Second-Order Uncertainty Quantification: A Distance-Based Approach

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

Abstract

In the past couple of years, various approaches to representing and quantifying different types of predictive uncertainty in machine learning, notably in the setting of classification, have been proposed on the basis of second-order probability distributions, i.e., predictions in the form of distributions on probability distributions. A completely conclusive solution has not yet been found, however, as shown by recent criticisms of commonly used uncertainty measures associated with second-order distributions, identifying undesirable theoretical properties of these measures. In light of these criticisms, we propose a set of formal criteria that meaningful uncertainty measures for predictive uncertainty based on second-order distributions should obey. Moreover, we provide a general framework for developing uncertainty measures to account for these criteria, and offer an instantiation based on the Wasserstein distance, for which we prove that all criteria are satisfied.

1. Introduction

The need for representing and quantifying uncertainty in machine learning (ML) – particularly in supervised learning scenarios – has become more and more obvious in the recent past (Hüllermeier & Waegeman, 2021). This is largely due to the increasing use of AI-driven systems in safety-critical real-world applications having stringent safety requirements, such as healthcare (Lambrou et al., 2010; Senge et al., 2014; Yang et al., 2009) and socio-technical systems (Varshney & Alemzadeh, 2017). Dealing appropriately with uncertainty is a fundamental necessity in all these domains.

Broadly, uncertainties are categorized as *aleatoric*, stemming from inherent data variability, and *epistemic*, which arises from a model's incomplete knowledge of the data-generating process. By its very nature, epistemic uncertainty (EU) – often being characterized as *reducible* – can be decreased with further information. In contrast, aleatoric uncertainty (AU), rooted in the data generating process itself, is fixed and cannot be mitigated (Hüllermeier & Waegeman, 2021). The distinction between these uncertainty types has been a subject of keen interest in recent ML and statistical research (Gruber et al., 2023), finding applications in areas such as Bayesian neural networks (Kendall & Gal, 2017), adversarial attack detection mechanisms (Smith & Gal, 2018), and data augmentation strategies in Bayesian classification (Kapoor et al., 2022).

Arguably, predictive uncertainty is the most studied form of uncertainty in both ML and statistics. It pertains prediction tasks such as those in supervised learning. In the latter, we consider a hypothesis space $\mathcal{H}$, where each hypothesis $h \in \mathcal{H}$ maps a query instance $\boldsymbol{x}_q \in \mathcal{X}$ to a probability measure p on $(\mathcal{Y}, \sigma(\mathcal{Y}))$, where $\mathcal{Y}$ denotes the outcome space, and $\sigma(\mathcal{Y})$ a suitable σ-algebra on $\mathcal{Y}$. By producing estimates of the ground-truth probability measure p^* on $(\mathcal{Y}, \sigma(\mathcal{Y}))$, this probabilistic approach encapsulates aleatoric uncertainty about the actual outcome $y \in \mathcal{Y}$. Since epistemic uncertainty is difficult to represent with conventional probability distributions (Hüllermeier & Waegeman, 2021), such predictions fail to capture the epistemic part of (predictive) uncertainty. In order to account for both types of uncertainty, machine learning methods founded on more general theories of probability such as imprecise probabilities or credal sets (Walley, 1991; Augustin et al., 2014) have been considered (Corani et al., 2012).

Another popular approach in this regard is to let the learner map a query instance $\boldsymbol{x}_q$ to a second-order distribution, i.e., a distribution on distributions, effectively assigning a probability to each candidate probability distribution p. Such an approach is realized, for example, by classical Bayesian inference (Gelman et al., 2013) or by the *Evidential Deep Learning* (EDL) paradigm, which has recently become increasingly popular (Ulmer et al., 2023). In the EDL paradigm, one essentially learns a model (usually a deep neural network) by empirical risk minimization, whose

[1]Institute of Informatics, LMU Munich, Munich, Germany [2]Munich Center for Machine Learning, Munich, Germany [3]Precise Center, University of Pennsylvania, Philadelphia, USA. Correspondence to: Yusuf Sale <yusuf.sale@ifi.lmu.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

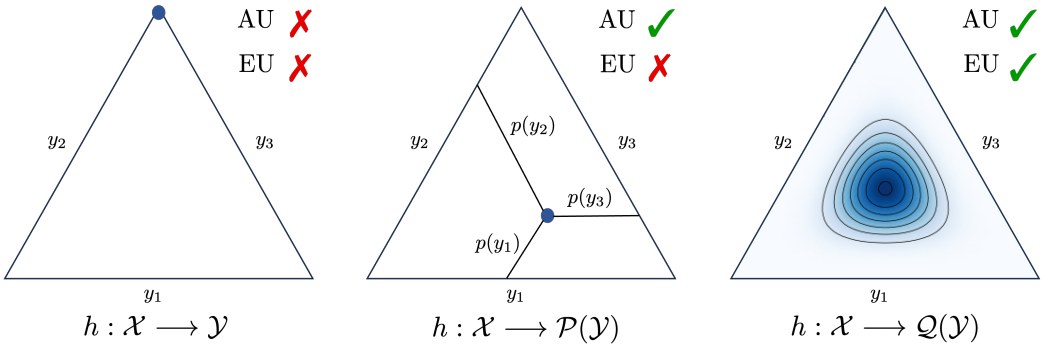

Figure 1: Uncertainty awareness in multi-class classification, illustrated on the probability simplex for $\mathcal{Y} = \{y_1, y_2, y_3\}$. From *left* to *right* increasing degrees of uncertainty awareness: Deterministic prediction (*no* uncertainty awareness), probabilistic prediction (AU, but *no* EU awareness), and second-order prediction (AU *and* EU awareness).

output for a query instance $\boldsymbol{x}_q$ are the parameters of a parameterized family of a second-order distribution.

So far, only the Dirichlet distribution has been used for classification, while the Normal-Inverse-Gamma distribution has been applied for univariate regression (Amini et al., 2020) and the Normal-Inverse-Wishart distribution for multivariate regression (Malinin et al., 2020; Meinert & Lavin, 2021). However, this approach is not without controversy, as it may lead to convergence issues of the empirical risk minimizer (Bengs et al., 2022; Meinert et al., 2023) and the predominantly used loss functions lack some desirable properties (Bengs et al., 2023).

Regardless of the specific design of the EDL approach, the concrete quantification of the total uncertainty (TU), aleatoric (AU), as well as epistemic (EU) associated with the second-order predictive distribution plays a central role in any case. For regression, essentially, the variances on the different levels of the second-order distribution are used for this purpose, while measures from information theory are applied for classification: Shannon entropy for TU, conditional entropy for AU, and mutual information for EU. Quite recently, Wimmer et al. (2023) criticized the latter for not complying with properties that one could naturally expect of uncertainty measures for second-order distributions. However, the authors do not provide an alternative for reasonable quantification either, which, of course, would be of great importance for practical ML purposes, especially in safety-critical applications.

Contributions. In this paper, we suggest an alternative way to obtain uncertainty measures in classification that overcome the drawbacks of the commonly used information-theory-based approach. To this end, we first propose a set of formal criteria that meaningful uncertainty measures for predictive uncertainty based on second-order distributions should obey. It extends the ones suggested by Wimmer et al. (2023). Moreover, we provide a general framework based on distances on the second-order probability level for developing uncertainty measures to account for these criteria. Using the Wasserstein distance, we instantiate this framework explicitly and prove that all criteria are met. Finally, we elaborate on these quantities when the second-order distribution is a Dirichlet distribution. All proofs of the theoretical statements are provided in the appendix.

2. Second-Order Uncertainty Quantification

In this section, we introduce the formal setting of supervised learning (throughout this paper we will exclusively deal with the case of classification) within which we establish further results. Let $(\mathcal{X}, \sigma(\mathcal{X}))$ and $(\mathcal{Y}, \sigma(\mathcal{Y}))$ be two measurable spaces. We will refer to $\mathcal{X}$ as *instance* (or input) space and to $\mathcal{Y}$ as *label* space, such that $|\,\mathcal{Y}\,| = K \in \mathsf{N}_{\geq 2}$. Further, we call the sequence $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ *training data*. For $i \in \{1, \ldots, n\}$, the pairs $(\boldsymbol{x}_i, y_i)$ are realizations of random variables (X_i, Y_i), which are independent and identically distributed (i.i.d.) according to some probability measure p on $(\mathcal{X} \times \mathcal{Y}, \sigma(\mathcal{X} \times \mathcal{Y}))$. Thus, each instance $\boldsymbol{x} \in \mathcal{X}$ is associated with a conditional distribution $p(\cdot \mid \boldsymbol{x})$ on $(\mathcal{Y}, \sigma(\mathcal{Y}))$, such that $p(y \mid \boldsymbol{x})$ is the probability to observe label $y \in \mathcal{Y}$ given $\boldsymbol{x} \in \mathcal{X}$.

