Introducing an Improved Information-Theoretic Measure of Predictive Uncertainty

Kajetan Schweighofer* **Lukas Aichberger*** **Mykyta Ielanskyi** **Sepp Hochreiter**

ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning,
Johannes Kepler University Linz, Austria
* Joint first authors

Abstract

Applying a machine learning model for decision-making in the real world requires to distinguish what the model knows from what it does not. A critical factor in assessing the knowledge of a model is to quantify its predictive uncertainty. Predictive uncertainty is commonly measured by the entropy of the Bayesian model average (BMA) predictive distribution. Yet, the properness of this current measure of predictive uncertainty was recently questioned. We provide new insights regarding those limitations. Our analyses show that the current measure erroneously assumes that the BMA predictive distribution is equivalent to the predictive distribution of the true model that generated the dataset. Consequently, we introduce a theoretically grounded measure to overcome these limitations. We experimentally verify the benefits of our introduced measure of predictive uncertainty. We find that our introduced measure behaves more reasonably in controlled synthetic tasks. Moreover, our evaluations on ImageNet demonstrate that our introduced measure is advantageous in real-world applications utilizing predictive uncertainty.

1 Introduction

Decision-making with machine learning models in the real world requires risk assessment based on the predictive uncertainty of the model to be actionable [1]. It is essential to avoid models making high-risk, uncertain decisions. Instead, such decisions should be deferred to human experts or default to a potentially sub-optimal but safe decision. Therefore, it is key to use grounded measures of predictive uncertainty and provide estimates for them when deploying machine learning models for decision-making in the real world [12, 19].

Predictive uncertainty is often categorized into two types, aleatoric and epistemic, according to the source of uncertainty [17]. We focus on measuring the predictive uncertainty by characterizing the predictive distribution $p(\boldsymbol{y} \mid \boldsymbol{x})$ of outcomes $\boldsymbol{y}$ for inputs $\boldsymbol{x}$. Here, *aleatoric* uncertainty refers to the inherent stochasticity of sampling outcomes from the predictive distribution, thus is irreducible. Further, *epistemic* uncertainty refers to the lack of knowledge about the true predictive distribution, thus is reducible. Commonly, those uncertainties are assumed to be additive, summing up to a *total* predictive uncertainty.

The currently prevalent measure of predictive uncertainty is the entropy of the Bayesian model average (BMA) predictive distribution, where the epistemic component is given by the *mutual information* I [15, 10, 8, 17, 27]. Recently, concerns about the properness of this current measure have been raised [41]. By further examining the current measure, we discern that the core issue lies in the assumption that the BMA predictive distribution is equivalent to the predictive distribution of the true model. Therefore, we propose to use a different information-theoretic measure that does not build upon the BMA, where the epistemic component is given by the *expected pairwise KL-divergence* K [24].

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

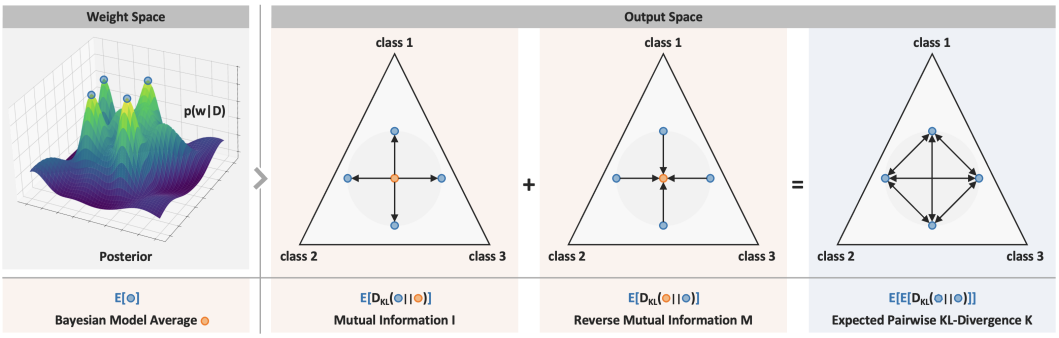

Figure 1: Relationship between epistemic components of the current and our introduced measure.

K is an upper bound of I, with the difference being termed *reverse mutual information* M [26] (Fig. 1). The aleatoric component is the same for both measures. Our main contributions are as follows:

- We provide new insights regarding the limitations of the current measure of predictive uncertainty.
- We introduce a theoretically grounded measure that addresses those limitations (Eq. (4)).
- We investigate the empirical benefits of using the theoretically grounded measure.

2 Analyzing the Current Measure of Predictive Uncertainty

Many machine learning models used for classification and regression yield the distribution parameters of a predictive distribution as their outputs. The softmax outputs of a classifier define a categorical distribution and the outputs of a regressor correspond to the mean (and variance, see e.g. [19, 20]) of a continuous (e.g. Gaussian) distribution. A machine learning model with model parameters $\boldsymbol{w}$ is used to obtain estimates of the distribution parameters of the true predictive distribution $p(\boldsymbol{y} \mid \boldsymbol{x})$ that generated the data. Therefore, the predictive distribution under the model is denoted as $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})$. We assume the dataset $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ is given, thus do not consider uncertainty due to how the dataset was sampled from $p(\boldsymbol{x}, \boldsymbol{y})$. Furthermore, we assume that the true model that created the dataset can be represented by the chosen model class. Those are common and often necessary assumptions [17]. From a Bayesian point of view, we can assign a posterior probability $p(\boldsymbol{w} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{w})p(\boldsymbol{w})$ of how probable it is that certain model parameters $\boldsymbol{w}$ are the true model parameters $\boldsymbol{w}^*$. For the true model parameters, $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}^*) = p(\boldsymbol{y} \mid \boldsymbol{x})$. The posterior distribution allows marginalization, yielding the BMA predictive distribution $p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) = \mathbb{E}_{p(\boldsymbol{w}\mid\mathcal{D})}\left[p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\right]$. This expectation is intractable and generally approximated by Monte Carlo sampling of model parameters using approximate inference techniques [23, 28, 39, 3, 11, 20, 40].

