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

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 fully explain y [Malinin, 2019]. Thus, it is not possible to predict the outcome with certainty even assuming perfect knowledge of the underlying data-generating process. Obviously, the learner does usually not have access to $p(\cdot \mid \boldsymbol{x})$ and instead 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 modern applications of AI, e.g., in order to invoke human intervention in multi-stage decision processes [Gal et al., 2017, Budd et al., 2021, Hüllermeier and Waegeman, 2021, Kapoor et al., 2022, 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, e.g., Houlsby et al. [2011] and Depeweg [2019], 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 metrics (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 tailored 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 metrics 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 often omitted third source of EU is misspecification with p lying outside $\mathcal{H}$ (e.g., Wilson [2020]).

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

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 (level-one) probability distributions over $\mathcal{Y}$ can be identified with the $(K-1)$-simplex $\Delta_K := \left\{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in [0,1]^K \mid \|\boldsymbol{\theta}\|_1 = 1\right\}$. 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 (level-two) probability distribution over (level-one) 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 level-two 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 $\boldsymbol{x}$ for which a prediction is sought.

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 higher concentration of Q indicates more certainty about the true distribution $\boldsymbol{\theta}^*$.

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 level-two distribution Q could be, for example, a Beta distribution (or, more generally, a Dirichlet distribution for $K \geq 2$).

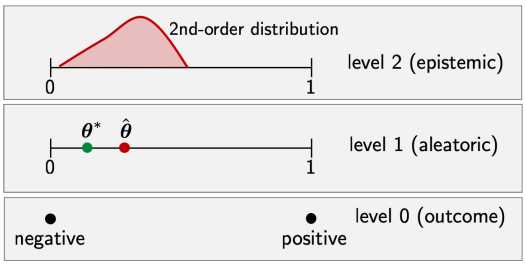


Figure 1: Bi-level uncertainty representation in binary classification. We pose a second-order distribution Q over Bernoulli distributions corresponding to the probability θ of the positive class. In Bayesian inference, Q is given by the posterior distribution (epistemic level), and a point prediction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

2.2 ENTROPY-BASED INFORMATION METRICS

Given a representation of an uncertain prediction in terms of a level-two 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 level-one 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 level-one 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 level-two distribution:

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

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 [Ash, 1965]. $I(Y, \Theta)$ between Y and Θ is equal to the Kullback-Leibler (KL) divergence between the joint distribution and the product of the marginals [Cover and Thomas, 2006], vanishing if both random variables are independent:

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

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 learners is common practice and known as *ensemble learning* (e.g., Breiman [2001], Lakshminarayanan et al. [2017]). For TU, the entropy over the approximate level-one 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 $\boldsymbol{\theta}^{(m)} = f_m(\boldsymbol{x}), m \in \{1, \ldots, M\}$, are the predictions of individual ensemble members. 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

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

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 level-one 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 we will discuss 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 metrics – to the best of our knowledge – have hardly drawn any criticism in the machine learning community. In the following, we point out numerous inconsistencies that cast doubt on the usefulness of the individual metrics and address the more general issue of additive decomposition.

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 (level-two) 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[2] 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[3] 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.

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-A4 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 level-two distributions Q together with the values they induce for TU, AU, and EU about the prediction of θ.

While properties A0, A2, and A4 are fulfilled for TU (in terms of entropy), property A3 is violated, 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. 2, top left), 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 $Q = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ (bottom right) 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 that is not symmetric around $\frac{1}{2}$ will have a lower TU, including, for example, $\mathcal{U}[0.45, 0.85]$ (bottom left), when it arguably represents a lower degree of informedness than the Dirac mixture.

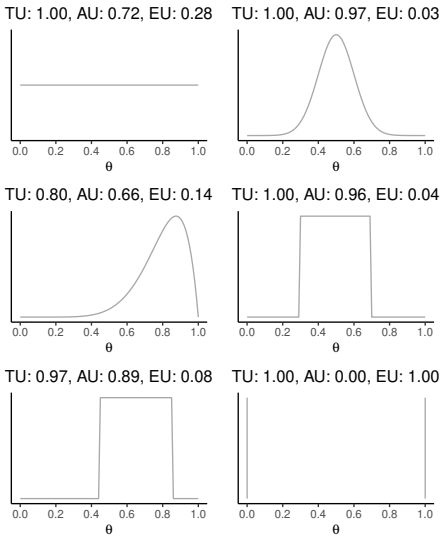


Figure 2: Different level-two distributions Q over the parameter θ of a Bernoulli distribution with associated uncertainty. Top: $\mathcal{U}[0, 1]$, $\mathcal{N}(0.5, 0.01)$; middle: $\text{Beta}(8, 2)$, $\mathcal{U}[0.3, 0.7]$; bottom: $\mathcal{U}[0.45, 0.85]$, $\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$.