To ease the notation, we will denote by $\mathcal{P}(\mathcal{Y})$ the set of all probability measures on the measurable space $(\mathcal{Y}, \sigma(\mathcal{Y}))$. Similarly, we write $\mathcal{Q}(\mathcal{Y})$ for the set of all probability measures on $(\mathcal{P}(\mathcal{Y}), \sigma(\mathcal{P}(\mathcal{Y}))$; we refer to $Q \in \mathcal{Q}(\mathcal{Y})$ as a *second-order distribution*.[1] While usually upper-case letters denote probability measures and lower-case letters their pdf/pmf, in this paper we use capital letters for second-order and lower-case letters for first-order distributions. The Dirac measure at $y \in \mathcal{Y}$ is denoted by $\delta_y \in \mathcal{P}(\mathcal{Y})$; likewise, $\delta_p \in \mathcal{Q}(\mathcal{Y})$ denotes the Dirac measure at $p \in \mathcal{P}(\mathcal{Y})$, where

[1] There is no general consensus on terminology, as terms such as level-2 or type-2 distributions are also encountered in the literature.

the underlying space of the Dirac measure should be clear from the context. Finally, $\mathrm{Unif}(\mathcal{Y})$ denotes the uniform distribution on $\mathcal{Y}$.

Given an instance $x \in \mathcal{X}$, let $Q \in \mathcal{Q}(\mathcal{Y})$ denote the learner's current probabilistic belief[2] about p, i.e., $Q(p)$ is the probability (density) of $p \in \mathcal{P}(\mathcal{Y})$. See Figure 1 for an illustration of the different degrees of uncertainty-aware predictions. As already mentioned in the introduction, there are two popular ways of obtaining such a second-order (predictive) distribution: by means of Bayesian inference or via Evidential Deep Learning. Throughout the rest of this paper, we assume such a second-order predictive distribution Q has been provided by a learner (though without being interested in how the prediction has been obtained). We raise the question of how to quantify the total amount of uncertainty (TU), as well as the aleatoric (AU) and epistemic (EU) uncertainties associated with Q.

2.1. Default Measures of Uncertainty

We begin by revisiting the arguably most common information-theoretic approach in machine learning for measuring predictive uncertainty in classification tasks. This approach exploits (Shannon) entropy and its link to mutual information and conditional entropy for specifying explicit quantities for the total (TU), aleatoric (AU), and epistemic (EU) uncertainties associated with a predictive second-order distribution $Q \in \mathcal{Q}(\mathcal{Y})$ (Houlsby et al., 2011; Gal, 2016; Depeweg et al., 2018; Mobiny et al., 2021).

The (Shannon) entropy (Shannon, 1948) of $p \in \mathcal{P}(\mathcal{Y})$ is defined as

$$\mathrm{H}(p) := \sum_{y \in \mathcal{Y}} p(y) \log_2 p(y). \tag{1}$$

We can analogously define the entropy of a (discrete) random variable $Y : \Omega \to \mathcal{Y}$ by

$$\mathrm{H}(Y) := \sum_{y \in \mathcal{Y}} p_Y(y) \log_2 p_Y(y), \tag{2}$$

where $p_Y \in \mathcal{P}(\mathcal{Y})$ is the corresponding push-forward measure on the measurable space $(\mathcal{Y}, 2^{\mathcal{Y}})$. The Shannon entropy has established itself as a standard measure of uncertainty due to its appealing theoretical properties and intuitive interpretation. Specifically, it measures the degree of uniformity of the distribution p_Y of a random variable Y, and corresponds to the log-loss of p_Y as a prediction of Y.

In the following, we assume $p_R \sim Q$; i.e., $p_R : \Omega' \to \mathcal{P}(\mathcal{Y})$ is a random first-order distribution distributed according to a second-order distribution Q and consequently taking values in the $(K-1)$-dimensional probability simplex. For $\omega' \in \Omega'$, we denote by $p = p_R(\omega')$ the realization of p_R, respectively.

[2] Although it would be more precise to let Q depend on x, for ease of notation we will simply write Q.

The core idea for obtaining uncertainty measures for a given second-order distribution Q is to consider the expectation of p_R with respect to Q given by

$$\bar{p} = \mathbb{E}_Q[p_R] = \int_{\mathcal{P}(\mathcal{Y})} p \, \mathrm{d}Q(p), \tag{3}$$

which yields a probability measure on $(\mathcal{Y}, \sigma(\mathcal{Y}))$; i.e., a first-order distribution.

With this, it seems natural to define the measure of total uncertainty as the entropy (1) of $\bar{p} \in \mathcal{P}(\mathcal{Y})$. More precisely, total uncertainty associated with a second-order distribution $Q \in \mathcal{Q}(\mathcal{Y})$ can be computed as

$$\mathrm{TU}(Q) = \mathrm{H}(\mathbb{E}_Q[p_R]). \tag{4}$$

In a similar fashion, one defines aleatoric uncertainty as conditional entropy

$$\mathrm{AU}(Q) = \mathbb{E}_Q[\mathrm{H}(Y \mid p_R)] = \int_{\mathcal{P}(\mathcal{Y})} \mathrm{H}(p) \, \mathrm{d}Q(p). \tag{5}$$

Further, the measure of epistemic uncertainty is in particular motivated by the well-known additive decomposition of entropy into conditional entropy and mutual information (Cover & Thomas, 1999, Equation (2.40)), i.e.,

$$\mathrm{H}(Y) = \mathrm{H}(Y \mid p_R) + \mathrm{I}(Y; p_R). \tag{6}$$

By rearranging (6) we get a measure of epistemic uncertainty

$$\mathrm{EU}(Q) = \mathrm{I}(Y; p_R) = \mathbb{E}_Q[D_{\mathrm{KL}}(p_R \,\|\, \bar{p})], \tag{7}$$

where $D_{\mathrm{KL}}(\cdot \,\|\, \cdot)$ denotes the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951).

Even though the individual measures, i.e., entropy, conditional entropy, and mutual information, have reasonable interpretations in terms of quantifying the respective uncertainty, which are particularly useful when applied to first-order predictive distributions, a different picture emerges for the above approach to second-order predictive distributions. Some issues regarding the quantification of the respective uncertainties have recently been intensively discussed by Wimmer et al. (2023), which we will take up and elaborate on in the following section. Essentially, the problem stems from TU in (4) and EU in (7) depending on the second-order predictive distribution Q only through their expectation $\bar{p}$ in (3).

2.2. Alternatives for the Default Measures

Recently, a variant of the above approach was proposed, which attempts to overcome the issues mentioned

(Schweighofer et al., 2023). For this purpose, the total uncertainty in (4) is rewritten as

$$\mathrm{TU}(Q) = E_Q[\mathrm{CE}(p_R; p)];$$

where $\mathrm{CE}(\cdot;\cdot)$ is the cross-entropy, i.e.,

$$\mathrm{CE}(p;q) := \sum_{y \in \mathcal{Y}} p(y) \log_2 q(y)$$

for $p, q \in \mathcal{P}(\mathcal{Y})$. Then, the alternative measure for total uncertainty suggested by the authors is

$$\mathrm{TU}(Q) = E_{Q;Q'}[\mathrm{CE}(p_R; p'_R)]; \tag{8}$$

where Q' is an i.i.d. copy of Q and p'_R Q'. Using again the decomposition in (6) and the resulting components as measures for aleatoric and epistemic uncertainty, one obtains the same aleatoric uncertainty measure as in (5), but the epistemic uncertainty measure changes to

$$\mathrm{EU}(Q) = E_{Q;Q'}[D_{KL}(p_R \,\|\, p'_R)]. \tag{9}$$

Thus, the proposed measures do not assume that the Bayesian model average predictive distribution is equivalent to the predictive distribution of the true data-generating process.

3. Novel Uncertainty Measures

3.1. Axiomatic Foundations

The criticism of the previous approach raised by Wimmer et al. (2023) is grounded in the postulation of criteria that measures of total, aleatoric, and epistemic uncertainties should naturally satisfy when used for quantifying predictive uncertainty associated with second-order distributions. This is similar to the literature on uncertainty quanti cation for other methods of representing uncertainty, such as belief functions or credal sets (Bronevich & Klir, 2008; Pal et al., 1993; Sale et al., 2023a). In the following, we build on – and extend – the criteria presented by Wimmer et al. (2023).

We begin by recalling some mathematical de nitions (see also Wimmer et al. (2023) and Sale et al. (2023b, p.4)).

De nition 3.1. Let $X \sim Q$; $X' \sim Q'$ be two random vectors, where we have that $Q, Q' \in \mathcal{Q}(\mathcal{Y})$. Denote by $\sigma(X)$ the σ-algebra generated by the random vector X. Then we call Q'

(i) a mean-preserving spread of Q, iff $X' \overset{d}{=} X + Z$, for some random vector Z with $E[Z \mid \sigma(X)] = 0$ almost surely (a.s.) and $\max_k \mathrm{Var}(Z_k) > 0$.

