Quantification of Credal Uncertainty in Machine Learning: A Critical Analysis and Empirical Comparison (supplementary material)

Eyke Hüllermeier[1] **Sebastian Destercke**[2] **Mohammad Hossein Shaker**[1]

[1]Institute of Informatics, University of Munich (LMU), Germany
[2]UMR CNRS 7253 Heudiasyc, Sorbonne Universités, Université de Technologie de Compiegne, France

1 PROOF OF THEOREM 1

Let us represent an interval $[a, b] \subseteq [0, 1]$ in the form $[\mu - \delta, \mu + \delta]$, where $0 \leq \mu \leq 1$ is the midpoint and $0 \leq \delta \leq \min(\mu, 1 - \mu)$ the width of the interval. There are two operations on such an interval that can increase (or vice versa decrease) uncertainty: shifting and widening. By shifting we mean moving the interval "closer to the middle", i.e., replacing $[\mu - \delta, \mu + \delta]$ by $[\mu' - \delta, \mu' + \delta]$ such that $|\mu' - 1/2| < |\mu - 1/2|$. By widening we mean increasing δ, i.e., replacing $[\mu - \delta, \mu + \delta]$ by $[\mu - \delta', \mu + \delta']$ such that $\delta' > \delta$.

To prove Theorem 1, note that every interval $[a, b]$ is equivalently expressed in terms of $[\mu - \delta, \mu + \delta]$, where $\mu = (a + b)/2$ and $\delta = (b - a)/2$. In the following, we use the uncertainty measure with both types of arguments, (a, b) and (μ, δ). To avoid confusion, we write U in the former and U' in the latter case, i.e., $U'(\mu, \delta) = U(\mu - \delta, \mu + \delta)$ and vice versa $U(a, b) = U((a + b)/2, (b - a)/2)$.

Due to the symmetry property A3, we can restrict our consideration to the subset $\mathbb{I}'$ of those intervals $[a, b]$ in $\mathbb{I}$ for which $1 - a \geq b$, and for which the transformation

$$\mathrm{TP}(a, b) := \frac{1}{1 + u(a, b)} = \min(1 - a, b)\,, \tag{1}$$

takes the value b. Indeed, suppose U is defined on $\mathbb{I}'$. Consider $[a, b] \notin \mathbb{I}$, i.e., $1 - a < b$, and let $a' = 1 - b$, $b' = 1 - a$. Then $1 - a' = b > 1 - a = b'$, hence $[a', b'] \in \mathbb{I}$ and $U(a, b) = U(a', b')$.

Note that $1 - a \geq b$ implies $\mu = (a + b)/2 \leq 1/2$. Thus, we need to consider U on those intervals $[\mu - \delta, \mu + \delta]$ for which $0 \leq \mu \leq 1/2$ and $0 \leq \delta \leq \mu$.

A4 implies that, for any $\delta' > 0$,

$$\frac{U'(\mu, \delta + \delta') - U'(\mu, \delta)}{\delta'}$$

is a constant, and hence (by letting $\delta \to 0$)

$$\frac{\partial U'(\mu, \delta)}{\partial \delta} = c'$$

for a constant $c' \geq 0$. Similarly, A5 implies that

$$\frac{\partial U'(\mu, 0)}{\partial \mu} = c$$

for a constant $c \geq 0$. As a consequence, together with $U(0, 0) = 0$ according to A1, U' is of the form

$$U'(\mu, \delta) = c \cdot \mu + c' \cdot \delta\,.$$

Moreover, since $U(0, 1) = U'(1/2, 1/2) = 1$ according to A2, $c/2 + c'/2 = 1$, and hence $c + c' = 2$. Finally, A6 implies $c = c'$, so that $U'(\mu, \delta) = \mu + \delta$, or equivalently,