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 it might seem

[2] 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.

[3] Q and Q' differ only in their respective means, where the mean of Q' is closer to the barycenter of $\Delta_K^{(2)}$.

that this is a sensible quantity, the relationship between this estimate and the true AU, $H(Y \mid \boldsymbol{\theta}^*)$, is by no means clear. In particular, it is neither a true expectation nor a lower or upper bound. The resulting point estimate suggests a degree of informedness that is hardly justified. In fact, by assuming a single ground truth, the learner even knows that its belief Q will be (almost-surely) wrong since the true level-two distribution is a Dirac distribution on $\boldsymbol{\theta}^*$[4]. 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 level-one 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 level-two 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 (see 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 bottom-right case, 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*[5] $Q' \in \Delta_K^{(2)}$ *of* $Q \in \Delta_K^{(2)}$ *such that* $EU(Q') \neq EU(Q)$.

For example, the middle-right and bottom-left distributions in Fig. 2 (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.

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

Mathematically, the numbers are clearly correct and meaningful: In the bottom-right case, 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 (see Section 2.3). Mutual information effectively quantifies the expected divergence of single hypotheses from the expected opinion given by integrating over all of them. Consequently, it is much higher for the Dirac mixture in Fig. 2, 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,

[4] If we assume the distribution of the predictor $F : \mathcal{X} \longrightarrow \Delta_K^{(2)}$ to be continuous, then $P(\{x \in \mathcal{X} : F(x) = \delta_{\boldsymbol{\theta}^*}\}) = 0$.

[5] Q, Q' differ only in their support.

additive representations of this kind appear natural and can also be found for other measures of uncertainty, such as variance (e.g., Depeweg et al. [2018], Egele et al. [2022], Rajaraman et al. [2022]). 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?

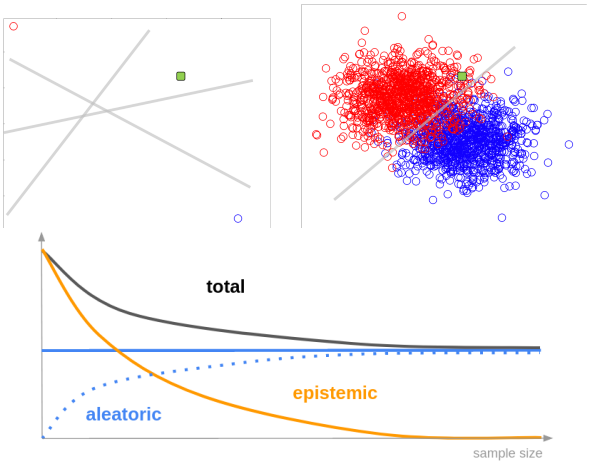


Figure 3: In the beginning, the learner is very uncertain about the true class probability for the (green) query point $\boldsymbol{x}$ (top-left), but this (epistemic) uncertainty disappears with increasing sample size (top-right). The bottom plot shows the expected qualitative behavior of uncertainty 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−EU should converge to the true AU (the former from above and the latter from below).

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. 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 to provide practical evidence about the incoherent behavior of entropy-based uncertainty measures. 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. In both, a fixed set of 10k samples serves as test data, compared to 60k training observations. For the training hyperparameters, we adopt standard choices (see supplementary material for the full code). We train `EfficientNet-B7` (approx. 64m parameters; Tan and Le [2019]) for `CIFAR10` and a small convolutional network (approx. 62k parameters) for `MNIST`, both with batch size 256. The number of training epochs is set to a maximum of 200 with early stopping if the validation loss (10% of training samples serving as a validation set) does not improve for ten/five (`CIFAR10`/`MNIST`) consecutive epochs. Furthermore, we employ an `SGD` optimizer with 0.9 momentum, using an initial learning rate of 10^{-2} that is adapted with cosine annealing [Loshchilov and Hutter, 2017], and apply weight decay equal to 5×10^{-4}. As a probabilistic learning algorithm, we employ deep ensembles, where M of the above networks, only differing in their random initialization, are trained in parallel [Lakshminarayanan et al., 2017], as well as a Laplace approximation [Daxberger et al., 2021] that models the Bayes posterior as a Gaussian at the location of the maximum-a-posteriori estimator. In the latter, the parameters of the Gaussian are computed over all weights in the small architecture and over the last layer in the `EfficientNet-B7` (e.g., Kristiadi

et al. [2020]). 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]), we use an ensemble size of $M = 5$.