(ii) a spread-preserving location shift of Q, iff $X' \overset{d}{=} X + z$, where $z \neq 0$ is a constant.

(iii) a spread-preserving center-shift of Q, iff it is a spread-preserving location shift with $E[X'] = E[X] + (1-\lambda)(1/K, \ldots, 1/K)^\top$ for some $\lambda \in (0,1)$.

For (ii) and (iii) it should be ensured that the shifted probability distribution Q' remains valid within its support.

In the following, we let TU, AU, and EU denote, respectively, measures $\mathcal{Q}(\mathcal{Y}) \to \mathbb{R}_{\geq 0}$ of total, aleatoric, and epistemic uncertainties associated with a second-order uncertainty representation $Q \in \mathcal{Q}(\mathcal{Y})$. If $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are partitions of $\mathcal{Y}$ and $Q \in \mathcal{Q}(\mathcal{Y})$, then we denote by $Q_{|\mathcal{Y}_i}$ the marginalized distribution on $\mathcal{Y}_i$. In the same spirit, we de ne $\mathrm{TU}_{\mathcal{Y}_i}$.

A0 TU, AU, and EU are non-negative.

A1 $\mathrm{AU}(\delta_{\mathrm{Unif}(\mathcal{Y})}) \leq \mathrm{AU}(\delta_p) \leq \mathrm{AU}(\delta_y) = 0$ holds for any $y \in \mathcal{Y}$ and any $p \in \mathcal{P}(\mathcal{Y})$.

A2 $\mathrm{EU}(Q) \leq \mathrm{EU}(\delta_p) = 0$ holds for any $Q \in \mathcal{Q}(\mathcal{Y})$; and any $p \in \mathcal{P}(\mathcal{Y})$. Further, for any $Q \in \mathcal{Q}(\mathcal{Y})$ with $\mathrm{AU}(Q) = 0$ we have $\mathrm{EU}(Q') \leq \mathrm{EU}(Q)$, where Q' is such that $Q'(\delta_y) = \frac{1}{K}$ for all $y \in \mathcal{Y}$.

A3 $\mathrm{AU}(Q) \leq \mathrm{TU}(Q)$ and $\mathrm{EU}(Q) \leq \mathrm{TU}(Q)$ holds for any $Q \in \mathcal{Q}(\mathcal{Y})$.

A4 $\mathrm{TU}(Q)$ is maximal for Q being the continuous second-order uniform distribution.

A5 If Q' is a mean-preserving spread of Q, then $\mathrm{EU}(Q') \geq \mathrm{EU}(Q)$ (weak version) or $\mathrm{EU}(Q') > \mathrm{EU}(Q)$ (strict version).

A6 If Q' is a spread-preserving location shift of Q, then $\mathrm{EU}(Q') = \mathrm{EU}(Q)$.

A7 $\mathrm{TU}_{\mathcal{Y}}(Q) \geq \mathrm{TU}_{\mathcal{Y}_1}(Q_{|\mathcal{Y}_1}) + \mathrm{TU}_{\mathcal{Y}_2}(Q_{|\mathcal{Y}_2})$.

A8 $\mathrm{TU}_{\mathcal{Y}}(Q_{|\mathcal{Y}_1} \otimes Q_{|\mathcal{Y}_2}) = \mathrm{TU}_{\mathcal{Y}_1}(Q_{|\mathcal{Y}_1}) + \mathrm{TU}_{\mathcal{Y}_2}(Q_{|\mathcal{Y}_2})$, where $\otimes$ denotes the product measure.

Before discussing each criterion[3], we first start with a joint and more in-depth discussion of A1 and A2, since they play a central role in the discussion of most of the other criteria.

A1 and A2: Since we are interested in second-order distributions Q for the purpose of predictive uncertainty, it is natural to speak of a state of absence of epistemic uncertainty if Q corresponds to a point mass of second order. This is re ected by the lower bound in A2 and is also a viewpoint shared in the literature (Bengs et al., 2022; Wimmer et al., 2023). Moreover, there is agreement in the literature that (i) the uniform distribution of rst order, i.e., $\mathrm{Unif}(\mathcal{Y})$, represents the case of highest outcome uncertainty, (ii) a degenerated rst-order distribution, i.e., a Dirac measure on a point $y \in \mathcal{Y}$, represents the case of lowest outcome uncertainty, and (iii) rst-order distributions between these extreme cases correspond to an outcome uncertainty that lays somewhere "in-between". In the absence of epistemic uncertainty in the second-order distribution, this should be re ected by the measure of aleatoric uncertainty (A1).

If the uncertainty is only epistemic in nature, that is, if according to A1 only rst-order Dirac measures remain as possible candidates, then the epistemic uncertainty should

[3] A0 is a trivial property and therefore not discussed.

be maximal when the ambiguity around the Diracs is maximal. This happens when the second-order distribution Q is a discrete uniform distribution on the first-order Dirac measures on the elements of Y (A2). Note that this view differs from that of Wimmer et al. (2023), which demands maximum epistemic and total uncertainties for the continuous second-order uniform distribution. However, our criteria are consistent w.r.t. the maximal total uncertainty (A4).

A3: As discussed in detail by Wimmer et al. (2023, Section 4.4), the aleatoric and epistemic uncertainties of a second-order predictive distribution are closely intertwined. Since total uncertainty subsumes both types of uncertainty simultaneously, it should be always an upper bound for AU and EU, respectively.

A5 and A6: These properties are again inspired by Wimmer et al. (2023). If two second-order distributions have the same expectation but differ in their dispersion or spread, the distribution with higher dispersion should be assigned higher epistemic uncertainty (A5). Similarly, with equal dispersion, epistemic uncertainty should be the same in all cases. Thus, if Q and Q^0 only differ in their respective means, epistemic uncertainty should be the same in both cases (A6).

A7 and A8: These criteria are inspired by those underlying Shannon entropy. Speci cally, these properties aim to ensure that the total uncertainty of a second-order predictive distribution does not exceed the total uncertainties over all its possible marginalizations with respect to the label space Y. Thus, a subadditivity property should also hold here (A7), with equality achieved when the marginalizations are independent (A8).

As shown by Wimmer et al. (2023), the measures for total, aleatoric, and epistemic uncertainties in (4-7) fail to satisfy A5 and A6 when it comes to second-order distributions. For the alternative version of these measures suggested by Schweighofer et al. (2023) it is not shown whether these properties are ful lled or not. However, total uncertainty in (8) will not be maximal for Q being the continuous second-order distribution, but for Q^0 as in A2, so violating A4. In addition, it is apparent from the de nition that both TU and EU in (8) and (9) can go to in nity. Thus, the measures are not naturally restricted to an interpretable range.

3.2. Distance-based Measures

We now introduce a general framework for deriving suitable measures for total, aleatoric, and epistemic uncertainties associated with a second-order distribution $Q \in \mathcal{Q}(\mathcal{Y})$. The main constituents of the framework are (i) a (suitable) distance $d_2(\cdot\,;\cdot)$ on $\mathcal{Q}(\mathcal{Y})$ and (ii) speci c reference sets of second-order distributions representative for TU, AU or EU, respectively, each lacking one or both types of uncertainties. Roughly speaking, each uncertainty measure (i.e., TU, AU or EU) of Q is de ned as the minimal distance of Q to the corresponding reference set. This approach is inspired by the field of optimal transport (Villani, 2009; 2021) and guided by the following question: "How much do we need to move Q to arrive at the nearest second-order distribution of the respective reference set for TU, AU or EU?" While the distance function – according to which Q moves in the space $\mathcal{Q}(\mathcal{Y})$ – is intentionally kept flexible in our framework, the reference sets are fixed and should naturally lead to the ful llment of A0–A8, ideally for a broad class of distances.

Total uncertainty. For the total uncertainty we suggest to use all second-order Dirac measures on the set of first-order Dirac measures as the reference set. More speci cally, total uncertainty is de ned as

$$\mathrm{TU}(Q) := \min_{y \in \mathcal{Y}} d_2(Q, \delta_{\delta_y}). \tag{10}$$

This choice of the reference set is natural as each element in this reference set represents the case of an absolutely certain prediction/decision, i.e., there is neither aleatoric (first-order) nor epistemic (second-order) uncertainty present. Thus, the farther Q is from such an element, the farther one is from making a decision without any kind of uncertainty, which is re ected by (10).

