# Credal Bayesian Deep Learning

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Uncertainty quantification and robustness to distribution shifts are important goals in machine learning and artificial intelligence. Although Bayesian Neural Networks (BNNs) allow for uncertainty in the predictions to be assessed, different sources of uncertainty are indistinguishable. We present Credal Bayesian Deep Learning (CBDL). Heuristically, CBDL allows to train an (uncountably) infinite ensemble of BNNs, using only finitely many elements. This is possible thanks to prior and likelihood finitely generated credal sets (FGCSs), a concept from the imprecise probability literature. Intuitively, convex combinations of a finite collection of prior-likelihood pairs are able to represent infinitely many such pairs. After training, CBDL outputs a set of posteriors on the parameters of the neural network. At inference time, such posterior set is used to derive a set of predictive distributions that is in turn utilized to distinguish between aleatoric and epistemic uncertainties, and to quantify them. The predictive set also produces either (i) a collection of outputs enjoying desirable probabilistic guarantees, or (ii) the single output that is deemed the best, that is, the one having the highest predictive lower probability – another imprecise-probabilistic concept. CBDL is more robust than single BNNs to prior and likelihood misspecification, and to distribution shift. We show that CBDL is better at quantifying and disentangling different types of uncertainties than single BNNs and ensemble of BNNs. In addition, we apply CBDL to two case studies to demonstrate its downstream tasks capabilities: one, for motion prediction in autonomous driving scenarios, and two, to model blood glucose and insulin dynamics for artificial pancreas control. We show that CBDL performs better when compared to an ensemble of BNNs baseline.

## 1 Introduction

One of the greatest virtues an individual can have is arguably being aware of their own ignorance, and acting cautiously as a consequence. Similarly, an autonomous system using neural networks (NNs) would greatly benefit from understanding the probabilistic properties of the NN's output (for example, its robustness to distribution shift), in order to incorporate them into any further decision-making. In this paper, we present a procedure that allows us to give a machine such a desirable quality.

In the last few years, there has been a proliferation of work on calibrating (classification) NNs, in order to estimate the confidence in their outputs (Guo et al., 2017) or to produce conformal sets that are guaranteed to contain the true label, in a probably approximately correct (PAC) sense (Park et al., 2020). While such methods are a promising first step, they require a calibration set (in addition to the original training set) and cannot be directly used on out-of-distribution data without further examples.

Bayesian neural networks (BNNs) offer one approach to overcome the above limitations. The Bayesian paradigm provides a rigorous framework to analyze and train uncertainty-aware neural networks, and more generally to support the development of learning algorithms (Jospin et al., 2022). In addition, it overcomes some of the drawbacks of deep learning models, namely that they are prone to overfitting, which adversely affects their generalization capabilities, and that they tend to be overconfident about their predictions when they provide a confidence interval. BNNs, though, are trained using a single prior, which may still suffer from miscalibration and robustness issues (Lenk & Orme, 2009).

In this work we introduce Credal Bayesian Deep Learning (CBDL), a procedure that draws on concepts from the imprecise probability (IP) literature (Augustin et al., 2014; Walley, 1991). Unlike other techniques in the fields of artificial intelligence (AI) and machine learning (ML) involving imprecise probabilities – that typically only focus on classification problems – CBDL can be used for both classification and regression. It captures the ambiguity the designer faces when selecting which prior to choose for the parameters of a neural network and which likelihood distribution to choose for the training data at hand.

CBDL can be thought of as a NN trained using prior and likelihood finitely generated credal sets (FGCSs), $\mathcal{P}_{\text{prior}}$ and $\mathcal{P}_{\text{lik}}$, respectively.[1] They are convex sets of probability measures having finitely many extreme elements (the elements that cannot be written as a convex combination of one another), $\text{ex}\mathcal{P}_{\text{prior}}$ and $\text{ex}\mathcal{P}_{\text{lik}}$, respectively (see also Remark 2). A very simple example of a prior FGCS $\mathcal{P}_{\text{prior}}$ is the collection of all the convex combinations of two one-dimensional Normal distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$, i.e.,

$$\mathcal{P}_{\text{prior}} = \{P : P = \beta\mathcal{N}(\mu_1, \sigma_1^2) + (1-\beta)\mathcal{N}(\mu_2, \sigma_2^2), \text{ for all } \beta \in [0,1]\}.$$

Consequently, in this toy example we have that $\text{ex}\mathcal{P}_{\text{prior}} = \{\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)\}$. FGCSs are further examined in section 2.2.

Given the use of finitely generated credal sets, CBDL can also be seen as a non-condensed (uncountably) infinite ensemble of BNNs – each BNN corresponding to a pair $(P, L)$ of prior $P$ from the prior FGCS $\mathcal{P}_{\text{prior}}$ and likelihood $L$ from the likelihood FGCS $\mathcal{P}_{\text{lik}}$. Such infinite ensemble, though, is carried out using only finitely many elements, that is, only pairs $(P^{\text{ex}}, L^{\text{ex}})$ of components of the extreme sets $\text{ex}\mathcal{P}_{\text{prior}}$ and $\text{ex}\mathcal{P}_{\text{lik}}$. This because every element $P$ of $\mathcal{P}_{\text{prior}}$ can be obtained from a convex combinations of the elements of $\text{ex}\mathcal{P}_{\text{prior}}$, and similarly for the likelihood FGCS. After training, CBDL produces a posterior FGCS $\mathcal{P}_{\text{post}}$ on the parameters of the neural network. At inference time, $\mathcal{P}_{\text{post}}$ is used to derive a predictive FGCS $\mathcal{P}_{\text{pred}}$, that is, given a new input, a set of plausible distributions over the space of outputs.[2] In turn, $\mathcal{P}_{\text{pred}}$ generates a set of outputs – or a single output, depending on the user's needs – that enjoys desirable probabilistic guarantees. CBDL also gives a way of quantifying and disentangling different types of uncertainties within $\mathcal{P}_{\text{pred}}$.

We use a credal set approach to overcome some of the drawbacks of single BNNs. In particular, CBDL allows to counter the criticism to the practice in (standard) Bayesian statistics of (i) using a single, arbitrary prior to represent the initial state of ignorance of the agent, (ii) using non-informative priors to model ignorance, and (iii) using a single, arbitrary likelihood to represent the agent's knowledge about the sampling model.[3] As a consequence, credal sets make the analysis more robust to prior and likelihood misspecification. In addition, they make it possible to quantify and distinguish between epistemic and aleatoric uncertainties (EU and AU, respectively). This is desirable in light of several areas of recent ML research, such as Bayesian deep learning (Depeweg et al., 2018; Kendall & Gal, 2017a), adversarial example detection (Smith & Gal, 2018), and data augmentation in Bayesian classification (Kapoor et al., 2022).

