Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning: Are Conditional Entropy and Mutual Information Appropriate Measures?

Lisa Wimmer[1,3] **Yusuf Sale**[2,3] **Paul Hofman**[2,3] **Bernd Bischl**[1,3] **Eyke Hüllermeier**[2,3]

[1]Department of Statistics, LMU Munich, Germany
[2]Institute of Informatics, LMU Munich, Germany
[3]Munich Center for Machine Learning (MCML), Germany

Abstract

The quantification of aleatoric and epistemic uncertainty in terms of conditional entropy and mutual information, respectively, has recently become quite common in machine learning. While the properties of these measures, which are rooted in information theory, seem appealing at first glance, we identify various incoherencies that call their appropriateness into question. In addition to the measures themselves, we critically discuss the idea of an additive decomposition of total uncertainty into its aleatoric and epistemic constituents. Experiments across different computer vision tasks support our theoretical findings and raise concerns about current practice in uncertainty quantification.

1 INTRODUCTION

Estimating, analyzing, and handling uncertainty has always been an integral part of classical statistics. More recently, the broader machine learning community has come to acknowledge its vital role in producing reliable estimates, which is, e.g., considered paramount in safety-critical applications [Senge et al., 2014, Kendall and Gal, 2017, Psaros et al., 2022]. In the context of supervised learning, the focus is typically on *predictive uncertainty*, i.e., the learner's uncertainty about the outcome $y \in \mathcal{Y}$ given a query instance $\boldsymbol{x}$ from an input space $\mathcal{X}$. It is often expedient to decompose the *total* amount of uncertainty (TU) into its *aleatoric* (AU) and *epistemic* (EU) parts [Hüllermeier and Waegeman, 2021]. The aleatoric component of TU arises from the stochastic nature of the mapping from $\mathcal{X}$ to $\mathcal{Y}$, which implies that the "ground truth" is a (generally non-Dirac) conditional probability distribution $p(\cdot \mid \boldsymbol{x})$ over $\mathcal{Y}$. Potential sources of AU include measurement errors, randomness in physical quantities, and the simple fact that $\boldsymbol{x}$ may not suffice to explain y [Malinin, 2019]. Thus, it is not possible to predict the outcome with certainty even assuming perfect knowledge of the data-generating process. In practice, the learner only finds an empirical estimate $\hat{p}(\cdot \mid \boldsymbol{x})$ from a pre-defined hypothesis space $\mathcal{H}$. Roughly speaking, EU then refers to a lack of knowledge about the discrepancy between p and $\hat{p}$ that can – in contrast to AU – be alleviated by gathering more training samples[1] [Senge et al., 2014]. Not least due to this potential reduction, the quantification of EU is helpful in many applications of AI, e.g., in order to invoke human intervention in multi-stage decision processes [Gal et al., 2017, Budd et al., 2021, Kirsch et al., 2022].

A specific set of measures for TU, AU, and EU has found broad consensus in the uncertainty quantification literature. Promoted by Houlsby et al. [2011] and Depeweg [2019], among others, the core idea revolves around *Shannon entropy* [Shannon, 1948] as a measure for TU, which can be shown to (additively) decompose into *conditional entropy*, interpreted as a measure of AU, and *mutual information*, resulting as a measure of EU [Kendall and Gal, 2017, Smith and Gal, 2018, Charpentier et al., 2022]. Various applications have recently used these measures (e.g., Michelmore et al. [2018], Shelmanov et al. [2021], Winkler et al. [2022]).

While the underlying ideas are well-established and mathematically well-founded, they originate from information theory and are not necessarily suited to uncertainty quantification in statistical learning. In fact, we find that the proposed entropy-related measures exhibit questionable behavior in a number of settings. We thus argue for a critical re-assessment of the established practice. Our contributions are as follows: We (i) point out numerous incoherencies of uncertainty measures based on decomposing Shannon entropy, (ii) shed light on why an additive disaggregation of TU might not be meaningful, and (iii) provide evidence for the relevance of these observations in practical applications.

[1]An additional source of EU is model misspecification, meaning that $p \notin \mathcal{H}$ (e.g., Wilson [2020]); although important, it is difficult to tackle formally and often omitted from the analysis.

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

2 BACKGROUND

2.1 UNCERTAINTY REPRESENTATIONS

In the following, we assume categorical target variables from a finite label space $\mathcal{Y} = \{y_1, \ldots, y_K\}, K \in \mathbb{N}$. The set $\mathbb{P}(\mathcal{Y})$ of (first-order) probability distributions over $\mathcal{Y}$ can be identified with the (K-1)-simplex $\Delta_K := \{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in [0,1]^K \mid \|\boldsymbol{\theta}\|_1 = 1\}$. For each $\boldsymbol{\theta} \in \Delta_K$, we can compute an associated degree of AU [Hüllermeier and Waegeman, 2021]. To capture EU, the learner must be able to express uncertainty about $\boldsymbol{\theta}$, which can be accomplished by a (second-order) probability distribution over (first-order) distributions $\boldsymbol{\theta}$.

The canonical approach to such a bi-level representation is Bayesian inference, where uncertainty about the (Bayes) predictor is translated into uncertainty about a prediction expressed in terms of the posterior predictive distribution [Gelman et al., 2021]. Another idea is to estimate EU directly along with the prediction of the outcome. For instance, it is possible to learn the parameters of a Dirichlet distribution with a concentration parameter expressing EU [Malinin and Gales, 2018, Charpentier et al., 2020, Huseljic et al., 2020, Kopetzki et al., 2021]. Either way, we arrive at a second-order or second-order predictor $F : \mathcal{X} \longrightarrow \Delta_K^{(2)}$, where $\Delta_K^{(2)} = \mathbb{P}(\mathbb{P}(\mathcal{Y}))$ denotes the set of second-order distributions over $\mathcal{Y}$. For the sake of simplicity, we subsequently omit the conditioning on the query instance $\boldsymbol{x}$.

Each probability vector $\boldsymbol{\theta} \in \Delta_K$ implies a categorical distribution on $\mathcal{Y}$, such that $\theta_k = p(y_k \mid \boldsymbol{\theta})$ is the probability of outcome y_k, and one of these distributions, $\boldsymbol{\theta}^*$, corresponds to the (unknown) ground truth. If $F(\boldsymbol{x}) = Q \in \Delta_K^{(2)}$ is the prediction for a query $\boldsymbol{x}$, then Q assigns a probability $Q(\boldsymbol{\theta})$ to each distribution $\boldsymbol{\theta} \in \Delta_K$. Q represents the epistemic state of the learner about $\boldsymbol{x}$, and a stronger concentration of Q indicates more certainty about the true distribution $\boldsymbol{\theta}^*$. Note that we immediately regard the output distribution and omit from our notation any relation to objects that play a role in estimating $\boldsymbol{\theta}$ (e.g., the Bayes posterior is often expressed as a distribution over model parameters given data).