Aleatoric uncertainty. The reference set for aleatoric uncertainty should be the set of all mixtures of second-order Dirac measures on first-order Dirac measures, i.e.,

$$\mathcal{M}_{\mathrm{m}} = \left\{ m \in \mathcal{Q}(\mathcal{Y}) : m = \sum_{y \in \mathcal{Y}} \alpha_y \delta_{\delta_y}; \sum_{y \in \mathcal{Y}} \alpha_y = 1 \right\}.$$

If we agree on A0–A8, each element in this set has no aleatoric uncertainty, so the assessment of a second-order distribution Q is solely in terms of its amount of aleatoric uncertainty. Accordingly, the measure of aleatoric uncertainty is de ned as

$$\mathrm{AU}(Q) := \min_{m \in \mathcal{M}_{\mathrm{m}}} d_2(Q, \delta_m). \tag{11}$$

Epistemic uncertainty. In the same spirit as (11), we want to assess Q solely in terms of its amount of epistemic uncertainty. Again, by agreeing on A0–A8, we naturally obtain as reference set the collection of all second-order Dirac measures on the probability simplex, since these have no epistemic uncertainty. If we denote the latter by $\mathcal{M}_{\mathrm{p}}$, we obtain for the measure of epistemic uncertainty

$$\mathrm{EU}(Q) := \min_{p \in \mathcal{M}_{\mathrm{p}}} d_2(Q, \delta_p). \tag{12}$$

It is worth noting that the entropy-based uncertainty measures in Section 2.1 can also be considered from the perspective of our distance-based framework. Indeed, the entropy

of a (discrete) distribution is related to the negative KL divergence (or KL distance) between the distribution and the uniform distribution (on the respective domain) (Cover & Thomas, 1999, Equation (2.107)). Thus, we could rewrite (4), (5), and (7) as

$$\begin{aligned} TU(Q) &= \log K - D_{KL}(E_Q[p_R] \,\|\, \mathrm{Unif}(Y)), \\ AU(Q) &= \log K - E_Q[D_{KL}(p_R \,\|\, \mathrm{Unif}(Y))], \\ EU(Q) &= -E_Q[D_{KL}(p_R \,\|\, p)]. \end{aligned} \tag{13}$$

With this representation, we see that the EU measure (7) has similarities to ours. More speci cally, it is obtained as a special case of (12) with $d_2(\cdot,\cdot)$ being the expected KL divergence (for which the minimum is obtained by $p = p$). Note, however, that the expected KL divergence is not a proper distance on $\mathcal{Q}(Y)$, wherefore (7) is not a special case of our framework in a strict sense. Moreover, the interpretation of the measures TU and AU is different from our measures (10) and (11), as both are measuring similarity (through the negated KL divergence) to the case of maximal uncertainty, namely the rst-order uniform distribution, instead of dissimilarity to a reference set of least uncertain distributions.

The alternative version (8–9) suggested by Schweighofer et al. (2023) does not have such an interpretation, except for the aleatoric uncertainty which remains the same. This is due to the lack of a reference set for TU and EU; so that both measures are more interpretable as a measure of the diversity of the second-order distribution. On a high level, the approach also follows the idea of including the entire characteristics of Q (rst- and second-order) in the respective uncertainty assessment, instead of narrowing down to the expected value $\bar{p}$ in (3) like the default case.

4. Wasserstein Instantiation

4.1. General Case

So far, we did not specify the distance $d_2 : \mathcal{Q}(Y) \times \mathcal{Q}(Y) \to \mathbb{R}_{\geq 0}$ on $\mathcal{Q}(Y)$. In the following, we will motivate one speci c choice, namely the Wasserstein distance (or Kantorovich–Rubinstein metric). For our discussion, we rst recall the concept of coupling, a term that is central to optimal transport theory (Villani, 2009, Chapter 1). Note that the de nition used in this paper is an adaptation of the standard one, as our focus is on second-order distributions.

De nition 4.1. We call the probability measure on $(\mathcal{P}(Y) \times \mathcal{P}(Y); \mathcal{B}(\mathcal{P}(Y) \times \mathcal{P}(Y)))$ coupling of $P, Q \in \mathcal{Q}(Y)$ iff for all $A, B \in \mathcal{B}(\mathcal{P}(Y))$ one has

$$\pi[A \times \mathcal{P}(Y)] = Q[A], \text{ and } \pi[\mathcal{P}(Y) \times B] = P[B].$$

Thus, π admits marginals P and Q.

Let $(\mathcal{P}(Y); d_1)$ be a metric space, where Y is de ned as before, and d_1 is a suitable metric on the space $\mathcal{P}(Y)$ (again, equipped with a suitable σ-algebra). Then, for $p \in [1, \infty]$ the (second-order) p-Wasserstein distance between two probability measures $P, Q \in \mathcal{Q}(Y)$ is de ned as

$$W_p(P, Q) = \inf_{\pi \in \Pi(P,Q)} \left(\int_{\mathcal{P}(Y) \times \mathcal{P}(Y)} d_1(p, q)^p \, \mathrm{d}\pi(p, q) \right)^{\frac{1}{p}} \tag{14}$$

where $\Pi(P, Q)$ denotes the set of all couplings between the probability measures P and Q (see De nition 4.1).

The choice of this metric for our purposes is quite natural based on its interpretation: The Wasserstein metric quanti es how much mass has to be moved around and how far in order to convert one distribution into another. This is perfectly in line with our view for the uncertainty measures in Section 3.2. In accordance with the literature, we will be exclusively concerned with the case $p = 1$ and omit in the following the subscript in $W_p(\cdot,\cdot)$. First, we show that $W(\cdot,\cdot)$ is indeed a well-de ned metric on $\mathcal{Q}(Y)$.

Lemma 4.2. The second-order Wasserstein distance $W : \mathcal{Q}(Y) \times \mathcal{Q}(Y) \to \mathbb{R}_{\geq 0}$ is a well-de ned metric on $\mathcal{Q}(Y)$.

Since in both (10) and (12) second-order Dirac measures are involved, we show now that the optimal coupling between a second-order distribution $Q \in \mathcal{Q}(Y)$ and a second-order Dirac measure δ_p, where $p \in \mathcal{P}(Y)$, is trivially given by the respective product measure. This simpli es corresponding computations.

Proposition 4.3. For any second-order Dirac measure $\delta_p \in \mathcal{Q}(Y)$, $p \in \mathcal{P}(Y)$, and any second-order distribution $Q \in \mathcal{Q}(Y)$, the optimal coupling between δ_p and Q is the product measure $\pi = Q \otimes \delta_p$.

This coupling is also frequently referred to as trivial coupling (Villani, 2009). Let us elaborate on the choice of the metric $d_1 : \mathcal{P}(Y) \times \mathcal{P}(Y) \to \mathbb{R}_{\geq 0}$ in (14). We will de ne this as the Wasserstein metric between two rst-order distributions induced by the trivial distance on the label space Y. Note that this is not xed by design, and without loss of generality other metrics on the label space (depending on the speci c problem at hand) can be considered. The trivial distance on Y is given for any $y, y' \in Y$ by

$$d_0(y, y') = \mathbb{1}_{\{y \neq y'\}}. \tag{15}$$

With the choice of the distance (15), for $p, q \in \mathcal{P}(Y)$ we obtain the following induced[4] rst-order distance d_1:

$$\begin{aligned} d_1(p, q) &= \inf_{\pi \in \Pi(p,q)} \int_{Y \times Y} d_0(y, y') \, \mathrm{d}\pi(y, y') \\ &= \inf_{\pi \in \Pi(p,q)} \sum_{\substack{y, y' \in Y \\ y \neq y'}} \pi(y, y') \end{aligned}$$

[4] Using the rst-order Wasserstein distance on $\mathcal{P}(Y) \times \mathcal{P}(Y)$.

$$= \inf_{\gamma \in \Gamma(p,q)} \left\{ 1 - \sum_{y \in Y} \gamma(y,y) \right\} : \quad (16)$$

Regarding the optimal coupling in (16), we can show the following.

Proposition 4.4. The coupling $\gamma \in \Gamma(p,q)$ minimizing the expression (16) is such that $\gamma(y,y) = \min\{p(y), q(y)\}$.

Proposition 4.4 yields

$$\begin{aligned} d_1(p,q) &= 1 - \sum_{y \in Y} \min\{p(y), q(y)\} \\ &= \frac{1}{2} \sum_{y \in Y} \max\{p(y), q(y)\} - \min\{p(y), q(y)\} \\ &= \frac{1}{2} \sum_{y \in Y} |p(y) - q(y)| =: \frac{1}{2} \|p - q\|_1 . \end{aligned}$$

In the context of usual probability measures, Proposition 4.4 is well-known in transportation theory, establishing a connection between the Wasserstein metric and the total variation distance. With this, the proposed uncertainty measures for $Q \in \mathcal{Q}(Y)$ in Section 3.2 simplify as follows.