The most common measure of predictive uncertainty is the Shannon-entropy $\mathrm{H}(\cdot)$ [34] of the BMA predictive distribution [15, 10, 35, 8, 17, 27], decomposing into aleatoric and epistemic components:

$$\underbrace{\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}))}_{\text{total}} = \underbrace{\mathbb{E}_{p(\boldsymbol{w}\mid\mathcal{D})}\left[\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}))\right]}_{\text{aleatoric}} + \underbrace{\mathrm{I}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D}))}_{\text{epistemic}} . \tag{1}$$

The aleatoric component best estimates the true aleatoric uncertainty $\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}^*))$, as it accounts for all possible models according to their posterior probability. The epistemic component is the *mutual information* $\mathrm{I}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D}))$, which measures the reduction of uncertainty about outcomes $\boldsymbol{y}$ through observing model parameters $\boldsymbol{w}$. Recently, the properness of the measure given by Eq. (1) has been questioned [41]. Rewriting Eq. (1) reveals (detailed steps given in Sec. A.1 in the appendix), that mutual information is equal to the expected KL-divergence $\mathrm{D}_{\mathrm{KL}}(\cdot \,||\, \cdot)$ between the predictive distributions of possibly selected models and the BMA [33, 41]. Furthermore, the entropy of the BMA predictive distribution is equal to the expected cross-entropy $\mathrm{CE}(\cdot, \cdot)$ between the predictive distributions of possibly selected models and the BMA:

$$\underbrace{\mathbb{E}_{p(\boldsymbol{w}\mid\mathcal{D})}\left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \;,\; p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}))\right]}_{\text{total}} \tag{2}$$

$$= \underbrace{\mathbb{E}_{p(\boldsymbol{w}\mid\mathcal{D})}\left[\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}))\right]}_{\text{aleatoric}} + \underbrace{\mathbb{E}_{p(\boldsymbol{w}\mid\mathcal{D})}\left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \,||\, p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}))\right]}_{\text{epistemic}}$$

Generally, the KL-divergence between two distributions $\mathrm{D}_{\mathrm{KL}}(p \mid\mid p^*)$ quantifies the additional surprisal when sampling according to p^* instead of p [6]. In this setting, the KL-divergence should quantify the epistemic uncertainty, resulting from sampling according to some model's predictive distribution $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})$ instead of the true model's predictive distribution $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}^*)$. However, the epistemic component in Eq. (2) uses the BMA predictive distribution as a surrogate for the predictive distribution of the true model. Those do not coincide in general! From a Bayesian point of view, any model parameters could be the true model parameters according to their posterior probability. Furthermore, there could be no model that has nonzero posterior probability and equivalent predictive distribution to the BMA.

3 Introducing a Theoretically Grounded Measure of Predictive Uncertainty

We introduce a theoretically grounded measure of predictive uncertainty that does not assume that the BMA predictive distribution is equivalent to the true model's predictive distribution. To assess predictive uncertainty, we want to characterize the stochasticity of sampling outcomes from the predictive distribution $p(\boldsymbol{y} \mid \boldsymbol{x})$ that generated the dataset. The most sensible information-theoretic approach is the entropy of the predictive distribution $\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}))$, capturing the true aleatoric uncertainty [17]. As stated in Sec. 2, the distribution parameters of the predictive distribution are usually estimated using a model with model parameters $\boldsymbol{w}$. Therefore, an estimate for the aleatoric uncertainty under a given model is $\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}))$.

This approximation gives rise to epistemic uncertainty, as the selected model parameters $\boldsymbol{w}$ are generally not the true model parameters $\boldsymbol{w}^*$. We want to measure the additional uncertainty due to the mismatch of predictive distributions under $\boldsymbol{w}$ and $\boldsymbol{w}^*$. This can be done using the KL-divergence $\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}^*))$, as elaborated for the epistemic component of Eq. (2). Unfortunately, we don't know the true model parameters $\boldsymbol{w}^*$ in general. However, the posterior distribution $p(\boldsymbol{w} \mid \mathcal{D})$ expresses how likely certain model parameters are the true model parameters. Therefore, we can perform an expectation over the model posterior: $\mathbb{E}_{p(\tilde{\boldsymbol{w}} \mid \mathcal{D})}\left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]$. This represents the expected mismatch between the predictive distribution under certain model parameters and all potential parameters, weighted by how likely they are the true parameters. Therefore, the predictive uncertainty of a specific model $\boldsymbol{w}$ is given by

$$\begin{aligned} &\underbrace{\mathbb{E}_{p(\tilde{\boldsymbol{w}} \mid \mathcal{D})}\left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \;,\; p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]}_{\text{total}} \\ &\quad = \underbrace{\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}))}_{\text{aleatoric}} + \underbrace{\mathbb{E}_{p(\tilde{\boldsymbol{w}} \mid \mathcal{D})}\left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]}_{\text{epistemic}} . \end{aligned} \tag{3}$$