AU refers to the uncertainty that is inherent to the data generating process; as such, it is irreducible. Think, for example, of a coin toss. No matter how many times the coin is tossed, the stochastic variability of the experiment cannot be eliminated. EU, instead, refers to the lack of knowledge about the data generating process; as such, it is reducible. It can be lessened on the basis of additional data. For example, after only a few tosses, we are unable to gauge whether a coin is biased or not, but if we repeat the experiment long enough, this type of uncertainty vanishes. We note in passing that EU cannot be captured using a single BNN (Hüllermeier & Waegeman, 2021). This because selecting a unique prior and a unique likelihood implicitly assumes perfect knowledge around the true prior and the true data generating process. In turn, a unique distribution is only able to retrieve AU. Methods that disentangle between the two types of uncertainties that are based on a single distribution are ad-hoc, but are not theoretically well-justified. EU can be typically reduced by retraining the model using an augmented training set (Lin et al., 2023) (e.g., via semantic preserving transformations (Kaur et al., 2023), Puzzle Mix (Kim et al., 2020), etc.). On the other hand,

[1] Intuitively, the larger these credal sets, the higher prior and likelihood ambiguity the user faces.

[2] As we shall see in section 3.1, since computing the posteriors and the predictive distributions is oftentimes an intractable problem, we approximate the elements of $\mathcal{P}_{\text{post}}$ and of $\mathcal{P}_{\text{pred}}$ using Variational Inference (VI). As a consequence, we denote the VI-approximated posterior and predictive credal sets as $\hat{\mathcal{P}}_{\text{post}}$ and $\hat{\mathcal{P}}_{\text{pred}}$, respectively.

[3] Criticisms (i) and (iii) are also pointed out in Manchingal & Cuzzolin (2022, Section 2.2).

since AU is irreducible, there is an increasing need for ML techniques that are able to detect an excess of AU and query for human help.

**Remark 1.** *EU should not be confused with the concept of epistemic probability (de Finetti, 1974; 1975; Walley, 1991). In the subjective probability literature, epistemic probability can be captured by a single distribution. Its best definition can be found in Walley (1991, Sections 1.3.2 and 2.11.2). There, the author specifies how epistemic probabilities model logical or psychological degrees of partial belief of the agent. We remark, though, how de Finetti and Walley work with finitely additive probabilities, while in this paper we use countably additive probabilities.*

The motivation for working with credal sets is threefold: (i) it allows to be robust against prior and likelihood misspecification; (ii) unlike when using a single probability distribution, it permits to represent ignorance in the sense of lack of knowledge; (iii) it allows to quantify and disentangle between EU and AU. A more in-depth discussion can be found in Appendix A. In addition, we point out that despite a hierarchical Bayesian model (HBM) approach seems to be a viable alternative to one based on credal sets, these latter are better justified philosophically, and do not suffer from the same theoretical shortcomings of HBM procedures (Bernardo, 1979; Hüllermeier & Waegeman, 2021; Jeffreys, 1946; Walley, 1991). A more detailed explanation can be found in Appendix B. Let us mention that, given our motivations, comparing CBDL with non-Bayesian techniques like the evidential-theory-based methodologies (Amini et al., 2020; Charpentier et al., 2020; Sensoy et al., 2018) or conformal prediction (Vovk et al., 2022; Shafer & Vovk, 2008; Gibbs et al., 2023; Barber et al., 2023) needs a separate treatment, and is beyond the current scope. We will delve into it in future work.

We summarize our contributions next: (1) We present CBDL, and develop the theoretical tools and the algorithm required to use it in practice. (2) We show that a CBDL approach is more robust than single BNNs to prior and likelihood misspecification, and to distribution shifts. We also explain how, during inference, the credal set of posteriors $\mathcal{P}_{\text{post}}$ on the network parameters obtained during training is used to derive a credal set of predictive distributions $\mathcal{P}_{\text{pred}}$, and in turn a set of outcomes – or a single outcome – that enjoys probabilistic guarantees.[4] (3) We show how CBDL is better at quantifying and disentangling AU and EU than single BNNs and ensemble of BNNs. We also apply CBDL to two safety critical systems to demonstrate its downstream tasks capabilities. One, motion prediction for autonomous driving, and two, the human insulin and blood glucose dynamics for artificial pancreas control. We demonstrate improvements in both these settings with respect to ensemble of BNNs methods, the reason being that better uncertainty quantification leads to a positive impact on the decisions made using them.[5]

Before moving on, let us point out that while CBDL pays a computational price coming from the use of credal sets, it is able to quantify both EU and AU, unlike single BNNs, and to do so in a principled manner, unlike ensemble of BNNs. In addition, it requires less stringent assumptions on the nature of the prior and likelihood ambiguity faced by the agent than other imprecise-probabilities-based techniques, as explained in Appendices B and L. We also stress that if the user prioritizes computational efficiency, they should use backpropagation-based methods, as they are much faster than Bayesian techniques. In safety-critical situations instead – where using uncertainty-informed methods is crucial for quantifying the types of uncertainties, but also for the outcome of the analysis at hand – CBDL is a natural choice.

**Structure of the paper.** Section 2 presents the needed preliminary concepts, followed by section 3 that introduces and discusses the CBDL algorithm, together with its theoretical properties. We present our experimental results in section 4, and we examine the related work in section 5. Section 6 concludes our work. In the appendices, we give further theoretical and philosophical arguments and we prove our claims.

[4]Once again, as we shall see in section 3.1, we approximate the elements of the posterior and the predictive credal sets using Variational Inference (VI).

[5]We note in passing how recently Mucsányi et al. (2024) show that (finite) deep ensembles are the current state-of-the-art methods for quantifying and disentangling different types of uncertainties on ImageNet. Our experiments show that CBDL improves on (finite) ensembles of BNNs, both in uncertainty quantification and disentanglement, and in downstream tasks performances. Together with the findings in Mucsányi et al. (2024), this is an additional argument in favor of the effectiveness of our methodology.