As an example, consider the case of binary classification ($K = 2$), where $\mathcal{Y} = \{y_0, y_1\}$ consists of only two classes, negative and positive (illustrated in Fig. 1). The conditional distribution $p(\cdot \mid \boldsymbol{x})$ is now determined by the probability vector $\boldsymbol{\theta} = (\theta_0, \theta_1) = (p(y_0 \mid \boldsymbol{x}),\ p(y_1 \mid \boldsymbol{x}))$. As the individual class probabilities sum up to one, prediction in $\boldsymbol{x}$ effectively comes down to inference about the ground-truth probability $\theta_1^* = p(y_1 \mid \boldsymbol{x})$ of the positive class, and hence about the success parameter θ of a Bernoulli distribution $\mathcal{B}(\theta)$. The corresponding second-order distribution Q could be, for example, a Beta distribution (or, more generally, a Dirichlet distribution for $K \geq 2$).

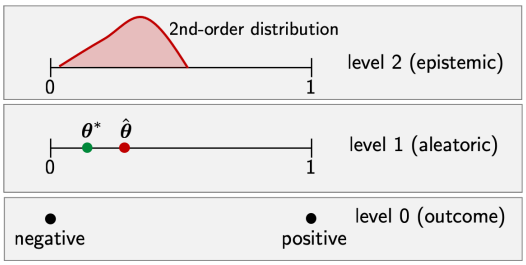


Figure 1: Bi-level uncertainty representation in binary classification. We pose a second-order distribution Q over Bernoulli distributions corresponding to the probability θ of the positive class. In Bayesian inference, Q is given by the posterior distribution (epistemic level), and a point prediction $\hat{\theta}$ (aleatoric level) is obtained by model averaging. On level zero, only binary class labels are observed.

2.2 ENTROPY-BASED INFORMATION MEASURES

Given a representation of an uncertain prediction in terms of a second-order distribution Q, one is often interested in quantifying the learner's uncertainty in terms of a single number, and decomposing the TU into its aleatoric and epistemic contributions. One approach that is now commonly adopted [Houlsby et al., 2011, Gal, 2016, Depeweg et al., 2018, Smith and Gal, 2018, Mobiny et al., 2021] is based on information-theoretic quantities derived from Shannon entropy. Let Y signify the random outcome variable with marginal probability

$$p(y) = \int_{\Delta_K} p(y \mid \boldsymbol{\theta}) \, \mathrm{d}\, Q(\boldsymbol{\theta}). \tag{1}$$

The (discrete) Shannon entropy of Y can be defined as

$$H(Y) = -\sum_{y \in \mathcal{Y}} p(y) \cdot \log p(y)\,, \tag{2}$$

where the logarithm is typically set to base 2. For continuous label spaces, we obtain the *differential* entropy analogously by replacing the sum in Eq. (2) with an integral over $\mathcal{Y}$ [Cover and Thomas, 2006].

Shannon entropy can be interpreted as the degree of uniformity in the distribution of a random variable, or the amount of information to be gained from observing its realization (further analogies exist in physics and source coding). It enjoys a number of desirable properties: it is non-negative, maximal for the uniform distribution, and invariant to permutations of classes in $\mathcal{Y}$ [Cover and Thomas, 2006].

Exploiting the fact that Y is the expectation of the conditional outcome (given $\boldsymbol{\theta}$) with respect to the second-order distribution Q, the Shannon entropy, as a measure of TU, can be computed as

$$H(Y) = H\big(\mathbb{E}_Q\left[Y \mid \boldsymbol{\theta}\right]\big). \tag{3}$$

A trivial result from information theory states that Shannon entropy additively decomposes into *conditional entropy* and *mutual information* (e.g., Ash [1965]):

$$H(Y) = H(Y \mid \Theta) + I(Y, \Theta). \tag{4}$$

Here, Θ denotes the random variable of first-order distributions $\boldsymbol{\theta} \sim Q$. Conditional entropy $H(Y \mid \Theta)$ expresses the uncertainty about Y that would remain if the realization of Θ were known, and is therefore taken to quantify AU. The motivation for expressing AU this way is as follows: By fixing a first-order distribution $\boldsymbol{\theta} \in \Delta_K$, all EU is essentially removed and only AU remains, such that $H(Y \mid \boldsymbol{\theta}) = -\sum_{k=1}^{K} \theta_k \log \theta_k$ is indeed a natural measure of AU. However, as $\boldsymbol{\theta}$ is not precisely known, we take the expectation with respect to the second-order distribution:

$$\begin{aligned} H(Y \mid \Theta) &= \mathbb{E}_Q \left[H\left(Y \mid \boldsymbol{\theta}\right) \right] \\ &= \int_{\Delta_K} \underbrace{\textstyle\sum_{y \in \mathcal{Y}} -p(y \mid \boldsymbol{\theta}) \log p(y \mid \boldsymbol{\theta})}_{H(Y \mid \boldsymbol{\theta})} \, \mathrm{d}\, Q(\boldsymbol{\theta}). \end{aligned} \tag{5}$$

Mutual information $I(Y, \Theta)$, proposed to embody EU, is a symmetric measure for the expected information gained about one variable through observing the other and vanishes if both variables are independent. Intuitively, it quantifies the potential reduction in uncertainty about Y through observing Θ[2] [Ash, 1965]:

$$\begin{aligned} I(Y, \Theta) &= H(Y) - H(Y \mid \Theta) \\ &= \mathbb{E}_Q \left[D_{\mathrm{KL}} \left(p(Y \mid \Theta) \parallel p(Y) \right) \right]. \end{aligned} \tag{6}$$

