An Axiomatic Assessment of Entropy- and Variance-based Uncertainty Quantification in Regression

Christopher Bülte[1,2] **Yusuf Sale**[1,2] **Timo Löhr**[1,2] **Paul Hofman**[1,2] **Gitta Kutyniok**[1,2,3,4] **Eyke Hüllermeier**[1,2,5]

[1]LMU Munich
[2]Munich Center for Machine Learning (MCML)
[3]DLR-German Aerospace Center
[4]University of Tromsø
[5]German Research Center for Artificial Intelligence (DFKI, DSA)

Abstract

Uncertainty quantification is crucial in machine learning, yet most (axiomatic) studies of uncertainty measures focus on classification, leaving a gap in regression settings with limited formal justification and evaluations. In this work, we provide a formal way of representing uncertainty in continuous space, using a general parametric formulation, allowing for tractable analysis and evaluation of uncertainty measures. Within this framework, we propose a set of axioms that enable rigorous assessment of total, aleatoric, and epistemic uncertainty measures. Together, this allows for a theoretical examination of uncertainty measures and their corresponding properties. As a specific example, we compare the widely used entropy- and variance-based measures with respect to established predictive models and analyze their limitations and challenges in uncertainty quantification. Our work provides a principled way to understand and develop uncertainty measures in supervised regression, offering theoretical insights and practical guidelines for reliable uncertainty assessment.

1 INTRODUCTION

Uncertainty quantification (UQ) has emerged as an active field in contemporary machine learning research and practice—particularly in supervised regression settings. Mostly due to the application of AI systems to safety-critical tasks, research has focused on quantifying uncertainty in domains such as weather forecasting [Rasp and Lerch, 2018, Bülte et al., 2026], energy systems [Miele et al., 2023], autonomous driving [Michelmore et al., 2018] or healthcare [Edupuganti et al., 2020, Löhr et al., 2024]. Having a thorough understanding of uncertainty is key in these domains.

In the aforementioned supervised learning tasks, the focus is usually on *predictive uncertainty*, which describes the uncertainty of the outcome $\boldsymbol{y} \in \mathcal{Y}$ given underlying data $\boldsymbol{x} \in \mathcal{X}$. There has been a particular emphasis on distinguishing *aleatoric* and *epistemic* uncertainty, as well as the decomposition of *total* uncertainty into its aleatoric and epistemic components [Hüllermeier and Waegeman, 2021]. Aleatoric uncertainty captures the inherent randomness in the data-generating process. As it represents variability that cannot be reduced even with more data, it is often referred to as *irreducible* uncertainty. Sources of aleatoric uncertainty include inherent randomness in physical systems or measurement errors. In contrast, epistemic uncertainty arises from a lack of knowledge or understanding about the underlying data-generating process, which can be reduced by acquiring additional data or improving the model itself. Therefore, epistemic uncertainty is also referred to as *reducible* uncertainty.

When analyzing uncertainty, it is of particular importance to distinguish between the *representation* and the *quantification* of uncertainty. The latter refers to the task of measuring (types of) uncertainty conveyed by a given representation. While aleatoric uncertainty can effectively be represented by probability distributions, representing epistemic uncertainty poses significant challenges. In fact, there is a general consensus that epistemic uncertainty *cannot* be represented by means of classical probability theory. Consequently, higher-order formalisms are adopted, such as second-order distributions (distributions of distributions) or credal sets (sets of probability distributions) [Levi, 1980, Walley, 1991].

Various approaches have been proposed in the literature, addressing the inherent difficulty of *evaluating* the quality of uncertainty measures [Hüllermeier and Waegeman, 2021, Ståhl et al., 2020]. One such prominent approach is the *axiomatic* evaluation of uncertainty quantification, which assesses the suitability of measures for total, aleatoric, and epistemic uncertainty based on a predefined set of axioms. However, evaluations primarily focus on the supervised classification setting [Hüllermeier et al., 2022, Wimmer et al., 2023, Sale et al., 2024b,a, 2023], although many ma-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:880–899.

chine learning problems are actually within the regression framework, from classical regression to inverse problems. In contrast to classification, the underlying outcome space of supervised regression is generally unbounded, preventing a direct transfer of results. Although some research has analyzed uncertainty quantification for the regression setting [Berry and Meger, 2025, Malinin et al., 2020, Amini et al., 2020], there is a noticeable lack of *formal* justification for many proposed approaches, as well as an absence of evaluations grounded in an axiomatic framework.

Contributions. In this paper, we propose a formal uncertainty representation framework that allows for tractable analysis, along with a set of formalized axioms that provide a way to evaluate uncertainty measures theoretically. To represent uncertainty, we propose a general parametric family that admits a natural characterization of epistemic uncertainty and contains many widely used predictive machine learning methods. We then provide a set of six axioms that uncertainty measures should satisfy, specifically designed and adapted to accommodate the continuous regression setting. Finally, we analyze the most commonly used uncertainty measures, based on entropy or variance, with respect to the proposed axioms, highlighting differences and identifying pitfalls. Our analysis reveals that both entropy- and variance-based measures, despite their widespread adoption, exhibit fundamental shortcomings: each violates specific axioms in ways that directly translate to undesirable and potentially misleading behavior regarding uncertainty quantification. Our framework provides a principled foundation for both evaluating existing uncertainty measures and guiding the design of new ones, directly addressing a gap that has left practitioners without formal criteria for this choice.

2 REPRESENTING AND QUANTIFYING UNCERTAINTY IN REGRESSION

In the following, we denote by $\mathcal{X} \subseteq \mathbb{R}^k$, $\mathcal{Y} \subseteq \mathbb{R}^d$, and $\Theta \subseteq \mathbb{R}^p$ the (real-valued) feature, target, and parameter space, respectively. Furthermore, define the probability measures $\mathcal{P}(\mathcal{X}), \mathcal{P}(\mathcal{Y}), \mathcal{P}(\Theta)$ over the corresponding measurable spaces. Assume we have training data $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, where for $i \in \{1, \ldots, n\}$, each pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is a realization of the i.i.d. random variables $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$, distributed according to some probability measure P. Consequently, a given feature vector $\boldsymbol{x} \in \mathcal{X}$ induces a conditional probability distribution $P(\cdot \mid \boldsymbol{x}) \in \mathcal{P}(\mathcal{Y})$.

2.1 UNCERTAINTY REPRESENTATION

In general, aleatoric uncertainty can naturally be represented by the induced conditional probability distribution $P(\cdot \mid \boldsymbol{x})$. Epistemic uncertainty, on the other hand, requires a higher-order representation, for which we adopt a second-order distribution [Hüllermeier and Waegeman, 2021]. In the classification setting, this definition is straightforward, as one is working with predictions in the $K-1$-dimensional probability simplex Δ_K, on top of which the second-order distribution Q is placed, i.e., $Q \in \mathcal{P}(\Delta_K)$.

However, in the regression setting, as the prediction space $\mathcal{Y}$ is continuous and unbounded, the corresponding definitions turn out to be more complex. While one can formally define the second-order distribution as an abstract distribution over probability measures, i.e., $Q \in \mathcal{P}(\mathcal{P}(\mathcal{Y}))$, this definition cannot be practically used to evaluate uncertainty measures. To make uncertainty quantification and uncertainty evaluation feasible, we propose to use parametric first-order distributions and place a second-order distribution over the corresponding parameter space Θ, allowing for operationalization of the second-order distribution.

In particular, we consider a parametric family of predictive distributions

$$\{P(\cdot \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\} \subseteq \mathcal{P}(\mathcal{Y}), \tag{1}$$

where the distributional family itself is fixed and all uncertainty resides in the parameter vector $\boldsymbol{\theta} \in \Theta$. The dependence on the covariates is then captured by a covariate-dependent second-order distribution

$$Q(\boldsymbol{x}) \in \mathcal{P}(\Theta), \tag{2}$$

representing the (epistemic) state of knowledge about the first-order parameters at a given instance $\boldsymbol{x} \in \mathcal{X}$.

In this setting, if $Q(\boldsymbol{x}) = \delta_{\boldsymbol{\theta}(\boldsymbol{x})}$, we recover the parametric predictive distribution $P(\cdot \mid \boldsymbol{x}) = P(\cdot \mid \boldsymbol{\theta}(\boldsymbol{x}))$ with covariate-dependent parameter vector $\boldsymbol{\theta}(\boldsymbol{x}) \in \Theta$. Here, the dependence on the covariates is typically modeled via $\theta^k(\boldsymbol{x}) = h_k(\eta_k(\boldsymbol{x}))$, $k = 1, \ldots, p$, where $\eta_k(\boldsymbol{x})$ are (parameter-specific) regression predictors and the response function $h_k : \mathbb{R} \to \mathbb{R}$ maps the predictors to the appropriate parameter space, for example with the exponential function for nonnegative parameters. This case includes generalized linear and generalized additive models [Kneib et al., 2023], but also neural network-based probabilistic methods, when $\eta_k(\boldsymbol{x}) = \mathcal{F}_\phi^k(\boldsymbol{x})$ is parameterized with a neural network $\mathcal{F}_\phi^k$ with weights (and biases) ϕ. In this setting, each distributional parameter θ^k is a (possibly nonlinear) function of the covariates $\boldsymbol{x}$, allowing for huge flexibility regarding the types and complexity of the target distribution, as the parameters can represent any distributional aspects, such as location, scale, or shape [Kneib et al., 2023]. A non-degenerate $Q(\boldsymbol{x})$, in contrast, additionally expresses epistemic uncertainty about the first-order parameters themselves, including conjugate prior settings such as in deep evidential regression [Amini et al., 2020].

More formally, let $\varphi : \Theta \to \mathcal{P}(\mathcal{Y}) : \varphi(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}} \coloneqq P(\cdot \mid \boldsymbol{\theta})$ be a measurable map, which essentially transforms the parameter vector $\boldsymbol{\theta}$ into a (specific) probability measure

$P_{\boldsymbol{\theta}}$. Given some $\boldsymbol{\theta} \in \Theta$, aleatoric uncertainty is naturally available via $P_{\boldsymbol{\theta}}$. Epistemic uncertainty, by contrast, requires a higher-order formalism, which we represent through the probability distribution $Q(\boldsymbol{x}) \in \mathcal{P}(\Theta)$ over Θ. By defining $Q^{\varphi}(\boldsymbol{x}) \coloneqq \varphi_{\#}Q(\boldsymbol{x}) \in \mathcal{P}(\mathcal{P}(\mathcal{Y}))$ as the push-forward of $Q(\boldsymbol{x})$ under φ (the law of the random first-order distributions $P_{\boldsymbol{\vartheta}}$, when $\boldsymbol{\vartheta} \sim Q(\boldsymbol{x})$), we then implicitly obtain the (second-order) probability distribution over the space $\mathcal{P}(\mathcal{Y})$.

Finally, we can define the predictive (mixture) distribution, as

$$\overline{P}(\boldsymbol{x}) \;=\; \int_{\Theta} P_{\boldsymbol{\theta}}\, Q(\boldsymbol{x})(d\boldsymbol{\theta}) \;\in\; \mathcal{P}(\mathcal{Y}). \tag{3}$$

Here, we assume that all first-order measures are dominated by the Lebesgue measure μ, i.e., for all $\boldsymbol{\theta} \in \Theta$ we have $P_{\boldsymbol{\theta}} \ll \mu$ (and hence $\overline{P}(\boldsymbol{x}) \ll \mu$), and obtain the corresponding densities, denoted as $p_{\boldsymbol{\theta}}$ and $\overline{p}(\cdot \mid \boldsymbol{x})$, respectively. As all subsequent definitions, axioms, and results are understood to hold pointwise for a given instance $\boldsymbol{x} \in \mathcal{X}$, we henceforth drop the dependence on the covariates $\boldsymbol{x}$ for ease of notation.

Note that in this setting, we treat the distributional family as fixed and assume that all uncertainty resides in the parameters $\boldsymbol{\theta}$. While this specification might initially seem restrictive, deriving theoretical statements about uncertainty measures requires a trade-off between tractability and generalization. Still, our setting allows for great flexibility regarding the distributional family (e.g. multivariate-, nonnegative- or skewed distributions) and has been utilized for regression tasks in various applications, such as geodesy [Kiani Shahvandi et al., 2025], forecasting of weather extremes [Schulz and Lerch, 2022, Friederichs and Thorarinsdottir, 2012], photometric redshift estimation [D'Isanto and Polsterer, 2018] or photovoltaic power forecasting [Mayer et al., 2026, Gneiting et al., 2023]. Furthermore, our setup also includes many commonly used machine learning approaches for uncertainty representation, including (deep) ensembles, distributional regression, or deep evidential regression.

2.2 UNCERTAINTY QUANTIFICATION

In this section, we recall the two predominant sets of measures used in the literature to *quantify* total (TU), aleatoric (AU), and epistemic uncertainty (EU), namely entropy-based and variance-based measures. In the following, consider a first-order distribution $P_{\boldsymbol{\vartheta}} \in \mathcal{P}(\mathcal{Y})$ with a random parameter vector $\boldsymbol{\vartheta} \sim Q$ and assume that Q satisfies the necessary integrability conditions ensuring the existence of all uncertainty measures defined below.