## 2 Background and Preliminaries

In this section, we present the background notions that are needed to understand our main results. In section 2.1 we introduce Bayesian neural networks. Section 2.2 discusses (finitely generated) credal sets, upper and lower probabilities, and imprecise highest density regions. Section 2.3 introduces the concepts of upper and lower entropy, which are used to quantify and disentangle AU and EU. The reader that is familiar with these concepts can skip to section 3.

### 2.1 Bayesian Neural Networks

In line with the recent survey on BNNs by Jospin et al. (2022), Bayes' theorem can be stated as $P(H \mid D) = [P(D \mid H)P(H)]/P(D) = P(D, H)/\int P(D, H')\mathrm{d}H'$, where $H$ is a hypothesis about which the agent holds some prior beliefs, and $D$ is the data the agent uses to update their initial opinion. Probability distribution $P(D \mid H)$ represents how likely it is to observe data $D$ if hypothesis $H$ were to be true, and is called *likelihood*, while probability distribution $P(H)$ represents the agent's initial opinion around the plausibility of hypothesis $H$, and is called *prior*. The *evidence* available is encoded in $P(D) = \int P(D, H')\mathrm{d}H'$, while *posterior* probability $P(H \mid D)$ represents the agent's updated opinion. Using Bayes' theorem to train a predictor can be understood as learning from data $D$: the Bayesian paradigm offers an established way of quantifying uncertainty in deep learning models.

BNNs are stochastic artificial neural networks (ANNs) trained using a Bayesian approach (Goan & Fookes, 2020; Jospin et al., 2022; Lampinen & Vehtari, 2001; Titterington, 2004; Wang & Yeung, 2021). The goal of ANNs is to represent an arbitrary function $y = \Phi(x)$. Let $\theta$ represent the parameters of the network, and call $\Theta$ the space $\theta$ belongs to. Stochastic neural networks are a type of ANN built by introducing stochastic components to the network. This is achieved by giving the network either a stochastic activation or stochastic weights to simulate multiple possible models with their associated probability distribution. This can be summarized as $\theta \sim p(\theta)$, $y = \Phi_\theta(x) + \varepsilon$, where $\Phi$ depends on $\theta$ to highlight the stochastic nature of the neural network, $p$ is the density of a probability measure $P$ on $\Theta$,[6] and $\varepsilon$ represents random noise to account for the fact that function $\Phi_\theta$ is just an approximation. The connection between the BNN notation and the general one of the Bayes' theorem is explained in the next two paragraphs.

To design a BNN, the first step is to choose a deep neural network *architecture*, that is, functional model $\Phi_\theta$. Then, the agent specifies the *stochastic model*, that is, a prior distribution over the possible model parametrization $p(\theta)$, and a prior confidence in the predictive power of the model $p(y \mid x, \theta)$. Given the usual assumption that multiple data points from the training set are independent, the product $\prod_{(x,y)\in D} p(y \mid x, \theta)$ represents the *likelihood* of outputs $y \in D_{\mathbf{y}}$ given inputs $x \in D_{\mathbf{x}}$ and parameter $\theta$, where (a) $D = D_{\mathbf{x}} \times D_{\mathbf{y}}$ is the training set; (b) $D_{\mathbf{x}} = \{x_i\}_{i=1}^n$ is the collection of training inputs, which is a subset of the space $\mathcal{X}$ of inputs; (c) $D_{\mathbf{y}} = \{y_i\}_{i=1}^n$ is the collection of training outputs, which is a subset of the space $\mathcal{Y}$ of outputs.

The model parametrization can be considered to be hypothesis $H$. Following Jospin et al. (2022), we assume independence between model parameters $\theta$ and training inputs $D_{\mathbf{x}}$, in formulas $D_{\mathbf{x}} \perp\!\!\!\perp \theta$. Hence, Bayes' theorem can be rewritten as

$$p(\theta \mid D) = \frac{p(D_{\mathbf{y}} \mid D_{\mathbf{x}}, \theta)p(\theta)}{\int_\Theta p(D_{\mathbf{y}} \mid D_{\mathbf{x}}, \theta')p(\theta')\mathrm{d}\theta'} \propto p(D_{\mathbf{y}} \mid D_{\mathbf{x}}, \theta)p(\theta).$$

Notice that the equality comes from having assumed $D_{\mathbf{x}} \perp\!\!\!\perp \theta$. Posterior density $p(\theta \mid D)$ is typically high dimensional and highly nonconvex (Izmailov et al., 2021c; Jospin et al., 2022), so computing it and sampling from it is a difficult task. The first issue is tackled using Variational Inference (VI) procedures, while Markov Chain Monte Carlo (MCMC) methods address the second challenge. Both are reviewed – in the context of machine learning – in Jospin et al. (2022, Section V), where the authors also inspect their limitations. BNNs can be used for both regression and classification (Jospin et al., 2022, Section II); besides having a solid theoretical justification, there are practical benefits from using BNNs, as presented in Jospin et al. (2022, Section III).

[6] We can write $p$ as the Radon-Nikodym derivative of $P$ with respect to some $\sigma$-finite dominating measure $\mu$, that is, $p = \mathrm{d}P/\mathrm{d}\mu$.

### 2.2 Imprecise Probabilities

As CBDL is rooted in the theory of imprecise probabilities (IPs), in this section we give a gentle introduction to the IP concepts we will use throughout the paper.

CBDL is based on the *Bayesian sensitivity analysis* (BSA) approach to IPs, that in turn is grounded in the *dogma of ideal precision* (DIP) (Berger, 1984), (Walley, 1991, Section 5.9). The DIP posits that in any problem there is an *ideal probability model* which is precise, but which may not be precisely known. We call this condition *ambiguity* (Ellsberg, 1961; Gilboa & Marinacci, 2013).

Facing ambiguity can be represented mathematically by a set $\mathcal{P}_{\text{prior}}$ of priors and a set $\mathcal{P}_{\text{lik}}$ of likelihoods that seem "plausible" or "fit" to express the agent's beliefs on the parameters of interest and their knowledge of the data generating process (DGP). Generally speaking, the farther apart the "boundary elements" of the sets (i.e., their infimum and supremum), the higher the agent's ambiguity. Of course, if $\mathcal{P}_{\text{prior}}$ and $\mathcal{P}_{\text{lik}}$ are singletons we go back to the usual Bayesian paradigm.

A procedure based on sets $\mathcal{P}_{\text{prior}}$ and $\mathcal{P}_{\text{lik}}$ yields results that are more robust to prior and likelihood misspecification than a regular Bayesian method. In the presence of prior ignorance and indecisiveness about the sampling model, it is better to give answers in the form of intervals or sets, rather than arbitrarily select a prior and a likelihood, and then update. Sets $\mathcal{P}_{\text{prior}}$ and $\mathcal{P}_{\text{lik}}$ allow to represent *indecision*, thus leading to less informative but more robust and valid conclusions.

