

# Ensemble-based Uncertainty Quantification: Bayesian versus Credal Inference

Mohammad Hossein Shaker<sup>1</sup>, Eyke Hüllermeier<sup>2</sup>

<sup>1</sup>Department of Computer Science  
Paderborn University  
E-Mail: mhshaker@mail.upb.de

<sup>2</sup>Institute of Informatics  
LMU Munich  
E-Mail: eyke@lmu.de

## 1 Introduction

A distinction between two different types of uncertainty, *aleatoric* and *epistemic* [6], has received increasing attention in the recent machine learning literature [8, 14]. While the former refers to statistical uncertainty in the sense of inherent randomness, the latter captures systematic uncertainty caused by a lack of knowledge.

In this paper, we consider ensemble-based approaches to uncertainty quantification, i.e., to derive meaningful measures of aleatoric and epistemic uncertainty in a prediction. In this regard, we propose a distinction between three types of uncertainty-aware learning algorithms: probabilistic agents, Bayesian agents, and Levi agents (Section 2). We address the question of how to quantify aleatoric and epistemic uncertainty in a formal way (Section 3), both for Bayesian and Levi agents, and how to approximate such quantities empirically using ensemble techniques (Section 4). Moreover, we analyze the effectiveness of corresponding measures in an empirical study on classification with a reject option (Section 5).

## 2 Representing Predictive Uncertainty

We consider a standard setting of supervised learning, in which a learner is given access to a set of (i.i.d.) training data  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$ , where  $\mathcal{X}$  is an instance space and  $\mathcal{Y}$  the set of outcomes that can be associated with an instance. In particular, we focus on the classification scenario, where  $\mathcal{Y} = \{y_1, \dots, y_K\}$  consists of a finite set of class labels, with binary classification ( $\mathcal{Y} = \{0, 1\}$ ) as an important special case.

Suppose a *hypothesis space*  $\mathcal{H}$  to be given, where a hypothesis  $h \in \mathcal{H}$  is a mapping  $\mathcal{X} \rightarrow \mathbb{P}(\mathcal{Y})$ , with  $\mathbb{P}(\mathcal{Y})$  the class of probability measures on  $\mathcal{Y}$ . Thus, a hypothesis maps instances  $\mathbf{x} \in \mathcal{X}$  to probability distributions on outcomes. The goal of the learner is to induce a hypothesis  $h^* \in \mathcal{H}$  with low risk (expected loss)

$$R(h) = \int_{\mathcal{X} \times \mathcal{Y}} \ell(h(\mathbf{x}), y) dP(\mathbf{x}, y) , \quad (1)$$

where  $P$  is the (unknown) data-generating process (a probability measure on  $\mathcal{X} \times \mathcal{Y}$ ), and  $\ell : \mathbb{P}(\mathcal{Y}) \times \mathcal{Y} \rightarrow \mathbb{R}$  a loss function.

Eventually, one is often interested in the *predictive uncertainty*, i.e., the uncertainty related to the prediction  $\hat{y}_q$  for a concrete query instance  $\mathbf{x}_q \in \mathcal{X}$ . Given such a query, different learning methods proceed on the basis of different types of information. Depending on how the uncertainty is represented as a basis for prediction and decision making, we propose to distinguish three types of learning methods, which we call, respectively, probabilistic, Bayesian, and Levi agents.

### 2.1 Probabilistic Agents

A common practice in machine learning is to consider learners that fully commit to a single hypothesis  $\hat{h} \in \mathcal{H}$  and use this hypothesis to make predictions. Such a learner will predict a single probability distribution

$$\mathbf{q} = \hat{h}(\mathbf{x}_q) = (q_1, \dots, q_K) \in \mathbb{P}(\mathcal{Y}) , \quad (2)$$

where  $q_k$  is the probability of the  $k^{th}$  class  $y_k$ . This prediction is considered as an estimation of the (true) conditional probability  $p(y|\mathbf{x}_q)$ . We call a learner of that kind a *probabilistic agent*. Such an agent’s uncertainty about the outcome  $y$  is purely aleatoric. At the level of the hypothesis space, the agent pretends full certainty, and hence the absence of any epistemic uncertainty about the best hypothesis  $h$ .