This measure was introduced in Schweighofer et al. [33] to quantify the predictive uncertainty of a given, pre-selected model. This is a critical aspect when deploying a machine learning model in real-world applications. However, it expresses a subjective uncertainty about the prediction of a specific model. In contrast, the measure given by Eq. (2) expresses the expected uncertainty when selecting a model according to the posterior. Thus, we propose to take a posterior expectation of Eq. (3), resulting in a measure of predictive uncertainty that does not share the limitations of Eq. (2):

$$\begin{aligned} &\underbrace{\mathbb{E}_{p(\boldsymbol{w} \mid \mathcal{D})}\left[\mathbb{E}_{p(\tilde{\boldsymbol{w}} \mid \mathcal{D})}\left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \;,\; p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]\right]}_{\text{total}} \\ &\quad = \underbrace{\mathbb{E}_{p(\boldsymbol{w} \mid \mathcal{D})}\left[\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}))\right]}_{\text{aleatoric}} + \underbrace{\mathbb{E}_{p(\boldsymbol{w} \mid \mathcal{D})}\left[\mathbb{E}_{p(\tilde{\boldsymbol{w}} \mid \mathcal{D})}\left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]\right]}_{\text{epistemic}} . \end{aligned} \tag{4}$$

The aleatoric component of Eq. (4) and Eq. (2) are the same. What differs is the epistemic component, where in Eq. (4) it is a pairwise comparison between the predictive distributions of possible models and weights according to their posterior probability. The epistemic component of Eq. (4) was discussed by Malinin et al. [24] as a measure of ensemble diversity, called the *expected pairwise KL-divergence* $\mathrm{K}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D}))$. However, in follow-up work, Malinin et al. [26] erroneously concluded that solely the mutual information 'cleanly' decomposes into total and aleatoric uncertainty.

Relation between measures. The expected pairwise KL-divergence is an upper bound of the mutual information by Jensen's inequality. The difference between those two is called the *reverse mutual information* [26], defined as $\mathrm{M}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) = \mathbb{E}_{p(\tilde{\boldsymbol{w}} \mid \mathcal{D})}\left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]$.

Table 1: AUROC using uncertainty measures as score to distinguish in-distribution (ImageNet-1K) and out-of-distribution samples (ImageNet-O) as well as natural adversarial examples (ImageNet-A).

Task	Method	Total		Aleatoric	Epistemic	
		Eq. (2)	Eq. (4)		Eq. (2)	Eq. (4)
ImageNet-O	cSG-HMC	$.609_{\pm.007}$	$\mathbf{.611}_{\pm.007}$	$.606_{\pm.007}$	$.675_{\pm.006}$	$\mathbf{.689}_{\pm.007}$
	MCD	$.631_{\pm.004}$	$\mathbf{.633}_{\pm.004}$	$.629_{\pm.004}$	$.682_{\pm.008}$	$\mathbf{.720}_{\pm.006}$
	DE (LL)	$\mathbf{.600}_{\pm.005}$	$\mathbf{.600}_{\pm.005}$	$.600_{\pm.005}$	$.561_{\pm.002}$	$\mathbf{.605}_{\pm.008}$
	DE (all)	$.703_{\pm.004}$	$\mathbf{.709}_{\pm.005}$	$.696_{\pm.004}$	$.717_{\pm.007}$	$\mathbf{.718}_{\pm.008}$
ImageNet-A	cSG-HMC	$.677_{\pm.001}$	$\mathbf{.687}_{\pm.001}$	$.666_{\pm.001}$	$.785_{\pm.000}$	$\mathbf{.792}_{\pm.000}$
	MCD	$.795_{\pm.002}$	$\mathbf{.797}_{\pm.002}$	$.794_{\pm.002}$	$.829_{\pm.001}$	$\mathbf{.860}_{\pm.001}$
	DE (LL)	$\mathbf{.819}_{\pm.003}$	$\mathbf{.819}_{\pm.003}$	$.818_{\pm.003}$	$.694_{\pm.003}$	$\mathbf{.813}_{\pm.002}$
	DE (all)	$.887_{\pm.002}$	$\mathbf{.892}_{\pm.002}$	$.879_{\pm.002}$	$\mathbf{.890}_{\pm.002}$	$.880_{\pm.002}$

Furthermore, the expected pairwise KL-divergence, the mutual information and the reverse mutual information satisfy $\mathrm{K}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) = \mathrm{I}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) + \mathrm{M}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D}))$ [26]. A proof is given in Sec. A.2 in the appendix. In Sec. 2 we concluded that the core problem of the current measure of predictive uncertainty is, that it assumes the BMA predictive distribution is equivalent to the predictive distribution under the true model. The reverse mutual information exactly accounts for this mismatch, by measuring the KL-divergence between the predictive distributions under all possible models weighted by their posterior probability and the BMA predictive distribution.

Related Work. We discuss other potential measures of uncertainty, not necessarily grounded in information theory in Sec. C in the appendix.

4 Experiments

Illustrative Example. First, we follow Wimmer et al. [41] and investigate an illustrative example of different posterior distributions of the parameter θ of a Bernoulli distribution. More details and results are given in section B.1 in the appendix. The results indicate, that the new measure (Eq. (4)) behaves more meaningfully than the current measure (Eq. (2)).