**Remark 2.** *Throughout the paper, we denote by* $\Pi = \{P_1, \dots, P_k\}$*,* $k \in \mathbb{N}$*, a finite set of probabilities on a generic space* $\Omega$*, such that for all* $j \in \{1, \dots, k\}$*,* $P_j$ *cannot be written as a convex combination of the other* $k-1$ *components of* $\Pi$*. We denote by* $\Pi'$ *its convex hull* $\Pi' \equiv Conv(\Pi)$*, i.e., the set of probabilities* $Q$ *on* $\Omega$ *that can be written as* $Q(A) = \sum_{j=1}^{k} \beta_j P_j(A)$*, for all* $A \subset \Omega$*, where the* $\beta_j$*'s are elements of* $[0, 1]$ *that sum up to* 1*. In the literature, it is referred to as a Finitely Generated Credal Set (FGCS, Levi (1980); Cozman (2000b)). Notice then that the extreme elements of* $\Pi'$ *correspond to the elements of* $\Pi$*, that is,* $ex\Pi' = \Pi$*. Simple graphical representations of finitely generated credal sets are given in Figures 1 and 2.*

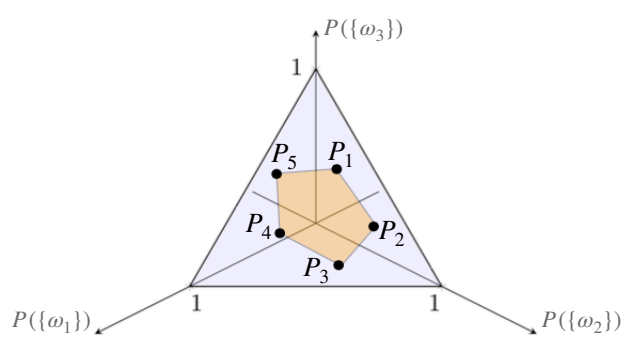



Figure 1: Suppose we are in a 3-class classification setting, so $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Then, any probability measure $P$ on $\Omega$ can be seen as a probability vector. For example, suppose $P(\{\omega_1\}) = 0.6$, $P(\{\omega_2\}) = 0.3$, and $P(\{\omega_3\}) = 0.1$. We have that $P \equiv (0.6, 0.3, 0.1)^\top$. Since its elements are positive and sum up to 1, probability vector $P$ belongs to the unit simplex, the purple triangle in the figure. Then, we can specify $\Pi = \{P_1, \dots, P_5\}$, and obtain as a consequence that $\Pi' = \text{Conv}(\Pi)$ is the orange pentagon. It is a convex polygon with finitely many extreme elements, and it is the geometric representation of a finitely generated credal set.

Let us now introduce the concepts of *lower* and *upper probabilities*. The lower probability $\underline{P}$ associated with $\Pi$ is given by $\underline{P}(A) = \inf_{P \in \Pi} P(A)$, for all $A \subset \Omega$. The upper probability $\overline{P}$ associated with $\Pi$ is defined as the conjugate to $\underline{P}$, that is, $\overline{P}(A) := 1 - \underline{P}(A^c) = \sup_{P \in \Pi} P(A)$, for all $A \subset \Omega$. These definitions hold even if $\Pi$ is not finite. Then, we have the following important result.

**Proposition 3.** $\overline{P}$ *is the upper probability for* $\Pi$ *if and only if it is also the upper probability for* $\Pi'$*. That is,* $\overline{P}(A) = \sup_{P \in \Pi} P(A) = \sup_{P' \in \Pi'} P'(A)$*, for all* $A \subset \Omega$*. The same holds for the lower probability.*

A version of Proposition 3 was proven in Dantzig (1963), while a variant for finitely additive probability measures can be found in Walley (1991, Section 3.6). A simple graphical representation of upper and lower probabilities for a set $A$ is given in Figure 2.

We now use lower probability $\underline{P}$ to define the *$\alpha$-level Imprecise Highest Density Region* (IHDR), for some $\alpha \in [0, 1]$.

**Definition 4.** *(Coolen, 1992, Section 2) Let $\alpha$ be any value in $[0, 1]$. Then, set $IR_\alpha(\Pi') \subset \Omega$ is called a $(1-\alpha)$-Imprecise Highest Density Region (IHDR) if*

1. $\underline{P}[\{\omega \in IR_\alpha(\Pi')\}] \geq 1 - \alpha$;
2. $\int_{IR_\alpha(\Pi')} d\omega$ *is a minimum. If $\Omega$ is at most countable, we replace $\int_{IR_\alpha(\Pi')} d\omega$ with $\#IR_\alpha(\Pi')$, where $\#$ denotes the cardinality operator.*

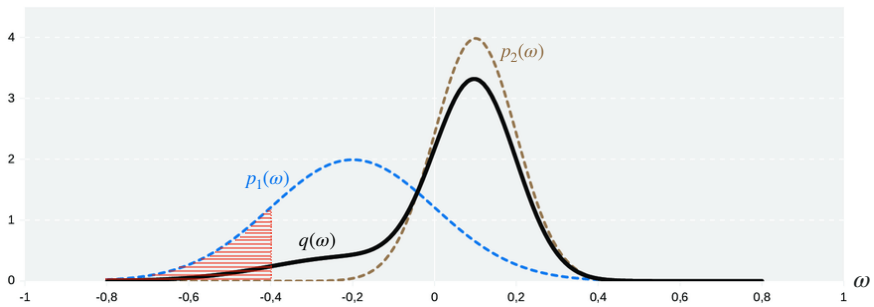



Figure 2: In this figure, a replica of Flint et al. (2017, Figure 1), $\Pi = \{P_1, P_2\}$, where $P_1$ and $P_2$ are two Normal distributions whose probability density functions (pdf's) $p_1$ and $p_2$ are given by the dashed blue and brown curves, respectively. Their convex hull is $\Pi' = \text{Conv}(\Pi) = \{Q : Q = \beta P_1 + (1-\beta)P_2$, for all $\beta \in [0, 1]\}$. The pdf $q$ of an element $Q$ of $\Pi'$ is depicted by a solid black curve. In addition, let $A = [-0.8, -0.4]$. Then, $\underline{P}(A) = \int_{-0.8}^{-0.4} p_2(\omega)\mathrm{d}\omega \approx 0$, while $\overline{P}(A)$ is given by the red shaded area under $p_1$, that is, $\overline{P}(A) = \int_{-0.8}^{-0.4} p_1(\omega)\mathrm{d}\omega$.