## 2.2 Bayesian Agents

Adhering to the principle of (strict) Bayesianism as advocated by statisticians such as De Finetti [4], a *Bayesian agent* will represent its belief about the best hypothesis in terms of a probability distribution on  $\mathcal{H}$ . Thus, instead of committing to a single hypothesis  $\hat{h}$ , the agent will assign a probability (density)  $p(h)$  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  $p(h|\mathcal{D})$ .

Since every  $h \in \mathcal{H}$  gives rise to a probabilistic prediction (2), a Bayesian agent’s belief about the outcome  $y_q$  is represented by a second-order probability: a probability distribution of probability distributions. If needed,  $p$  can be “collapsed” into a single distribution  $q$  on  $\mathcal{Y}$ . This is typically accomplished by inducing  $q$  from  $p$  (or, more generally, a corresponding measure  $P$ ) via Bayesian model averaging (BMA):

$$q = \text{bma}(p) = \int_{\mathcal{H}} h(\mathbf{x}_q) dP(h) \quad (3)$$

## 2.3 Levi Agents

As a further generalization, instead of committing to a single probability distribution  $p \in \mathbb{P}(\mathcal{H})$  on the hypothesis space, the learner may work with a *set*  $\mathcal{Q}' \subseteq \mathbb{P}(\mathcal{H})$  of such distributions, all of which are deemed plausible candidates. Each distribution  $p \in \mathcal{Q}'$  again gives rise to a probability distribution

according to (3). Eventually, the relevant representation of the learner is a set of probability distributions

$$\mathcal{Q} = \{ \text{bma}(p) \mid p \in \mathcal{Q}' \} \subseteq \mathbb{P}(\mathcal{Y}). \quad (4)$$

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 Levi [11, 12]. Correspondingly, we call a learner of this kind a *Levi agent*. The set  $\mathcal{Q}'$  (and thereby the set  $\mathcal{Q}$ ) can be produced in different ways, for example as a *credal set* in the context of imprecise probability theory [16].

### 3 Uncertainty Quantification

According to our discussion so far, different types of learners represent their information or “belief” about the outcome  $y_q$  for a query instance  $\mathbf{x}_q$  in different ways. What we are mainly interested in is a quantification of these learner’s epistemic and aleatoric uncertainty, i.e., we are seeking a measure of epistemic uncertainty, EU, and a measure of aleatoric uncertainty, AU.

For ease of notation, we subsequently omit the conditioning on the query instance  $\mathbf{x}_q$ , i.e., all probabilities of outcomes should be understood as conditional probabilities given  $\mathbf{x}_q$  (for example, we write  $p(y)$  instead of  $p(y \mid \mathbf{x}_q)$  and  $p(y \mid h)$  instead of  $p(y \mid h, \mathbf{x}_q)$ ). We denote the set of all probability distributions (probability vectors)  $q = (q_1, \dots, q_K) \in [0, 1]^K$  by  $\Delta_K$ .

#### 3.1 Probabilistic Agents: Entropy

The most well-known measure of uncertainty of a single probability distribution is the (Shannon) entropy, which, in the case of discrete  $\mathcal{Y}$ , is given as

$$S(q) = - \sum_{y \in \mathcal{Y}} q(y) \log_2 q(y), \quad (5)$$

where  $0 \log 0 = 0$  by definition. This measure can be justified axiomatically, and different axiomatic systems have been proposed in the literature [3]. It is the most obvious candidate to quantify the aleatoric uncertainty of a probabilistic agent, i.e.,  $\text{AU}(q) = S(q)$ . As such an agent pretends to have precise knowledge of the predictive distribution, the epistemic uncertainty is 0.