ImageNet. Second, we investigated the common tasks of out-of-distribution (OOD) detection and adversarial example detection, using the predictive uncertainty as a scoring function [29, 9, 2, 27, 33]. This large-scale experiment was conducted on ImageNet-1K [7], using ImageNet-O [14] for OOD detection and ImageNet-A [14] for adversarial example detection. This experiment utilizes versions of the EfficientNet models [36]. We compare cyclical Stochastic Gradient Hamiltonian Monte Carlo (cSG-HMC) [42], Monte Carlo Dropout (MCD) [11] and Deep Ensembles [20], ensembling only the last layer (DE (LL)), as well as ensembling over different pre-trained models (DE (all)). Details and further experiments are given in Sec. B.2 in the appendix. The results are given in Tab. 1.

The results show, that using the new measure of (total) predictive uncertainty (Eq. (4)) is superior to using the current measure (Eq. (2)) for three out of four methods, and equally good for the fourth method. Also, using the epistemic component of the new measure (Eq. (4)) is, except for DE (all) on ImageNet-A, superior to using the epistemic component of the current measure (Eq. (2)).

5 Conclusion and Future Work

We analyzed the limitations of current measures of predictive uncertainty, giving the new insight that their deficiency comes from assuming the BMA predictive distribution is the true predictive distribution. Therefore, we introduced a principled measure of predictive uncertainty that address those limitations. We showed that this new set of measures exhibits more sensible behavior and improves performance on common tasks where measures of predictive uncertainty are utilized. Future work should shed light on how those measures compare in active learning settings, where epistemic uncertainty is used to select the most informative datapoints.

Acknowledgements

The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria. We thank the projects AI-MOTION (LIT-2018-6-YOU-212), DeepFlood (LIT-2019-8-YOU-213), Medical Cognitive Computing Center (MC3), INCONTROL-RL (FFG-881064), PRIMAL (FFG-873979), S3AI (FFG-872172), DL for GranularFlow (FFG-871302), EPILEPSIA (FFG-892171), AIRI FG 9-N (FWF-36284, FWF-36235), AI4GreenHeatingGrids(FFG- 899943), INTEGRATE (FFG-892418), ELISE (H2020-ICT-2019-3 ID: 951847), Stars4Waters (HORIZON-CL6-2021-CLIMATE-01-01). We thank Audi.JKU Deep Learning Center, TGW LOGISTICS GROUP GMBH, Silicon Austria Labs (SAL), FILL Gesellschaft mbH, Anyline GmbH, Google, ZF Friedrichshafen AG, Robert Bosch GmbH, UCB Biopharma SRL, Merck Healthcare KGaA, Verbund AG, GLS (Univ. Waterloo) Software Competence Center Hagenberg GmbH, TÜV Austria, Frauscher Sensonic, TRUMPF and the NVIDIA Corporation.

References

[1] George Apostolakis. The concept of probability if safety assessments of technological systems. *Science*, 250(4986):1359–1364, 1990.

[2] Neil Band, Tim G. J. Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. Benchmarking bayesian deep learning on diabetic retinopathy detection tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 07–09 Jul 2015. PMLR.

[4] John Bradshaw, Alexander G. de G. Matthews, and Zoubin Ghahramani. Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. *arXiv*, 1707.02476, 2017.

[5] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, page 3–14, New York, NY, USA, 2017. Association for Computing Machinery.

[6] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1184–1193. PMLR, 10–15 Jul 2018.

[9] Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv*, 1912.10481, 2019.

[10] Yarin Gal. *Uncertainty in deep learning*. PhD thesis, University of Cambridge, 2016.

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

[12] Jon C. Helton. Risk, uncertainty in risk, and the EPA release limits for radioactive waste disposal. *Nuclear Technology*, 101(1):18–39, 1993.

[13] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv*, 1610.02136, 2018.

[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, June 2021.

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

[16] Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 548–557. PMLR, 01–05 Aug 2022.

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

[18] Edwin Thompson Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

[19] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[21] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1):17816, Dec 2017.

[22] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7498–7512. Curran Associates, Inc., 2020.

[23] David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, 05 1992.

[24] Andrey Malinin. *Uncertainty estimation in deep learning with application to spoken language assessment*. PhD thesis, University of Cambridge, 2019.

[25] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[26] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

[27] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H.S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24384–24394, June 2023.

[28] Radford M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.

[29] 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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[31] Janis Postels, Mattia Segù, Tao Sun, Luca Daniel Sieber, Luc Van Gool, Fisher Yu, and Federico Tombari. On the practicality of deterministic epistemic uncertainty. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17870–17909. PMLR, 17–23 Jul 2022.

[32] Yusuf Sale, Michele Caprio, and Eyke Höllermeier. Is the volume of a credal set a good measure for epistemic uncertainty? In Robin J. Evans and Ilya Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1795–1804. PMLR, 31 Jul–04 Aug 2023.

[33] Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, Günter Klambauer, and Sepp Hochreiter. Quantification of uncertainty with adversarial models. *arXiv*, 2307.03217, 2023.

[34] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[35] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. In A. Globerson and R. Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 560–569. AUAI Press, 2018.

[36] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[37] Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. On feature collapse and deep kernel learning for single forward pass uncertainty. *arXiv*, 2102.11409, 2022.