Definition 4 holds also if $\Pi = \text{ex}\Pi'$ is not finite. Notice that condition 2 is needed so that $IR_\alpha(\Pi')$ is the subset of $\Omega$ having the lowest possible cardinality, which still satisfies condition 1. By the definition of lower probability, Definition 4 implies that $P'[\{\omega \in IR_\alpha(\Pi')\}] \geq 1 - \alpha$, for all $P' \in \Pi'$. Here lies the appeal of the IHDR concept. Let us give a simple example, borrowed from Caprio et al. (2024a). Suppose $\Omega = \{\omega_1, \ldots, \omega_5\}$, $\Pi = \{P_1, P_2, P_3\}$, and $\alpha = 0.1$. The numerical values for $P_s(\{\omega_j\})$ are given in Table 1, for all $s \in \{1, 2, 3\}$ and all $j \in \{1, \ldots, 5\}$. Then, from Proposition 3 and Definition 4, we have that $IR_\alpha(\Pi') = \{\omega_1, \omega_2, \omega_3\}$.

| | $\omega_1$ | $\omega_2$ | $\omega_3$ | $\omega_4$ | $\omega_5$ |
|---|---|---|---|---|---|
| $P_1$ | 0.7 | 0.25 | 0.03 | 0.01 | 0.01 |
| $P_2$ | 0.6 | 0.2 | 0.1 | 0.05 | 0.05 |
| $P_3$ | 0.5 | 0.3 | 0.15 | 0.025 | 0.025 |

Table 1: Numerical values for our example. It is easy to see that the smallest subset of $\Omega$ that is assigned a probability of at least 0.9 by all the elements of $\Pi$ is $\{\omega_1, \omega_2, \omega_3\}$.

An operative way of building the IHDR is to consider the union of the (precise) Highest Density Regions (HDRs) of the elements of $\Pi = \text{ex}\Pi'$.[7] Let us define them formally.

**Definition 5.** *(Coolen, 1992, Section 1) Pick any $P_j \in \Pi$, $j \in \{1, \ldots, k\}$. Let $\alpha$ be any value in $[0, 1]$. Then, set $R_\alpha(P_j) \subset \Omega$ is called a $(1-\alpha)$-Highest Density Region (HDR) for $P_j$ if*

$$P_j[\{\omega \in R_\alpha(P_j)\}] \geq 1 - \alpha \quad \textit{and} \quad \int_{R_\alpha(P_j)} d\omega \textit{ is a minimum.}$$

[7]Here, by "operative" we mean that this procedure to compute the IHDR is easy to carry out in practice.

*If $\Omega$ is at most countable, we replace $\int_{R_\alpha(P_j)} d\omega$ with $\#R_\alpha(P_j)$. Equivalently (Hyndman, 1996),*

$$R_\alpha(P_j) = \{\omega \in \Omega : p_j(\omega) \geq p_j^\alpha\} \subset \Omega,$$

*where $p_j$ is the pdf or the probability mass function (pmf) of $P_j$, and $p_j^\alpha$ is a constant value. In particular, it is the largest constant such that $P_j[\{\omega \in R_\alpha(P_j)\}] \geq 1 - \alpha$.*

In dimension 1, $R_\alpha(P_j)$ can be interpreted as the smallest collection of elements of $\Omega$ (interval or union of intervals) to which distribution $P_j$ assigns probability of at least $1 - \alpha$. As we can see, HDRs are a Bayesian counterpart of confidence intervals.[8] We give a simple visual example in Figure 3.

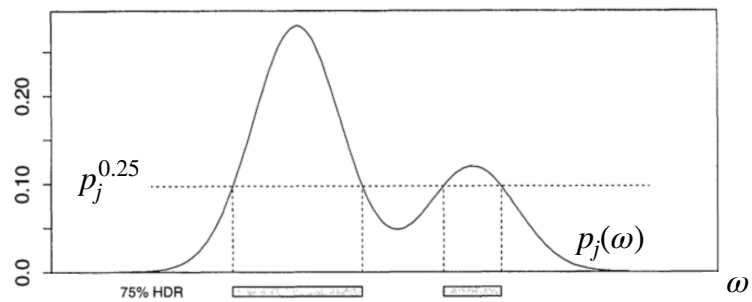



Figure 3: The 0.25-HDR from a Normal Mixture density. This picture is a replica of Hyndman (1996, Figure 1). The geometric representation of "75% probability according to $P_j$" is the area between the pdf curve $p_j(\omega)$ and the horizontal bar corresponding to $p_j^{0.25}$. A higher probability coverage (according to $P_j$) would correspond to a lower constant, so $p_j^\alpha < p_j^{0.25}$, for all $\alpha < 0.25$. In the limit, we recover 100% coverage at $p_j^0 = 0$.

As we mentioned earlier, an operative way of obtaining IHDR $IR_\alpha(\Pi')$ is by putting $IR_\alpha(\Pi') = \cup_{j=1}^k R_\alpha(P_j)$. Thanks to Proposition 3, by taking the union of the HDRs, we ensure that all the probability measures in the credal set $\Pi' = \text{Conv}(\Pi)$ assign probability of at least $1 - \alpha$ to the event $\{\omega \in IR_\alpha(\Pi')\}$. In turn, this implies that $\underline{P}[\{\omega \in IR_\alpha(\Pi')\}] = \min_{j\in\{1,\dots,k\}} P_j[\{\omega \in IR_\alpha(\Pi')\}] \geq 1 - \alpha$. We also have that the difference between the upper and lower probabilities of $\{\omega \in IR_\alpha(\Pi')\}$ is bounded by $\alpha$. To see this, notice that $\overline{P}[\{\omega \in IR_\alpha(\Pi')\}] \leq 1$, so $\overline{P}[\{\omega \in IR_\alpha(\Pi')\}] - \underline{P}[\{\omega \in IR_\alpha(\Pi')\}] \leq \alpha$.