Proposition 4.5. Using $d_1(\cdot,\cdot)$ as above, the measures of uncertainty in (10), (11), and (12) simplify to

$$TU(Q) = 1 - \max_{y \in Y} E_{Q_y}[p(y)], \quad (17)$$

$$AU(Q) = 1 - E_Q[\max_{y \in Y} p(y)], \quad (18)$$

$$EU(Q) = \frac{1}{2} \min_{q \in \mathcal{P}(Y)} E_{p \sim Q}[\|p - q\|_1]. \quad (19)$$

Here, Q_y denotes the marginal distribution associated with $Q \in \mathcal{Q}(Y)$ for some $y \in Y$.

The following proposition elaborates on the ranges of the proposed measures of uncertainty. Although the results appear natural, they yield interesting findings from an uncertainty quantification perspective.

Proposition 4.6. With the choice of $d_1(\cdot,\cdot)$ as distance on $\mathcal{P}(Y)$, we have that for all $Q \in \mathcal{Q}(Y)$ it holds that

i.) $TU(Q) \leq \frac{K-1}{K}$, where the upper bound is reached for $Q^0 \in \mathcal{Q}(Y)$ such that $E_{Q^0}[p] = \text{Unif}(Y)$.
ii.) $AU(Q) \leq \frac{K-1}{K}$, where the upper bound is reached for $Q^0 = \text{Unif}(\mathcal{P}(Y))$.
iii.) $EU(Q) \leq \frac{K-1}{K}$, where the upper bound is reached for any $Q^0 \in \mathcal{Q}(Y)$ such that $Q^0(\{p \in \mathcal{P}(Y) : p(y) = \frac{1}{K}\}) = 1$, for all $y \in Y$.

The property from Proposition 4.6 is desirable for two reasons. On the one hand, the value range grows with increasing complexity of the classification problem in terms of the number of labels K. This is similar to the entropy, see (13). On the other hand, the value ranges are "normalizing themselves" with increasing complexity. More precisely, for $K \to \infty$, the maximum of the uncertainty measures converges (with respect to the standard Euclidean metric on $\mathbb{R}$) to 1. Needless to say, the upper bounds of the value ranges can also be used to normalize the uncertainty measures a priori by multiplying them with $K/(K-1)$.

A direct consequence of Proposition 4.6 is that maximum epistemic uncertainty can be achieved only when there is no aleatoric uncertainty and vice versa.

Corollary 4.7. For any $Q \in \mathcal{Q}(Y)$, it holds that $EU(Q) = \frac{K}{K-1}$ if and only if $AU(Q) = 0$.

Finally, we show that the proposed uncertainty measures with the Wasserstein distance instantiations fulfill the criteria specified in Section 3.1.

Theorem 4.8. Uncertainty measures (10-12) with the Wasserstein distance instantiation satisfy Axioms A0 – A8.

4.2. Dirichlet Distribution

Owing to its key role as a conjugate prior for a categorical distribution, the Dirichlet distribution is arguably the most important family of parameterized (second-order) distributions employed in various areas of theoretical and applied research. In Bayesian inference and Evidential Deep Learning, the Dirichlet distribution has become the gold standard. Accordingly, in this section we focus on the computation of the proposed uncertainty measures with the Wasserstein distance initialization for the case of Dirichlet distributions. We start with a brief introduction to the Dirichlet distribution and identify without loss of generality each element in the label space Y with an integer, i.e., $Y = \{1, 2, \ldots, K\}$.

Let $\boldsymbol{\pi}$ denote a K-dimensional probability vector, and assume it is distributed according to a Dirichlet distribution, that is, $\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\alpha})$. The Dirichlet distribution $\text{Dir}(\boldsymbol{\alpha})$ is supported on the $(K-1)$-dimensional unit simplex, and it is parameterized by $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top$, a K-dimensional vector whose entries are such that $\alpha_j > 0$, for all $j \in Y$. Its probability density function (pdf) is given by

$$\frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{K} \pi_j^{\alpha_j - 1},$$

where $B(\boldsymbol{\alpha})$ denotes the multivariate Beta function. We can interpret the j-th entry α_j of $\boldsymbol{\alpha}$ as pseudo-counts: α_j represents the virtual observations that we have for label j. It captures the agent's (i.e., machine learning algorithm's) knowledge around label j that comes e.g. from previous or similar experiments. The expected value of $\text{Dir}(\boldsymbol{\alpha})$ is given by $E(\pi_j) = \bar{\pi}_j = \alpha_j / \alpha_0$, $j \in Y$, and it expresses the belief that j is the "true label". This is due to the fact that the marginals π_j of the Dirichlet distribution are distributed according to a Beta distribution with parameters α_j and $\alpha_0 - \alpha_j$, with $\alpha_0 = \sum_{i=1}^{K} \alpha_i$.

Dirichlet distributions are second-order distributions since their support is the $(K-1)$-dimensional simplex, i.e., $\mathcal{P}(Y)$. That is, they can be thought of as distributions over the

actual probability measures that generated the data.

In the following, we assume that our current probabilistic knowledge is given by $Q \sim \mathrm{Dir}(\alpha)$, so that the marginal distributions are Beta distributions, i.e. $Q_i \sim \mathrm{Beta}(\alpha_i, \alpha_0 - \alpha_i)$ with $\alpha_0 = \sum_{j=1}^{K} \alpha_j$ for each $i \in \mathcal{Y}$. Using the closed-form for the expectation of the marginals, we obtain

$$\mathrm{TU}(Q) = 1 - \max_{y \in \mathcal{Y}} \frac{\alpha_y}{\alpha_0} \tag{20}$$

for the total uncertainty in (17). Unfortunately, it is difficult to derive a closed form for the expression in (18). However, the expected value is easily approachable through Monte Carlo simulations.

Finally, for EU in (19), we are dealing with a constrained optimization problem, which, however, has appealing properties. Indeed, given a Dirichlet distribution, we seek to solve the following constraint optimization problem for (19):

$$\underset{q \in (0,1)^K}{\text{minimize}} \quad h(q) := \frac{1}{2} \sum_{i=1}^{K} \mathrm{E}_{p_i \sim Q_i}[\,|p_i - q_i|\,] \tag{21}$$

$$\text{subject to} \quad c(q) := \sum_{i=1}^{K} q_i - 1 = 0 \tag{22}$$

By further evaluation of the sum of expectations involved in (21) and using the method of Lagrange multipliers, we obtain the following result.

Proposition 4.9. The convex constrained optimization problem (21–22) has a unique solution.

Accordingly, EU for a given Dirichlet distribution can be computed quickly using a common optimization method.

Figure 2 displays, for $|\mathcal{Y}| = 3$, some exemplary Dirichlet distributions with different α parameters over a 2-simplex, along with their corresponding normalized[5] values for TU, AU, and EU in (10-12). We observe that the desired properties are captured as follows: First, AU and respectively EU is always smaller or equal to TU. Moreover, TU is maximal for the uniform distribution, as shown in Fig. 2a. EU also attains its maximum under other parameter conditions, but with varying aleatoric and epistemic contributions: This can occur with a high AU value, stemming from a high concentration around the first-order uniform distribution (see Fig. 2b, c). Alternatively, a high EU value can drive this, due to a strong similarity to the discrete uniform distribution on the first-order Dirac measures, namely the vertices (see Fig. 2f). Additionally, we observe that EU strictly increases for mean-preserving spreads (Fig. 2b, c). Fig. 2e depicts a Dirichlet distribution which is quite confident about one of the actual outcomes. This is reflected accordingly in low values of the uncertainty measures. These observations based on Dirichlet distributions align with our theoretical analysis of the proposed distance-based uncertainty measures.

[5] The values are normalized by multiplying them with $K = (K - 1)$; see discussion after Proposition 4.6.

Figure 2: Dirichlet distributions with different choices of α with normalized values for TU, AU, and EU.

Finally, we also consider the same exemplary second-order distributions Q used by Wimmer et al. (2023) to illustrate the issues of the entropy-based uncertainty measure (Figure 3 in Appendix B). In line with our theoretical results, our Wasserstein metric-induced measures behave as desired with respect to the axioms.

5. Conclusion

Recent criticisms have pointed to limitations in widely accepted uncertainty measures for second-order distributions, primarily due to certain unfavorable theoretical properties. Responding to this criticism, we presented a set of formal criteria that any uncertainty measure should fulfill. Additionally, we introduced a distance-based approach to obtain uncertainty measures for total, aleatoric, and epistemic uncertainties tailored towards obeying the criteria. On the basis of the Wasserstein metric, we demonstrated that this approach is fruitful and practical, especially for the often-used Dirichlet distributions.

The motivation for adopting a distance-based method for uncertainty quantification stems from the intuitive and geometric interpretation of uncertainty in probability spaces. Traditionally, uncertainty measures such as entropy-based

ones provide insight into the spread or unpredictability of a distribution. However, they do not always capture subtleties of second-order distributions effectively. We address this gap by leveraging a method that quanti es the distance between probabilistic beliefs (represented by second-order distributions). Our approach also closely aligns with a statistical viewpoint: In statistics it is quite natural to assess the discrepancy between probability distributions using distances. Such distances are well-established for (rst-order) probability distributions, with prominent examples including the Wasserstein distance and the Kullback-Leibler divergence, among others.