### 3.2 Bayesian Agents: Entropy and Mutual Information

A principled approach to measuring and separating aleatoric and epistemic uncertainty on the basis of classical information-theoretic measures of entropy is proposed by [5]. This approach is developed in the context of neural networks for regression, but the idea as such is more general and can also be applied to other settings. A similar approach was recently adopted by [13].

More specifically, the idea is to exploit the following information-theoretic separation of the total uncertainty in a prediction, measured in terms of the (Shannon) entropy of the predictive posterior distribution (in the case of discrete  $\mathcal{Y}$  given by (5)): Considering the outcome as a random variable  $Y$  and the hypothesis as a random variable  $H$ , we have

$$S(Y) = I(Y, H) + S(Y | H),$$

where  $I(Y, H)$  is the mutual information between hypotheses and outcomes (i.e., the Kullback-Leibler divergence between the joint distribution of outcomes and hypotheses and the product of their marginals):

$$I(Y, H) = \mathbf{E}_{p(y, h)} \left\{ \log_2 \left( \frac{p(y, h)}{p(y)p(h)} \right) \right\}. \quad (6)$$

This term qualifies as a measure of epistemic uncertainty, as it captures the dependency between the probability distribution on  $\mathcal{Y}$  and the (uncertain) hypothesis  $h$ .

Finally, the conditional entropy is given by

$$\begin{aligned} S(Y|H) &= \mathbf{E}_{p(h|\mathcal{D})} \{S(p(y|h))\} = \\ &= - \int_{\mathcal{H}} p(h|\mathcal{D}) \left( \sum_{y \in \mathcal{Y}} p(y|h) \log_2 p(y|h) \right) dh \end{aligned} \quad (7)$$

This measure qualifies as a measure of aleatoric uncertainty: By fixing a hypothesis  $h \in \mathcal{H}$ , the epistemic uncertainty is essentially removed. Thus, the entropy  $S(p(y|h))$ , i.e., the entropy of the conditional distribution on  $\mathcal{Y}$  predicted by  $h$  (for the query  $\mathbf{x}_q$ ) is a natural measure of the aleatoric uncertainty. However, since  $h$  is not precisely known, aleatoric uncertainty is measured in terms of the expectation of this entropy with regard to the posterior probability  $p(h|\mathcal{D})$ .

### 3.3 Levi Agents: Uncertainty Measures for Credal Sets

In the case of a Levi agent, uncertainty degrees ought to be specified for a set of probability distributions  $Q \subseteq \Delta_K$ . In the literature, such sets are also referred to as *credal sets* [16]. There is quite some work on defining uncertainty measures for credal sets and related representation, such as Dempster-Shafer evidence theory [15], asking for a generalized representation

$$U(Q) = AU(Q) + EU(Q), \quad (8)$$

where  $U$  is a measure of total (aggregate) uncertainty,  $AU$  a measure of aleatoric uncertainty (a generalization of the Shannon entropy), and  $EU$  a measure of epistemic uncertainty.

As for the latter, the following generalization of the Hartley measure, a well-established measure of uncertainty for sets, has been proposed by various authors [2]:

$$GH(Q) = \sum_{A \subseteq \mathcal{Y}} m_Q(A) \log(|A|), \quad (9)$$

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

$$\nu_Q(A) = \inf_{q \in Q} q(A) \quad (10)$$

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

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

This measure is “well-justified” in the sense of possessing a sound axiomatic basis and obeying a number of desirable properties [9].

Regarding  $AU(Q)$ , an extension of Shannon entropy, “well-justified” in the same sense, has not been found so far. As a possible way out, it was suggested to define a meaningful measure of total or aggregate uncertainty  $U(Q)$ , and to *derive* a generalized measure of aleatoric uncertainty via *disaggregation*, i.e., in terms of the difference between this measure and the measure of epistemic uncertainty (Hartley), or vice versa, to derive a measure of epistemic uncertainty as the difference between total uncertainty and a meaningful measure of aleatoric uncertainty.