### 2.3 Quantifying and Disentangling Aleatoric and Epistemic Uncertainties

Recall that, given a probability measure $P$ on a generic space $\Omega$, the (Shannon) entropy of $P$ is defined as $H(P) := \mathbb{E}[-\log p] = -\int_\Omega \log[p(\omega)]P(\mathrm{d}\omega)$ if $\Omega$ is uncountable, where $p$ denotes the pdf of $P$. If $\Omega$ is at most countable, we have that $H(P) = -\sum_{\omega\in\Omega} P(\{\omega\}) \log[P(\{\omega\})]$. As pointed out by Dubois & Hüllermeier (2007); Hüllermeier & Waegeman (2021), the entropy primarily captures the shape of the distribution, namely its "peakedness" or non-uniformity, and hence informs about the predictability of the outcome of a random experiment: the higher its value, the lower the predictability. Then, we can define the imprecise versions of the Shannon entropy as proposed by Abellán et al. (2006); Hüllermeier & Waegeman (2021), $\overline{H}(P) := \sup_{P\in\Pi'} H(P)$ and $\underline{H}(P) := \inf_{P\in\Pi'} H(P)$, called the *upper* and *lower Shannon entropy*, respectively.[9] Notice that these definitions hold for all sets of probabilities, not just for (finitely generated) credal sets. The upper entropy is a measure of total uncertainty since it represents the minimum level of predictability associated with the elements of $\Pi'$. In Abellán et al. (2006); Hüllermeier & Waegeman (2021), the authors posit that it can be decomposed as a sum of aleatoric and epistemic uncertainties, and that this latter can be specified as the difference between upper and lower entropy, thus obtaining

$$\underbrace{\overline{H}(P')}_{\text{TU}(\Pi')} = \underbrace{\underline{H}(P')}_{\text{AU}(\Pi')} + \underbrace{\left[\overline{H}(P') - \underline{H}(P')\right]}_{\text{EU}(\Pi')},$$

where $\text{TU}(\Pi')$ denotes the total uncertainty associated with set $\Pi'$, $\text{AU}(\Pi')$ is the AU associated with $\Pi'$, and $\text{EU}(\Pi')$ represents the EU associated with $\Pi'$. We have the following proposition.

[8] Hence standard choices for the value of $\alpha$ are 0.1, 0.05, 0.01.

[9] In Appendix E, we provide bounds to the values of upper and lower entropy.

**Proposition 6.** *Let $\Pi, \Pi'$ be sets of probability measures as the ones considered in Remark 2. Then,* $\sup_{P\in\Pi} H(P) = \overline{H}(P) \leq \overline{H}(P') = \sup_{P'\in\Pi'} H(P')$ *and* $\inf_{P\in\Pi} H(P) = \underline{H}(P) = \underline{H}(P') = \inf_{P'\in\Pi'} H(P')$.

Proposition 6 tells us that the upper entropy of the extreme elements in $\Pi = \text{ex}\Pi'$ is a lower bound for the upper entropy of the whole credal set $\Pi'$, and that the lower entropy of the extreme elements in $\Pi$ is equivalent to the lower entropy of the whole credal set $\Pi'$. These facts imply that $\text{AU}(\Pi') = \text{AU}(\Pi)$, and that $\text{EU}(\Pi') \geq \text{EU}(\Pi)$. In addition, as a consequence of (Smieja & Tabor, 2012, Theorem III.1), we have that $\text{EU}(\Pi') \leq \text{EU}(\Pi) + \log(\#\Pi)$. In turn, we have that $\text{EU}(\Pi') \in [\text{EU}(\Pi), \text{EU}(\Pi) + \log(\#\Pi)]$.

## 3 Our Procedure and Its Properties

This is the main portion of the paper. In section 3.1, we present and discuss the CBDL algorithm. Its theoretical properties are derived in section 3.2.

### 3.1 CBDL algorithm

Recall that $D = D_{\mathbf{x}} \times D_{\mathbf{y}}$ denotes the training set, where $D_{\mathbf{x}} = \{x_i\}_{i=1}^n \subset \mathcal{X}$ is the collection of training inputs, $D_{\mathbf{y}} = \{y_i\}_{i=1}^n \subset \mathcal{Y}$ is the collection of training outputs, and $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output spaces, respectively. We then denote by $P$ a generic prior on the parameters $\theta \in \Theta$ of a BNN having pdf $p$, and by $L \equiv L_{x,\theta}$ a generic likelihood on the space $\mathcal{Y}$ of outputs having pdf $\ell \equiv \ell_{x,\theta}$. The act of computing the posterior from prior $P$ and likelihood $L$ using a BNN is designated by $\mathsf{post}[P, L]$. The CBDL procedure is presented in Algorithm 1, and it is discussed in the following paragraphs.

**Algorithm 1** Credal Bayesian Deep Learning (CBDL) – Training and Inference

*During Training*
**Step 1** Specify $K$ priors $\text{ex}\mathcal{P}_{\text{prior}} = \{P_k^{\text{ex}}\}_{k=1}^K$
**Step 2** Specify $S$ likelihoods $\text{ex}\mathcal{P}_{\text{lik}} = \{L_s^{\text{ex}}\}_{s=1}^S$
**Step 3** Compute $P_{k,s}(\cdot \mid D) = \mathsf{post}[P_k^{\text{ex}}, L_s^{\text{ex}}]$, for all $k$ and all $s$ ▷ % Approximated via VI by $\breve{P}_{k,s}$%
*During Inference*
**Inputs:** New input $\tilde{x} \in \mathcal{X}$
**Parameters:** Confidence parameter $\alpha \in [0, 1]$
**Outputs:** Aleatoric and Epistemic Uncertainties, $\alpha$-level IHDR
**Step 4** Compute $P_{k,s}^{\text{pred}} = \mathsf{pred}[\breve{P}_{k,s}, L_s^{\text{ex}}]$, for all $k$ and all $s$ ▷ % Approximated via VI by $\hat{P}_{k,s}^{\text{pred}}$%
**Step 5** Compute and return $\text{AU}(\hat{\mathcal{P}}_{\text{pred}})$ and the bounds for $\text{EU}(\hat{\mathcal{P}}_{\text{pred}})$
**Step 6** Compute and return the $(1-\alpha)$-IHDR $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$

During training, in **Step 1** the user specifies $K$ priors on the parameters of the neural network, that constitute the extrema $\text{ex}\mathcal{P}_{\text{prior}}$ of the prior FGCS $\mathcal{P}_{\text{prior}}$. Similarly, in **Step 2** they elicit $S$ likelihoods capturing the possible architectures of the neural network, that correspond to the extrema $\text{ex}\mathcal{P}_{\text{lik}}$ of the likelihood FGCS $\mathcal{P}_{\text{lik}}$. Let us give an example. Outlined in Jospin et al. (2022, Sections IV-B and IV-C1), for classification, the standard process for BNNs involves