Our results open several venues for future work. First, it would be interesting to instantiate the proposed uncertainty measures with metrics on probability measures other than the Wasserstein metric and verify that the proposed criteria are met. In that respect, it would be interesting to work out general properties that a metric must satisfy in order for the criteria to be met. Although the focus of our work is on the theoretical aspects of uncertainty measures, a systematic experimental comparison in the context of evidential deep learning would be intriguing.

Impact Statement

This paper presents work whose goal is to advance the eld of Machine Learning. There are many potential societal consequences of our work, none which we feel must be speci cally highlighted here.

Acknowledgements

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

References

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. In Proc. NeurIPS, 33rd Advances in Neural Information Processing Systems, volume 33, pp. 14927–14937, 2020.

Augustin, T., Coolen, F. P., De Cooman, G., and Troffaes, M. C. Introduction to Imprecise Probabilities. John Wiley & Sons, 2014.

Bengs, V., Hüllermeier, E., and Waegeman, W. Pitfalls of epistemic uncertainty quanti cation through loss minimisation. In Proc. NeurIPS, 35th Advances in Neural Information Processing Systems, volume 35, pp. 29205–29216, 2022.

Bengs, V., Hüllermeier, E., and Waegeman, W. On second-order scoring rules for epistemic uncertainty quanti cation. In Proc. ICML, 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pp. 2078–2091. PMLR, 2023.

Bronevich, A. and Klir, G. J. Axioms for uncertainty measures on belief functions and credal sets. Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS), pp. 1–6. IEEE, 2008.

Corani, G., Antonucci, A., and Zaffalon, M. Bayesian networks with imprecise probabilities: Theory and application to classi cation. Data Mining: Foundations and Intelligent Paradigms: Volume 1: Clustering, Association and Classi cation, pp. 49–93, 2012.

Cover, T. and Thomas, J. A. Elements of Information Theory. John Wiley & Sons, 1999.

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., and Udluft, S. Decomposition of uncertainty in Bayesian deep learning for ef cient and risk-sensitive learning. In Proc. ICML, 35th International Conference on Machine Learning, pp. 1184–1193. PMLR, 2018.

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. Bayesian Data Analysis. CRC Press, 2013.

Gruber, C., Schenk, P. O., Schierholz, M., Kreuter, F., and Kauermann, G. Sources of uncertainty in machine learning – a statisticians' view. arXiv preprint arXiv:2305.16703, 2023.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classi cation and preference learning. arXiv preprint arXiv:1112.5745, 2011.

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

Kapoor, S., Maddox, W. J., Izmailov, P., and Wilson, A. G. On uncertainty, tempering, and data augmentation in Bayesian classi cation. In Proc. NeurIPS, 35th Advances in Neural Information Processing Systems, volume 35, pp. 18211–18225, 2022.

Kendall, A. and Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In Proc. NeurIPS, 30th Advances in Neural Information Processing Systems, volume 30, pp. 5574–5584, 2017.

Kullback, S. and Leibler, R. A. On information and sufficiency. The Annals of Mathematical Statistics, 22(1): 79–86, 1951.

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

Malinin, A., Chervontsev, S., Provilkov, I., and Gales, M. Regression prior networks. arXiv preprint arXiv:2006.11590, 2020.

Meinert, N. and Lavin, A. Multivariate deep evidential regression. arXiv preprint arXiv:2104.06135, 2021.

Meinert, N., Gawlikowski, J., and Lavin, A. The unreasonable effectiveness of deep evidential regression. In Proc. AAAI, Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pp. 9134–9142, 2023.

Mobiny, A., Yuan, P., Moulik, S. K., Garg, N., Wu, C. C., and Van Nguyen, H. DropConnect is effective in modeling uncertainty of Bayesian deep networks. Scientific Reports, 11:5458, 2021.

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

Sale, Y., Caprio, M., and Hüllermeier, E. Is the volume of a credal set a good measure for epistemic uncertainty? In Proc. UAI, 39th Conference on Uncertainty in Artificial Intelligence, pp. 1795–1804. PMLR, 2023a.

Sale, Y., Hofman, P., Wimmer, L., Hüllermeier, E., and Nagler, T. Second-order uncertainty quantification: Variance-based measures. arXiv preprint arXiv:2401.00276, 2023b.

Schweighofer, K., Aichberger, L., Ielanskyi, M., and Hochreiter, S. Introducing an improved information-theoretic measure of predictive uncertainty. In NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning, 2023.

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

Shannon, C. E. A mathematical theory of communication. The Bell System Technical Journal, 27(3):379–423, 1948.

Smith, L. and Gal, Y. Understanding measures of uncertainty for adversarial example detection. In Proc. UAI, 34th Conference on Uncertainty in Artificial Intelligence, pp. 560–570, 2018.

Ulmer, D., Hardmeier, C., and Frellsen, J. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. Transaction of Machine Learning Research, 2023.

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

Villani, C. Optimal Transport: Old and New, volume 338. Springer Science & Business Media, 2009.

Villani, C. Topics in Optimal Transportation, volume 58. American Mathematical Society, 2021.

Walley, P. Statistical Reasoning with Imprecise Probabilities. Chapman & Hall, 1991.

Wimmer, L., Sale, Y., Hofman, P., Bischl, B., and Hüllermeier, E. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In Proc. UAI, 39th Conference on Uncertainty in Artificial Intelligence, pp. 2282–2292. PMLR, 2023.

Yang, F., Wang, H.-Z., Mi, H., Lin, C.-D., and Cai, W.-W. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC Bioinformatics, 10(1):1–14, 2009.

A. Proofs

Proof of Lemma 4.2

Since $\mathcal{Y} = \{y_1, \ldots, y_K\}$ is nite, this means that a probability measure $p \in \mathcal{P}(\mathcal{Y})$ can be seen as a K-dimensional probability vector. In symbols, $p \leftrightarrow (p(y_1), \ldots, p(y_K))^\top$. The latter is a vector belonging to the $(K-1)$-unit simplex Δ_{K-1}. As a consequence, a second-order distribution $Q \in \mathcal{P}(\mathcal{P}(\mathcal{Y}))$ can be seen as a rst-order distribution on the $(K-1)$-unit simplex Δ_{K-1}. In symbols, $Q(\mathcal{P}) \leftrightarrow \mathcal{P}(\Delta_{K-1})$. This, together with the rst-order Wasserstein distance being a well-de ned metric on $\mathcal{P}(\mathcal{Y})$ (Villani, 2009; 2021), allows us to conclude that the second-order Wasserstein distance is itself a well-de ned metric. □

Proof of Proposition 4.3

Let $\mathbb{P}_p \in \mathcal{Q}(\mathcal{P})$ be a second-order Dirac measure, where $p \in \mathcal{P}(\mathcal{Y})$, and $Q \in \mathcal{Q}(\mathcal{P})$ a second-order distribution. We show that any coupling $\pi \in \Pi(\mathcal{P}(\mathcal{Y}) \sim Q, \mathcal{P}(\mathcal{Y}) \sim \mathbb{P}_p)$ has to be necessarily given by $\pi = Q \otimes \mathbb{P}_p$.

Thus, for any $A, B \in \sigma(\mathcal{P}(\mathcal{Y}))$ we show that

$$\pi(A, B) = \begin{cases} Q(A), & \text{if } p \in B, \\ 0, & \text{else} \end{cases}$$

Let $p \in B$, then we have $\pi(A \times B^c) = 0$. Hence, this implies $\pi(A \times B) = \pi(A \times \mathcal{P}(\mathcal{Y})) - \pi(A \times B^c) = Q(A)$. This shows the rst case. Assume $p \notin B$, then $\pi(A \times B) \leq \pi(\mathcal{P}(\mathcal{Y}) \times B) = \mathbb{P}_p(B) = 0$, showing the second case. □

Proof of Proposition 4.4

Let $p, q \in \mathcal{P}(\mathcal{Y})$. Note that $\gamma(y, y) = \min\{p(y), q(y)\}$ is trivially a coupling, hence $\gamma \in \Gamma(p, q)$. For the corresponding marginals we have $p(y) = \sum_{y' \in \mathcal{Y}} \gamma(y, y')$ and $q(y) = \sum_{y' \in \mathcal{Y}} \gamma(y', y)$, thus $\gamma(y, y) \leq \min\{p(y), q(y)\}$. This implies directly that $\min\{p(y), q(y)\}$ maximizes $\sum_{y \in \mathcal{Y}} \gamma(y, y)$, and therefore minimizes the distance $d_1(p, q) = \inf_{\gamma \in \Gamma(p,q)} \left[1 - \sum_{y \in \mathcal{Y}} \gamma(y, y)\right]$. □