The upper and lower Shannon entropy play an important role in this regard:

$$S^*(Q) = \max_{q \in Q} S(q), \quad S_*(Q) = \min_{q \in Q} S(q) \quad (11)$$

Based on these measures, the following disaggregations of total uncertainty (8) have been proposed [1]:

$$S^*(Q) = (S^*(Q) - GH(Q)) + GH(Q) \quad (12)$$

$$S^*(Q) = S_*(Q) + (S^*(Q) - S_*(Q)) \quad (13)$$

In both cases, upper entropy serves as a measure of total uncertainty  $U(Q)$ , 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, lower entropy is used as a (well-justified) measure

of aleatoric uncertainty, and epistemic uncertainty is derived in terms of the difference between upper and lower entropy.

## 4 Ensemble-Based Uncertainty Quantification

Ensemble-based approaches to uncertainty quantification have recently been advocated by several authors [10]. Adopting a Bayesian perspective, the variance of the predictions produced by an ensemble is inversely related to the “peakedness” of a posterior distribution  $p(h|\mathcal{D})$ . Thus, an ensemble can be considered as an approximate representation of a second-order distribution  $p(h|\mathcal{D})$  in a Bayesian setting.

Given this motivation, we address the question of how the measures of uncertainty introduced above can be realized by means of ensemble techniques, i.e., how they can be computed (approximately) on the basis of a finite ensemble of hypotheses  $H = \{h_1, \dots, h_M\}$ , which can be thought of as a sample from the posterior distribution  $p(h|\mathcal{D})$ . More specifically, we consider this question for the case of a Bayesian and a Levi agent. The following notation will be used:

- $p_{k,m} = p(y_k | h_m, \mathbf{x}_q)$  is the probability predicted for class  $y_k$  by hypothesis  $h_m$  for query  $\mathbf{x}_q$ , i.e.,  $(p_{1,m}, \dots, p_{K,m}) = p(\cdot | h_m, \mathbf{x}_q)$ ;
- $l_m = p(\mathcal{D} | h_m)$  denotes the likelihood of  $h_m$ ;
- $q_k = \sum_{m=1}^M p(h_m | \mathcal{D}) p_{k,m}$  is the posterior probability estimate for class  $y_k$  produced by the ensemble through weighted averaging.

### 4.1 Bayesian Agents

Recalling the approach presented in Section 3.2, it is obvious that (6) and (7) cannot be computed efficiently, because they involve an integration over the



hypothesis space  $\mathcal{H}$ . Based on an ensemble  $H = \{h_1, \dots, h_M\}$ , an approximation of (7) can be obtained by

$$\text{AU}(\mathbf{x}_q) = - \sum_{m=1}^M p(h_m | \mathcal{D}) \sum_{k=1}^K p_{k,m} \log_2 p_{k,m}, \quad (14)$$

an approximation of total uncertainty, i.e., Shannon entropy (5), by

$$\text{U}(\mathbf{x}_q) = - \sum_{k=1}^K q_k \log_2 q_k, \quad (15)$$

and finally an approximation of (6) by  $\text{EU}(\mathbf{x}_q) = \text{U}(\mathbf{x}_q) - \text{AU}(\mathbf{x}_q)$ . Assuming a uniform prior, which is quite natural in the case of ensembles, the posterior probability of hypotheses can be obtained from  $p(h_m | \mathcal{D}) \propto l_m$ .