2.3 FINITE-ENSEMBLE APPROXIMATION

In practice, it will hardly be possible to compute exact expectations with respect to Q due to the intractability of the corresponding integrals, so we must permit some kind of approximation. The standard approach to this problem is Monte Carlo integration with a finite number $M \in \mathbb{N}$ of samples $\boldsymbol{\theta} \sim Q$, whose induced mean, by the law of large numbers, almost surely converges to the true integral [Andrieu et al., 2003]. In machine learning, this way of combining the predictions of multiple models is common practice and known as *ensemble learning* (e.g., Breiman [2001], Lakshminarayanan et al. [2017]). For TU, the entropy over the approximate first-order distribution becomes

$$H(Y) = H\left(\tfrac{1}{M} \textstyle\sum_{m=1}^{M} p\left(y \mid \boldsymbol{\theta}^{(m)} \right) \right), \tag{7}$$

where $\boldsymbol{\theta}^{(m)}$, $m \in \{1, \ldots, M\}$, is the prediction of the m-th ensemble member. Analogously, we obtain the ensemble conditional entropy as a measure of AU:

$$H(Y \mid \Theta) = \tfrac{1}{M} \textstyle\sum_{m=1}^{M} H\left(p\left(y \mid \boldsymbol{\theta}^{(m)} \right) \right). \tag{8}$$

The discrete EU is the Jensen-Shannon divergence

$$\tfrac{1}{M} \textstyle\sum_{m=1}^{M} D_{\mathrm{KL}} \left(p\left(y \mid \boldsymbol{\theta}^{(m)} \right) \,\middle\|\, \tfrac{1}{M} \textstyle\sum_{j=1}^{M} p\left(y \mid \boldsymbol{\theta}^{(j)} \right) \right). \tag{9}$$

Finite-ensemble EU still results as the difference between Eq. (7) and Eq. (8) under the conditions of Section 2.2. Obviously, ensemble learning provides a rather coarse approximation to the expectation $\mathbb{E}_Q(\cdot)$, especially if the size M of the ensemble is small. Consider, for example, deep ensembles [Lakshminarayanan et al., 2017], which have become a *de facto* gold standard in probabilistic machine learning [Ovadia et al., 2019, Ashukha et al., 2020, Psaros et al., 2022]. The cost of training each of their neural network members is typically so high that M remains in the realm of five to ten (e.g., Abe et al. [2022], Turkoglu et al. [2022]). We must then expect certain (low-density) regions in the space of first-order distributions to be systematically undersampled, affecting the estimate of EU in particular.

3 RELATED WORK

The decomposition in Eq. (4) is widely applied and often used in conjunction with Bayesian learning, e.g., to obtain more robust predictions by explicitly modeling both components of TU [Kendall and Gal, 2017], to design better optimization procedures by using the decomposition in stochastic differential equations [Winkler et al., 2022], or to improve the general understanding of neural networks [Woo, 2022]. In natural language processing, Wu et al. [2020] filter unreliable predictions in spoken language assessment caused by high EU, while Shelmanov et al. [2021] perform active learning with EU-based acquisition for label-efficient sequence tagging. Further fields of application include autonomous driving, where EU is found to be predictive of forthcoming accidents [Michelmore et al., 2018], and medical diagnostics, with a focus on identifying inherently ambiguous cases [Mobiny et al., 2021].

We find, however, that the entropy decomposition may lead to unintended results discussed in the subsequent section. While we do not dispute the mathematical correctness of Eq. (4)-(6), we argue that the individual quantities and their additive aggregation are not suitable for evaluating predictive uncertainty. Although previous work uncovered similar shortcomings of measures in related uncertainty frameworks (e.g., Pal et al. [1992] on evidential reasoning), the entropy-based measures – to the best of our knowledge – have drawn hardly any criticism in the machine learning community. In the following, we point out numerous inconsistencies that cast doubt on the usefulness of the individual measures and address the more general issue of additive decomposition.

[2] It should be noted that Houlsby et al. [2011] proposed the entropy decomposition to inform uncertainty sampling in active learning, where mutual information has a slightly different purpose, namely reducing uncertainty about Θ, while we use $I(Y, \Theta)$ to quantify uncertainty about Y (which the symmetric nature of $I(\cdot)$ enables us to do).

4 CRITICAL ASSESSMENT

Let TU, AU, and EU denote, respectively, measures $\Delta_K^{(2)} \longrightarrow \mathbb{R}$ of total, aleatoric, and epistemic uncertainty associated with a (second-order) uncertainty representation $Q \in \Delta_K^{(2)}$. In the literature, it is common to define the fundamental properties of uncertainty measures through an axiomatic approach (see also, for example, Bronevich and Klir [2008], Pal et al. [1993]. In the following, we define some properties these measures should naturally satisfy.

- A0 TU, AU, and EU are non-negative.
- A1 EU vanishes for Dirac measures $Q = \delta_{\boldsymbol{\theta}}$.
- A2 EU and TU are maximal for Q being the uniform distribution on $\Delta_K^{(2)}$.
- A3 If Q' is a mean-preserving spread[3] of Q, then $\text{EU}(Q') \geq \text{EU}(Q)$ (weak version) or $\text{EU}(Q') > \text{EU}(Q)$ (strict version); the same holds for TU.
- A4 If Q' is a center-shift[4] of Q, then $\text{AU}(Q') \geq \text{AU}(Q)$ (weak version) or $\text{AU}(Q') > \text{AU}(Q)$ (strict version); the same holds for TU.
- A5 If Q' is a spread-preserving location shift[5] of Q, then $\text{EU}(Q') = \text{EU}(Q)$.

4.1 TOTAL UNCERTAINTY

Although Shannon entropy is widely established as a measure of uncertainty associated with a random variable, one may wonder about its appropriateness for quantifying total predictive uncertainty in the scenario outlined above. As an illustration, we pick up our previous example of binary classification (note that properties A0-A5 apply to $Q, Q' \in \Delta_K^{(2)}$ but violations for $K = 2$ suffice to conclude that they do not hold in general). Fig. 2 shows some exemplary second-order distributions Q together with the values they induce for TU, AU, and EU about the prediction of θ.

We first note that property A3 is violated by TU in terms of entropy, at least in its strict version, because (3) depends on Q only through its expectation.

Proposition 1. *Total uncertainty defined in terms of Shannon entropy (3) violates the strict version of A3.*

For example, in the special case $Q \in \Delta_2^{(2)}$, $H(Y)$ is maximal as soon as Q is symmetric around $\theta = \frac{1}{2}$. Consequently, the uniform distribution on the unit interval, $Q = \mathcal{U}[0, 1]$ (Fig. 2a), which is commonly considered a representation of *complete ignorance* (justified by the "principle of indifference" invoked by Laplace, or by referring to the principle of maximum entropy), has the same TU as a mixture of Dirac distributions $Q = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ (Fig. 2f) dividing all probability mass between the outcomes $\theta = 0$ and $\theta = 1$. While one may argue that the uncertainty about the target variable Y is indeed the same in both cases, it seems counter-intuitive to obtain identical results when in one scenario, the learner (supposedly) knows nothing, whereas, in the other, it has substantial knowledge about the ground-truth distribution θ^*. Likewise, it appears strange that any distribution Q not symmetric around $\frac{1}{2}$ will have a lower TU, including, for example, $\mathcal{U}[0.45, 0.85]$ (Fig. 2e), when it arguably represents a lower degree of informedness than the Dirac mixture.