Proof of Proposition 4.5

Let the distance on $\mathcal{P}(\mathcal{Y})$ be given by $d_1(p, q) = \frac{1}{2}\|p - q\|_1$, where $p, q \in \mathcal{P}(\mathcal{Y})$. Now, for any $Q \in \mathcal{Q}(\mathcal{P})$ the proposed uncertainty measures simplify as follows:

$$\mathrm{TU}(Q) = \min_{y \in \mathcal{Y}} W(Q, \mathbb{P}_y) = \min_{y \in \mathcal{Y}} \mathbb{E}_{p \sim Q}\left[\frac{1}{2}\|p - \mathbb{P}_y\|_1\right] \tag{23}$$

$$= \min_{y \in \mathcal{Y}} \mathbb{E}_{p \sim Q}[1 - p(y)] \tag{24}$$

$$= 1 - \max_{y \in \mathcal{Y}} \mathbb{E}_{p \sim Q}[p(y)]. \tag{25}$$

Note that for (24) we used the fact that $\frac{1}{2}\|p - \mathbb{P}_y\|_1 = \frac{1}{2}(1 - p(y) + \sum_{y' \neq y} p(y')) = 1 - p(y)$.

Further, we have

$$\mathrm{AU}(Q) = \min_{m \in \mathcal{M}} W(Q, \mathbb{P}_m) \tag{26}$$

$$= \min_{m \in \mathcal{M}} \inf_{\pi \in \Pi(Q, \mathbb{P}_m)} \int_{\mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y})} d_1(p, q)\, d\pi(p, q) \tag{27}$$

$$= \min_{m \in \mathcal{M}} \inf_{\pi \in \Pi(Q, \mathbb{P}_m)} \int_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}} 1 - p(y)\, d\pi(p, q) \tag{28}$$

$$= \int_{\mathcal{P}(\mathcal{Y})} 1 - \max_{y \in \mathcal{Y}} p(y)\, dQ(p) \tag{29}$$

$$= E_{p \sim Q}[1 - \max_y p(y)]; \tag{30}$$

where we used $y = \operatorname{argmax}_{y' \in Y} q(y')$. Equality in (29) is reached for the Dirac mixture with $_m \in {}_m$ with ${}_m(\ _y) = Q(y = \operatorname{argmax}_{y' \in Y} p(y'))$. Thus, we have the following:

$$\begin{aligned}
AU(Q) &= \inf_{ \in (Q;\ _m)} \int_{P(Y)} \Big(1 - \sum_{y \in Y} p(y)q(y)\Big)\, d\ (p;q) \\
&= \inf_{ \in (Q;\ _m)} \int_{P(Y)} \Big\{ 1 - \sum_{q \in \ _y} \Big[\sum_{y \in Y} p(y)q(y)\Big]\ (q|p) \Big\}\, dQ(p) \\
&= \inf_{ \in (Q;\ _m)} \int_{P(Y)} \Big\{ 1 - \sum_{y \in Y} p(y)\ (\ _y|p) \Big\}\, dQ(p) \\
&= \int_{P(Y)} \Big(1 - \max_{y \in Y} p(y)\Big)\, dQ(p).
\end{aligned}$$

The conditional probability measure $(\ _y|p) = 1_{\{y = \operatorname{argmax}_{y' \in Y} p(y')\}}$ is valid, since

$$_m(\ _y) = \int_{P(Y)} (\ _y|p)\, dQ(p) = Q(y = \operatorname{argmax}_{y' \in Y} p(y')).$$

Finally, we also have the following:

$$\begin{aligned}
EU(Q) &= \min_{p \in \ _p} W(Q;\ _p) \\
&= \min_{p \in \ _p} \int_{P(Y)} \int_{P(Y)} d_1(p;q)\, d\ _p(p)\, dQ(q) \\
&= \frac{1}{2} \min_{p \in P(Y)} \int_{P(Y)} \|q - p\|_1\, dQ(q) \\
&= \frac{1}{2} \min_{p \in P(Y)} E_{q \sim Q}[\|q - p\|_1].
\end{aligned}$$

This concludes the proof. □

Proof of Proposition 4.6

Let $Q \in \mathcal{Q}(Y)$, then we have:

i.) $TU(Q) = 1 - \max_{y \in Y} E_Q[p(y)] \geq 1 - \frac{1}{K} = \frac{K-1}{K}$, where the inequality is a direct consequence of $\sum_{y \in Y} p(y) = 1$ for any $p \in P(Y)$ which implies that $\max_{y \in Y} p(y) \geq 1/K$. It is clear that this upper bound is reached for $Q^0 \in \mathcal{Q}(Y)$ such that $E_{Q^0}[p] = \mathrm{Unif}(Y)$.

ii.) $AU(Q) = 1 - E_Q[\max_{y \in Y} p(y)] \geq 1 - \frac{1}{K} = \frac{K-1}{K}$. Clearly the upper bound is reached for $Q^0 \in \mathcal{Q}(Y)$ such that $Q^0 = \mathrm{Unif}(Y)$.

iii.) For $EU(Q)$ we obtain

$$EU(Q) = \frac{1}{2} \min_{p \in P(Y)} E_{q \sim Q}[\|q - p\|_1] \tag{31}$$

$$\frac{1}{2} E_{q \sim Q}[\|q - E_Q[p]\|_1] \tag{32}$$

$$\frac{1}{2} E_{q \sim \ _m}[\|q - E_Q[p]\|_1] \tag{33}$$

$$= \sum_{y \in Y} E_Q[p(y)](1 - E_Q[p(y)]) \tag{34}$$

$$= 1 - \sum_{y \in Y} E_Q[p(y)]^2 \tag{35}$$

$$1 - \frac{1}{K} = \frac{K - 1}{K}; \tag{36}$$

where (33) follows from the Dirac mixture $_m \in _m$ being a mean-preserving spread of Q. Inequality (36) is a consequence of the Cauchy-Schwarz inequality and the linearity of expectation. The upper bound is reached for Q^0, which is such that $Q^0(_y) = \frac{1}{K}$ for all $y \in Y$.

This concludes the proof. □

Proof of Corollary 4.7

Corollary 4.7 is an immediate consequence of Proposition 4.6. □

Proof of Theorem 4.8

We show that the Wasserstein distance instantiated measures (10) - (12) satisfy Axioms A0 - A8 discussed in Section 3.1.

A0: Since the proposed measures are distance-based this property holds trivially true.

A1: Let $p \in P(Y)$ and $y \in Y$, then we have

$$\begin{aligned} AU(_{Unif(Y)}) &= 1 - \max_{y \in Y} Unif(Y)(y) \\ &= \frac{K - 1}{K} \\ &\quad 1 - \max_{y \in Y} p(y) \\ &= AU(_p) \\ &\quad 0 \\ &= 1 - \max_{y^0 \in Y} {}_y(y^0) \\ &= AU(_y): \end{aligned}$$

The first inequality is a direct consequence of Proposition 4.6.

A2: For $p \in P(Y)$ and $Q \in Q(Y)$ we have immediately by de nition $EU(Q) - EU(_p) = 0$. The other inequality in this axiom follows directly from Proposition 4.6 iii.).

A3: Since $\bigcup_{y \in Y} \{_y\} \subseteq$, it follows $EU(Q) = \min_{p \in} W(Q; _p) - \min_{y \in Y} W(Q; _y) = TU(Q)$, for any $Q \in Q(Y)$. Similarly, since $\bigcup_{y \in Y} \{_y\} \subseteq _m$ we obtain $AU(Q) = \min_{p \in _m} W(Q; _m) - \min_{y \in Y} W(Q; _y) = TU(Q)$, for any $Q \in Q(Y)$.

A4: This follows from Proposition 4.6, since we have $E_Q[p] = Unif(Y)$ for Q being the continuous second-order uniform distribution.