## 4.2 Levi Agents

How could the idea of a Levi agent be implemented on the basis of an ensemble approach? As explained above, credal inference yields a set of probability estimates, each of which is obtained by Bayesian model averaging according to a different prior. Thus, instead of assuming a uniform prior  $p(h_m) \equiv 1/M$ , we should now proceed from a set of priors. A simple example is the family

$$S_\delta = \left\{ \mathbf{s} = (s_1, \dots, s_M) \mid \frac{1}{\delta M} \leq s_m \leq \frac{\delta}{M}, \sum_{m=1}^M s_m = 1 \right\} \quad (16)$$

of distributions  $\delta$ -close to uniform, where  $\delta \geq 1$  is a (hyper-)parameter. Thus, compared to the uniform prior, the probability of a single hypothesis can now be decreased or increased by a factor of at most  $\delta$ . The set of posterior probabilities is then given by

$$\left\{ p(h_m | \mathcal{D}) = \frac{s_m l_m}{\sum_{i=1}^M s_i l_i} \mid \mathbf{s} \in S_\delta \right\},$$

and hence the credal set on  $\mathcal{Y}$  by

$$Q = \left\{ q = \sum_{m=1}^M s_m l_m h_m(\mathbf{x}_q) / \sum_{m=1}^M s_m l_m \mid \mathbf{s} \in S_\delta \right\}$$

To compute the decompositions (12) and (13) for  $Q$ , we need to compute the measures  $S^*$ ,  $S_*$ , GH. According to (9), the computation of the measure GH requires the capacity (10), i.e., the lower probability  $v_Q(A)$  of each subset of classes  $A \subseteq \mathcal{Y}$ . For  $A = \{y_j\}_{j \in J}$  identified by an index set  $J \subseteq [K]$ , the latter is given by

$$v_Q(A) = \min_{q \in Q} q(A) = \min_{\mathbf{s} \in S_\delta} \frac{\sum_{j \in J} \sum_{m=1}^M s_m l_m p_{j,m}}{\sum_{m=1}^M s_m l_m}.$$

Thus, finding  $v_Q(A)$  comes down to solving a linear-fractional programming problem (for which standard solvers can be used). Moreover, finding  $S^*$  comes down to solving

$$\max_{\mathbf{s} \in S_\delta} \sum_{k=1}^K \frac{\sum_{m=1}^M s_m l_m p_{k,m}}{\sum_{m=1}^M s_m l_m} \log \frac{\sum_{m=1}^M s_m l_m p_{k,m}}{\sum_{m=1}^M s_m l_m},$$

and similarly for  $S_*$  (with max replaced by min).

## 5 Experiments

Predicted uncertainties are often evaluated indirectly, for example by assessing their usefulness for improved prediction and decision making, because the data does normally not contain information about any sort of “ground truth” uncertainties. Here, we conducted such an evaluation by producing *accuracy-rejection curves*, which depict the accuracy of a predictor as a function of the percentage of rejections [7]: a learner, which is allowed to abstain on a certain percentage  $p$  of predictions, will predict on those  $(1 - p)\%$  on which it feels most certain. Being able to quantify its own uncertainty well, it should improve its accuracy with increasing  $p$ , hence the accuracy-rejection curve should be monotone increasing (unlike a flat curve obtained for random abstention).

## 5.1 Data Sets and Experimental Setting

We compare the Bayesian agent with different variants of the Levi agent in terms of their ability to quantify aleatoric and epistemic uncertainty. The Bayesian agent quantifies these uncertainties according to (14) and (15). The Levi agent is implemented as described in Section 4.2. Uncertainty is quantified based on the generalized Hartley measure (Levi-GH) according to (12), or based on upper and lower entropy (Levi-Ent) according to (13). In this experiment, we set the hyper-parameter  $\delta = 2$ .

We performed experiments on various well-known data sets from the UCI repository<sup>1</sup>. The data sets are randomly split into 70% for training and 30% for testing, and accuracy-rejection curves are produced on the latter. Each experiment is repeated and averaged over 100 runs. We create ensembles using the Random Forest Classifier from SKlearn. The number of trees within the ensemble is set to 10. Each tree can grow to a maximum of 10 splits. Probabilities are estimated by (Laplace-corrected) relative frequencies in the leaf nodes of a tree.