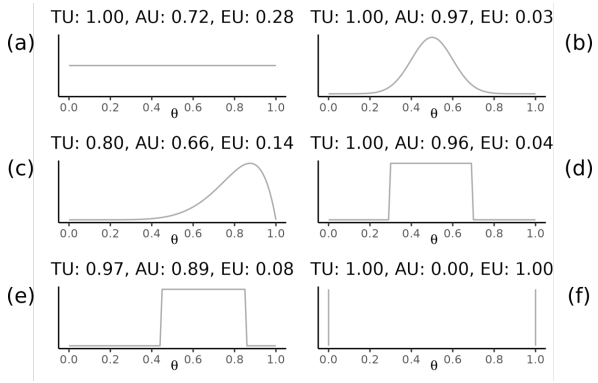


Figure 2: Different second-order distributions Q over the parameter θ of a Bernoulli distribution with associated uncertainty. (a) $\mathcal{U}[0, 1]$; (b) $\mathcal{N}(0.5, 0.01)$; (c) $\text{Beta}(8, 2)$; (d) $\mathcal{U}[0.3, 0.7]$; (e) $\mathcal{U}[0.45, 0.85]$; (f) $\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$.

The root cause of these peculiarities is the role of the distribution Q in the computation of TU – effectively, Q is marginalized over. In this sense, the reason for having maximal TU in the case of $\mathcal{U}[0, 1]$ is not that it (supposedly) encodes the greatest lack of knowledge, but only its symmetry around $\theta = \frac{1}{2}$. Information about the second-order distribution is thus compressed to a single number in the form of its expectation. On the other hand, the concentration of Q is hardly reflected in the TU, although it should be an important indicator of the learner's state of knowledge.

4.2 ALEATORIC UNCERTAINTY

Noting that, in Eq. (5), Q represents the learner's *subjective belief* about a supposedly objective ground-truth $\boldsymbol{\theta}^*$ associated with query $\boldsymbol{x}$, $H(Y \mid \Theta)$ can be seen as the learner's (current) best guess of the true AU. Although this might seem a sensible quantity, the relationship between this estimate and the true AU, $H(Y \mid \boldsymbol{\theta}^*)$, is by no means clear.

[3] Let $X \sim Q, X' \sim Q'$ be two random variables, where $Q, Q' \in \Delta_K^{(2)}$. Then, Q' is called a mean-preserving spread of Q iff $X' \stackrel{d}{=} X + Z$, for some random variable Z with $\mathbb{E}[Z|X = x] = 0 \; \forall x$ in the support of X.

[4] Q and Q' differ only in their respective means, where the mean of Q' is closer – in terms of $L1$ distance in Cartesian coordinates – to the barycenter of $\Delta_K^{(2)}$.

[5] Q and Q' differ only in their respective means.

The problem is that AU estimates – even if the additive nature of the decomposition might suggest otherwise – cannot be separated from the learner's epistemic state. In the finite-sample case, the learner must fall short of perfect knowledge, meaning that its belief Q will not coincide with the Dirac distribution on $\boldsymbol{\theta}^*$ that represents the true second-order distribution. This, however, directly translates to the estimate of AU that is derived by taking the expectation over Q. Whenever there is EU, we must expect the AU estimation to be compromised, and the absence of a ground-truth precludes us from knowing how it is affected exactly.

In particular, conditional entropy is neither a true expectation nor a lower or upper bound on the true AU, and the resulting point estimate suggests a degree of informedness that is hardly justified. The limited ability of probability distributions to properly distinguish between different subjective beliefs adds to the problem. For example, if a uniform Q is meant to represent complete ignorance, then all the learner can infer is that the true AU is between zero (for $\theta \in \{0, 1\}$) and one (for $\theta = \frac{1}{2}$). This is clearly a different situation than having perfect knowledge about all first-order distributions being equally probable. Ultimately, however, both lead to the same average over the support of Q. Therefore, we must conclude that AU is computed with respect to a belief that the second-order distribution may fail to express, and that must be expected to be wrong.

4.3 EPISTEMIC UNCERTAINTY

The definition of mutual information (Eq. (6)) may appear intuitively plausible at first sight: If the observation of Y reveals a sizeable amount of information about Θ, then the learner's uncertainty about Θ must be high (*et vice versa*). Likewise, a high divergence in the finite-ensemble version (Eq. (9)), i.e., in the predictions of different ensemble members, can be viewed as indicating a high EU.

That said, the measures do not necessarily behave as one may expect. Looking at Fig. 2, a first observation is that EU tends to be relatively low, which is especially noticeable in the case of the standard uniform distribution (upper left). As already mentioned, $\mathcal{U}[0, 1]$ is usually meant to represent complete ignorance, suggesting that the learner is as uninformed as it could be, so we would expect a measure of EU to fulfill property A2 (i.e., be maximal) in this case.

Proposition 2. *EU in terms of mutual information $I(Y, \Theta)$ violates property A2.*

This violation holds in general, and can be seen very easily for the case $K = 2$ in Fig. 2. In Bayesian inference, for example, the uniform distribution is commonly adopted as a prior, representing knowledge before seeing any data. In such a state, one would assume all uncertainty to epistemic, and hence the measure of EU to be very high. On the contrary, however, AU (at 0.72) is well above EU (at 0.28). Actually, the maximal EU (1.00) is reached only for the case of Fig. 2f, in which Q is a mixture of two Dirac measures, one placed at $\theta = 0$ and the other at $\theta = 1$. This distribution suggests a deterministic dependency, where the outcome is either certainly positive or certainly negative – a case in which the learner undoubtedly knows more than nothing, so again, the level of EU does not seem appropriate.

Also, note that EU is not invariant against a location shift of the distribution Q, even though it would be reasonable to expect that EU remains constant when the spread of Q does not change.