- A Normal prior with zero mean and diagonal covariance $\sigma^2 I$ on the coefficients of the network, that is, $p(\theta) = \mathcal{N}(0, \sigma^2 I)$. In the context of CBDL, we could specify, e.g., $\text{ex}\mathcal{P}_{\text{prior}} = \{P : p(\theta) = \mathcal{N}(\mu, \sigma^2 I),\ \mu \in \{\mu_-, \mathbf{0}, \mu_+\},\ \sigma^2 \in \{3, 7\}\}$. That is, the extreme elements of the prior credal set are five independent Normals having different levels of "fatness" of the tails, and centered at a vector $\mu_+$ having positive entries, a vector $\mu_-$ having negative entries, and a vector $\mathbf{0}$ having entries equal to 0. They capture the ideas of positive bias, negative bias, and no bias of the coefficients, respectively. This is done to hedge against possible prior misspecification.

- A Categorical likelihood, $p(y \mid x, \theta) = \text{Cat}(\Phi_\theta(x))$, whose parameter is given by the output of a functional model $\Phi_\theta$. In the context of CBDL, we could specify the set of extreme elements of the likelihood credal set as $\text{ex}\mathcal{P}_{\text{lik}} = \{L : \ell_{x,\theta}(y) = \text{Cat}(\Phi_{s,\theta}(x)), s \in \{1, \ldots, S\}\}$. Specifying set

$\text{ex}\mathcal{P}_{\text{lik}}$, then, corresponds to eliciting a finite number $S$ of possible (parametrized) architectures for the neural network $\Phi_{s,\theta}$, $s \in \{1, \dots, S\}$, and obtain, as a consequence, $S$ categorical distributions $\{\text{Cat}(\Phi_{s,\theta}(x))\}_{s=1}^{S}$. This captures the ambiguity around the true data generating process faced by the agent, and allows them to hedge against likelihood misspecification.

More in general, we can use the priors and likelihoods that better fit the type of analysis we are performing.[10] For example, for the choice of the priors we refer to Fortuin et al. (2021), where the authors study the problem of selecting the right type of prior for BNNs.

**Step 3** performs an element-wise application of Bayes' rule for all the elements of $\text{ex}\mathcal{P}_{\text{prior}}$ and $\text{ex}\mathcal{P}_{\text{lik}}$. Each posterior is approximated using the Variational Inference (VI) method (Jospin et al., 2022, Section V). That is, we project every posterior $P_{k,s}(\cdot \mid D)$ onto a set $\mathbb{S}$ of "well-behaved" distributions (e.g., Normals) using the KL divergence. In formulas, $\breve{P}_{k,s} = \arg\min_{Q \in \mathbb{S}} KL[Q \| P_{k,s}(\cdot \mid D)]$. By "well-behaved", we mean that they have to satisfy the conditions in Zhang & Gao (2020, Sections 2 and 3).[11] This ensures that as the sample size goes to infinity, the approximated posteriors converge to the true data generating process. We also point out how despite using a VI approximation for the exact posteriors, as we shall see in the next section, the credal set of approximated posteriors is closer to the "oracle" posterior $P^o(\cdot \mid D)$ than any of its elements. As a consequence, working with credal sets leads to VI posterior approximations that are better than the ones resulting from a single BNN, or an ensemble of BNNs, where several BNNs are combined into one.

**Remark 7.** *Although highly unlikely in practice, it is theoretically possible that the VI approximation of the (finite) set $\{P_{k,s}(\cdot \mid D)\}_{k,s}$ of posteriors is a singleton, see Figure 4. While the conditions in Zhang & Gao (2020) guarantee that asymptotically the approximated posteriors coincide with the true data generating process, this typically does not happen with finite-dimensional datasets. As a consequence, obtaining a singleton when projecting $\{P_{k,s}(\cdot \mid D)\}_{k,s}$ onto $\mathbb{S}$ may result in an underestimation of the uncertainties faced by the user. In that case, we either consider a different set – whose elements still satisfy the conditions in Zhang & Gao (2020, Sections 2 and 3) – on which to project $\{P_{k,s}(\cdot \mid D)\}_{k,s}$ according to the KL divergence, or we use a different "projection operator", that is, a divergence different from the KL. For example, Rényi and $\chi^2$ divergences, or Hellinger and total variation metrics are suggested by Zhang & Gao (2020). Alternatively, we can consider a different approximation strategy altogether, for instance the Laplace approximation (Ritter et al., 2018).*

After **Step 3**, we obtain a finite set $\{\breve{P}_{k,s}\}_{k,s}$ of VI approximation of the posteriors on the network parameters, whose cardinality is $K \times S$. Its convex hull constitutes $\breve{\mathcal{P}}_{\text{post}}$, that is, the VI-approximated posterior FGCS. We assume that $\text{ex}\breve{\mathcal{P}}_{\text{post}} = \{\breve{P}_{k,s}\}_{k,s}$. This is an assumption because it may well be that – due to the approximation procedure – some of the elements of $\{\breve{P}_{k,s}\}_{k,s}$ are not independent of one another. We defer to future work the design of a procedure that finds the elements of $\{\breve{P}_{k,s}\}_{k,s}$ that cannot be written as a convex combination of one another.

Being a combinatorial task, **Step 3** is a computational bottleneck of Algorithm 1. We have to calculate $K \times S$ VI approximations to as many posteriors, but this allows us to forego any additional assumptions on the nature of the lower and upper probabilities that are oftentimes required by other imprecise-probabilities-based techniques.[12] Clearly, CBDL is simplified if either $\mathcal{P}_{\text{prior}}$ or $\mathcal{P}_{\text{lik}}$ are singletons. Notice that in the case that $\mathcal{P}_{\text{prior}}$ and $\mathcal{P}_{\text{lik}}$ are both not singletons, for all $A \subset \Theta$ the interval $[\underline{\breve{P}}(A), \overline{\breve{P}}(A)]$ is wider than the case when one or the other is a singleton. In the limiting case where both are singletons, we retrieve the usual Bayesian updating, so the interval shrinks down to a point.

Before moving to inference time, let us remark a difference between Bayesian Model Averaging (BMA) and CBDL. In BMA, the user specifies a distribution on the models. Translated in the notation we use in this work, this means having a discrete distribution $Q$ over the $K \times S$ prior-likelihood combinations, that is, over the elements of $\text{ex}\breve{\mathcal{P}}_{\text{post}}$. Such $Q$ is then used to select an element $\breve{P}^\star$ from the (VI-approximated) posterior

[10] Choosing 2 to 5 priors and likelihoods is usually enough to safely hedge against prior and likelihood misspecification.