$$U(a, b) = \frac{a + b}{2} + \frac{b - a}{2} = b = \min(1 - a, b)\,.$$

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

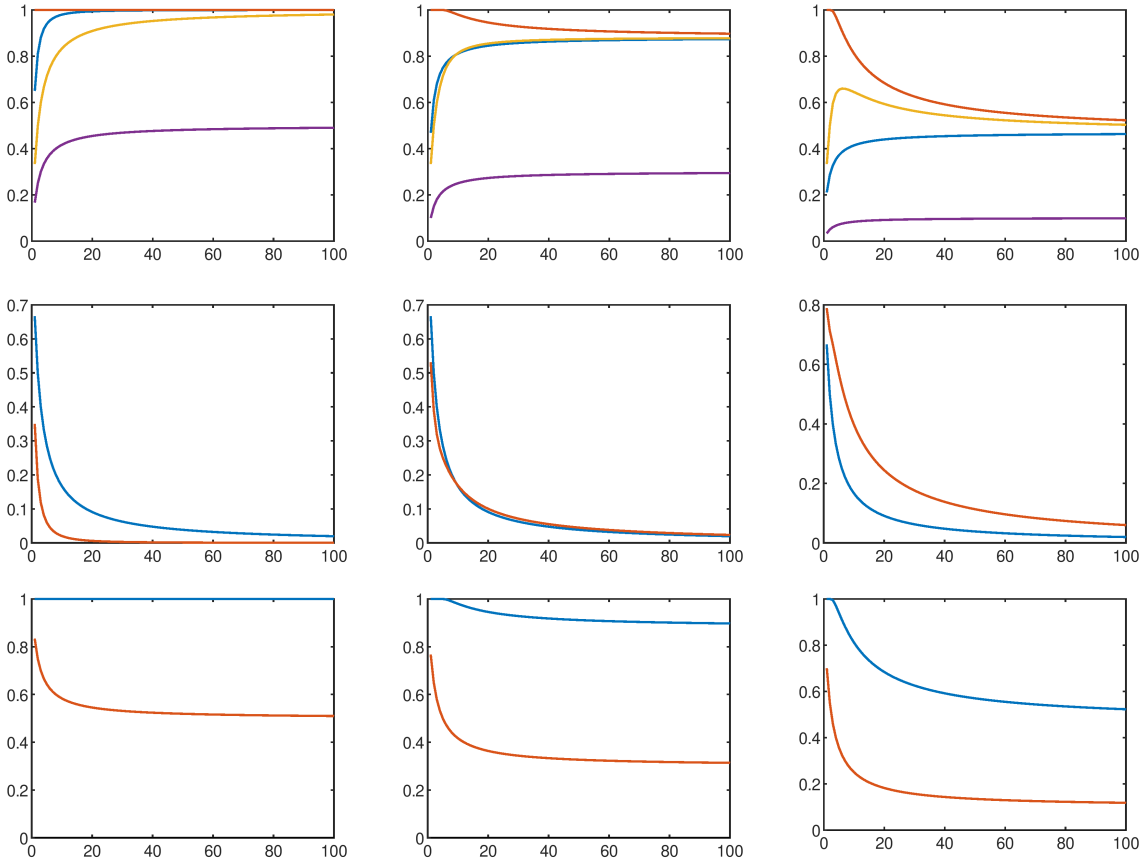

Figure 1: Upper panel: Aleatoric uncertainties AU (red), AL (blue), AD (yellow), AP (purple) for $\theta = 0.5$ (left), $\theta = 0.7$ (middle), $\theta = 0.9$ (right). Middle panel: Epistemic uncertainties EH (red), ES (blue) for $\theta = 0.5$ (left), $\theta = 0.7$ (middle), $\theta = 0.9$ (right). Lower panel: Total uncertainties TA (blue), TP (red) for $\theta = 0.5$ (left), $\theta = 0.7$ (middle), $\theta = 0.9$ (right).

Likewise, again exploiting symmetry, we find that $U(a, b) = 1 - a$ in the case $1 - a < b$ (for which $\mu > 1/2$). Therefore, $U(a, b) = \min(1 - a, b)$ for all $[a, b] \in \mathbb{I}$.

2 ILLUSTRATION: COIN TOSSING

As an illustration, let us consider (biased) coin tossing as a simple example: Given a sequence of outcomes so far, the problem is to predict the outcome of the next toss. Essentially, this is a problem of learning the parameter θ of a Bernoulli distribution, which corresponds to the bias of the coin. Although it may look like a toy example, this problem is actually quite relevant for ML, namely for estimating the probability θ of the positive class in binary classification, assuming this probability to be constant in a certain region of the instance space (like in nearest neighbor classification or decision tree learning).

Note that $\theta \in [0, 1]$ specifies the probability distribution $(\theta, 1 - \theta) \in \mathbb{P}(\{0, 1\})$, so that a credal set (a subset of $\mathbb{P}(\{0, 1\})$) can simply be represented by an interval $C \subseteq [0, 1]$. Credal inference can be done with the imprecise Dirichlet model, which, after n coin tosses, leads to a credal set of the form

$$C = \left[\frac{a}{n + s}, \frac{a + s}{n + s} \right] , \tag{2}$$

where a is the number of times the positive class occurred (coin landed heads up) and $s > 0$ is a parameter of the imprecise Dirichlet model (we take $s = 2$, which is often recommended in the literature).