[38] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9690–9700. PMLR, 13–18 Jul 2020.

[39] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688, Madison, WI, USA, 2011. Omnipress.

[40] Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4697–4708. Curran Associates, Inc., 2020.

[41] 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.

[42] Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. In *International Conference on Learning Representations*, 2020.

A Theoretical Results

A.1 Equivalence of Eq. (1) and Eq. (2)

We want to show that Eq. (1) and Eq. (2) are equivalent. The aleatoric component is already the same for both. Therefore, we need to show that the total components are equivalent:

$$\begin{aligned}
\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})) &= \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})}\left[-\log p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})\right] && (5)\\
&= \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})}\left[-\log p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})\right]\right] && (6)\\
&= \mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\ ,\ p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}))\right] && (7)
\end{aligned}$$

Furthermore, we need to show that the epistemic components are equivalent:

$$\begin{aligned}
\mathrm{I}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) &= \mathbb{E}_{p(\boldsymbol{y},\boldsymbol{w}|\boldsymbol{x},\mathcal{D})}\left[\log \frac{p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})}{p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})\, p(\boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})}\right] && (8)\\
&= \mathbb{E}_{p(\boldsymbol{y},\boldsymbol{w}|\boldsymbol{x},\mathcal{D})}\left[\log \frac{p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\, p(\boldsymbol{w} \mid \mathcal{D})}{p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})\, p(\boldsymbol{w} \mid \mathcal{D})}\right] && (9)\\
&= \mathbb{E}_{p(\boldsymbol{y},\boldsymbol{w}|\boldsymbol{x},\mathcal{D})}\left[\log \frac{p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})}{p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})}\right] && (10)\\
&= \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})}\left[\log \frac{p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})}{p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})}\right]\right] && (11)\\
&= \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}))\right] && (12)
\end{aligned}$$

This holds, as $\boldsymbol{y}$ depends on $\boldsymbol{w}$, i.e. the model $\boldsymbol{w}$ has to be selected before being able to draw $\boldsymbol{y}$ from its predictive distribution. Consequently, Eq. (1) and Eq. (2) are equivalent.

□

A.2 Proof of Additive Decomposition of Expected Pairwise KL-Divergence into Mutual Information and Reverse Mutual Information

Given are the definitions of the expected pairwise KL-divergence

$$\mathrm{K}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) \;=\; \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]\right]\ , \tag{13}$$

the mutual information

$$\mathrm{I}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) \;=\; \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})}\left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}))\right]\ , \tag{14}$$

and the reverse mutual information

$$\mathrm{M}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) \;=\; \mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}\left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]\ . \tag{15}$$

We want to show that

$$\mathrm{K}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) = \mathrm{I}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) \;+\; \mathrm{M}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D}))\ . \tag{16}$$

Note that $p(\tilde{\boldsymbol{w}} \mid \mathcal{D}) = p(\boldsymbol{w} \mid \mathcal{D})$ is used redundantly to keep track of integration variables. The proof is as follows.

$$\mathrm{K}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) \tag{17}$$

$$= \mathrm{I}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) \; + \; \mathrm{M}(p(\boldsymbol{y}, \boldsymbol{w} \mid \boldsymbol{x}, \mathcal{D})) \tag{18}$$

$$= \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}))\right] \; + \; \mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right] \tag{19}$$

$$= \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})} \left[\log \frac{p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})}{p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})}\right]\right] \; + \; \mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})} \left[\log \frac{p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})}{p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})}\right]\right] \tag{20}$$

$$= \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\right] \; - \; \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})\right]\right] \; + \tag{21}$$
$$\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})\right] \; - \; \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right]\right]$$

$$= \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\right]\right] \; - \; \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})\right] \; + \tag{22}$$
$$\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D})\right] \; - \; \mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right]\right]$$

$$= \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\right]\right] \; - \tag{23}$$
$$\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right]\right]\right]$$

$$= \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\right]\right]\right] \; - \tag{24}$$
$$\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})} \left[\log p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})\right]\right]\right]$$

$$= \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})} \left[\log \frac{p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})}{p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}})}\right]\right]\right] \tag{25}$$

$$= \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]\right] \; , \tag{26}$$

which is exactly the definition of K in (13). The step from (21) to (22) is due to additivity and linearity of expectations. The step from (23) to (24) is due to the fact that we can insert the expectation $\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})}$ in the first term as it does not depend on $\tilde{\boldsymbol{w}}$ and due to the fact that $p(\tilde{\boldsymbol{w}} \mid \mathcal{D}) = p(\boldsymbol{w} \mid \mathcal{D})$.

□

A.3 Overview of Measures of Uncertainty

In the following, we give an overview of the measures of uncertainty discussed in the main paper. As the aleatoric components are equivalent, we categorize them by the epistemic component. The current measure of predictive uncertainty, where the epistemic component is the mutual information is given by

$$\underbrace{\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \; , \; p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}))\right]}_{\text{total}} \tag{27}$$
$$= \underbrace{\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}))\right]}_{\text{aleatoric}} + \underbrace{\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}))\right]}_{\text{epistemic}} \; .$$

Our new measure of predictive uncertainty, where the epistemic component is the expected pairwise KL-divergence is given by

$$\underbrace{\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \; , \; p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]\right]}_{\text{total}} \tag{28}$$
$$= \underbrace{\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathrm{H}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}))\right]}_{\text{aleatoric}} + \underbrace{\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]\right]}_{\text{epistemic}} \; .$$