## 5.2 Results

Fig. 1 shows the accuracy-rejection curves for the different learners, separated into epistemic uncertainty (EU) in the left, aleatoric uncertainty (AU) in the middle, and total uncertainty (TU) on the right column. Due to space restrictions, we only show the results for five data sets, noting that the results for other data sets are very similar. The following observations can be made.

- As suggested by the shape of the accuracy-rejection curves, both the Bayesian and the Levi agent perform quite well in general. On total uncertainty, they are basically indistinguishable, which is almost a bit surprising, given that these uncertainties are quantified on the basis of different principles.

---

<sup>1</sup> <http://archive.ics.uci.edu/ml/index.php>

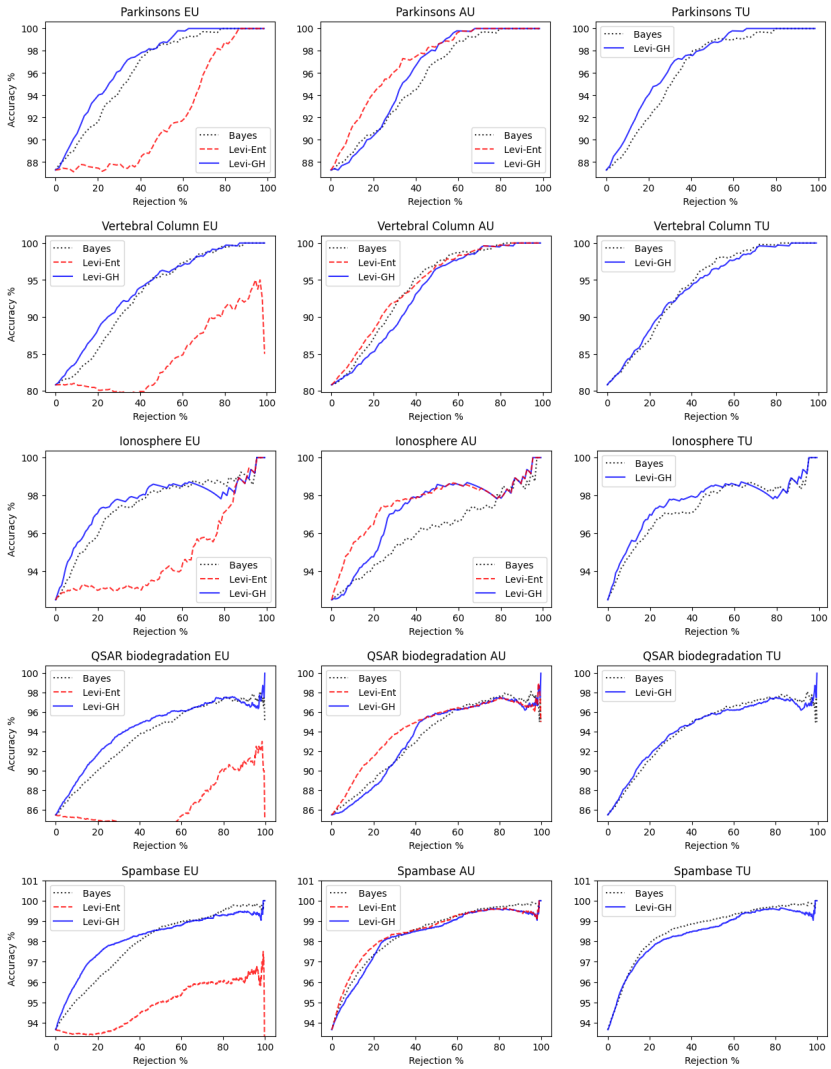


Figure 1: Accuracy-rejection curves for the Bayesian and the Levi agent.

