of epistemic and aleatoric uncertainty in machine learning systems. Moreover, the corresponding geometric representation of credal sets as d-dimensional polytopes enables a thoroughly intuitive view of uncertainty representation and quantification.

In this paper, we showed that the volume of a credal set is a sensible measure of epistemic uncertainty in the context of binary classification, as it enjoys many desirable properties suggested in the existing literature. On the other side, the volume forfeits these properties in a multi-class classification setting, despite its intuitive meaningfulness.

In addition, this work stimulates a fundamental question as to what extent a geometric approach to uncertainty quantification (in ML) is sensible.

[10] Because $\mathcal{Q}\subsetneq\mathcal{P}$ and $d_H(\mathcal{P},\mathcal{Q})=\epsilon$, packing using balls of radius $r-\epsilon$ is a sensible choice.

[11] We state (a version of) Carl-Pajor's Theorem in Appendix **??**.

This is the first step toward studying the geometric properties of (epistemic) uncertainty in AI and ML. In the future, we plan to explore the geometry of aleatoric uncertainty and introduce techniques from high-dimensional geometry and high-dimensional probability to enhance and deepen the study of EU and AU in the contexts of AI and ML.

Author Contributions

Yusuf Sale and Michele Caprio contributed equally to this paper.

Acknowledgements

Michele Caprio would like to acknowledge partial funding by the Army Research Office (ARO MURI W911NF2010080). Yusuf Sale is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

References

Joaquín Abellán and Serafín Moral. A non-specificity measure for convex sets of probability distributions. *International journal of uncertainty, fuzziness and knowledge-based systems*, 8(03):357–367, 2000.

Joaquín Abellán and Serafín Moral. Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems*, 18(12):1215–1225, 2003.

Joaquin Abellan and Serafin Moral. Maximum of entropy for credal sets. *International journal of uncertainty, fuzziness and knowledge-based systems*, 11(05):587–597, 2003.

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

Mathieu Anel. *The Geometry of Ambiguity: An Introduction to the Ideas of Derived Geometry*, volume 1, pages 505–553. Cambridge University Press, 2021.

Thomas Augustin, Frank PA Coolen, Gert De Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

Keith Ball and Alain Pajor. Convex bodies with few faces. *Proceedings of the American Mathematical Society*, 110 (1):225–231, 1990.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in Neural Information Processing Systems*, 2022.

Isabelle Bloch. Some aspects of Dempster-Shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account. *Pattern Recognition Letters*, 17(8):905–919, 1996.

Andrey Bronevich and George J Klir. Axioms for uncertainty measures on belief functions and credal sets. In *NAFIPS 2008-2008 Annual Meeting of the North American Fuzzy Information Processing Society*, pages 1–6. IEEE, 2008.

Andrey Bronevich and George J Klir. Measures of uncertainty for imprecise probabilities: an axiomatic approach. *International journal of approximate reasoning*, 51(4): 365–390, 2010.

Yuhang Cai and Lek-Heng Lim. Distances between probability distributions of different dimensions. *IEEE Transactions on Information Theory*, 2022.

Michele Caprio, Souradeep Dutta, Radoslav Ivanov, Kuk Jang, Vivian Lin, Oleg Sokolsky, and Insup Lee. Imprecise Bayesian Neural Networks. *arXiv preprint arXiv:2302.09656*, 2023a.

Michele Caprio, Souradeep Dutta, Kaustubh Sridhar, Kuk Jang, Vivian Lin, Oleg Sokolsky, and Insup Lee. EpiC INN: Epistemic Curiosity Imprecise Neural Network. Technical report, University of Pennsylvania, Department of Computer and Information Science, 01 2023b.

Henry Cohn, Abhinav Kumar, Stephen D. Miller, Danylo Radchenko, and Maryna S. Viazovska. The sphere packing problem in dimension 24. *Annals of Mathematics*, 185(3):1017–1033, 2017.

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

Giorgio Corani, Alessandro Antonucci, and Marco Zaffalon. Bayesian networks with imprecise probabilities: Theory and application to classification. In *Data Mining: Foundations and Intelligent Paradigms*, pages 49–93. Springer, 2012.

Inés Couso, Serafín Moral, and Peter Walley. Examples of independence for imprecise probabilities. In *Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications (ISIPTA 1999)*, pages 121–130, 1999.

Fabio Cuzzolin. *The Geometry of Uncertainty*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer Nature Switzerland, 2021.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR, 2018.

Mhand Hifi and Labib Yousef. A local search-based method for sphere packing problems. *European Journal of Operational Research*, 247:482–500, 2019.

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

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

Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Uncertainty in Artificial Intelligence*, pages 548–557. PMLR, 2022.

Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? *Available at arxiv:2209.03302*, 2022.

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

Sanyam Kapoor, Wesley J Maddox, Pavel Izmailov, and Andrew Gordon Wilson. On uncertainty, tempering, and data augmentation in bayesian classification. *arXiv preprint arXiv:2203.16481*, 2022.

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

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

Isaac Levi. *The Enterprise of Knowledge*. London : MIT Press, 1980.

Karl Mosler. Zonoids and lift zonoids. In *Multivariate Dispersion, Central Regions, and Depth: The Lift Zonoid Approach*, volume 165 of *Lecture Notes in Statistics*, pages 25–78. New York : Springer, 2002.

Nikhil R Pal, James C Bezdek, and Rohan Hemasinha. Uncertainty measures for evidential reasoning i: A review. *International Journal of Approximate Reasoning*, 7(3-4): 165–183, 1992.

Nikhil R Pal, James C Bezdek, and Rohan Hemasinha. Uncertainty measures for evidential reasoning ii: A new measure of total uncertainty. *International Journal of Approximate Reasoning*, 8(1):1–16, 1993.

Teddy Seidenfeld, Mark J. Schervish, and Joseph B. Kadane. Forecasting with imprecise probabilities. *International Journal of Approximate Reasoning*, 53(8):1248–1261, 2012. Imprecise Probability: Theories and Applications (ISIPTA'11).

Robin Senge, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, 2014.

Burr Settles. Active Learning Literature Survey. Technical report, University of Wisconsin-Madison, Department of Computer Sciences, 2009.

Mohammad Hossein Shaker and Eyke Hüllermeier. Aleatoric and epistemic uncertainty with random forests. In *Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings 18*, pages 444–456. Springer, 2020.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

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

Kush R Varshney. Engineering safety in machine learning. In *2016 Information Theory and Applications Workshop (ITA)*, pages 1–5. IEEE, 2016.

Kush R Varshney and Homa Alemzadeh. On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, 5(3):246–255, 2017.

Maryna S. Viazovska. The sphere packing problem in dimension 8. *Annals of Mathematics*, 185(3):991–1015, 2017.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. London : Chapman and Hall, 1991.

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

Fan Yang, Hua-zhen Wang, Hong Mi, Cheng-de Lin, and Wei-wen Cai. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC bioinformatics*, 10(1):1–14, 2009.

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