Entropy-based measures. The most widely adopted measures for quantifying total, aleatoric, and epistemic uncertainty associated with a second-order distribution in the classification setting rely on Shannon entropy and its decomposition into conditional entropy and mutual information

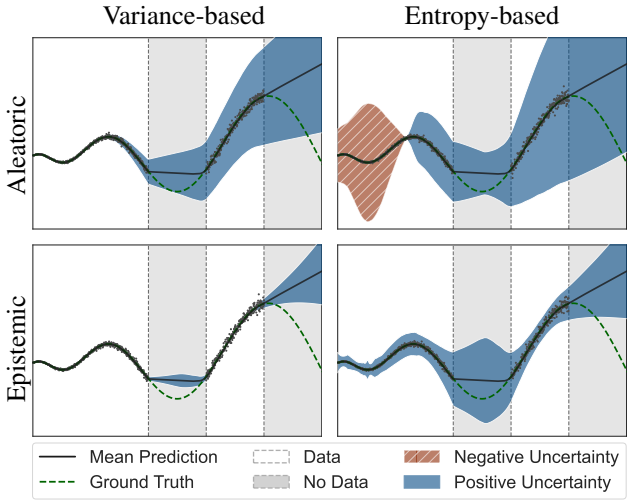

Figure 1: Comparison of variance and entropy-based uncertainty measures for deep ensembles on a synthetic example. Note in particular the occurrence of negative uncertainty values for the entropy-based measure.

[Houlsby et al., 2011, Kendall and Gal, 2017, Charpentier et al., 2022]. This decomposition can be directly transferred to the regression setting by considering differential entropy [Malinin et al., 2020, Postels et al., 2021], which however has structurally different properties.

Let $\boldsymbol{Y}$ denote the outcome with marginal density $p(\boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y} \mid \boldsymbol{\theta})\, dQ(\boldsymbol{\theta})$. Then, its differential entropy is given by $H(\boldsymbol{Y}) = -\int_{\mathcal{Y}} p(\boldsymbol{y}) \log p(\boldsymbol{y})\, d\boldsymbol{y}$. Following Houlsby et al. [2011], we can utilize conditional differential entropy to establish the following well-known decomposition:

$$\underbrace{H(\boldsymbol{Y})}_{\text{Entropy}} = \underbrace{H(\boldsymbol{Y} \mid P_{\boldsymbol{\vartheta}})}_{\text{conditional entropy}} + \underbrace{I(\boldsymbol{Y}, P_{\boldsymbol{\vartheta}})}_{\text{mutual information}}. \tag{4}$$

Then, the uncertainty measures can be defined as

$$\mathrm{TU}(Q) = H(\overline{P}), \tag{5}$$
$$\mathrm{AU}(Q) = \mathbb{E}_Q[H(\boldsymbol{Y} \mid P_{\boldsymbol{\vartheta}})], \tag{6}$$
$$\mathrm{EU}(Q) = \mathbb{E}_Q[D_{\mathrm{KL}}(P_{\boldsymbol{\vartheta}} \| \overline{P})], \tag{7}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ denotes the Kullback-Leibler (KL) divergence and $\overline{P}$ is the previously introduced predictive mixture distribution.

Intuitively, AU can be assessed by fixing a first-order distribution $P_{\boldsymbol{\theta}}$, thus removing all epistemic uncertainty. However, as $\boldsymbol{\theta}$ is not exactly known, we take the expectation over the second-order distribution Q. Epistemic uncertainty, defined as the mutual information, quantifies the potential reduction in uncertainty of $\boldsymbol{Y}$ through observing $P_{\boldsymbol{\vartheta}}$ [Ash, 1990]. The entropy-based measures require absolutely continuous distributions with respect to the Lebesgue measure, i.e., $P_{\boldsymbol{\theta}} \ll \mu$, $\forall \boldsymbol{\theta} \in \Theta$, which we assume throughout.

Variance-based measures. In classification, another set of commonly used uncertainty measures is based on the law

of total variance [Sale et al., 2024a]. Even more naturally, this decomposition extends to the real-valued case and is widely applied in regression settings [Amini et al., 2020, Valdenegro-Toro and Mori, 2022, Laves et al., 2021].

To accommodate for the (possibly) multivariate setting, we use the scalar generalization of the variance via the trace of the covariance matrix, i.e., $\mathbb{V}(\boldsymbol{Y}) \coloneqq \mathbb{E}\left[\|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\|_2^2\right] = \mathrm{tr}(\mathrm{Cov}(\boldsymbol{Y}))$. From this, we can leverage the law of total variance to decompose total uncertainty into its aleatoric and epistemic components:

$$\underbrace{\mathbb{V}(\boldsymbol{Y})}_{\text{total}} = \underbrace{\mathbb{E}_Q[\mathbb{V}(\boldsymbol{Y} \mid \boldsymbol{\vartheta})]}_{\text{aleatoric}} + \underbrace{\mathbb{V}_Q(\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{\vartheta}])}_{\text{epistemic}}. \tag{8}$$

In the multivariate case, the uncertainty measures correspond to the sum of all marginal uncertainties; in the univariate case, this corresponds to the commonly-used variance-based decomposition [Amini et al., 2020].

Similar to the entropy-based measures, AU is the variance in the outcome variable when removing epistemic uncertainty by fixing a (known) parameter $\boldsymbol{\vartheta}$. However, as $\boldsymbol{\vartheta}$ is unknown, we take the expectation with respect to Q. Epistemic uncertainty, on the other hand, is given by the variance of the mean prediction, e.g., the uncertainty induced by the second-order distribution Q. The variance-based measures require finite second moments, i.e., $\mathbb{E}[\|\boldsymbol{Y}\|_2^2] < \infty$ to exist, which, in contrast to the classification setting, is not necessarily fulfilled, for example, for heavy-tailed distributions.

While both types of measures provide meaningful estimates of the different kinds of uncertainty and fulfill an additive decomposition, i.e., $\mathrm{TU} = \mathrm{AU} + \mathrm{EU}$, they differ significantly in their definition, assumptions, and properties, as visualized in Figure 1.

2.3 UNCERTAINTY EVALUATION

Although many uncertainty measures have been proposed, it remains an open question how they can be properly justified or evaluated. Generally speaking, such a justification could be either of a theoretical or empirical nature. The question of how to empirically evaluate (aleatoric and epistemic) uncertainty in practice is highly nontrivial, in particular since there is typically no directly observable ground truth for uncertainty. Unlike standard predictive tasks, where predictions can be compared against actual observed outcomes, no such ground truth is available when evaluating uncertainty. As a consequence, uncertainty evaluation typically relies on proxies or indirect validation strategies. In particular, the focus is often on downstream tasks such as selective prediction, active learning, and out-of-distribution detection. While these evaluation protocols have become standard, they are not without criticism [Li et al., 2025].

Given the inherent difficulties of empirical evaluation, a complementary evaluation approach is of a theoretical nature, namely, assessing whether a given uncertainty measure satisfies qualitative properties one would expect from its intended semantics. The importance of this (axiomatic) perspective is highlighted by recent results indicating that prominent uncertainty measures do not always satisfy even basic desiderata [Wimmer et al., 2023]. In this spirit, our paper is centered around an axiomatic evaluation approach.

3 FORMAL PROPERTIES FOR UNCERTAINTY MEASURES

Evaluating measures based on a set of axioms has been standard practice in the (uncertainty) literature [Pal et al., 1993, Bronevich and Klir, 2008] and has been adopted by the machine learning community [Hüllermeier et al., 2022].

In the context of supervised classification, recent works have proposed axiomatic frameworks for second-order uncertainty [Wimmer et al., 2023, Sale et al., 2024a] with subsequent contributions extending these foundations [Sale et al., 2023] or proposing novel types of uncertainty measures [Kotelevskii et al., 2025, Hofman et al., 2024]. Moving beyond classification, we propose a set of axioms specifically tailored to the regression setting and the continuous structure of the target space $\mathcal{Y}$. The transfer of existing axioms (A0–A3) is inherently asymmetric: classification admits a natural notion of maximal epistemic uncertainty through the uniform distribution, which has no counterpart in regression, requiring a mathematical reformulation. Our proposed novel axioms regarding translation invariance (A4) and scale-order equivariance (A5) further exploit the metric and algebraic structure of $\mathbb{R}^d$, which is absent on a discrete label space. Before introducing this set of axioms, we first establish the relevant notation and preliminaries.

Definition 3.1 (Translation-invariance)**.** *Assume that the family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is closed under translation, i.e., for every $\boldsymbol{c} \in \mathbb{R}^d$ there exists a parametrization $\tau_{\boldsymbol{c}} : \Theta \to \Theta,\ \boldsymbol{\theta} \mapsto \boldsymbol{\theta}_{\boldsymbol{c}}$ such that $p_{\tau_{\boldsymbol{c}}(\boldsymbol{\theta})}(\boldsymbol{y}) = p_{\boldsymbol{\theta}}(\boldsymbol{y} - \boldsymbol{c}),\ \forall \boldsymbol{y} \in \mathcal{Y}$. Define $Q_{\boldsymbol{c}}$ as the pushforward of $Q \in \mathcal{P}(\Theta)$ under this mapping, i.e., $Q_{\boldsymbol{c}} \coloneqq \tau_{\boldsymbol{c}\#}Q$. Then, an uncertainty measure M is called **translation-invariant** if*

$$M(Q) = M(Q_{\boldsymbol{c}}), \quad \forall Q \in \mathcal{P}(\Theta),\ \forall \boldsymbol{c} \in \mathbb{R}^d.$$

This definition formalizes the common notion of translation-invariance to a second-order uncertainty measure and implies that the corresponding measure remains unchanged under a constant location shift in the first-order distribution.

Definition 3.2 (Scale-order equivariance)**.** *For $a > 0$, define the scaling map S_a implicitly via*

$$Y \sim P \implies aY \sim S_a P.$$

Assume that the predictive family $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is closed under scaling, i.e., there exists a measurable map $\tau_a : \Theta \to \Theta$

such that $P_{\tau_a(\boldsymbol{\theta})} = S_a P_{\boldsymbol{\theta}}$. Define Q^a as the pushforward of $Q \in \mathcal{P}(\Theta)$ under this mapping, i.e., $Q^a := \tau_{a\#}Q$. Then, an uncertainty measure M is called ***scale-order equivariant*** *if*

$$M(Q^a) = g(a)M(Q) + c(a), \quad a > 0,$$

for functions $g : \mathbb{R}_+ \to \mathbb{R}_+$, $c : \mathbb{R}_+ \to \mathbb{R}$, with g, c nondecreasing in a.

Similarly, this definition formalizes the notion of scale-invariance. However, here we only require monotonicity, i.e., that the ordering of the uncertainty measure remains unchanged under a scaling of the first-order distribution.

Further, let $\delta_{\boldsymbol{\theta}} \in \mathcal{P}(\Theta)$ denote the Dirac measure at $\boldsymbol{\theta} \in \Theta$. For $P_{\boldsymbol{\theta}_u}, P_{\boldsymbol{\theta}_l}, \in \Theta$ let $\leq_{\text{cx}}$ denote the convex order, meaning that $P_{\boldsymbol{\theta}_l} \leq_{\text{cx}} P_{\boldsymbol{\theta}_u} \iff \mathbb{E}_{X\sim P_{\boldsymbol{\theta}_l}}[\phi(X)] \leq \mathbb{E}_{Y\sim P_{\boldsymbol{\theta}_u}}[\phi(Y)]$ for all convex functions $\phi : \mathcal{Y} \to \mathbb{R}$. In particular for $P_{\boldsymbol{\theta}_l} \leq_{\text{cx}} P_{\boldsymbol{\theta}_u}$ it holds that $\mathbb{E}_{X\sim P_{\boldsymbol{\theta}_l}}[X] = \mathbb{E}_{Y\sim P_{\boldsymbol{\theta}_u}}[Y]$ and $\mathbb{V}_{X\sim P_{\boldsymbol{\theta}_l}}(X) \leq \mathbb{V}_{Y\sim P_{\boldsymbol{\theta}_u}}(Y)$, since the convex order is a measure of variability of a distribution [Shaked and Shanthikumar, 2007]. Similarly, for $Q_l^{\varphi}, Q_u^{\varphi} \in \mathcal{P}(\mathcal{P}(\mathcal{Y}))$, let $\leq_{\text{cx}}^2$ denote the convex order with respect to all convex functionals $\Phi : \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$. We define the following properties that a suitable uncertainty measure should satisfy:

A0: TU, AU, and EU should be non-negative.
A1: $\text{EU}(Q) = 0$ if and only if $Q = \delta_{\boldsymbol{\theta}}$.
A2: $\text{EU}(\delta_{\boldsymbol{\theta}}) \leq \text{EU}(Q_l) \leq \text{EU}(Q_u)$, $\forall Q_l^{\varphi} \leq_{\text{cx}}^2 Q_u^{\varphi}$.
A3: $\text{AU}(\delta_{\boldsymbol{\theta}_l}) \leq \text{AU}(\delta_{\boldsymbol{\theta}_u})$, $\forall P_{\boldsymbol{\theta}_l} \leq_{\text{cx}} P_{\boldsymbol{\theta}_u}$.
A4: TU, AU, and EU should be translation invariant.
A5: TU, AU, and EU should be scale-order equivariant.

Axiom A0. Non-negativity is a justifiable assumption in the setting of the uncertainty measures, as it allows for interpretability and, together with an additive decomposition, implies that $\text{TU} \geq \text{EU}$ and $\text{TU} \geq \text{AU}$.

Axioms A1 & A2. Intuitively, the smallest value of EU should be attained for a distribution with no variability at all; a (second-order) Dirac distribution $\delta_{\boldsymbol{\theta}}$. In addition, the converse should hold as well (A1). Further, when Q_l^{φ} is less variable than Q_u^{φ}, as implied by the convex order, the corresponding measure of epistemic uncertainty should assign a smaller value to Q_l (A2). Since the uncertainty measures might depend on $\boldsymbol{\theta}$ in a nonlinear way, we assume the convex order on the push-forward distribution Q^{φ}.