The (expected) curves for the different uncertainty measures, i.e., the value of the respective measure (y-axis) as a function

of the number of trials (x-axis), are shown in Figures 1. Basically, the results confirm our expectation, though a few notable observations can be made.

As for aleatoric uncertainty, upper entropy is a monotone decreasing and lower entropy a monotone increasing function, forming an interval for the "true" entropy. Interestingly, the derived measure AD does not necessarily behave monotonically. For example, in the case $\theta = 0.9$, it first increases and then decreases, although both epistemic and total uncertainty are monotonically decreasing. This may appear somewhat questionable, just like the (semantic) interpretation of the measure itself. The measure AP takes smaller values overall and, as discussed above, cannot exceed the value $1/2$. As can be seen in the case $\theta = 0.5$, this value is assumed for the interval $[0.5, 0.5]$.

As for epistemic uncertainty, the curves are monotonically decreasing, which is clearly expected. Remarkably, however, according to the derived measure ES, the epistemic uncertainty is overall higher for $\theta = 0.9$ than for $\theta = 0.7$, which in turn is higher than the uncertainty for $\theta = 0.5$. Again, this appears to be somewhat counter-intuitive. This problem is obviously related to the lack of shift-invariance of upper entropy (differences in entropy are smaller around $1/2$ and bigger in the boundary regions).

As shown by our discussion so far, every attempt at combining the classical measures of aleatoric (conflict) and epistemic uncertainty (non-specificity) for credal sets turns out to be problematic. This is true for both AD as a derived measure of aleatoric uncertainty as well as for ES as derived measure of epistemic uncertainty. We consider this as an affirmation of our conjecture that the two types of uncertainty are of different nature and hence not fully compatible — implying that a decomposition of total into aleatoric and epistemic uncertainty may simply not work. We are left with GH as a natural measure of epistemic uncertainty and upper entropy as a meaningful measure of total uncertainty.

Apparently, the decomposition of the measure TP into AP and EP is more meaningful.

3 SUMMARY OF MEASURES

In the following, we summarize the measures that have been proposed. We also provide expressions for the Bernoulli case, i.e., where uncertainty about a binary outcome is represented in terms of an interval $[a, b]$ for the probability of the positive class. In this case, S^* corresponds to the entropy

$$S(q) := -\sum_{y \in \mathcal{Y}} q(y) \log_2 q(y)\,, \tag{3}$$

for the distribution q such that

$$q(+1) = \begin{cases} b & \text{if } b < 1/2 \\ a & \text{if } a > 1/2 \\ 1/2 & \text{otherwise} \end{cases},$$

i.e.,

$$S^* = \begin{cases} S(1/2, 1/2) = \log(2) & \text{if } 1/2 \in [a, b] \\ \max(S(a, 1-a), S(b, 1-b)) & \text{otherwise} \end{cases}$$

Likewise, S_* corresponds to (3) for the distribution q such that

$$q(+1) = \begin{cases} a & \text{if } b < 1/2 \\ b & \text{if } a > 1/2 \\ a & \text{if } 1/2 \in [a, b], 1/2 - a > b - 1/2 \\ b & \text{if } 1/2 \in [a, b], 1/2 - a \leq b - 1/2 \end{cases},$$

i.e., $S_* = \min(S(a, 1-a), S(b, 1-b))$.

Aleatoric uncertainty:

Lower:	AL	$=$	S_*
Upper:	AU	$=$	S^*
Derived:	AD	$=$	$S^* - \text{GH} \;=\; S^* - b + a$
Predictive:	AP	$=$	$\min(a, 1-b)$

Epistemic uncertainty:

Based on Hartley:	EH	$=$	$\mathrm{GH} = b - a$
Based on Shannon:	ES	$=$	$S^* - S_*$

Total uncertainty:

Axiomatic:	TA	$=$	S^*
Predictive:	TP	$=$	$\min(1-a, b)$

Note that the upper entropy S^* occurs twice, namely as an upper bound on the aleatoric uncertainty (AU) and as an axiomatically justified measure of total uncertainty (TA).

As an aside, let us note that one may also think of the sum

$$S_* + \mathrm{GH} \;=\; S_* + (b - a) \tag{4}$$

as a measure of total uncertainty. Indeed, given S_* and GH as well-justified measures of aleatoric and epistemic uncertainty, respectively, (4) appears quite natural. Besides, this measure could also be motivated by the decomposition

$$\begin{aligned} \mathrm{TP}(a,b) &= \underbrace{\min(1-a, b)}_{\text{total}} \\ &= \underbrace{\min(a, 1-b)}_{\text{aleatoric (AP)}} + \underbrace{(b-a)}_{\text{epistemic (EP)}} \quad . \end{aligned} \tag{5}$$

In fact, both expressions proceed from the generalized Hartley measure $b - a$ as a natural measure of epistemic uncertainty, to which they add an "optimistic" measure of aleatoric uncertainty: As explained above, just like the lower entropy S_*, the term $\min(a, 1-b)$ in (5) specifies a lower bound, as it corresponds to the "best" case in the sense of the most extreme among all conceivable probabilities. However, one can easily verify that (4), unlike (5), is not monotone, which is very undesirable for a measure of total uncertainty.