In addition, we synthesize images of geometric shapes (train/test size: 6k/1k) to obtain a setting over which we can exert full control. This dataset comprises two classes of rectangles (white on black): one characterized by vertical extension, i.e., height $>$ width, and the other *vice versa*. We use the same training configurations as for `MNIST` here.

As before, since entropy is naturally upper-bounded by $\log_2 K$, the reported measures are normalized to a maximum of one. For reproducibility we include our code in the supplementary material.

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 gradually increase the training dataset from 1% to 100% of available samples (randomly selected), and employ both Laplace approximation (Fig. 4) and deep ensembles (Fig. 5) as probabilistic learners. TU, AU, and EU are depicted as curves over increasing sample size, averaged over all test samples. We further include accuracy (ACC) and expected calibration error (ECE; Naeini et al. [2015]) as bars to put provide context on predictive performance.

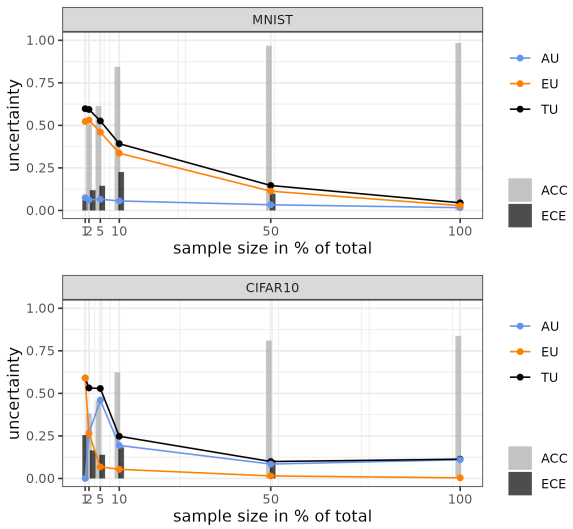


Figure 4: Entropy-based uncertainty for increasing sample size (Laplace approximation).

Ideally, we would obtain a picture similar to Fig. 3. The actual result for Laplace approximation reflects our discussion in Section 4.4: for `CIFAR10` in particular, high EU at small sample sizes impacts the estimate for AU. EU keeps decreasing with more training samples, which is in line with our expectations, but its absolute level seems somewhat arbitrary: it reduces only slowly for `MNIST`, even though accuracy reaches 84% already for 10% of samples, but decays much more quickly for `CIFAR10` than the poor early-on performance would advise. The deep ensemble appears able to offset this underestimation to some extent, which underlines our stance that mutual information is mainly a measure of conflict. For `MNIST`, the low uncertainty levels seem justified as the model average achieves very high performance from the beginning.

As expected, inconsistencies emerge primarily in the initial stages of the learning process when the learner's lack of knowledge has negative impact on the quality of uncertainty estimates, and are more pronounced in complex settings for the same reason. Considering the fact that many learning problems need to be solved in the light of quite limited data, this raises concerns for practical applications.

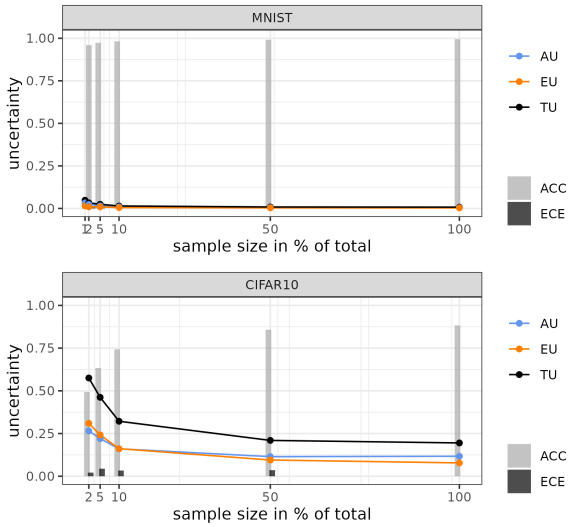


Figure 5: Entropy-based uncertainty for increasing sample size (deep ensemble).

5.2.2 Increasing Data Noise

In a second experiment, we modify the level of noise in the data, for which we use varying image resolution of the `MNIST` data as a proxy. Intuitively, lower resolution should increase the data-inherent uncertainty and thus push conditional entropy as a measure of AU upwards.