Axiom A3. Similarly, if the second-order distribution is degenerate, and $P_{\boldsymbol{\theta}_l}$ is less variable than $P_{\boldsymbol{\theta}_u}$, the corresponding measure of AU should assign smaller values as well.

Axiom A4. A measure of uncertainty should be invariant to a translation shift in the predictive distribution, i.e., it should not matter whether the first-order distribution is shifted by some constant value. This is formalized in Definition 3.1 and should hold for all types of uncertainty.

Axiom A5. Finally, all measures of uncertainty should be scale-order equivariant, since a nonnegative scaling of the outcome Y should not affect the order of the assigned uncertainties. Definition 3.2 formalizes this notion of scale-order equivariant with respect to two scaling functions g and c. Note that for the case $g(a) = 1, c(a) = 0$ we obtain *scale equivariance*, which would imply that the assigned uncertainty values do not change at all with a scaling of the outcome variable. While the latter is helpful for interpretability, it might be too strict as a requirement, as it would force uncertainty measures to be unitless—an unrealistic constraint given that uncertainty could naturally change with the underlying unit.

Overall, our axioms are not intended as an exhaustive set of requirements for uncertainty measures in regression, but provide a principled baseline of intuitive properties against which any candidate measure should be evaluated. Yet even this basic set already exposes fundamental differences between the most widely used uncertainty measures.

4 COMPARISON OF UNCERTAINTY QUANTIFICATION METHODS

In this section, we demonstrate how our uncertainty representation framework includes commonly-used predictive methods, in particular deep ensembles [Lakshminarayanan et al., 2017] and deep evidential regression [Amini et al., 2020]. Furthermore, we highlight how the variance- and entropy-based measures differ across the two representation methods regarding their relative assessment of aleatoric and epistemic uncertainty. More details and derivations of the measures, an analysis of additional representation methods, as well as visualizations are available in Appendix B.

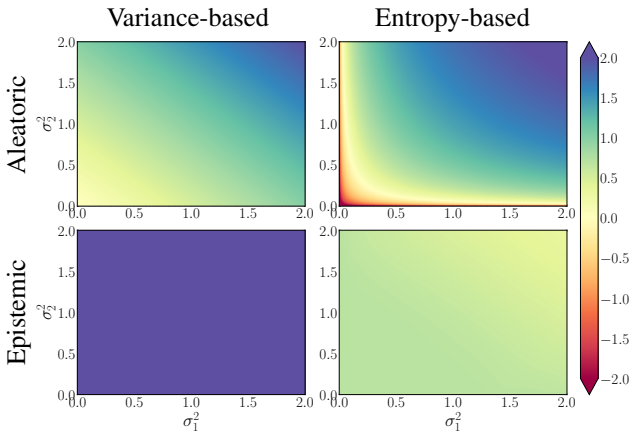

Figure 2: Comparison of the different measures for aleatoric and epistemic uncertainty for a two-member deep ensemble with $\mu_1 = 2, \mu_2 = -2$. The variance-based measure for epistemic uncertainty is invariant against changes in σ^2.

Deep Ensembles. Consider a (univariate) deep ensemble

[Lakshminarayanan et al., 2017], which utilizes a predictive Gaussian distribution, i.e. $P_{\boldsymbol{\theta}} = \mathcal{N}(\cdot\,; \mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2)^\top \in \mathbb{R} \times \mathbb{R}_{>0}$. Here, M networks are trained in a stochastic manner, leading to a discrete set of predictions $\boldsymbol{\theta}_m = (\mu_m, \sigma_m^2)^\top$, $m = 1, \ldots, M$. The ensemble acts as an empirical distribution $Q = \frac{1}{M}\sum_{m=1}^{M} \delta_{\boldsymbol{\theta}_m}$, e.g., a mixture of (equally weighted) Dirac measures over the parameter space. The second-order distribution is then given as $Q^\varphi = \frac{1}{M}\sum_{m=1}^{M} \delta_{P_{\boldsymbol{\theta}_m}}$ and the predictive mixture is given as a Gaussian mixture distribution $\overline{P} = \frac{1}{M}\sum_{m=1}^{M} \mathcal{N}(\cdot\,, \mu_m, \sigma_m^2)$.

For the variance-based measures, we obtain

$$\mathrm{AU}(Q) = \frac{1}{M}\sum_{m=1}^{M} \sigma_m^2,$$

$$\mathrm{EU}(Q) = \frac{1}{M}\sum_{m=1}^{M} \mu_m^2 - \left(\frac{1}{M}\sum_{m=1}^{M} \mu_m\right)^2,$$

while the entropy-based measures are given as

$$\mathrm{AU}(Q) = \frac{1}{2M}\sum_{m=1}^{M} \log(2\pi e \sigma_m^2),$$

$$\mathrm{EU}(Q) = \frac{1}{M}\sum_{m=1}^{M} \int_{\mathbb{R}} \phi(t \mid \boldsymbol{\theta}_m) \log \frac{\phi(t \mid \boldsymbol{\theta}_m)}{\frac{1}{M}\sum_{l=1}^{M} \phi(t \mid \boldsymbol{\theta}_l)}\, dt,$$

where $\phi(\cdot \mid \boldsymbol{\theta}_m)$ denotes the density of a Gaussian with parameters $\boldsymbol{\theta}_m$.

Clearly, both measures behave quite differently. In particular, the entropy-based measure does not admit a closed-form expression for epistemic uncertainty. For the variance-based measure, EU depends only on the first moment and AU only on the second moment of the underlying distribution. Therefore, the measure cannot distinguish between distributions where only higher-order moments are different, leading to a violation of (A1), as highlighted in Figure 2.

Deep Evidential Regression. Consider the deep evidential regression framework [Amini et al., 2020], which also assumes a predictive Gaussian distribution. Here, second-order uncertainty is incorporated by specifying a Normal Inverse-Gamma (NIG) prior over the Gaussian parameters $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$. Applying this to our setting, we obtain: $Q = q(\mu, \sigma^2) = q(\mu)\, q(\sigma^2) = \mathcal{N}(\mu; \gamma, \sigma^2 \upsilon^{-1})\, \Gamma^{-1}(\sigma^2; \alpha, \beta)$ with $\gamma \in \mathbb{R}, \upsilon > 0, \alpha > 1, \beta > 0$. While the push-forward Q^φ is only defined implicitly with no analytic expression, the predictive mixture is given via a Student's t distribution $\overline{P} = t_{2\alpha}\left(\cdot\,; \gamma, \frac{\beta(1+\upsilon)}{\alpha\upsilon}\right)$ with 2α degrees of freedom.

In the deep evidential regression setting, the variance-based measures are given as

$$\mathrm{AU}(Q) = \frac{\beta}{\alpha - 1}, \qquad \mathrm{EU}(Q) = \frac{\beta}{\upsilon(\alpha - 1)},$$

while for the entropy-based measures, we obtain

$$\mathrm{AU}(Q) = \frac{1}{2}\left(\log(2\pi e) + \log(\beta) - \psi(\alpha)\right),$$

$$\mathrm{EU}(Q) = H\left(t_{2\alpha}\left(\gamma, \frac{\beta(1+\upsilon)}{\alpha\upsilon}\right)\right) - \mathrm{AU}(Q),$$

where ψ denotes the digamma function. Again, both measures differ significantly in their behavior with respect to the parameters of the second-order distribution. In particular, as visualized in Figure 3, both measures behave very differently for decreasing parameters α, υ. Intuitively, for the NIG distribution, the mean can be interpreted as being estimated from υ virtual observations, while its variance is estimated from α virtual observations [Amini et al., 2020]. Therefore, the limiting cases $\alpha \downarrow 1$ and $\upsilon \downarrow 0$ describe the behavior of the model under decreasing evidence. While the entropy-based measures remain well-defined for $\alpha \downarrow 1$, the variance-based measure diverges. For υ, both measures behave similarly, with AU remaining bounded and EU diverging, since with zero evidence υ, there is maximal uncertainty about the model parameters. This directly relates to the existence assumption of the measures; since the variance based measures requires finite second-moments, it diverges in the limiting case. Further visualizations of the behavior of the different measures for varying α, β are available in Appendix B.

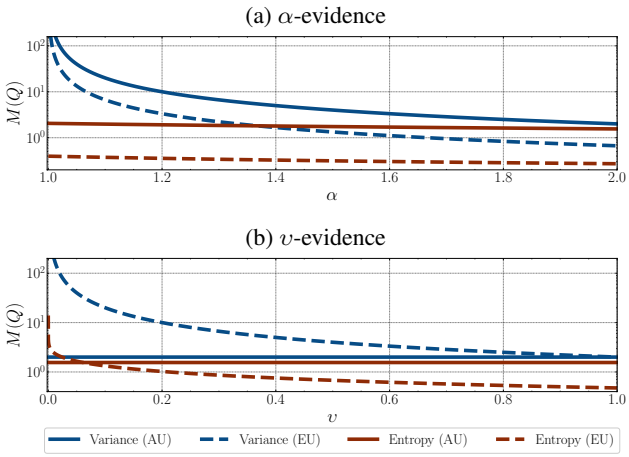

Figure 3: Comparison of the asymptotic behavior of the deep evidential regression parameters α, υ under the two different uncertainty measures. Although both measures behave similarly for υ, the variance-based measures diverge for $\alpha \downarrow 1$, while the entropy-based measure remains bounded.

5 ASSESSMENT OF PROPERTIES

In this section, we demonstrate how our proposed axioms can be used to assess existing or novel uncertainty measures theoretically and can ultimately guide practitioners in selecting or designing measures with specific characteristics,

depending on the underlying task. Here, we specifically analyze the previously introduced variance- and entropy-based measures, with respect to the proposed axioms. As already highlighted in Figure 1 and Figure 2, the same prediction can lead to substantially different estimations of uncertainties. This also transfers to the axioms, where we prove that both measures fulfill, but also violate some of the axioms, highlighting individual strengths and weaknesses. All corresponding proofs can be found in Appendix A.

Proposition 5.1. *The variance-based measures violate A1.*

The variance-based measure can only measure epistemic uncertainty in terms of the mean of the predictive distribution, and therefore cannot distinguish between two different second-order distributions that have the same marginal distribution over the mean. For instance, for a predictive Gaussian $p(y \mid \boldsymbol{\vartheta}) = \mathcal{N}(\mu, \sigma^2)$, the epistemic uncertainty reduces to $\mathrm{EU}(Q) = \mathbb{V}_Q(\mu)$ and is therefore insensitive to the marginal distribution of σ^2. Consequently, two second-order distributions that share the same marginal over μ but differ in their distribution over σ^2 are indistinguishable—compare the deep-ensemble example in Figure 2.

Proposition 5.2. *The variance-based measures fulfill A0, A2, A3, A4, A5.*

By nonnegativity of the variance, A0 follows immediately. A2 and A3 follow from the connection between convex order and increasing variance. Furthermore, A4 also holds, as the variance is translation-invariant. Finally, the variance-based measures are scale-order equivariant with $g(a) = a^2$ and $c(a) = 0$. Therefore, the variance-based measures fulfill the majority of the proposed axioms.

Proposition 5.3. *The entropy-based measure violates A0 for AU and TU.*

As highlighted in Figure 1, one major drawback of the differential entropy and the entropy-based measure is that although EU is nonnegative, AU (and therefore TU) can be assigned negative values. A natural definition of the absence of uncertainty is to have zero uncertainty, or in other words, complete knowledge about the underlying process. Negative uncertainty, on the other hand, does not admit a natural interpretation and makes the entropy-based measure difficult to evaluate in practice and impossible to compare to other measures. Since the lower bound for AU regarding entropy is negative infinity, this problem cannot be adjusted through a linear transformation. Furthermore, this implies that TU is not necessarily larger than EU, which is counterintuitive regarding the very definitions of the underlying uncertainties.

Proposition 5.4. *The entropy-based measure fulfills A2, A4, and A5.*

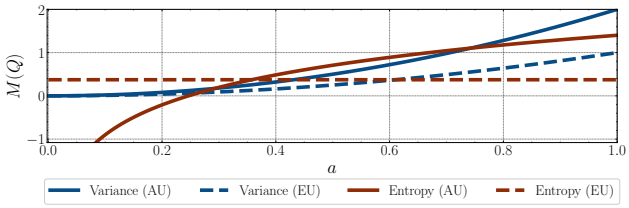

Figure 4: Comparison of the scale-order equivariance for the case of univariate deep evidential regression. While each of the entropy-based measures is equivariant when scaling the underlying data with a factor a, their ordering depends on the chosen scaling. In particular, AU is smaller (larger) than EU for $a \approx 0$ ($a \approx 1$), which can lead to problems as the ordering of the uncertainties should not depend on the scaling of the data. For the variance-based measure this problem does not occur.

A2 follows from the definition of convex order and convexity of the Kullback-Leibler divergence in P. A4 follows since all involved quantities are translation-invariant. Finally, A5 follows by properties of the differential entropy with $g(a) = 1,\ c(a) = d\log(a)$ for TU, AU and by properties of the KL-divergence with $g(a) = 1,\ c(a) = 0$ for EU. Interestingly, while the entropy-based measures fulfill scale-order equivariance for each individual measure, as the shift function c is different across EU and AU, their ordering changes with the scaling factor a. In fact, since EU is scale equivariant and nonnegative, it is always possible to find $a_l, a_u > 0$ such that $\mathrm{AU}(Q^{a_l}) < \mathrm{EU}(Q^{a_l}) = \mathrm{EU}(Q^{a_u}) \leq \mathrm{AU}(Q^{a_u})$, as visualized in Figure 4. From an interpretability perspective, this is problematic, as the scaling of the data impacts the connection of EU and AU, in particular, which of the two is larger. This is directly related to the entropy-based measure violating (A0).