Proposition 3. *There exist location shifts $Q' \in \Delta_K^{(2)}$ of $Q \in \Delta_K^{(2)}$ such that property A5 is violated.*

For example, the distributions in Fig. 2d and Fig. 2e (case $K = 2$) are both uniform on an interval of length 0.4, suggesting the same level of informedness about the ground truth θ. However, the respective degrees of EU (0.04 vs. 0.08) differ. Of course, one may argue that the uncertainty will depend not only on the shape but also on the location of the distribution. Then, however, $\mathcal{U}[0.3, 0.7]$ should arguably exhibit higher uncertainty, as it puts more probability mass close to $\theta = \frac{1}{2}$ than $\mathcal{U}[0.45, 0.85]$. This is indeed the case for TU and AU but exactly reversed for EU.

Lastly, we would expect EU to increase when Q's mass is spread over a larger support (all else being equal), such that more hypotheses are deemed likely. Such spreading happens, for example, in moving from $\frac{1}{2}\delta_{0.45} + \frac{1}{2}\delta_{0.95}$ to $\mathcal{U}[0, 1]$, while EU actually gets higher rather than lower.

Proposition 4. *EU in terms of mutual information $I(Y, \Theta)$ violates property A3.*

Mathematically, the numbers are clearly correct and meaningful: In Fig. 2f, for example, the ground-truth distribution θ is uniquely determined as soon as the outcome Y has been observed, so mutual information should be maximal. What must be concluded is rather that mutual information may not be the right measure of EU. Indeed, mutual information is arguably more a measure of *divergence* or *conflict* than of ignorance, as the common understanding of EU would suggest. This is quite obvious from Eq. (6) and even easier to see in the discrete version (Eq. (9)) for finite ensembles. Mutual information effectively quantifies the expected divergence of single hypotheses from the opinion given by integrating over all of them. Consequently, it is much higher for the Dirac mixture in Fig. 2f, with maximal divergence of both hypotheses from the average $\theta = \frac{1}{2}$, than for $\mathcal{U}[0, 1]$, which also assigns probability mass to many alternatives that diverge only little from the expectation. Thus, mutual information behaves as expected but does not seem well-suited to measuring a quantity taken to represent *ignorance*.

4.4 ADDITIVE DECOMPOSITION

Let us conclude this section with a few remarks on another potential problem of uncertainty quantification, which is related, though not restricted, to the quantification discussed in this paper. According to Eq. (4), TU (entropy) decomposes *additively* into AU (conditional entropy) and EU (mutual information). When combining distinct sources of uncertainty, additive representations of this kind appear natural and can also be found for other measures of uncertainty, such as variance (see, e.g., Depeweg et al. [2018] for a discussion of entropy and variance decomposition). However, when considering Eq. (4) in a machine learning setting, where the (total) uncertainty is not fixed but decreases with increasing sample size, additivity might be less obvious because the aleatoric part does not correspond to the *true* AU, but only to an *estimate* thereof. Moreover, this estimate depends on the EU, such that the two measures are tightly interwoven and certainly not independent of each other. So, what is the justification for simply adding them up?

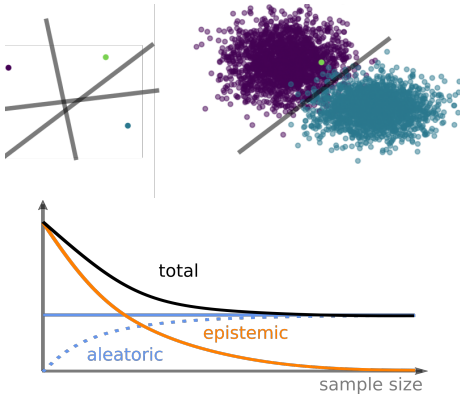


Figure 3: In the beginning, the learner is very uncertain about the true class probability for the (green) query point $\boldsymbol{x}$ (top-left), but this (epistemic) uncertainty disappears with increasing sample size (top-right). The bottom plot shows the expected qualitative behavior of uncertainty components. The dotted line marks the difference of TU and EU (which equals AU when stipulating additivity).

The dependence between the measures can be seen most clearly at the beginning of a learning process, during which the number of observed samples increases when the learner has seen very little or even no data. For example, take the case of binary classification, where the learner is still fully ignorant about the probability $\theta = p(y_1 \mid \boldsymbol{x})$ of the positive class in a certain point $\boldsymbol{x}$: it might be a clear positive case ($\theta = 1$), a clear negative case ($\theta = 0$), but also everything in between. In this situation, an ideal measure of TU should arguably assume its maximum value. Likewise, an ideal measure of EU should be maximal because the learner is as uninformed about θ as it can be (note this is *not* the case for mutual information when starting with a uniform prior). Then, however, additivity implies that AU must be zero, at least if the measures share a common scale with the same maximum value (which holds for entropy and mutual information). In a sense, the aleatoric part is already comprised in the epistemic part, suggesting that the difference between TU and EU can, at best, be interpreted as a *lower bound* to the true AU. This is depicted graphically in Fig. 3: With a growing sample size, we expect EU to decrease and eventually vanish because the learner will gain full knowledge about the data-generating process, while both TU and TU minus EU should converge to the true AU (the former from above and the latter from below).

The above can be summarized as follows:

Proposition 5. *If EU and TU attain their respective maxima at the beginning of learning, and they are constructed to be on the same scale, then TU cannot decompose additively into EU and AU if AU is positive.*

What these considerations suggest is that, in the finite-sample (machine learning) regime, additivity might inevitably be violated by "ideal" measures of TU, EU, and (ground-truth) AU (again, note that entropy and mutual information with a uniform prior are *not* ideal in this sense, which is what we criticized them for). Instead, additivity may only hold for the decomposition of TU into EU and a *lower bound* to AU, or might be relaxed to sub-additivity along the lines of Dubois et al. [1996]. Alternatively, additivity must be preserved by giving up other assumptions, e.g., that all measures share the same scale. Recently, for example, Hüllermeier et al. [2022] proposed a decomposition for the uncertainty of sets of probability distributions (credal sets), in which EU can become twice as high as AU.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We conduct a number of experiments that provide practical evidence about the incoherent behavior of entropy-based uncertainty measures. For further details on training configurations, see supplementary material. The experimental code is available in full via a public repository[6].

Datasets We consider image classification tasks for two real-world datasets, `CIFAR10` [Krizhevsky, 2009] and `MNIST` [LeCun et al., 1998], containing ten classes of color and gray-scale images, respectively. For a setting over which we can exert full control, we further synthesize black-and-white images of rectangles, where one class is characterized by vertical extension (i.e., height $>$ width) and the other *vice versa*, and simulate a bivariate classification problem with tabular features and four classes.