Furthermore, it is possible to decompose our new measure of predictive uncertainty, such that the reverse mutual information is the epistemic component:

$$\underbrace{\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \; , \; p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]\right]}_{\text{total (Eq. (28))}} \tag{29}$$
$$= \underbrace{\mathbb{E}_{p(\boldsymbol{w}|\mathcal{D})} \left[\mathrm{CE}(p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}) \; , \; p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}))\right]}_{\text{total (Eq. (27))}} + \underbrace{\mathbb{E}_{p(\tilde{\boldsymbol{w}}|\mathcal{D})} \left[\mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) \mid\mid p(\boldsymbol{y} \mid \boldsymbol{x}, \tilde{\boldsymbol{w}}))\right]}_{\text{epistemic}} \; .$$

However, this does not yield an exclusively aleatoric component, but the (total) predictive uncertainty of Eq. (27). This is also signified by the coloring of the respective terms in Eq. (29).

B Experimental Results

B.1 Illustrative Example: Bernoulli Distribution

Setup. Following Wimmer et al. [41], we analyze the behaviour of different measures of uncertainty using a posterior distribution over the distribution parameter $\theta \in [0, 1]$ of a Bernoulli distribution. In this setting, *no model is involved*, thus the distribution parameter θ is directly sampled from the posterior distribution. Furthermore, we drop the dependence on a specific input $\boldsymbol{x}$, assuming that the input gives no information about the outcome. For the sake of an example, consider a machine where one can press a button and one of two signals, red (0) or green (1) will light up. The task is to predict which signal will light up.

We analyze the behavior of the measures of uncertainty under several closed-form posterior distributions, such as uniform, beta and a mixture of delta distributions. Those correspond to different knowledge about which signal will light up. For example, a delta distribution at $\theta = 0$ would correspond to that it is known the outcome is either green or red with equal probability. Thus the outcome is maximally uncertain, due to aleatoric uncertainty. A different case would be a uniform posterior distribution, thus every parameter $\theta \in [0, 1]$ is equally probable. There should be less aleatoric uncertainty, as parameters that lead to Bernoulli distributions certain in their prediction are also possible. However, epistemic uncertainty must be higher, as all parameters are equally probable. Where should we expect the highest epistemic uncertainty? One point could be made that we should expect the highest epistemic uncertainty under the uniform distribution, due to the principle of maximum entropy [18]. Therefore, if all parameters are equally probable, epistemic uncertainty should be highest. However, this captures uncertainty in the parameter space, not in the output space, which is what we are concerned about for predictive uncertainty. We don't care, how many models are possible, but about how different their predictions are. Therefore, we argue that a mixture of two delta distributions at the two extreme parameters $\theta = 0$ and $\theta = 1$ should correspond to the highest epistemic uncertainty in this example.

How could such a situation occur? This reflects the prior belief that the machine in our example is deterministic in nature, thus no matter how often the button is pressed, the same signal will light up. However, before trying it once, it is unknown which signal will light up. Therefore, when choosing the parameter corresponding to the green light, there would be maximal surprisal when observing the red light. Yet after observing a single outcome, the posterior will update, collapsing to a delta distribution at a single parameter, removing all uncertainty about the prediction. Still, before observing a single outcome, the epistemic uncertainty should be maximal.

Results. Different posterior distributions and associated total (TU), aleatoric (AU) and epistemic uncertainty (EU) under the considered measures of uncertainty are depicted in Fig. 2.

For the current measure of uncertainty given by Eq. (2), results are given in the first row above each plot, written in red. We find that total uncertainty is maximal if the Bernoulli parameter of the BMA is $\theta = 0.5$. This does not depend on how the probability mass of the posterior is distributed, but only on the expectation. For the introduced measure of uncertainty given by Eq. (4), results are given in the second row above each plot, written in black. We find, that the measure for epistemic uncertainty gives infinity, and consequently also the measure for total uncertainty. This corresponds exactly to the expected behavior of an information-theoretic measure of uncertainty about the outcome. We also consider the reverse mutual information as a standalone measure of epistemic uncertainty [26]. Results are given in the third row above each plot, written in violet. We find, that the behavior of this measure of epistemic uncertainty is similar to the epistemic component of the measure given by Eq. (4).

Another Limitation of Current Measure of Uncertainty. Another limitation of the current measure of uncertainty given by Eq. (2) is exemplified in Fig. 3. Here, all three posterior distributions lead to equal total, aleatoric, and epistemic uncertainty. Such examples are easy to construct. The total uncertainty given by Eq. (2) (better seen in the equivalent formulation in Eq. (1)) only depends on the expected value (the BMA) of the Bernoulli parameter θ. By fixing the expected value to e.g. $\theta = 0.5$, it is trivial to search for distribution parameters that lead to equal aleatoric and, due to additivity, also to equal epistemic uncertainty. Using the measure of uncertainty given by Eq. (2) resolves the ambiguity, assigning different epistemic and therefore also total uncertainty to the three posterior distributions.