Proposition 5.5. *The entropy-based measure fulfills A1 if the parametric family is identifiable and A3 if the parametric family is log-concave.*

In general, the differential entropy and KL-divergence can depend on the parameter $\boldsymbol{\theta}$ in any complex and nonlinear way. Therefore, for these axioms to hold, additional assumptions are required. A1 holds if the parametric family is identifiable, which is a reasonable and commonly used assumption. A3 holds if the parametric family is log-concave, which includes distributions such as the normal or Gamma distribution. Both assumptions are fulfilled for the previously introduced uncertainty representation methods. However, for both axioms, it can be shown that they do not hold without the additional assumptions discussed above.

6 RELATED WORK

Many works focus on analyzing different measures of uncertainty and their corresponding properties in the classifica-

tion setting. Specifically, axiomatic evaluation has been put forward in the machine learning community [Hüllermeier et al., 2022] and built the foundation for the development of many new measures of uncertainty, for example, based on proper scoring rules [Kotelevskii et al., 2025, Hofman et al., 2024], Wasserstein distance [Sale et al., 2024a] or credal sets [Hofman et al., 2024]. Despite this focus on an axiomatic understanding of uncertainty measures for classification, in the regression case, research mainly focuses on uncertainty representation and only little on a formal understanding of uncertainty measures and their underlying properties.

Amini et al. [2020], Meinert and Lavin [2022] propose the (multivariate) deep evidential regression, as a generalization of the evidential learning framework. To assess aleatoric and epistemic uncertainty, they utilize the variance-based decomposition. Laves et al. [2021] propose a calibration method that removes bias from an estimate of aleatoric uncertainty by correcting the scale of the first-order predictive variance. They also utilize the variance-based measure to decompose the different types of uncertainty. Valdenegro-Toro and Mori [2022] compare several different methods for uncertainty representation, such as deep ensembles or dropout, with respect to their capacity to measure and decompose epistemic and aleatoric uncertainty. For their assessment, they utilize a first-order predictive Gaussian and the variance-based uncertainty decomposition. For the selected methods, they critique the lack of disentanglement between aleatoric and epistemic uncertainty across different prediction tasks.

Moving to entropy-based methods, Berry and Meger [2025] propose a combination of deep ensembles and normalizing flows to generate a non-parametric model to estimate uncertainties. In their approach, aleatoric uncertainty is measured via conditional entropy with respect to the output distribution of the normalizing flow. Furthermore, they provide so-called Pairwise-Distance Estimators to provide bounds on the differential entropy and therefore allow for representing the epistemic uncertainty in the base distribution of the normalizing flow in an analytical way. Malinin et al. [2020] introduce regression prior networks, which are similar to multivariate deep evidential regression [Meinert and Lavin, 2022] and utilize a prior distribution over the parameters of a predictive multivariate Gaussian. To assess uncertainties, they utilize the variance- and entropy-based decomposition, as well as an additional measure called the expected pairwise KL-divergence, which is an upper bound on mutual information and is analytically available for a larger class of models. Postels et al. [2021] consider an approach that utilizes a latent representation of a neural network to measure epistemic and aleatoric uncertainty. By decomposing the network into a latent representation for each of the L layers, they can utilize bounds on differential entropy to provide estimates for both types of uncertainty. Using a quite different approach, Lahlou et al. [2023] propose to characterize epistemic and aleatoric uncertainty in terms of the Bayes risk on an estimator. By utilizing out-of-sample errors of a trained model, they obtain an estimate for the excess risk, which represents epistemic uncertainty. For aleatoric uncertainty, they propose several heuristics, such as estimation based on intrinsic data variation.

While all of the aforementioned approaches provide novel and well-founded ways to measure epistemic and aleatoric uncertainty, the focus is usually set on a specific predictive model and corresponding uncertainty representation. To the best of our knowledge, there exists no work that considers a general uncertainty representation framework, formalizes the corresponding evaluation procedure in a theoretical and axiomatic manner in general regression tasks, and compares the properties of different measures in a theoretical and axiomatic manner.

7 DISCUSSION

This work establishes a formal foundation for uncertainty representation, quantification, and evaluation in the supervised regression setting. We introduce a parametric uncertainty representation framework that allows for operationalization via a tractable second-order distribution, induced by the distribution over the parameter space and encompassing many well-known predictive machine learning methods, thereby enabling the first systematic cross-method comparison in regression. Furthermore, we develop an axiomatic methodology, comparing both classification-adapted and regression-specific axioms, that allows for analyzing given uncertainty measures with respect to their fundamental properties, strengths, and weaknesses. We demonstrate the practicality of our approach by evaluating the commonly used entropy- and variance-based uncertainty measures, exposing the strengths and weaknesses of each. In particular, we show that entropy-based measures can yield negative values in certain settings, which directly impacts scale equivariance across AU and EU and hinders natural interpretability. In contrast, the variance-based approach satisfies the majority of the proposed axioms but lacks a unique minimum for epistemic uncertainty, rendering it insensitive to uncertainty in higher-order moments, and requires stricter assumptions for its existence. Beyond the specific measures analyzed here, our axiomatic framework provides practitioners and researchers with a principled methodology for evaluating existing uncertainty measures and provides a principled guideline for designing novel uncertainty measures.

Limitations While choosing a predictive model with uncertainty only in its parameters is very general and includes commonly used approaches, as highlighted in this paper, it might be too restrictive in certain settings, for example, when Gaussian processes or sample-based generative models, such as diffusion models. In addition, although

the proposed axioms provide a principled baseline, the selection is not exhaustive; alternative or additional axioms, specifically designed for the regression domain, could reveal further distinctions between uncertainty measures not captured here. Finally, while the axioms capture theoretically desirable properties, satisfying or violating them does not directly guarantee good or poor empirical performance in any particular downstream task. The relationship between axiomatic compliance and practical utility—for example, in uncertainty-guided decision making or out-of-distribution detection—remains an open question.

Future work Our work opens several venues for future research. The most immediate open problem is to search for novel uncertainty measures that satisfy the proposed axioms in their entirety—or to establish whether such measures exist in the first place. For the classification setting, new measures have been proposed based on Wasserstein distance [Sale et al., 2024a] or score-based entropies Kotelevskii et al. [2025], Hofman et al. [2024], and an extension to the regression setting, as well as a corresponding theoretical analysis, seems promising. Here, our proposed axioms can provide a clear guidance on the design of new measures, for example, measures based on statistical divergences seem to be a natural candidate to fulfill (A1) and (A2). In addition, one may think of further analyzing and extending the set of axioms itself to accommodate other desirable properties of uncertainty measures, such as robustness or interpretability. Furthermore, validating the axioms in real-world regression settings could lead to a deeper understanding of the strengths and weaknesses of individual approaches. In particular, it would be worth analyzing whether the implications of the axioms persist across different data modalities. Finally, generalizing the framework to a more general class of predictive distributions is a direction worth pursuing. This could allow for analyzing the axioms for arbitrary predictive distributions, such as nonparametric methods or generative models.

Acknowledgements

C. Bülte, Y. Sale, G. Kutyniok, and E. Hüllermeier acknowledge support by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Research, Technology and Space.

C. Bülte and G. Kutyniok acknowledge support by the German Research Foundation under the grant DFG-SPP-2298.

E. Hüllermeier acknowledges support by the German Research Foundation under the grant GRK 3081 (project number 534429653).

G. Kutyniok also acknowledges support by the gAIn project, which is funded by the Bavarian Ministry of Science and the Arts (StMWK Bayern) and the Saxon Ministry for Science, Culture and Tourism (SMWK Sachsen). Furthermore, G. Kutyniok is supported by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder as well as by the Hightech Agenda Bavaria.

References

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems*, volume 33, pages 14927–14937. Curran Associates, Inc., 2020.

R.B. Ash. *Information Theory*. Dover books on advanced mathematics. Dover Publications, 1990.

Lucas Berry and David Meger. Epistemic uncertainty estimation in regression ensemble models with pairwise epistemic estimators. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Andrey Bronevich and George J Klir. Axioms for uncertainty measures on belief functions and credal sets. In *NAFIPS 2008 – 2008 Annual Meeting of the North American Fuzzy Information Processing Society*, pages 1–6. IEEE, 2008.

Thomas Buddenkotte, Lorena Escudero Sanchez, Mireia Crispin-Ortuzar, Ramona Woitek, Cathal McCague, James D. Brenton, Ozan Öktem, Evis Sala, and Leonardo Rundo. Calibrating ensembles for scalable uncertainty quantification in deep learning-based medical image segmentation. *Computers in Biology and Medicine*, 163: 107096, 2023. doi:10.1016/j.compbiomed.2023.107096.

Christopher Bülte, Nina Horat, Julian Quinting, and Sebastian Lerch. Uncertainty quantification for data-driven weather models. *Artificial Intelligence for the Earth Systems*, 2026. doi:10.1175/AIES-D-24-0049.1.

Bertrand Charpentier, Ransalu Senanayake, Mykel Kochenderfer, and Stephan Günnemann. Disentangling epistemic and aleatoric uncertainty in reinforcement learning, 2022. arXiv:2206.01558.

Antonio D'Isanto and Kai L. Polsterer. Photometric redshift estimation via deep learning – generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy & Astrophysics*, 609:A111, 2018. doi:10.1051/0004-6361/201731326.

Vineet Edupuganti, Morteza Mardani, and Shreyas Vasanawala. Uncertainty quantification in deep MRI reconstruction. *IEEE Transactions on Medical Imaging*, PP, 2020. doi:10.1109/TMI.2020.3025065.

Petra Friederichs and Thordis L. Thorarinsdottir. Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23(7):579–594, 2012. doi:10.1002/env.2176.

M.A. Ganaie, Minghui Hu, A.K. Malik, M. Tanveer, and P.N. Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022. doi:10.1016/j.engappai.2022.105151.

Tilmann Gneiting, Sebastian Lerch, and Benedikt Schulz. Probabilistic solar forecasting: Benchmarks, post-processing, verification. *Solar Energy*, 252:72–80, 2023. doi:10.1016/j.solener.2022.12.054.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty: A credal approach. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning, 2011. arXiv:1112.5745.

M. Huet-Dastarac, N.M.C. van Acht, F.C. Maruccio, J.E. van Aalst, J.C.J. van Oorschodt, F. Cnossen, T.M. Janssen, C.L. Brouwer, A. Barragan Montero, and C.W. Hurkmans. Quantifying and visualising uncertainty in deep learning-based segmentation for radiation therapy treatment planning: What do radiation oncologists and therapists want? *Radiotherapy and Oncology*, 201:110545, 2024. doi:10.1016/j.radonc.2024.110545.

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

Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Uncertainty in Artificial Intelligence*, pages 548–557. PMLR, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017.

Mostafa Kiani Shahvandi, Siddhartha Mishra, and Benedikt Soja. Laplacian deep ensembles: Methodology and application in predicting dut1 considering geophysical fluids. *Computers & Geosciences*, 196:105818, 2025. doi:10.1016/j.cageo.2024.105818.

Thomas Kneib, Alexander Silbersdorff, and Benjamin Säfken. Rage against the mean–a review of distributional regression approaches. *Econometrics and Statistics*, 26: 99–123, 2023.

Nikita Kotelevskii, Vladimir Kondratyev, Martin Takáč, Eric Moulines, and Maxim Panov. From risk to uncertainty: Generating predictive uncertainty measures via bayesian estimation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor I Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Max-Heinrich Laves, Sontje Ihler, Jacob F. Fast, Lüder A. Kahrs, and Tobias Ortmaier. Recalibration of aleatoric and epistemic regression uncertainty in medical imaging. *Machine Learning for Biomedical Imaging*, 1:1–26, 2021. MIDL 2020 special issue; doi:10.59275/j.melba.2021-a6fd.

Isaac Levi. *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT Press, 1980.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim GJ Rudner, C Bayan Bruss, and Andrew Gordon Wilson. Position: Supervised classifiers answer the wrong questions for ood detection. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Timo Löhr, Michael Ingrisch, and Eyke Hüllermeier. Towards aleatoric and epistemic uncertainty in medical image classification. In *International Conference on Artificial Intelligence in Medicine*, pages 145–155. Springer, 2024.

Andrey Malinin, Sergey Chervontsev, Ivan Provilkov, and Mark Gales. Regression prior networks, 2020. arXiv:2006.11590.

Martin János Mayer, Ágnes Baran, Sebastian Lerch, Nina Horat, Dazhi Yang, and Sándor Baran. Post-processing of ensemble photovoltaic power forecasts with distributional and quantile regression methods. *Solar Energy*, 307: 114361, 2026. doi:10.1016/j.solener.2026.114361.

Hendrik A. Mehrtens, Alexander Kurz, Tabea-Clara Bucher, and Titus J. Brinker. Benchmarking common uncertainty estimation methods with histopathological images under domain shift and label noise. *Medical Image Analysis*, 89:102914, 2023. doi:10.1016/j.media.2023.102914.

Nis Meinert and Alexander Lavin. Multivariate deep evidential regression, 2022. arXiv:2104.06135.