A5: Further, let $Q^0 \in Q(Y)$ be a mean-preserving spread of $Q \in Q(Y)$, i.e., let $X - Q; X^0 - Q^0$ be two random variables such that $X^0 \overset{d}{=} X + Z$, for some random variable Z with $E[Z|X = x] = 0$, for all x in the support of X. Then, we have

$$EU(Q^0) = \frac{1}{2} \min_{p \in P(Y)} E[k(X + Z) - pk_1]$$

$$= \frac{1}{2} \min_{p \in P(Y)} \sum_{i=1}^{K} E(|X_i + Z_i - p_i|);$$

where $X_1, \ldots, X_K$ are the marginals of X and $Z_1, \ldots, Z_K$ the marginals of Z, respectively. From this, we further infer that for any $p = (p_1, \ldots, p_k) \in P(Y)$ and any x in the support of X that

$$\begin{aligned}
EU(Q') &= \frac{1}{2} \min_{p \in P(Y)} \sum_{i=1}^{K} E(|X_i - p_i| + Z_i(1_{X_i > p_i} - 1_{X_i < p_i})) \\
&= \frac{1}{2} \min_{p \in P(Y)} \sum_{i=1}^{K} E_{Q_i}(E_{Z_i}(|X_i - p_i| + Z_i(1_{X_i > p_i} - 1_{X_i < p_i}) \mid X_i = x_i)) \\
&= \frac{1}{2} \min_{p \in P(Y)} \sum_{i=1}^{K} E_{Q_i}(|X_i - p_i| + E(Z_i \mid X_i = x_i)(1_{X_i > p_i} - 1_{X_i < p_i})) \\
&= \frac{1}{2} \min_{p \in P(Y)} \sum_{i=1}^{K} E_{Q_i}(|X_i - p_i|) \\
&= EU(Q).
\end{aligned}$$

A6: Let Q be a second-order distribution with mean $p \in P(Y)$. Assume that $Q' \in Q(Y)$ is a spread-preserving shift of Q, shifted along the vector z with $\sum_{i=1}^{K} z_i = 0$. Then, we obtain

$$\begin{aligned}
EU(Q') &= \frac{1}{2} \min_{(p+z) \in P(Y)} E_{q \sim Q'}[\|q - (p+z)\|_1] \\
&= \frac{1}{2} \min_{(p+z) \in P(Y)} E_{q \sim Q}[\|(q+z) - (p+z)\|_1] \\
&= \frac{1}{2} \min_{p \in P(Y)} E_{q \sim Q}[\|q - p\|_1] \\
&= EU(Q).
\end{aligned}$$

A7: Let Y_1 and Y_2 be partitions of Y and $Q \in Q(Y)$. Further, denote by $Q_{|Y_i}$ the marginalized distribution on Y_i. First, we observe that $E_Q[p] = (E_{Q|Y_1}[p], E_{Q|Y_2}[p])^{\top}$. This observations yields

$$\begin{aligned}
TU(Q) &= 1 - \max_{y \in Y} E_{p \sim Q_y}[p(y)] \\
&= 1 - \max\left\{ \max_{y_1 \in Y_1} E_{Q|Y_1}[p(y_1)], \max_{y_2 \in Y_2} E_{Q|Y_2}[p(y_2)] \right\}.
\end{aligned}$$

From this, we immediately see that

$$TU(Q) \geq 1 - \max_{y_1 \in Y_1} E_{Q|Y_1}[p(y_1)] = TU_{Y_1}(Q_{|Y_1})$$

as well as

$$TU(Q) \geq 1 - \max_{y_2 \in Y_2} E_{Q|Y_2}[p(y_2)] = TU_{Y_2}(Q_{|Y_2}).$$

This implies $TU(Q) \geq TU_{Y_1}(Q_{|Y_1}) + TU_{Y_2}(Q_{|Y_2})$ as asserted.

A8: This property follows immediately, since TU only depends on the mean of the respective second-order distribution $Q \in Q(Y)$.

This concludes the proof. □

Proof of Proposition 4.9

Assume that $Q \in \mathcal{Q}(Y)$ is given by a Dirichlet distribution $\mathrm{Dir}(\ )$, so that the marginal distributions are Beta distributions, i.e., $Q_i \sim \mathrm{Beta}(\ _i;\ _0 - \ _i)$ with $\ _0 = \sum_{j=1}^{K} \ _j$ for each $i \in Y$. Hence, we seek to solve the following constraint optimization problem:

$$\underset{q \in (0;1)^K}{\text{minimize}} \quad h(q) := \frac{1}{2} \sum_{i=1}^{K} E_{p_i \sim Q_i}[|p_i - q_i|] \tag{37}$$

$$\text{subject to} \quad c(q) := \sum_{i=1}^{K} q_i - 1 = 0 \tag{38}$$

Further evaluation of the terms in the objective function yields:

$$\begin{aligned}
E_{Q_i}(|p_i - q_i|) &= \int_0^1 |p_i - q_i| f(p_i)\,dp_i \\
&= \int_{q_i}^1 (p_i - q_i) f(p_i)\,dp_i + \int_0^{q_i} (q_i - p_i) f(p_i)\,dp_i \\
&= \int_{q_i}^1 p_i f(p_i)\,dp_i - \int_0^{q_i} p_i f(p_i)\,dp_i + q_i \int_0^{q_i} f(p_i)\,dp_i - q_i \int_{p_i}^1 f(p_i)\,dp_i \\
&= \int_{q_i}^1 p_i f(p_i)\,dp_i - \int_0^{q_i} p_i f(p_i)\,dp_i + q_i F(q_i) - q_i(1 - F(q_i)) \\
&= \int_{q_i}^1 p_i f(p_i)\,dp_i - \int_0^{q_i} p_i f(p_i)\,dp_i + q_i(2F(q_i) - 1).
\end{aligned}$$

It is easy to evaluate the involved integrals, since we know

$$\begin{aligned}
\int_a^b p_i f(p_i)\,dp_i &= \frac{1}{B(\ ;\ )} \int_a^b p_i (1 - p_i)^{\ - 1}\,dp_i \\
&= \frac{B(\ + 1;\ )}{B(\ ;\ )} \int_a^b \frac{1}{B(\ + 1;\ )} p_i^{(\ + 1) - 1} (1 - p_i)^{\ - 1}\,dp_i \\
&= \frac{\ }{\ + } P_{\ + 1;\ }(a \le p_i \le b).
\end{aligned}$$

Together, this yields

$$\begin{aligned}
EU(Q) &= \frac{1}{2} E_Q(\|p - q\|_1) \\
&= \frac{1}{2} \sum_{i=1}^{K} E_{Q_i}(|p_i - q_i|) \\
&= \frac{1}{2} \sum_{i=1}^{K} \frac{\ _i}{\ _0} (1 - 2F_{\ _i + 1;\ _0}(q_i)) + q_i (2F_{\ _i;\ _0}(q_i) - 1)\ .
\end{aligned}$$

Further, the Lagrangian function is given by

$$L(q;\ ) = h(q) + \ c(q); \tag{39}$$

where $\in \mathbb{R}$. The extreme points of the Lagrangian (39) are the solutions of the equations

$$\nabla h(q) = \ \nabla c(q) \tag{40}$$

$$\sum_{i=1}^{K} q_i - 1 = 0. \tag{41}$$

The solution to (40) is given by

$$F_{i;\alpha_0}(q_i) = \frac{1}{2} \quad \forall i \in \{1, \ldots, K\}$$

$$\Rightarrow \quad q_i = F^{-1}_{i;\alpha_0}\left(\frac{1}{2}\right) \quad \forall i \in \{1, \ldots, K\}.$$

Since the cdf is strictly monotone increasing, there can be no more than one solution under the constraint (38). The quantile function is continuous and for $\to 0.5$ it becomes 0 and for $\to 0.5$ it goes towards 1, hence there must be exactly one $\in \mathbb{R}$ such that both equations (37) and (38) are ful lled. Clearly, the bordered Hessian is given by:

$$H(q;) = \begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & f_{1;\alpha_0}(q_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & f_{K;\alpha_0}(q_K) \end{bmatrix} \tag{42}$$

Consequently, for the determinant of the bordered Hessian we get

$$\det(H(q;)) = \frac{\prod_i f_{i;\alpha_0}(q_i)}{\sum_i f_{i;\alpha_0}(q_i)}.$$

Since there is only one solution and the determinant of the bordered Hessian is negative, we conclude that the minimum is unique. Thus, the constraint optimization problem has a unique solution. □

Computational aspects

Here, we brie y discuss the computational aspects of the proposed uncertainty measures. Notably, the computational challenges typically associated with traditional optimal transport problems do not apply in this context because:

- TU, Eq. (17): All it takes here is to calculate the mean of the Dirichlet Distribution, which has an analytical solution.
- AU, Eq. (18): The expected value of the maximum of a Dirichlet distribution does not have an analytical solution. However, it can be easily estimated using Monte Carlo samples, which is computationally ef cient and straightforward.
- EU, Eq. (19): To solve Eqs. (40) and (41) of the optimization problem, must lie between $1/2$ and $1/2$. By solving q_i for all $i \in \{1, \ldots, K\}$ in terms of , we observe that (41) is strictly monotonically decreasing in . Therefore, starting with $= 0$, we can iteratively evaluate the left side of Eq. (41) to determine whether it is greater or less than 1, thereby halving the search space for . In the worst case, it will take only 30 iterations to get within 10^{-10} of its true value. We can terminate the procedure when (41) is suf ciently close to 0.

In conclusion, although we cannot speak for other instantiations of our proposed uncertainty measures, the speci c instantiation we present does not exhibit any computational limitations.