[6] `https://github.com/lisa-wm/entropybaseduq`

Probabilistic Learners For the computer vision tasks, we employ deep ensembles of M neural networks with varying random initialization [Lakshminarayanan et al., 2017], as well as a Laplace approximation [Daxberger et al., 2021] that fits a Gaussian approximate posterior to the location of the maximum-a-posteriori estimator and draws M samples from this distribution. Furthermore, we train a random forest [Breiman, 2001] and an ensemble of single-hidden-layer feedforward neural networks (MLPs) for the tabular classification problem. Ensemble predictions are computed by averaging over the outputs of individual members. We set $M = 10$. While this may seem small, note that this is in accordance with common practice (e.g., Beluch et al. [2018], Lee et al. [2021], Abe et al. [2022], Kristiadi et al. [2022], Turkoglu et al. [2022]) because larger ensembles are rarely affordable in the deep learning with architectures typically numbering in the millions. Our ablations suggest that the observed qualitative behavior is fairly robust with respect to ensemble size (see supplementary material).

Evaluation The figures in the following paragraphs visualize estimates for AU, EU, and TU (normalized to values in $[0, 1]$) alongside predictive performance in terms of accuracy (ACC) and expected calibration error (ECE; Naeini et al. [2015]). The uncertainty estimates are depicted as lines (averaged over all test samples, left y-axis), while the bars represent performance (right y-axis). All results are aggregated over three independent runs, with error bars on the uncertainty curves indicating one standard deviation.

5.2 RESULTS AND DISCUSSION

5.2.1 Increasing Sample Size

We first study how the components of total Shannon entropy evolve with sample size, where we admit increasingly larger training datasets from 1% to 100% of available samples (randomly selected; the classes are balanced).

Expected Behavior TU and EU start from their highest level and decrease gradually, while AU approaches its true constant value from below (due to the additivity assumption). The reported uncertainty is higher when accuracy is low.

Observed Behavior We observe the expected downward slope of TU and EU as a general pattern, but the evidence is otherwise not quite consistent. Most strikingly, the AU estimates also assume high initial values and decrease for larger sample size in three out of four cases. For `CIFAR10`, this behavior is clearly visible (Fig. 5), underlining our concern that AU estimates are unreliable in the light of limited knowledge. In the case of `MNIST` and the deep ensemble, uncertainty is ultra-low from the beginning, but a close-up in the small overlaying plot of Fig. 4 (top) reveals the same trend. Reported EU for small sample sizes seems quite low overall, despite limited information and consequently poor performance. A notable exception is the Laplace approximation for `MNIST` (Fig. 4, bottom), which resembles most closely what we would have expected, though it appears more susceptible to random changes (see close-up).

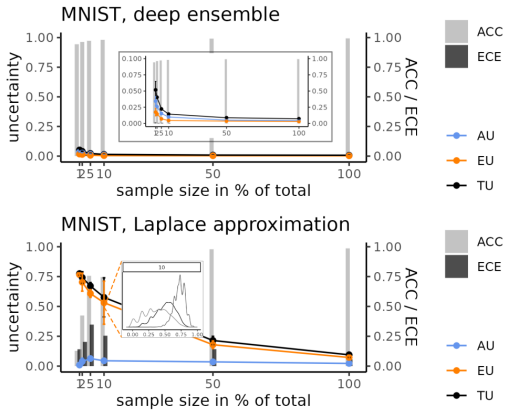


Figure 4: Entropy-based uncertainty for increasing sample size (`MNIST`). Overlaying plot (top): re-scaled left y-axis. Overlaying plot (bottom): empirical densities of observed instance-wise EU values (each curve represents one run of the experiment) with 10% sample size.

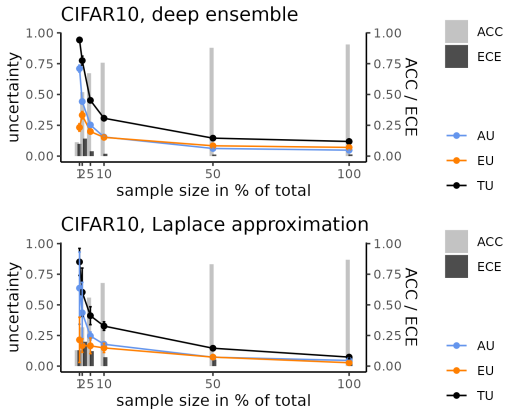


Figure 5: Entropy-based uncertainty for increasing sample size (`CIFAR10`).

5.2.2 Increasing Data Noise

Next, we modify the level of noise in the data, for which we use different proxies. First, we vary image resolution in the computer vision tasks by downscaling (original sizes: 28×28 / 32×32 for `MNIST` / `CIFAR10`); second, we shrink the relative class distance for the tabular data and thus enforce stronger overlap (Fig. 6); and lastly, we randomly change class labels for a varying share of observations in the same dataset (see supplementary material).

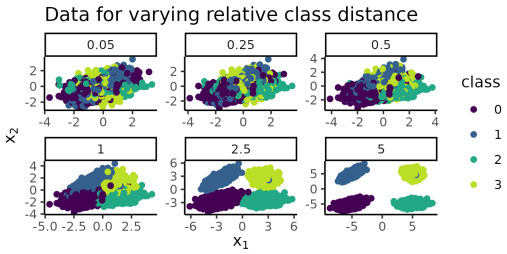


Figure 6: Tabular data with two features and four classes for increasing class overlap.

Expected Behavior AU picks up with increasing noise level. Since learner capacity remains fixed, it is reasonable to assume that EU also rises to some extent when the decision boundaries become more complex with mounting degree of dataset contamination.

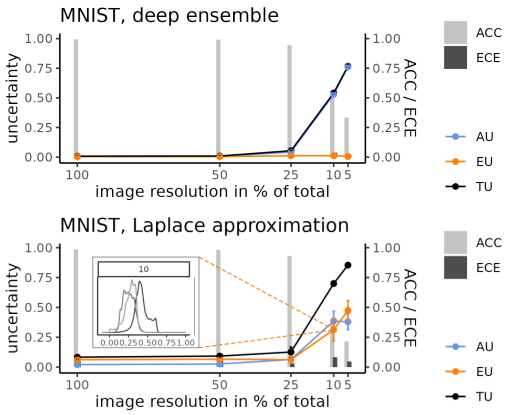


Figure 7: Entropy-based uncertainty for decreasing image resolution (MNIST). Overlaying plot: empirical densities of observed instance-wise EU values (each curve represents one run of the experiment) with 10% resolution.