Rhiannon Michelmore, Marta Kwiatkowska, and Yarin Gal. Evaluating uncertainty quantification in end-to-end autonomous driving control, 2018. arXiv:1811.06817.

Eric Stefan Miele, Nicole Ludwig, and Alessandro Corsini. Multi-horizon wind power forecasting using multi-modal spatio-temporal neural networks. *Energies*, 16(8):3522, 2023. doi:10.3390/en16083522.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

Janis Postels, Hermann Blum, Yannick Strümpler, Cesar Cadena, Roland Siegwart, Luc Van Gool, and Federico Tombari. The hidden uncertainty in a neural networks activations, 2021. arXiv:2012.03082.

Stephan Rasp and Sebastian Lerch. Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900, 2018. doi:10.1175/MWR-D-18-0187.1.

Yusuf Sale, Michele Caprio, and Eyke Hüllermeier. Is the volume of a credal set a good measure for epistemic uncertainty? In *Uncertainty in Artificial Intelligence*, pages 1795–1804. PMLR, 2023.

Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. Second-order uncertainty quantification: A distance-based approach. In *Forty-first International Conference on Machine Learning*, pages 43060–43076, 2024a.

Yusuf Sale, Paul Hofman, Timo Löhr, Lisa Wimmer, Thomas Nagler, and Eyke Hüllermeier. Label-wise aleatoric and epistemic uncertainty quantification. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024b.

Benedikt Schulz and Sebastian Lerch. Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, 150(1):235–257, 2022. doi:10.1175/MWR-D-21-0150.1.

Benedikt Schulz, Lutz Köhler, and Sebastian Lerch. Aggregating Distribution Forecasts from Deep Ensembles. *Machine Learning*, 115(5):120, May 2026. doi:10.1007/s10994-025-06946-3.

Moshe Shaked and J. George Shanthikumar. *Stochastic Orders*. Springer, 2007. doi:10.1007/978-0-387-34675-5.

Niclas Ståhl, Göran Falkman, Alexander Karlsson, and Gunnar Mathiason. Evaluation of uncertainty quantification in deep learning. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 556–568, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-50146-4_41.

V. Strassen. The Existence of Probability Measures with Given Marginals. *The Annals of Mathematical Statistics*, 36(2):423 – 439, 1965. doi:10.1214/aoms/1177700153.

Arthur Thuy and Dries F. Benoit. Fast and reliable uncertainty quantification with neural network ensembles for industrial image classification. *Annals of Operations Research*, 353(2):517–543, December 2024. doi:10.1007/s10479-024-06440-4.

Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1508–1516. IEEE Computer Society, 2022. doi:10.1109/CVPRW56347.2022.00157.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1991.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in Artificial Intelligence*, pages 2282–2292. PMLR, 2023.

An Axiomatic Assessment of Entropy- and Variance-based Uncertainty Quantification in Regression (Supplementary Material)

Christopher Bülte[1,2] **Yusuf Sale**[1,2] **Timo Löhr**[1,2] **Paul Hofman**[1,2] **Gitta Kutyniok**[1,2,3,4] **Eyke Hüllermeier**[1,2,5]

[1]LMU Munich
[2]Munich Center for Machine Learning (MCML)
[3]DLR-German Aerospace Center
[4]University of Tromsø
[5]German Research Center for Artificial Intelligence (DFKI, DSA)

A PROOFS OF PROPOSITIONS

A.1 VARIANCE-BASED

Proof of Proposition 5.1. We prove that the variance-based measure violates A1. Consider $P_1 = \mathcal{N}(0, \sigma_1^2), P_2 = \mathcal{N}(0, \sigma_2^2)$ with $\sigma_1^2 \neq \sigma_2^2$ and a second-order distribution, specified as a Dirac mixture, i.e. $Q = \frac{1}{2}\delta_{\sigma_1^2} + \frac{1}{2}\delta_{\sigma_2^2}$. Recall that for the variance-based measure, we have $\mathrm{EU}(Q) = \mathbb{V}_Q[\mathbb{E}[Y \mid \boldsymbol{\vartheta}]]$, with $\boldsymbol{\vartheta} \sim Q$. In addition, we obtain $\overline{P} = \frac{1}{2}P_1 + \frac{1}{2}P_2$ and $\mathbb{E}_{Y' \sim \overline{P}}[Y'] = 0$. Then we obtain

$$\begin{aligned}\mathrm{EU}(Q) &= \mathbb{V}_Q[\mathbb{E}[Y \mid \boldsymbol{\vartheta}]] = \mathbb{E}_{\boldsymbol{\vartheta}\sim Q}[(\mathbb{E}_{Y\sim P_{\boldsymbol{\vartheta}}}[Y] - \underbrace{\mathbb{E}_{Y'\sim\overline{P}}[Y']}_{=0})^2] = \mathbb{E}_{\boldsymbol{\vartheta}\sim Q}[(\mathbb{E}_{Y\sim P_{\boldsymbol{\vartheta}}}[Y])^2] \\ &= \frac{1}{2}(\underbrace{\mathbb{E}_{Y\sim P_1}[Y]}_{=0})^2 + \frac{1}{2}(\underbrace{\mathbb{E}_{Y\sim P_2}[Y]}_{=0})^2 = 0.\end{aligned}$$

Therefore, we obtain $\mathrm{EU}(Q) = 0$ although $Q \neq \delta_{\boldsymbol{\theta}}$. □

Proof of Proposition 5.2. We prove that the variance-based measures fulfill A0, A2, A3, A4, and A5.

A0: A0 follows immediately from the variance being nonnegative.

A2: By definition of the convex order and the push-forward measure, we have

$$\mathbb{E}_{Q_l^\varphi}[\Phi(P)] = \mathbb{E}_{\boldsymbol{\vartheta}\sim Q_l}[\Phi(P_{\boldsymbol{\vartheta}})] \leq^2_{\mathrm{cx}} \mathbb{E}_{\boldsymbol{\vartheta}\sim Q_u}[\Phi(P_{\boldsymbol{\vartheta}})] = \mathbb{E}_{Q_u^\varphi}[\Phi(P)],$$

for all convex functionals $\Phi : \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$. By definition, convex order preserves expectations, which implies equal predictive mixtures, i.e., $\overline{P}_{Q_l} = \overline{P}_{Q_u} \coloneqq \overline{P}$.

Define the random vectors $\boldsymbol{M}_l \coloneqq \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{\vartheta}],\ \boldsymbol{\vartheta} \sim Q_l$ and $\boldsymbol{M}_u \coloneqq \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{\vartheta}],\ \boldsymbol{\vartheta} \sim Q_u$. Then, for the variance-based measure, we have $\mathrm{EU}(Q_l) = \mathbb{V}(\boldsymbol{M}_l), \quad \mathrm{EU}(Q_u) = \mathbb{V}(\boldsymbol{M}_u)$.

Now, consider the functional $\Phi(P) \coloneqq g(\mathbb{E}_{\boldsymbol{Y}\sim P}[\boldsymbol{Y}])$, which is convex on $\mathcal{P}(\mathcal{Y})$ for any convex $g : \mathbb{R}^d \to \mathbb{R}$. Using the definition of the convex order, we obtain

$$\mathbb{E}_{\boldsymbol{\vartheta}\sim Q_l}[g(\mathbb{E}_{\boldsymbol{Y}\sim P_{\boldsymbol{\vartheta}}})] \leq \mathbb{E}_{\boldsymbol{\vartheta}\sim Q_u}[g(\mathbb{E}_{\boldsymbol{Y}\sim P_{\boldsymbol{\vartheta}}})] \quad \forall \text{ convex } g,$$

which, by definition, is

$$\mathbb{E}_{Q_l}[g(\boldsymbol{M}_l)] \leq \mathbb{E}_{Q_u}[g(\boldsymbol{M}_u)] \quad \forall \text{ convex } g,$$

and implies $\boldsymbol{M}_l \leq_{\text{cx}} \boldsymbol{M}_u$[1]. From the convex order it follows immediately that $\mathbb{E}[\boldsymbol{M}_l] = \mathbb{E}[\boldsymbol{M}_u]$ and $\mathbb{V}(\boldsymbol{M}_l) \leq \mathbb{V}(\boldsymbol{M}_u)$, which directly implies

$$\text{EU}(Q_l) \leq \text{EU}(Q_u).$$

A3: Recall that we have $P_{\boldsymbol{\theta}_l} \leq_{\text{cx}} P_{\boldsymbol{\theta}_u},\ \boldsymbol{\theta}_l, \boldsymbol{\theta}_u \in \Theta$ and $\text{AU}(\delta_{\boldsymbol{\theta}}) = \mathbb{E}_{\delta_{\boldsymbol{\theta}}}[\mathbb{V}(P_{\boldsymbol{\theta}})] = \mathbb{V}(P_{\boldsymbol{\theta}})$. By definition of the convex order we have

$$P_{\boldsymbol{\theta}_l} \leq_{\text{cx}} P_{\boldsymbol{\theta}_u} \implies \mathbb{V}(P_{\boldsymbol{\theta}_l}) \leq \mathbb{V}(P_{\boldsymbol{\theta}_u}),$$

and therefore

$$\text{AU}(\delta_{\boldsymbol{\theta}_l}) \leq \text{AU}(\delta_{\boldsymbol{\theta}_u}).$$

A4: Let $Y \mid \boldsymbol{\vartheta} \sim P_{\boldsymbol{\vartheta}}$ and $Y_c \mid \boldsymbol{\vartheta}' \sim P_{\boldsymbol{\vartheta}'}$ with $\boldsymbol{\vartheta}' = \tau_c(\boldsymbol{\vartheta})$ with $\boldsymbol{\vartheta} \sim Q$. Then, by definition, we have $Y \overset{d}{=} Y + c$. By the properties of the variance, we obtain

$$\text{AU}(Q_c) = \mathbb{E}_{Q_c}[\mathbb{V}(Y \mid \boldsymbol{\vartheta}')] = \mathbb{E}_Q[\mathbb{V}(Y + c \mid \tau_c(\boldsymbol{\vartheta}))] = \mathbb{E}_Q[\mathbb{V}(Y \mid \boldsymbol{\vartheta})] = \text{AU}(Q),$$

and

$$\text{EU}(Q_c) = \mathbb{V}_{Q_c}(\mathbb{E}[Y \mid \boldsymbol{\vartheta}']) = \mathbb{V}_Q(\mathbb{E}[Y + c \mid \tau_c(\boldsymbol{\vartheta})]) = \mathbb{V}_Q(\mathbb{E}[Y \mid \boldsymbol{\vartheta}]) = \text{EU}(Q).$$

TU then follows from the additive decomposition.

A5: By definition, we have $P_{\tau_a(\boldsymbol{\vartheta})} = S_a P_{\boldsymbol{\vartheta}}$ for $\boldsymbol{\vartheta} \sim Q$. If $Y \mid \boldsymbol{\vartheta} \sim P_{\boldsymbol{\vartheta}}$, then $Y^a \mid \boldsymbol{\vartheta}' \sim S_a P_{\boldsymbol{\vartheta}'}$ is the law of aY with $\boldsymbol{\vartheta}' = \tau_a(\boldsymbol{\vartheta}) \sim Q^a$. By properties of the variance, we obtain

$$\text{AU}(Q^a) = \mathbb{E}_{Q^a}[\mathbb{V}(Y \mid \boldsymbol{\vartheta}')] = \mathbb{E}_Q[\mathbb{V}(Y^a \mid \boldsymbol{\vartheta}')] = \mathbb{E}_Q[a^2 \mathbb{V}(Y \mid \boldsymbol{\vartheta})] = a^2 \,\text{AU}(Q),$$

and

$$\text{EU}(Q^a) = \mathbb{V}_{Q^a}(\mathbb{E}[Y \mid \boldsymbol{\vartheta}']) = \mathbb{V}_Q(\mathbb{E}[Y^a \mid \boldsymbol{\vartheta}']) = \mathbb{V}_Q(a\mathbb{E}[Y \mid \boldsymbol{\vartheta}]) = a^2 \,\text{EU}(Q).$$

TU follows from the additive decomposition and we obtain $g(a) = a^2$ and $c(a) = 0$. □

A.2 ENTROPY-BASED

Proof of Proposition 5.3. We prove that the entropy-based measure violates A0 for AU and TU. For EU, as the KL-divergence is nonnegative by definition, EU is also nonnegative. AU can be negative since differential entropy can be negative. As an example, consider the exponential distribution with entropy $H(\text{EXP}(\lambda)) = 1 - \log(\lambda)$, which is negative for $\lambda > e$. Taking $Q = \delta_{2e}$ as a second-order distribution then leads to $\text{AU}(Q) < 0$. Due to its additive structure, total uncertainty $\text{TU} = \text{EU} + \text{AU}$ can also be negative. □

Proof of Proposition 5.4. We prove that the entropy-based measures fulfill A2, A4, and A5.

A2: By definition of the convex order and the push-forward measure, we have

$$\mathbb{E}_{Q_l^\varphi}[\Phi(P)] = \mathbb{E}_{\boldsymbol{\vartheta} \sim Q_l}[\Phi(P_{\boldsymbol{\vartheta}})] \leq^2_{\text{cx}} \mathbb{E}_{\boldsymbol{\vartheta} \sim Q_u}[\Phi(P_{\boldsymbol{\vartheta}})] = \mathbb{E}_{Q_u^\varphi}[\Phi(P)],$$

for all convex functionals $\Phi : \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$. By definition, convex order preserves expectations, which implies equal predictive mixtures, i.e., $\overline{P}_{Q_l} = \overline{P}_{Q_u} \coloneqq \overline{P}$.