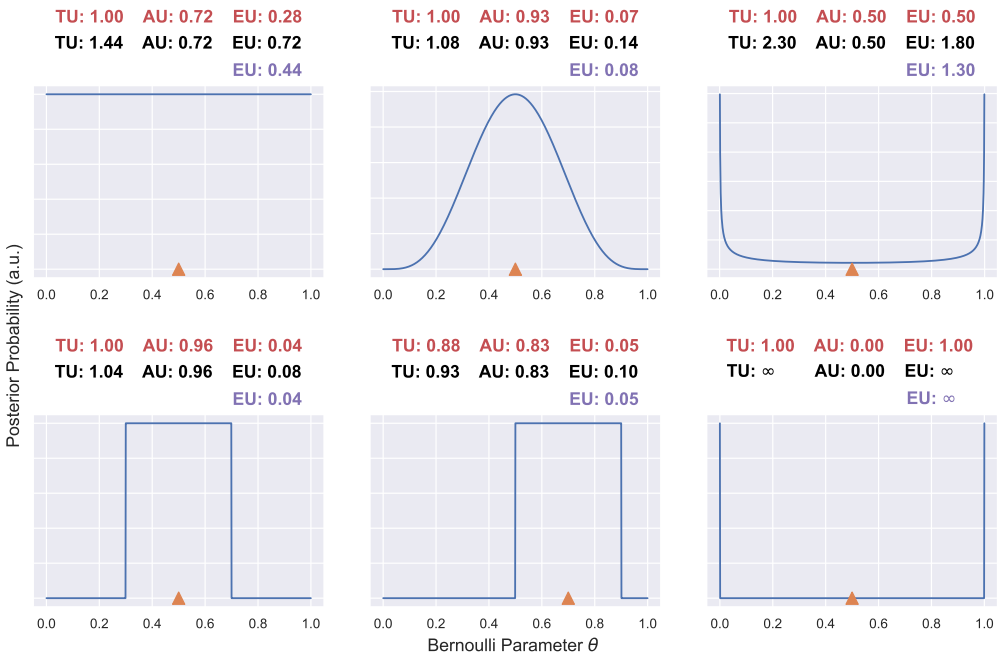

Figure 2: Different posterior distributions $p(\theta \mid \mathcal{D})$ for the parameter θ of a Bernoulli distribution. Uncertainties using the current measure of predictive uncertainty (Eq. (2)), the introduced measure of predictive uncertainty (Eq. (4)), and the reverse mutual information as measure for epistemic uncertainty (Eq. (29)). Orange triangles denote the expected value of θ under the posterior distribution. Posterior distributions from left to right, per row: $\mathcal{U}[0, 1]$, $Beta(5, 5)$, $Beta(0.4, 0.4)$, $\mathcal{U}[0.3, 0.7]$, $\mathcal{U}[0.5, 0.9]$, $\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$.

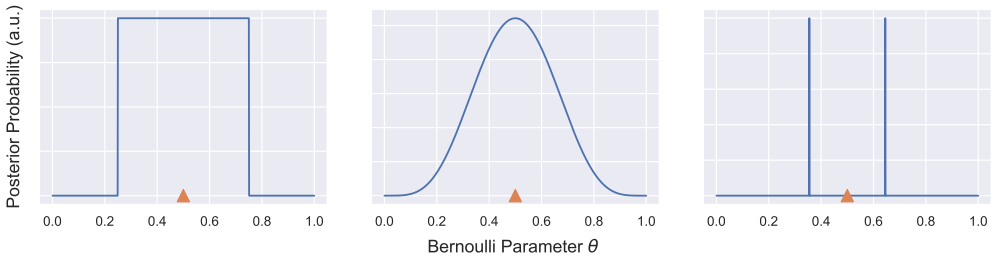

Figure 3: Three different posterior distributions $p(\theta \mid \mathcal{D})$ (Uniform, Beta, Mixture of delta distributions) for the parameter θ of a Bernoulli distribution. All of them have the same total, aleatoric, and epistemic uncertainty according to Eq. (2). Those can easily be found, as the total uncertainty only depends on the expected value of θ under the posterior distribution, depicted by the orange triangle. Fixing the expected value of the posterior distribution, the distribution parameters are trivial to tune to exhibit the same aleatoric and consequently the same epistemic uncertainty. However, those posterior distributions exhibit different total and epistemic uncertainty according to Eq. (4).

Table 2: AUROC using different uncertainty measures as a score to distinguish between correctly and misclassified samples on the ImageNet-1K validation set. For selective prediction, the AUC of accuracy vs. fraction of most certain samples is reported.

Task	Method	Total		Aleatoric	Epistemic	
		Eq. (2)	Eq. (4)		Eq. (2)	Eq. (4)
Misclass.	cSG-HMC	$.700_{\pm.003}$	$\mathbf{.705}_{\pm.003}$	$.693_{\pm.003}$	$.758_{\pm.010}$	$\mathbf{.767}_{\pm.009}$
	MCD	$\mathbf{.867}_{\pm.007}$	$\mathbf{.867}_{\pm.007}$	$.866_{\pm.007}$	$.791_{\pm.012}$	$\mathbf{.831}_{\pm.011}$
	DE (LL)	$\mathbf{.910}_{\pm.007}$	$\mathbf{.910}_{\pm.007}$	$.910_{\pm.007}$	$.661_{\pm.002}$	$\mathbf{.824}_{\pm.006}$
	DE (all)	$\mathbf{.879}_{\pm.004}$	$.869_{\pm.005}$	$.883_{\pm.004}$	$\mathbf{.808}_{\pm.011}$	$.795_{\pm.013}$
Select. Pred.	cSG-HMC	$.911_{\pm.005}$	$\mathbf{.913}_{\pm.005}$	$.910_{\pm.005}$	$\mathbf{.926}_{\pm.007}$	$\mathbf{.926}_{\pm.006}$
	MCD	$\mathbf{.959}_{\pm.003}$	$\mathbf{.959}_{\pm.003}$	$.958_{\pm.003}$	$.934_{\pm.009}$	$\mathbf{.945}_{\pm.008}$
	DE (LL)	$\mathbf{.971}_{\pm.002}$	$\mathbf{.971}_{\pm.002}$	$.971_{\pm.002}$	$.892_{\pm.005}$	$\mathbf{.926}_{\pm.005}$
	DE (all)	$\mathbf{.968}_{\pm.001}$	$.966_{\pm.001}$	$.969_{\pm.001}$	$\mathbf{.953}_{\pm.003}$	$.950_{\pm.003}$