- Levi-GH seems to have an advantage over the Bayesian agent on epistemic uncertainty, providing evidence for the generalized Hartley measure as a reasonable measure of epistemic uncertainty.
- Levi-Ent seems to have an advantage over the Bayesian agent on aleatoric uncertainty, providing evidence for the lower entropy as a reasonable measure of aleatoric uncertainty.
- The “derived” measures,  $S^*(Q) - \text{GH}(Q)$  for aleatoric and  $S^*(Q) - S_*(Q)$  for epistemic uncertainty, both perform quite poorly.

## 6 Conclusion

We proposed a distinction between different types of uncertainty-aware learning algorithms, discussed measures of total, aleatoric and epistemic uncertainty of such learners, and developed ensemble-methods for approximating these measures. In particular, we compared the classical Bayesian approach with what we call a Levi agent, which makes predictions in terms of credal sets.

In an experimental study on uncertainty-based abstention, both methods show strong performance. While the Bayesian and the Levi agent are on a par for total uncertainty, improvements of the Bayesian approach can be achieved for the two types of uncertainty separately: The generalized Hartley measure appears to be superior for epistemic and the lower entropy for aleatoric uncertainty quantification. On the other side, the alternative measures of aleatoric and epistemic uncertainty obtained through disaggregation perform quite poorly. These results can be seen as an interesting empirical complement to the theoretical (axiomatic) research on uncertainty measures for credal sets.

In future work, we seek to further deepen our understanding of ensemble-based uncertainty quantification and elaborate on the approach presented in this paper. An interesting problem, for example, is the tuning of the (hyper-)parameter  $\delta$  in (16), for which we simply took a default value in the experiments. Obviously, this parameter has an important influence on the uncertainty of the Levi agent. Besides, we also plan to develop alternative approaches for constructing

ensembles. Last but not least, going beyond abstention and accuracy-rejection curves, we plan to apply and analyze corresponding methods in the context of other types of uncertainty-aware decision problems.

## References

- [1] J. Abellan, J. Klir, and S. Moral. Disaggregated total uncertainty measure for credal sets. *International Journal of General Systems*, 35(1), 2006.
- [2] J. Abellan and S. Moral. A non-specificity measure for convex sets of probability distributions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8:357–367, 2000.
- [3] I. Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10:261–273, 2008.
- [4] B. De Finetti. Foresight: It’s logical laws, it’s subjective sources. In H.E. Kyburg and H.E. Smokler, editors, *Studies in Subjective Probability*. R.E. Krieger, New York, 1980.
- [5] S. Depeweg, J.M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proc. ICML*, Stockholm, Sweden, 2018.
- [6] S.C. Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering and System Safety*, 54(2–3):217–223, 1996.
- [7] J. Hühn and E. Hüllermeier. FR3: A fuzzy rule learner for inducing reliable classifiers. *IEEE Transactions on Fuzzy Systems*, 17(1):138–149, 2009.
- [8] A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proc. NIPS*, pages 5574–5584, 2017.
- [9] G.J. Klir and M. Mariano. On the uniqueness of possibilistic measure of uncertainty and information. *Fuzzy Sets and Systems*, 24(2):197–219, 1987.

- [10] B. Lakshminarayanan, A. Pritzel, C. and Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. NeurIPS*, 2017.
- [11] I. Levi. On indeterminate probabilities. *Journal of Philosophy*, 71:391–418, 1974.
- [12] I. Levi. *The Enterprise of Knowledge*. MIT Press, Cambridge, 1980.
- [13] A. Mobiny, H.V. Nguyen, S. Moulik, N. Garg, and C.C. Wu. DropConnect is effective in modeling uncertainty of Bayesian networks. *CoRR*, abs/1906.04569, 2017.
- [14] R. Senge, S. Bösner, K. Dembczynski, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, and E. Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Inf. Sciences*, 255:16–29, 2014.
- [15] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.
- [16] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.