For the entropy-based measure we have $\text{EU}(Q) = \mathbb{E}_Q[\Phi(P_{\boldsymbol{\vartheta}})]$, where $\Phi(P) \coloneqq D_{\text{KL}}(P \| \overline{P})$ is a convex functional, due to the convexity of the KL-divergence in the first argument. Therefore, by the assumption of $Q_l^\varphi \leq^2_{\text{cx}} Q_u^\varphi$, we obtain

$$\text{EU}(Q_l) = \mathbb{E}_{\boldsymbol{\vartheta} \sim Q_l}[D_{\text{KL}}(P_{\boldsymbol{\vartheta}} \mid \overline{P})] \leq \mathbb{E}_{\boldsymbol{\vartheta} \sim Q_u}[D_{\text{KL}}(P_{\boldsymbol{\vartheta}} \mid \overline{P})] = \text{EU}(Q_u).$$

A4: Let $\boldsymbol{\vartheta}' = \tau_c(\boldsymbol{\vartheta})$ with $\boldsymbol{\vartheta} \sim Q$. By the shift invariance of the differential entropy $H(\cdot)$, we obtain

$$\text{AU}(Q_c) = \mathbb{E}_{Q_c}[H(P_{\boldsymbol{\vartheta}'})] = \mathbb{E}_Q[H(P_{\tau_c(\boldsymbol{\vartheta})})] = \mathbb{E}_Q[H(P_{\boldsymbol{\vartheta}}(\cdot - c))] = \mathbb{E}_Q[H(P_{\boldsymbol{\vartheta}})] = \text{AU}(Q),$$

[1]Here we slightly overload the notation and use $\leq_{\text{cx}}$ for the convex order, as introduced earlier, but comparing random vectors instead of probability measures.

and

$$\begin{aligned}\mathrm{EU}(Q_c) &= H(\mathbb{E}_{Q_c}[P_{\boldsymbol{\vartheta}'}]) - \mathrm{AU}(Q_c) = H(\mathbb{E}_Q[P_{\boldsymbol{\vartheta}}(\cdot - c)]) - \mathrm{AU}(Q) \\ &= H(\overline{P}_{\boldsymbol{\vartheta}}(\cdot - c)) - \mathrm{AU}(Q) = H(\overline{P}_{\boldsymbol{\vartheta}}) - \mathrm{AU}(Q) = \mathrm{EU}(Q).\end{aligned}$$

TU then follows from the additive decomposition.

A5: By definition, we have $P_{\tau_a(\boldsymbol{\vartheta})} = S_a P_{\boldsymbol{\vartheta}}$ for $\boldsymbol{\vartheta} \sim Q$. If $Y \mid \boldsymbol{\vartheta} \sim P_{\boldsymbol{\vartheta}}$, then $Y^a \mid \boldsymbol{\vartheta}' \sim S_a P_{\boldsymbol{\vartheta}'}$ is the law of aY with $\boldsymbol{\vartheta}' = \tau_a(\boldsymbol{\vartheta}) \sim Q^a$. By properties of the differential entropy $H(\cdot)$, we obtain

$$\mathrm{AU}(Q^a) = \mathbb{E}_{Q^a}[H(P_{\boldsymbol{\vartheta}'})] = \mathbb{E}_{Q^a}[H(S_a P_{\boldsymbol{\vartheta}})] = \mathbb{E}_{Q^a}[H(P_{\boldsymbol{\vartheta}}) + d \log a] = \mathrm{AU}(Q) + d \log a.$$

Furthermore, we have

$$H(\mathbb{E}_{Q^a}[P_{\boldsymbol{\vartheta}'}]) = H(\mathbb{E}_Q[P_{\tau_a(\boldsymbol{\vartheta})}]) = H(\mathbb{E}_Q[S_a P_{\boldsymbol{\vartheta}}]) = H(S_a \overline{P}) = H(\overline{P}) + d \log a,$$

and therefore

$$\mathrm{EU}(Q^a) = H(\mathbb{E}_{Q^a}[P_{\boldsymbol{\vartheta}'}]) - \mathrm{AU}(Q^a) = (H(\overline{P}) + d \log a) - (\mathrm{AU}(Q) + d \log a) = \mathrm{EU}(Q).$$

With this, we obtain $g(a) = 1$, $c(a) = d \log a$ for AU, TU and $c(a) = 0$ for EU. □

Proof of Proposition 5.5. Here, we prove that the entropy-based measure fulfills A1 if the parametric family is identifiable and A3 if the parametric family is log-concave.

First, we show that the entropy-based measure fulfills A1 if the parametric family is identifiable.

Let $Q \in \mathcal{P}(\Theta)$, $\bar{P} = \int P_{\boldsymbol{\theta}} \, dQ(\boldsymbol{\theta})$, and define $Q^{\varphi} = \phi_{\#} Q \in \mathcal{P}(\mathcal{P}(\mathcal{Y}))$ as the push-forward of Q under φ. Then

$$\mathrm{EU}(Q) = \mathbb{E}_Q[D_{\mathrm{KL}}(P_{\boldsymbol{\vartheta}} \mid \overline{P})] = \int D_{\mathrm{KL}}(P \| \overline{P}) \, Q^{\varphi}(dP) \geq 0.$$

In particular, we obtain $\mathrm{EU}(Q) = 0$ iff $Q^{\varphi} = \delta_{\bar{P}}$. Under identifiability of φ, this is equivalent to $Q = \delta_{\boldsymbol{\theta}}$.

Second, we show that the entropy-based measure fulfills A3 if the parametric family is log-concave. A probability distribution has log-concave density if the density can be expressed as $p(x) \equiv \exp(\varphi(x))$ for a concave function $\varphi(x)$. Since by assumption we have $P_{\boldsymbol{\theta}_l} \leq_{\mathrm{cx}} P_{\boldsymbol{\theta}_u}$ and $p_{\boldsymbol{\theta}_u}$ is log-concave, one obtains $\mathrm{supp}(P_{\boldsymbol{\theta}_l}) \subseteq \mathrm{supp}(P_{\boldsymbol{\theta}_u})$ [Strassen, 1965] such that the cross-entropy defined below is finite. Now, recall that differential entropy, which can be expressed as

$$\mathrm{AU}(\delta_{\boldsymbol{\theta}}) = H(P_{\boldsymbol{\theta}}) \coloneqq -\int p_{\boldsymbol{\theta}}(x) \log p_{\boldsymbol{\theta}}(x) d\mu(x) = \mathbb{E}_{P_{\boldsymbol{\theta}}}[-\log p_{\boldsymbol{\theta}}(X)].$$

Then, for a log-concave density, we have that $\phi(x) \coloneqq -\log p_{\boldsymbol{\theta}_u}(x)$ is a convex function in x. By convex order, we then have

$$\mathbb{E}_{X \sim P_{\boldsymbol{\theta}_l}}[-\log p_{\boldsymbol{\theta}_u}(X)] = \mathbb{E}_{X \sim P_{\boldsymbol{\theta}_l}}[\phi(X)] \leq \mathbb{E}_{Y \sim P_{\boldsymbol{\theta}_u}}[\phi(Y)] = \mathrm{AU}(\boldsymbol{\theta}_u).$$

The left-hand side is the cross-entropy of $P_{\boldsymbol{\theta}_l}, P_{\boldsymbol{\theta}_u}$, which, by definition, can be decomposed into

$$\mathbb{E}_{X \sim P_{\boldsymbol{\theta}_l}}[-\log p_{\boldsymbol{\theta}_u}(X)] = H(P_{\boldsymbol{\theta}_l}) + D_{\mathrm{KL}}(P_{\boldsymbol{\theta}_l} \| P_{\boldsymbol{\theta}_u}) \geq H(P_{\boldsymbol{\theta}_l}),$$

where the inequality follows from the KL-divergence being nonnegative. Combining the above gives

$$\mathrm{AU}(\delta_{\boldsymbol{\theta}_l}) \leq \mathbb{E}_{X \sim P_{\boldsymbol{\theta}_l}}[-\log p_{\boldsymbol{\theta}_u}(X)] = \mathbb{E}_{X \sim P_{\boldsymbol{\theta}_l}}[\phi(X)] \leq \mathbb{E}_{Y \sim P_{\boldsymbol{\theta}_u}}[\phi(Y)] = \mathrm{AU}(\delta_{\boldsymbol{\theta}_u}).$$

□

Proposition A.1. *The entropy-based measure violates A1 if the model is not identifiable.*

Proof. Consider a non-identifiable model $P_{\boldsymbol{\theta}}$ with $P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}$ and $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$. Then $Q = \frac{1}{2}\delta_{\boldsymbol{\theta}_1} + \frac{1}{2}\delta_{\boldsymbol{\theta}_2}$ is not the Dirac measure, but we obtain $\overline{P} = \frac{1}{2}P_{\boldsymbol{\theta}_1} + \frac{1}{2}P_{\boldsymbol{\theta}_2} = P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}$, and therefore

$$\mathrm{EU}(Q) = \mathbb{E}_{\boldsymbol{\vartheta} \sim Q}[D_{\mathrm{KL}}(P_{\boldsymbol{\vartheta}} \mid \overline{P})] = 0.$$

□

Proposition A.2. *Without additional assumptions, the entropy-based measure violates A3.*

Proof. Consider the (symmetric) Beta distribution $Y_\alpha \sim P_\alpha = \mathrm{Beta}(\alpha, \alpha),\ \alpha > 0$, which has mean $\mathbb{E}_{P_\alpha}[Y_\alpha] = 1/2$ and a second-order Dirac distribution $Q = \delta_\alpha$. Consider two special cases:

- $Y_1 \sim \mathrm{Beta}(1,1) = \mathrm{Unif}(0,1)$ with CDF $P_1(x) = x,\ x \in [0,1]$.
- $Y_{1/2} = \mathrm{Beta}(1/2, 1/2)$ (arcsine distribution) with cdf $P_{1/2}(x) = \frac{2}{\pi}\arcsin\left(\sqrt{x}\right),\ x \in (0,1)$.

First, we check the convex order. From [Shaked and Shanthikumar, 2007, Theorem 3.A.5] we know that for two variables X, Y with equal means and CDF F and G, respectively, we have

$$X \leq_{\mathrm{cx}} Y \iff \int_0^p F^{-1}(u)\,du \geq \int_0^p G^{-1}(u)\,du, \quad \forall p \in [0,1].$$

In our case, we have

$$\int_0^p P_1^{-1}(u)\,du = \int_0^p u\,du \geq \int_0^p \sin^2\left(\frac{\pi u}{2}\right)\,du = \int_0^p P_{1/2}^{-1}(u)\,du, \quad \forall p \in [0,1],$$

which can be verified by solving the integrals. Therefore, we have

$$P_1 \leq_{\mathrm{cx}} P_{1/2}.$$

On the other hand, using the differential entropy, we obtain $H(P_1) = \log(1) = 0$ and $H(P_{1/2}) = \log\left(\frac{\pi}{4}\right) \approx -0.24 < 0$. Therefore, we have $P_1 \leq_{\mathrm{cx}} P_{1/2}$, but $\mathrm{AU}(\delta_1) = 0 > -0.24 = \mathrm{AU}(\delta_{1/2})$, which violates A3. □

B ANALYSIS OF UNCERTAINTY REPRESENTATION METHODS

In this section, we analyze the already introduced uncertainty representation methods, i.e., deep ensembles and deep evidential regression, in more detail and provide an analysis for additional variants.

B.1 GAUSSIAN DEEP ENSEMBLE

(Gaussian) deep ensembles, as popularized by Lakshminarayanan et al. [2017] provide a way to improve the performance of a neural network by simply training an ensemble of networks in a stochastic manner, therefore obtaining a more robust solution. When combined with a first-order predictive distribution, this leads to possible assessment of aleatoric- as well as epistemic uncertainty. Deep ensembles have been a gold standard for the past years [Ovadia et al., 2019, Ganaie et al., 2022] and have been applied to numerous tasks, such as computer vision [Thuy and Benoit, 2024, Ovadia et al., 2019], image segmentation [Buddenkotte et al., 2023, Mehrtens et al., 2023, Huet-Dastarac et al., 2024] or weather prediction [Rasp and Lerch, 2018, Schulz et al., 2026].