B.2 ImageNet Experiments

ImageNet experiments were conducted using the codebase of [33] with their default hyperparameters.

Additionally to the out-of-distribution (OOD) detection and adversarial example detection tasks reported in Tab. 1 of the main paper, we investigated the common tasks of misclassification detection and selective prediction, again using the predictive uncertainty as a scoring function [29, 9, 2, 27, 33]. We conducted those experiments on the official validation set of ImageNet-1K.

For each of the four experiments, we use pre-trained EfficientNet [36] architectures available through PyTorch [30]. To approximate the posterior expectations in Eq. (2) and Eq. (4), we utilize the standard approach of MC integration thus approximating the integral by the average over functions of model parameters (approximately) drawn from the posterior distribution. As an example, the BMA predictive distribution is approximated by

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) = \mathbb{E}_{p(\boldsymbol{w}\mid\mathcal{D})}\left[p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\right] \approx \frac{1}{N}\sum_{i=1}^{N} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w}_i) \;, \tag{30}$$

where $\boldsymbol{w}_i \sim p(\boldsymbol{w} \mid \mathcal{D})$. To sample different model parameters, we utilize cSG-HMC [42], MC dropout [11] and versions of Deep Ensembles [20]. cSG-HMC and MC dropout were performed on the last layer of the EfficientNetV2-S architecture with 2000 samples each. Deep Ensembles were used in two different versions. Once by ensembling 10 different pre-trained EfficientNet models with different architecture (DE (all)) and once by ensembling 10 different last layers of the EfficientNetV2-S architecture (DE (LL)).

To provide confidence intervals, we performed all experiments on three distinct dataset splits. Regarding OOD detection, each split consists of all 2000 Imagenet-O samples and 2000 unique ImageNet-1K samples. Regarding adversarial example detection, each split consists of all 7000 Imagenet-A samples and 7000 unique ImageNet-1K samples. Regarding misclassification detection and selective prediction, each spit consists of 7000 unique ImageNet-1K samples. ImageNet-1K samples were randomly drawn from the official validation set.

The additional results for misclassification detection and selective prediction are given in Tab. 2. The results show, that using the new measure of (total) predictive uncertainty given by Eq. (4), is always at least equally good and most of the time better then using the current measure of (total) predictive uncertainty given by Eq. (2). The only exception is DE (all) for both tasks. Results for the epistemic components of both measures are similar.

C Related Work

Mutual information as a measure of epistemic uncertainty (about the parameters of the model) and the respective decomposition into total and aleatoric uncertainty was introduced by Houlsby et al. [15]. Those measures have remained popular ever since [10, 35, 8, 17, 27]. Smith et al. [35] analyzed measures of uncertainty for adversarial example detection. Furthermore, they illustrate the relation of mutual information to another ad-hoc measure for uncertainty in classification settings, the softmax variance [21, 5]. Another ad-hoc measure of uncertainty in classification settings is the maximum softmax value of a classifier [20, 13, 10].

Depeweg et al. [8] considers the measures in Eq. (1), but also proposes to use the variance as a measure of uncertainty. Based upon the law of total variance, the variance of the BMA predictive distribution is decomposed to a posterior expectation over the variance of the predictive distributions given by individual models (the aleatoric component) and the posterior variance over the expected value of predictive distributions given by individual models (the epistemic component).

Malinin et al. [25] introduces measures similar to Eq. (1), but modeling the mutual information between the prediction and the data distribution. This considers epistemic uncertainty arising from distributional mismatch between the data distribution during training and inference. Furthermore, they consider the differential entropy of their introduced Dirichlet Prior Network, which captures the entropy of the predicted Dirichlet distribution.

Measuring predictive uncertainty through the information-theoretic measures discussed in this work are based upon posterior expectations, which are approximated by Monte-Carlo sampling in practice. This requires to sample multiple models and obtain predictions for them, making it expensive at inference time. Therefore, methods to estimate epistemic uncertainty with just a single pass through the network were considered [4, 22, 38, 37, 27]. They measure epistemic uncertainty via feature-space distance or feature-space densities. Thus, epistemic uncertainty is high if a new input is far from samples in the training set or the density of the training set is low. However, this requires that the feature space is meaningful, where feature collapse might severely impact the quality of the uncertainty estimate [37, 31].

Apart from distributional representations of uncertainty, set-based formalisms are alternative ways to express uncertainty, e.g. using credal sets [17, 16]. Sale et al. [32] investigated the volume of the credal set as a measure for epistemic uncertainty, but found it is only meaningful in the case of binary classification and less so for multiclass-classification.