Fig. 6 reveals that this is indeed the case (depicted: deep ensemble, for Laplace, see supplementary material). Moving from 50% to 5% resolution, we observe a sharp uptick in AU that is complemented by deteriorating accuracy. EU, on the other hand, remains ultra-low even when the data become so noisy that the learner effectively resorts to guessing. Since we would expect the induced class overlap to also have

an impact on the learner's knowledge about the decision boundaries, this behavior is again not entirely plausible.

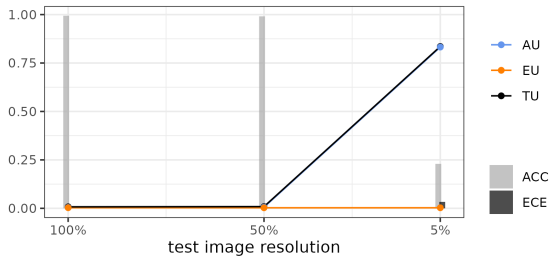
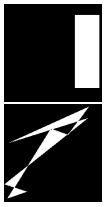


Figure 6: Entropy-based uncertainty measures for decreasing image resolution (`MNIST`, deep ensemble).

5.2.3 Synthetic Data



Besides real image data we also consider synthetic examples with full control over the present level of uncertainty. Images of rectangles (left, top) serve as training data, where class assignment is determined by the rectangles' side-length ratio (below/above one for vertical/horizontal extension).

Varying this ratio in the test data, we would expect uncertainty, and AU in particular, to rise as the rectangles resemble squares more and more closely, thereby simulating class interpolation. We do see such an increase in AU (albeit not exactly linear). Yet, similar to the previous experiment, there is virtually no change in the epistemic component (Fig. 8, left; depicted: deep ensemble, for Laplace, see supplementary material).

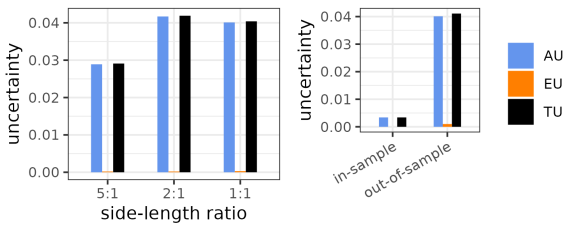


Figure 8: Entropy-based uncertainty for synthetic data for increasing class interpolation (left) and in-sample vs out-of-sample prediction (right; synthetic data, deep ensemble).

Lastly, we expose the synthetic-rectangle data to out-of-sample prediction, where the test instances are randomly drawn, non-convex polygons with 3–5 vertices as depicted at the beginning of this section. This task should spike an increase in EU, commonly taken as an indicator for an instance being out-of-distribution. Fig. 8 (right) clearly shows that no such signaling occurs – on the contrary, it seems that the additional uncertainty is borne almost entirely by the aleatoric component. We also notice that uncertainty overall remains on a very low level that is hardly justified for such a strong distribution shift.

In summary, there is substantial doubt that the entropy-based uncertainty measures discussed in Section 4 can generally be relied upon in practical application with limited data.

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 do neither behave 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 combination of (i) a ground-truth scenario, which corresponds to 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.

Adhering to the basic (Bayesian) representational framework with second-order distributions for the epistemic part of uncertainty, one 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]).

Yet another alternative is to change, not only the measures for quantification, but also the underlying representation, for example by going beyond standard 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 was 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 standard probability is the right formalism for modeling the epistemic part. On the other side, measuring uncertainty and disaggregating total predictive uncertainty into its aleatoric and epistemic constitutents 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.

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.

R. Egele, R. Maulik, K. Raghavan, B. Lusch, I. Guyon, and P. Balaprakash. AutoDEUQ: Automated Deep Ensemble with Uncertainty Quantification, 2022.

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.

S. Kapoor, W. J. Maddox, P. Izmailov, and A. G. Wilson. On Uncertainty, Tempering, and Data Augmentation in Bayesian Classification, 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 Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks. In *ICML*. arXiv, 2020.

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.

I. Loshchilov and F. Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *ICLR*, 2017.

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

S. Rajaraman, G. Zamzmi, F. Yang, Z. Xue, S. Jaeger, and S. K. Antani. Uncertainty Quantification in Segmenting Tuberculosis-Consistent Findings in Frontal Chest X-rays. *Biomedicines*, 10(6):1323, 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.