In the (univariate) deep ensemble setting [Lakshminarayanan et al., 2017], we have $Y \sim P_{\boldsymbol{\vartheta}} = \mathcal{N}(\mu, \sigma^2)$, with $\boldsymbol{\vartheta} \sim Q$. Here, M different networks are trained in a stochastic manner, leading to a set of predictions $\boldsymbol{\theta}_m = (\mu_m, \sigma_m^2),\ m = 1, \ldots, M$. The ensemble acts as an empirical distribution $Q = \frac{1}{M}\sum_{m=1}^M \delta_{\boldsymbol{\theta}_m}$, e.g., a mixture of (equally weighted) Dirac measures. For the predictive mixture, we obtain the Gaussian mixture

$$\overline{P} = \frac{1}{M}\sum_{m=1}^M \mathcal{N}(\cdot\,; \mu_m, \sigma_m^2).$$

Variance-based measures: For the variance-based measures, we obtain

$$\begin{aligned}
\mathrm{AU}(Q) &= \mathbb{E}_Q[\mathbb{V}(Y \mid \boldsymbol{\vartheta})] = \frac{1}{M}\sum_{m=1}^M \sigma_m^2, \\
\mathrm{EU}(Q) &= \mathbb{V}_Q(\mathbb{E}[Y \mid \boldsymbol{\vartheta}]) = \frac{1}{M}\sum_{m=1}^M \mu_m^2 - \left(\frac{1}{M}\sum_{m=1}^M \mu_m\right)^2, \\
\mathrm{TU}(Q) &= \mathbb{V}(Y) = \frac{1}{M}\sum_{m=1}^M \left(\sigma_m^2 + \mu_m^2\right) - \left(\frac{1}{M}\sum_{m=1}^M \mu_m\right)^2.
\end{aligned}$$

Entropy-based measures: For the entropy-based measures, by the definition of differential entropy and the KL-divergence, we obtain

$$\mathrm{AU}(Q) = \mathbb{E}_Q[H(Y \mid P_{\boldsymbol{\vartheta}})] = \frac{1}{2M} \sum_{m=1}^{M} \log(2\pi e \sigma_m^2),$$

$$\mathrm{EU}(Q) = \mathbb{E}_Q[D_{\mathrm{KL}}(P_{\boldsymbol{\vartheta}} \| \overline{P})] = \frac{1}{M} \sum_{m=1}^{M} \int_{-\infty}^{\infty} p(t \mid \boldsymbol{\theta}_m) \log \frac{p(t \mid \boldsymbol{\theta}_m)}{\frac{1}{M} \sum_{l=1}^{M} p(t \mid \boldsymbol{\theta}_l)} \, dt,$$

$$\mathrm{TU}(Q) = H(\overline{P}) = \mathrm{AU}(Q) + \mathrm{EU}(Q),$$

where $p(x \mid \boldsymbol{\theta}_m) = \mathcal{N}(x; \mu_m, \sigma_m^2)$. Here, $\mathrm{EU}(Q)$ (and therefore $\mathrm{TU}(Q)$) does not admit a closed-form solution, but can be approximated numerically.

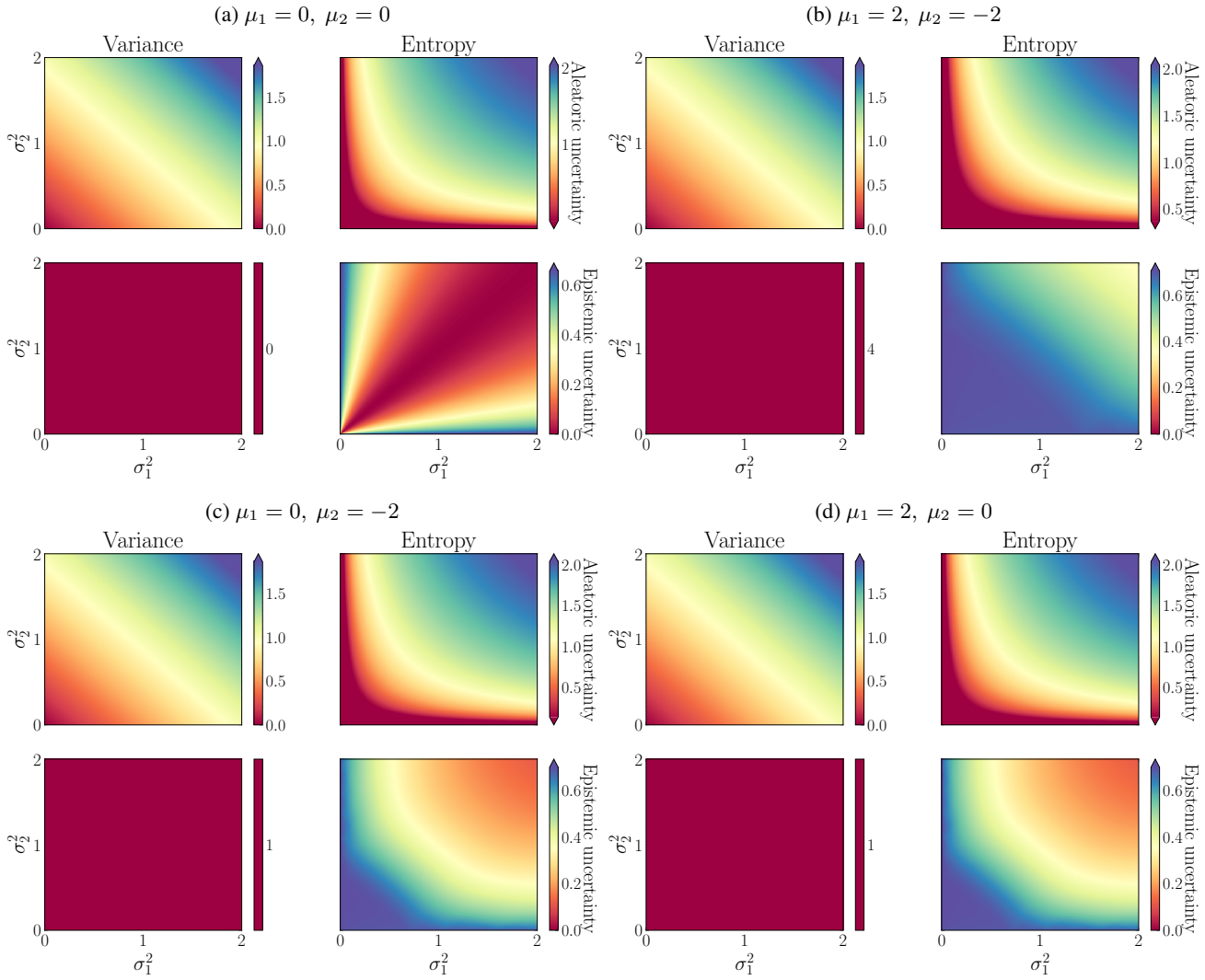

Figure 5: The figure shows the variance- and entropy-based different measures of uncertainty for a deep Gaussian ensemble $(\mu_m, \sigma_m^2)_{m=1}^2$ with two members across varying parameters.

Figure 5 shows the closed-form solutions of the variance- and entropy-based measures (EU and AU) for a two-member Gaussian deep ensembles. It is visible that both measures differ significantly. Most notable, the variance-based measure for EU does not change across σ_1^2, σ_2^2, as it only measures uncertainty based on the first-order moments, which is a direct implication of Proposition 5.1. In addition, both measures provide very different estimates for AU. For the variance-based measure the estimate depends, by definition, on σ_1^2, σ_2^2 in a linear way, while for the entropy-based measure the dependence is in a significantly different, nonlinear way, as visualized in the first row of each subplot in Figure 5.

B.2 LAPLACIAN DEEP ENSEMBLE

While research and applications have mainly focused on the Gaussian case [Ganaie et al., 2022], for the deep ensemble one can essentially consider any learnable parametric first-order distribution. The variance- and entropy-based uncertainty measures transfer in the same way, although closed-form solutions might not necessarily be available. As an example, we analyze how the different measures behave for a deep ensemble with a first-order Laplace distribution, which was used by Kiani Shahvandi et al. [2025] to model and predict deviation of time differences between earth's rotation time and coordinated universal time.

The setting is the same as the Gaussian deep ensemble setting, but with a predictive Laplace distribution, i.e., $Y \sim P_{\boldsymbol{\vartheta}} = \text{Laplace}(\mu, \eta), \boldsymbol{\vartheta} \sim Q$ with density

$$p(y \mid \boldsymbol{\theta}) = \frac{1}{2\eta} \exp\left(-\frac{|y-\mu|}{\eta}\right), \ \boldsymbol{\theta} = (\mu, \eta)^\top \in \mathbb{R} \times \mathbb{R}_{>0}.$$

For the Laplace distribution it holds that $\mathbb{E}[\text{Laplace}(\mu, \eta)] = \mu$ and $\mathbb{V}(\text{Laplace}(\mu, \eta)) = 2\eta^2$. For the predictive mixture, we obtain

$$\overline{P} = \frac{1}{M} \sum_{m=1}^{M} \text{Laplace}(\cdot\,; \mu_m, \eta_m).$$

Variance-based measures: For the variance-based measures, we obtain

$$\begin{aligned}
\text{AU}(Q) &= \mathbb{E}_Q[\mathbb{V}(Y \mid \boldsymbol{\vartheta})] = \frac{1}{M} \sum_{m=1}^{M} 2\eta_m^2, \\
\text{EU}(Q) &= \mathbb{V}_Q(\mathbb{E}[Y \mid \boldsymbol{\vartheta}]) = \frac{1}{M} \sum_{m=1}^{M} \mu_m^2 - \left(\frac{1}{M} \sum_{m=1}^{M} \mu_m\right)^2, \\
\text{TU}(Q) &= \mathbb{V}(Y) = \frac{1}{M} \sum_{m=1}^{M} \left(2\eta_m^2 + \mu_m^2\right) - \left(\frac{1}{M} \sum_{m=1}^{M} \mu_m\right)^2.
\end{aligned}$$

Entropy-based measures: For the entropy-based measures, by the definition of differential entropy and the KL-divergence, we obtain

$$\begin{aligned}
\text{AU}(Q) &= \mathbb{E}_Q[H(Y \mid P_{\boldsymbol{\vartheta}})] = 1 + \frac{1}{M} \sum_{m=1}^{M} \log(2\eta_m), \\
\text{EU}(Q) &= \mathbb{E}_Q[D_{\text{KL}}(P_{\boldsymbol{\vartheta}} \| \overline{P})] = \frac{1}{M} \sum_{m=1}^{M} \int_{-\infty}^{\infty} p(t \mid \boldsymbol{\theta}_m) \log \frac{p(t \mid \boldsymbol{\theta}_m)}{\frac{1}{M} \sum_{l=1}^{M} p(t \mid \boldsymbol{\theta}_l)} \, dt, \\
\text{TU}(Q) &= H(\overline{P}) = \text{AU}(Q) + \text{EU}(Q).
\end{aligned}$$

Here, $\text{EU}(Q)$ (and therefore $\text{TU}(Q)$) does not admit a closed-form solution, but can be approximated numerically.

Figure 6 shows the closed-form solutions of the variance- and entropy-based measures (EU and AU) for a two-member Laplacian deep ensembles. It is evident that both measures differ heavily. Similar to the Gaussian case, the variance-based measure of EU does not change across η_1, η_2. However, even for AU, both measures differ significantly. For both, AU depends on η_1, η_2 in a nonlinear way, however the corresponding curvatures are different, as visible in the two first rows of each plot.

B.3 UNIVARIATE DEEP EVIDENTIAL REGRESSION

Deep evidential regression, introduced by Amini et al. [2020] is a widely used method to provide uncertainty estimates and is based on the conjugate prior of a normal distribution. Recall that in this setting, we have $P_{\boldsymbol{\vartheta}} = \mathcal{N}(\mu, \sigma^2)$ with $\boldsymbol{\vartheta} \sim Q$ and $q(\mu, \sigma^2) = q(\mu)\, q(\sigma^2) = \mathcal{N}(\mu; \gamma, \sigma^2 \upsilon^{-1})\, \Gamma^{-1}(\sigma^2; \alpha, \beta)$. The predictive mixture is given via a Student's t distribution

$$\overline{P} = t_{2\alpha}\left(\cdot\,; \gamma, \frac{\beta(1+\upsilon)}{\alpha\upsilon}\right).$$

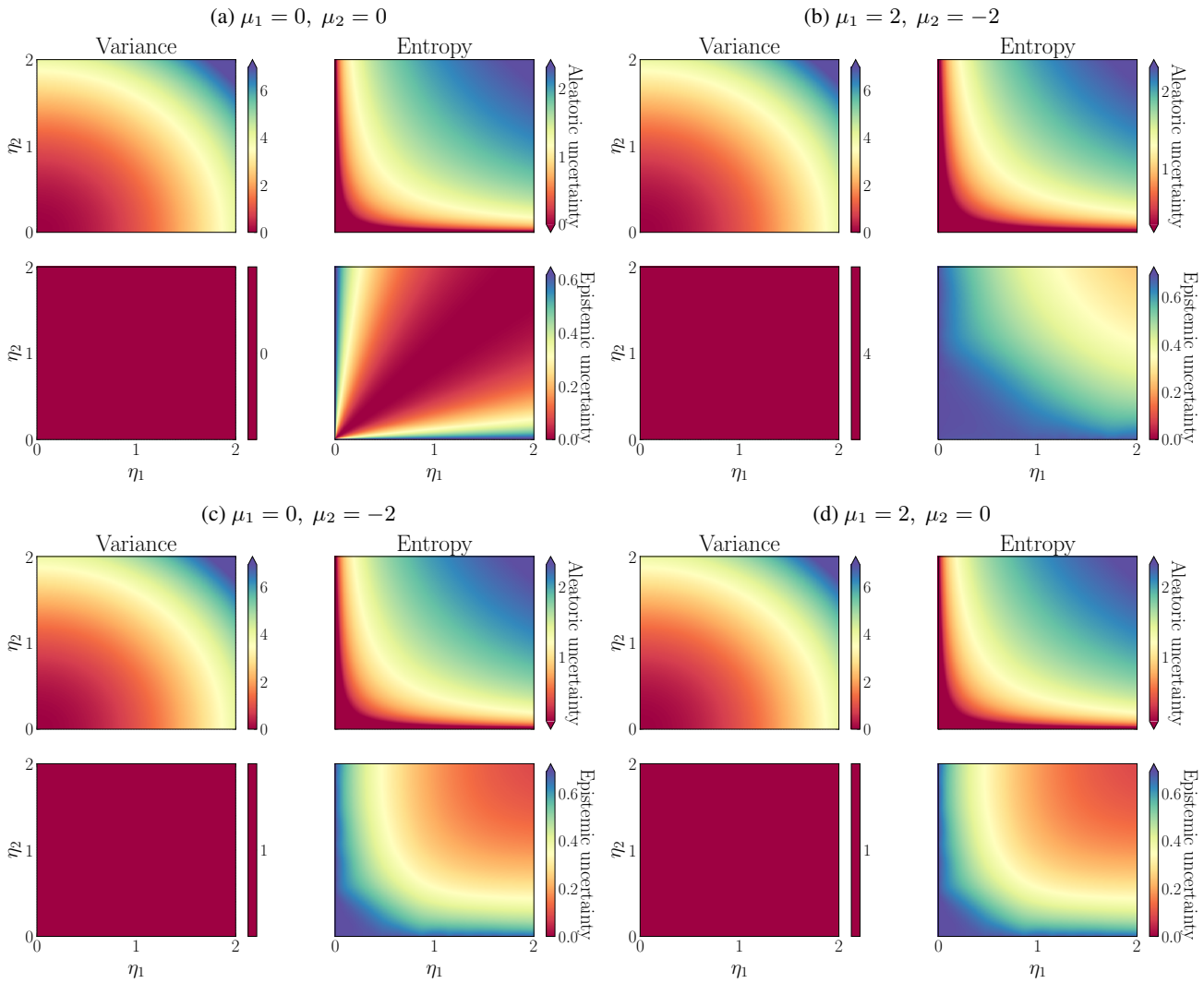

Figure 6: The figure shows the variance- and entropy-based different measures of uncertainty for a deep Laplacian ensemble $(\mu_m, \sigma_m^2)_{m=1}^2$ with two members across varying parameters.