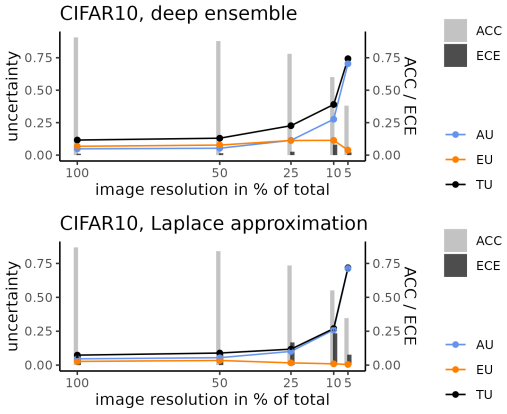


Figure 8: Entropy-based uncertainty for decreasing image resolution (CIFAR10).

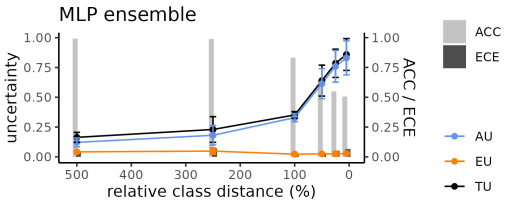
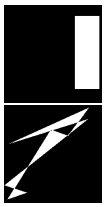


Figure 9: Entropy-based uncertainty for for increasing class overlap (tabular data).

Observed Behavior More noise indeed prompts a rise in AU in all cases (Fig. 7-9). However, EU remains basically constant and very low, even when the data are so noisy that performance plummets. Note that, for instance, 10% resolution corresponds to a compression into 3×3 pixels for both MNIST and CIFAR10, and how the decision boundaries become quite blurred in the tabular dataset when classes overlap more strongly (Fig. 6). Again, the Laplace approximation for MNIST is something of an exception: here, EU increases for the lowest resolution values, but leaves AU remarkably small and even falling from 10% to 5% resolution. We find thus that the uncertainty measures behave, if not entirely implausible, again somewhat inconsistently.

5.2.3 Test Distribution Shift



Following a slightly different idea than in the previous experiment, we now train the learner on a well-separated dataset of synthetic rectangles and seek predictions for instances which interpolate between the classes. This is achieved by decreasing the rectangles' side-length ratio, effectively pushing them towards a square.

We further present the learner with samples that are from an entirely different class, namely random non-convex polygons with 3–5 vertices, and thus out-of-distribution (OOD) relative to the training data. Both settings create a form of distribution shift between training and test data that is associated with the presence of predictive uncertainty.

Expected Behavior Class interpolation mainly affects AU through instances located close to the learned decision boundaries. In OOD detection, demanding predictions for samples from a previously unseen class should spike an increase in EU.

Observed Behavior The deep ensemble and Laplace approximation act quite differently in the presence of class interpolation (Fig. 10). While the former behaves roughly in line with our expectation, the Laplace learner attributes most of the uncertainty to the epistemic component when samples become increasingly quadratic, leaving AU almost

unchanged. In OOD detection, both classifiers react appropriately in principle by reporting higher uncertainty for the polygon images. However, they allocate the majority of the uncertainty premium differently: Laplace almost exclusively raises the epistemic component, the deep ensemble signals both higher AU and EU. In both ablations, a vast gap in overall levels of uncertainty is evident (e.g., for the OOD case, Laplace reports roughly five times higher TU).

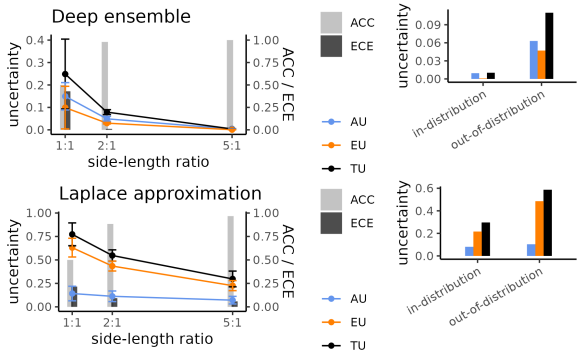


Figure 10: Entropy-based uncertainty for synthetic data with increasing class interpolation (left) and OOD detection (random polygons; right).

5.2.4 Discussion

Our experiments confirm empirically that the entropy-based uncertainty measures exhibit some – sometimes subtle – inconsistencies in different learning situations. While TU as a whole behaves fairly sensibly, the attribution to its aleatoric and epistemic components often appears implausible. We further identify very low levels of EU as a global pattern. That said, we also find that the design of the probabilistic learner used can affect the results considerably, even when predictive performance remains similar. In particular, deep ensembles draw from the opinions of various base learners when they estimate predictive uncertainty, as opposed to Laplace approximation, where a single base network determines uncertainty quantification. Consequently, the behavior of explicit ensembles also depends on the complexity of their base learners in that AU estimates decrease considerably when each ensemble member is granted more capacity (see supplementary material). This relates to another insight relevant for this kind of empirical studies: reasoning about uncertainty estimates must always occur in conjunction with predictive accuracy and calibration, since we can hardly expect sensible uncertainty behavior when the learner fails on the basic underlying task. Numerous practical aspects thus interact to elicit or mitigate shortcomings of the discussed measures and warrant caution in interpreting any predictive uncertainty estimate one may obtain in practice. Worryingly, this becomes all the more important with limited data, a situation often encountered in, e.g., medical applications.

6 CONCLUDING REMARKS

Despite the common use of Shannon entropy, conditional entropy, and mutual information for the quantification of predictive uncertainty, we observed and empirically confirmed that these measures behave neither coherently nor always as expected. First, each of the measures itself can be criticized for (the lack of) certain properties. For example, we argued that mutual information is rather a measure of divergence than a measure of ignorance. Second, the discrepancy between (i) the ground-truth scenario concerning the true data-generating process with an objective ground-truth AU and zero EU, and (ii) the finite-sample setting in which AU is *derived* from (and hence depends on) subjective EU, leads to incoherencies and calls additivity of the decomposition into question. With this in mind, we encourage a critical assessment of the existing framework and cautious use of these measures.