4 ACCURACY-REJECTION PLOTS

As mentioned in the experimental part of the paper, the evaluation of the uncertainty measures are done indirectly with the use of accuracy-rejection curves. The idea is to sort all of test instances based on their uncertainty and start rejecting the prediction from the most uncertain to the least uncertain, and then calculate the accuracy of the model on the test data-points that remain. Ideally, we expect the highly certain predictions to be correct most of the times. Fig. 2 shows the AR curves for the five UCI data sets that are not shown in the main paper.

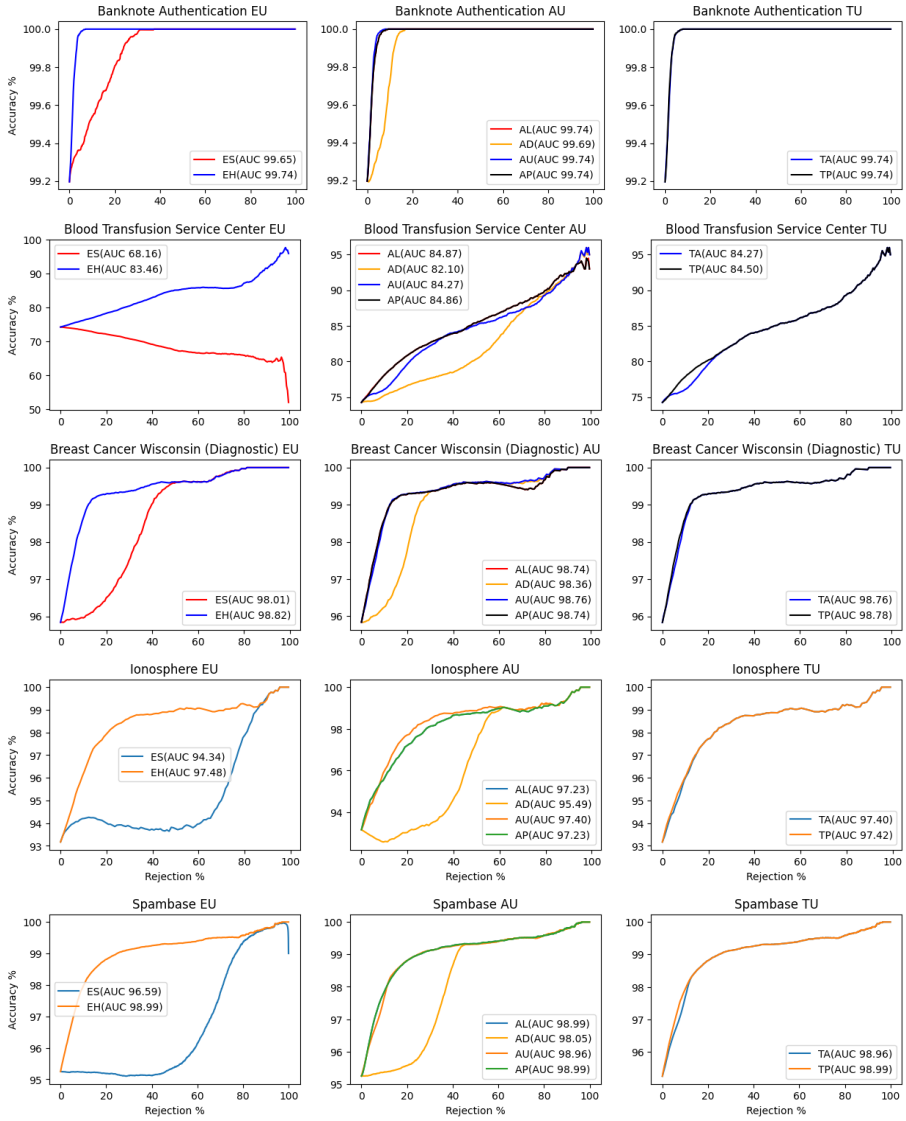

Figure 2: Accuracy-rejection curves for all the uncertainty measures summarized in Section 3, separated into epistemic uncertainty on the left, aleatoric uncertainty in the middle, and total uncertainty on the right.