Variance-based measures: [Amini et al., 2020] show that by calculating the corresponding moments, the variance-based uncertainty measures are given as

$$\begin{aligned}
\mathrm{AU}(Q) &= \mathbb{E}_Q[\mathbb{V}(Y \mid \boldsymbol{\vartheta})] = \frac{\beta}{\alpha - 1}, \\
\mathrm{EU}(Q) &= \mathbb{V}_Q(\mathbb{E}[Y \mid \boldsymbol{\vartheta}]) = \frac{\beta}{\upsilon(\alpha - 1)}, \\
\mathrm{TU}(Q) &= \mathbb{V}(Y) = \frac{\beta}{\alpha - 1}(1 + \upsilon^{-1}).
\end{aligned}$$

Entropy-based measures: For the entropy-based measures, we obtain

$$\begin{aligned}
\mathrm{AU}(Q) &= \frac{1}{2}(\log(2\pi e) + \log(\beta) - \psi(\alpha)), \\
\mathrm{EU}(Q) &= \frac{2\alpha + 1}{2}\psi\left(\alpha + \frac{1}{2}\right) - \alpha\psi\left(\alpha\right) - \frac{1}{2}\log(\pi e) + \log B\left(\alpha, \frac{1}{2}\right) + \frac{1}{2}\log\left(\frac{1 + \upsilon}{\upsilon}\right), \\
\mathrm{TU}(Q) &= \frac{2\alpha + 1}{2}\left(\psi\left(\alpha + \frac{1}{2}\right) - \psi\left(\alpha\right)\right) + \log B\left(\alpha, \frac{1}{2}\right) + \frac{1}{2}\log\left(\frac{2\beta(1 + \upsilon)}{\upsilon}\right).
\end{aligned}$$

This can be derived by the properties of the Normal-Inverse-Gamma distribution. For aleatoric uncertainty, we have

$$\begin{aligned}\mathrm{AU}(Q) &= \mathbb{E}_{q(\boldsymbol{\vartheta})}[H(P_{\boldsymbol{\vartheta}})] = \mathbb{E}_{q(\boldsymbol{\vartheta})}\left[\frac{1}{2}\log(2\pi e\sigma^2)\right] = \mathbb{E}_{\sigma^2\sim\Gamma^{-1}(\alpha,\beta)}\left[\frac{1}{2}\log(2\pi e\sigma^2)\right] \\ &= \frac{1}{2}\log(2\pi e) + \frac{1}{2}\mathbb{E}_{\sigma^2\sim\Gamma^{-1}(\alpha,\beta)}\left[\log(\sigma^2)\right] = \frac{1}{2}(\log(2\pi e) + \log(\beta) - \psi(\alpha)),\end{aligned}$$

where we used that for $S \sim \Gamma^{-1}(\alpha, \beta)$ we have $\mathbb{E}[\log(S)] = \log(\beta) - \psi(\alpha)$, where ψ denotes the digamma function. For total uncertainty, we use

$$\begin{aligned}\mathrm{TU}(Q) &= H(\overline{P}) = H\left(t_{2\alpha}\left(\gamma, \frac{\beta(1+\upsilon)}{\alpha\upsilon}\right)\right) \\ &= \frac{2\alpha+1}{2}\left(\psi\left(\alpha+\frac{1}{2}\right) - \psi\left(\alpha\right)\right) + \log B\left(\alpha, \frac{1}{2}\right) + \frac{1}{2}\log\left(\frac{2\beta(1+\upsilon)}{\upsilon}\right),\end{aligned}$$

where $B(\cdot, \cdot)$ is the Beta function. Then epistemic uncertainty can be obtained via $\mathrm{EU}(Q) = \mathrm{TU}(Q) - \mathrm{AU}(Q)$. Figure 7

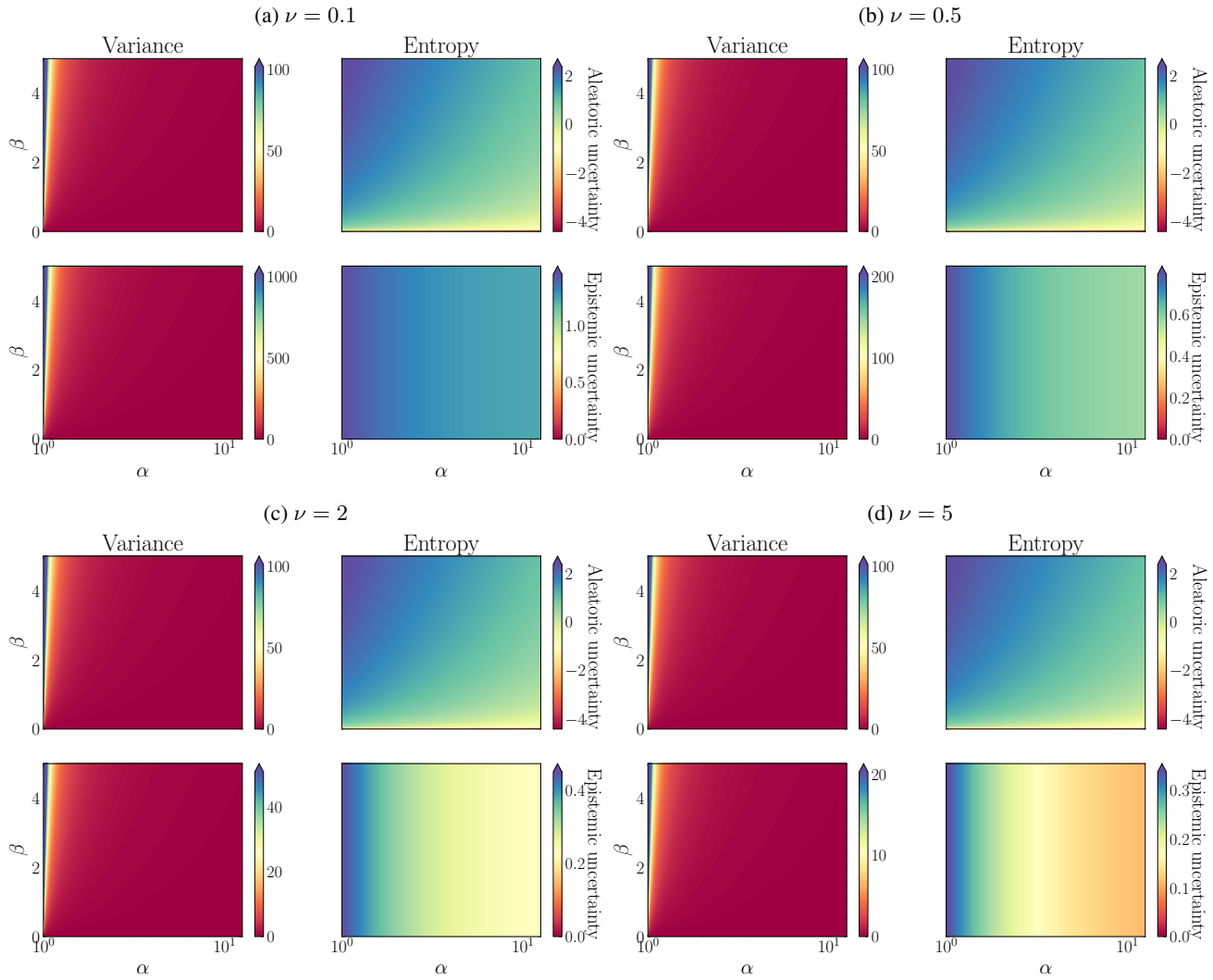

Figure 7: The figure shows the variance- and entropy-based measures of uncertainty for the deep evidential regression setting, across varying parameters.

shows the closed-form solutions of the variance- and entropy-based measures (EU and AU) for the deep evidential regression setting. It is clear that both measures behave very differently. In particular, for the variance-based measures EU and AU

diverge for $\alpha \downarrow 1$, while the entropy-based measures are well defined for all $\alpha > 0$. Furthermore, the entropy-based EU estimate does not depend on β, which corresponds to the scale parameter of the first-order noise.

B.4 MULTIVARIATE DEEP EVIDENTIAL REGRESSION

Meinert and Lavin [2022] extend the deep evidential regression to the multivariate setting, by considering the corresponding conjugate prior of a multivariate normal, which is given by a Normal-Inverse-Wishart distribution. In this setting, we have $\boldsymbol{Y} \in \mathbb{R}^d$ and $\boldsymbol{Y} \sim P_{\boldsymbol{\vartheta}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\vartheta} \sim Q$. The corresponding densities are given as $q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = q(\boldsymbol{\mu})q(\boldsymbol{\Sigma})$ with

$$q(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa), \qquad q(\boldsymbol{\Sigma}) = \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu),$$

where $\mathcal{W}^{-1}$ denotes the inverse-Wishart distribution and we have $\kappa > 0, \nu > d + 1, \boldsymbol{\Psi} \in \mathbb{R}^{d \times d}$. The predictive mixture is given via a multivariate t-distribution with $\nu - d + 1$ degrees of freedom, i.e.,

$$\overline{P} = t_{\nu - d + 1}\left(\cdot\,; \boldsymbol{\mu}_0, \frac{1 + \kappa}{\kappa(\nu - d + 1)}\boldsymbol{\Psi}\right).$$

Variance-based measures: Similar to the univariate case, the variance-based uncertainty measures can be calculated from the moments of the underlying (Normal-Inverse-Wishart) distribution as

$$\begin{aligned}
\mathrm{AU}(Q) &= \mathbb{E}_Q[\mathbb{V}(\boldsymbol{Y} \mid \boldsymbol{\vartheta})] = \mathbb{E}_Q[\mathrm{tr}(\mathrm{Cov}(\boldsymbol{Y} \mid \boldsymbol{\vartheta}))] = \frac{\mathrm{tr}(\boldsymbol{\Psi})}{\nu - d - 1}, \\
\mathrm{EU}(Q) &= \mathbb{V}_Q(\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{\vartheta}]) = \mathrm{tr}(\mathrm{Cov}(\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{\vartheta}])) = \frac{\mathrm{tr}(\boldsymbol{\Psi})}{\kappa(\nu - d - 1)}, \\
\mathrm{TU}(Q) &= \mathbb{V}(\boldsymbol{Y}) = \frac{\mathrm{tr}(\boldsymbol{\Psi})}{\nu - d - 1}\left(1 + \frac{1}{\kappa}\right).
\end{aligned}$$

Entropy-based measures: For the entropy-based measures, we obtain

$$\begin{aligned}
\mathrm{AU}(Q) &= \frac{d}{2}\log(\pi e) + \frac{1}{2}\log(|\boldsymbol{\Psi}|) - \frac{1}{2}\psi_d\left(\frac{\nu}{2}\right), \\
\mathrm{EU}(Q) &= \frac{d}{2}\log\left(\tfrac{1+1/\kappa}{e}\right) - \log\Gamma\left(\frac{\nu+1}{2}\right) + \log\Gamma\left(\frac{\nu-d+1}{2}\right) \\
&\quad + \frac{\nu+1}{2}\left[\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu-d+1}{2}\right)\right] + \frac{1}{2}\psi_d\left(\frac{\nu}{2}\right), \\
\mathrm{TU}(Q) &= \frac{1}{2}\log|\Psi| + \frac{d}{2}\log\left(\pi\left(1 + \frac{1}{\kappa}\right)\right) - \log\Gamma\left(\frac{\nu+1}{2}\right) + \log\Gamma\left(\frac{\nu-d+1}{2}\right) \\
&\quad + \frac{\nu+1}{2}\left[\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu-d+1}{2}\right)\right],
\end{aligned}$$

where $\Gamma(\cdot)$ denotes the Gamma function, $\psi(\cdot)$ the digamma function, and $\psi_d(\cdot)$ the multivariate digamma function. The derivations are similar to the univariate case and based on the differential entropy of a multivariate t-distribution.