When remaining committed to the basic (Bayesian) representational framework with second-order distributions for the epistemic part of uncertainty, an obvious alternative is to look for measures with better properties, and indeed, some proposals can be found in the recent literature (e.g., Sicking et al. [2020]). Another option is to change not only the measures for quantification, but also the underlying representation, for example by moving beyond classical probability. In fact, the suitability of probability distributions to represent ignorance in the sense of a lack of knowledge has been questioned by various scholars [Dubois et al., 1996], who advocate the use of generalized, more expressive uncertainty formalisms. In particular, it has been argued that the uniform distribution is not a suitable representation of complete ignorance, as it fails to distinguish the epistemic state of zero knowledge and the state of perfect information, where the learner knows that all outcomes are equiprobable. Indeed, the averaging over conditional entropies, which is part of the problems with the aleatoric component, is arguably meaningful in the latter case but much less so in the former. Seen from this perspective, one may indeed wonder whether classical probability is the right paradigm for modeling the epistemic part. On the other side, measuring uncertainty and disaggregating total predictive uncertainty into its aleatoric and epistemic components do not necessarily become simpler for generalized formalisms [Hüllermeier et al., 2022]. Clearly, we are not yet at the end of the path toward a truly meaningful uncertainty representation and quantification.

Acknowledgements

This work was supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research (LW, YS).

References

T. Abe, E. K. Buchanan, G. Pleiss, R. Zemel, and J. P. Cunningham. Deep Ensembles Work, But Are They Necessary?, 2022.

C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50:5–43, 2003.

R. B. Ash. *Information Theory*. Dover Publications, 1965.

A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov. Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning. In *ICLR*, 2020.

W. H. Beluch, T. Genewein, A. Nurnberger, and J. M. Kohler. The Power of Ensembles for Active Learning in Image Classification. In *CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

L. Breiman. Random Forests. *Machine Learning*, 45(1): 5–32, 2001.

A. Bronevich and G. J. Klir. Axioms for uncertainty measures on belief functions and credal sets. In *NAFIPS*, pages 1–6. IEEE, 2008.

S. Budd, E. C. Robinson, and B. Kainz. A Survey on Active Learning and Human-in-the-Loop Deep Learning for Medical Image Analysis. *Medical Image Analysis*, 71, 2021.

B. Charpentier, D. Zügner, and S. Günnemann. Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts. In *NeurIPS*, pages 1103–1130, 2020.

B. Charpentier, R. Senanayake, M. Kochenderfer, and S. Günnemann. Disentangling Epistemic and Aleatoric Uncertainty in Reinforcement Learning, 2022.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig. Laplace Redux – Effortless Bayesian Deep Learning, 2021.

S. Depeweg. *Modeling Epistemic and Aleatoric Uncertainty with Bayesian Neural Networks and Latent Variable*. PhD thesis, TU München, 2019.

S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning, 2018.

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

Y. Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Y. Gal, R. Islam, and Z. Ghahramani. Deep Bayesian Active Learning with Image Data, 2017.

A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*. CRC Press, 2021.

N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian Active Learning for Classification and Preference Learning, 2011.

D. Huseljic, B. Sick, M. Herde, and D. Kottke. Separation of Aleatoric and Epistemic Uncertainty in Deterministic Deep Neural Networks. In *ICPR*, 2020.

E. Hüllermeier and W. Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, 2021.

E. Hüllermeier, S. Destercke, and M. H. Shaker. Quantification of Credal Uncertainty in Machine Learning: A Critical Analysis and Empirical Comparison. In *UAI*, 2022.

A. Kendall and Y. Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *NeurIPS*, 2017.

A. Kirsch, J. Kossen, and Y. Gal. Marginal and Joint Cross-Entropies & Predictives for Online Bayesian Inference, Active Learning, and Active Sampling, 2022.

A.-K. Kopetzki, B. Charpentier, D. Zügner, S. Giri, and S. Günnemann. Evaluating Robustness of Predictive Uncertainty Estimation: Are Dirichlet-based Models Reliable? In *ICML*, pages 5707–5718, 2021.

A. Kristiadi, M. Hein, and P. Hennig. Being a Bit Frequentist Improves Bayesian Neural Networks. In *AISTATS*, 2022.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *NeurIPS*, 2017.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. In *Proceedings of the IEEE*, 1998.

K. Lee, M. Laskin, A. Srinivas, and P. Abbeel. SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. In *ICML*, 2021.

Lightning AI. PyTorch Lightning: Train and Deploy PyTorch at Scale, 2023.

A. Malinin. *Uncertainty Estimation in Deep Learning with Application to Spoken Language Assessment*. PhD thesis, University of Cambridge, 2019.

A. Malinin and M. Gales. Predictive Uncertainty Estimation via Prior Networks. In *NeurIPS*, 2018.

R. Michelmore, M. Kwiatkowska, and Y. Gal. Evaluating Uncertainty Quantification in End-to-End Autonomous Driving Control, 2018.

A. Mobiny, P. Yuan, S. K. Moulik, N. Garg, C. C. Wu, and H. Van Nguyen. DropConnect is effective in modeling uncertainty of Bayesian deep networks. *Scientific Reports*, 11, 2021.

M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

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

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

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A. F. Psaros, X. Meng, Z. Zou, L. Guo, and G. E. Karniadakis. Uncertainty Quantification in Scientific Machine Learning: Methods, Metrics, and Comparisons, 2022.

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

C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 1948.

A. Shelmanov, D. Puzyrev, L. Kupriyanova, D. Belyakov, D. Larionov, N. Khromov, O. Kozlova, E. Artemova, D. V. Dylov, and A. Panchenko. Active Learning for Sequence Tagging with Deep Pre-trained Models and Bayesian Uncertainty Estimates. In *Proceedings of the 16th Conference of the EACL*, pages 1698–1712. Association for Computational Linguistics, 2021.

J. Sicking, M. Akila, M. Pintz, T. Wirtz, A. Fischer, and S. Wrobel. Second-moment loss: A novel regression objective for improved uncertainties. *CoRR*, abs/2012.12687, 2020.

L. Smith and Y. Gal. Understanding Measures of Uncertainty for Adversarial Example Detection, 2018.

M. Tan and Q. V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *ICML*, 2019.

M. O. Turkoglu, A. Becker, H. A. Gündüz, M. Rezaei, B. Bischl, R. C. Daudt, S. D'Aronco, J. D. Wegner, and K. Schindler. FiLM-Ensemble: Probabilistic Deep Learning via Feature-wise Linear Modulation. Accepted at the Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS), 2022.

A. G. Wilson. The Case for Bayesian Deep Learning, 2020.

L. Winkler, C. Ojeda, and M. Opper. Stochastic Control for Bayesian Neural Network Training. *Entropy*, 24, 2022.

J. O. Woo. Analytic Mutual Information in Bayesian Neural Networks, 2022.

X. Wu, K. M. Knill, M. J. Gales, and A. Malinin. Ensemble Approaches for Uncertainty in Spoken Language Assessment. In *Interspeech 2020*, pages 3860–3864. ISCA, 2020.