[11] We assume that the conditions on the priors and the likelihoods given in Zhang & Gao (2020) are satisfied.

[12] If we are willing to make such assumptions, Theorem 10 in Appendix D shows how to compute the upper posterior using only upper prior and upper likelihood.

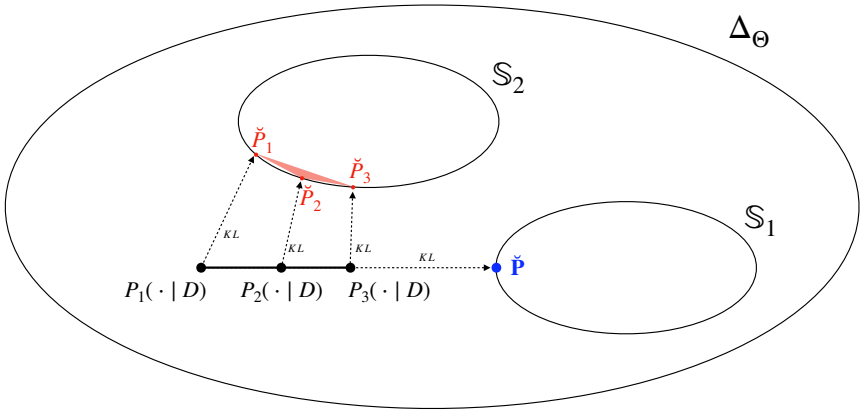



Figure 4: Let $\Delta_\Theta$ denote the space of probability measures on $\Theta$. Suppose that in the analysis at hand we specified three priors and only one likelihood, so $S = 1$ and we can drop the $s$ index. Let $\{P_k(\cdot \mid D)\}_{k=1}^3$ be the collection of exact posteriors, so that the black segment represents the exact posterior FGCS. Then, if we project the elements of $\{P_k(\cdot \mid D)\}_{k=1}^3$ onto $\mathbb{S}_1$ via the KL divergence, we obtain the same distribution $\breve{\mathbf{P}}$. This is detrimental to the analysis because such an approximation underestimates epistemic (and possibly also aleatoric) uncertainty faced by the agent. Then, the user should specify a different set $\mathbb{S}_2$ of "well-behaved" distributions onto which project the elements of $\{P_k(\cdot \mid D)\}_{k=1}^3$. In the figure, we see that they are projected onto $\mathbb{S}_2$ via the KL divergence to obtain $\breve{P}_1$, $\breve{P}_2$, and $\breve{P}_3$. The convex hull of these latter, captured by the red shaded triangle, represents the variational approximation of the exact posterior FGCS.

FGCS $\breve{\mathcal{P}}_{\text{post}}$ as

$$\breve{\mathcal{P}}_{\text{post}} \ni \breve{P}^\star = \sum_{k,s} Q(\{\breve{P}_{k,s}\})\breve{P}_{k,s}, \tag{1}$$

where $Q(\{\breve{P}_{k,s}\}) \in [0,1]$ for all $k$ and all $s$, and $\sum_{k,s} Q(\{\breve{P}_{k,s}\}) = 1$. Instead, CBDL does not select a unique distribution from $\breve{\mathcal{P}}_{\text{post}}$. Rather, distributions are kept separate so to be able to derive a predictive FGCS in **Step 4** of Algorithm 1, that is in turn used to quantify and disentangle predictive uncertainties, and to compute the predictive IHDR, that is, a collection of outputs having a high probability of being the correct ones for a new input $\tilde{x}$.

During inference, a new input $\tilde{x}$ is provided. In **Step 4**, every element of $\text{ex}\breve{\mathcal{P}}_{\text{post}}$ is used to derive a predictive distribution $P_{k,s}^{\text{pred}} = \mathsf{pred}[\breve{P}_{k,s}, L_s^{\text{ex}}]$ on the output space $\mathcal{Y}$. In particular, for every $k$ and every $s$, the pdf $p_{k,s}^{\text{pred}}$ of $P_{k,s}^{\text{pred}}$ is obtained as

$$\begin{aligned} p_{k,s}^{\text{pred}}(\tilde{y} \mid \tilde{x}, x_1, y_1, \ldots, x_n, y_n) &= \int_\Theta \ell_s^{\text{ex}}(\tilde{y} \mid \theta, \tilde{x}) \cdot p_{k,s}(\theta \mid x_1, y_1, \ldots, x_n, y_n)\mathrm{d}\theta \\ &\approx \int_\Theta \ell_s^{\text{ex}}(\tilde{y} \mid \theta, \tilde{x}) \cdot \breve{p}_{k,s}(\theta)\mathrm{d}\theta, \end{aligned}$$

where $\ell_s^{\text{ex}}$ is the pdf of likelihood $L_s^{\text{ex}}$, $p_{k,s}$ is the pdf of the true posterior $P_{k,s}(\cdot \mid D)$, $\breve{p}_{k,s}$ is the pdf of the VI-approximated posterior $\breve{P}_{k,s}$, and $\tilde{y}$ is the output associated to the new input $\tilde{x}$ (see Appendix F for more details). Each predictive distribution $P_{k,s}^{\text{pred}}$ is approximated to $\hat{P}_{k,s}^{\text{pred}}$ (e.g., using Normals) via Variational Inference. The convex hull of the collection $\{\hat{P}_{k,s}^{\text{pred}}\}_{k,s}$ having cardinality $K \times S$ constitutes the VI-approximated predictive FGCS $\hat{\mathcal{P}}_{\text{pred}}$. Similarly to what we did for $\text{ex}\breve{\mathcal{P}}_{\text{post}}$, we assume that $\text{ex}\hat{\mathcal{P}}_{\text{pred}} = \{\hat{P}_{k,s}^{\text{pred}}\}_{k,s}$.

In **Step 5**, building on the results in section 2.3, and in particular on Proposition 6, we compute and return $\text{AU}(\hat{\mathcal{P}}_{\text{pred}})$ and the bounds for $\text{EU}(\hat{\mathcal{P}}_{\text{pred}})$. In particular, we have

$$\text{AU}(\mathcal{P}_{\text{pred}}) \approx \text{AU}(\hat{\mathcal{P}}_{\text{pred}}) = \underline{H}(\hat{P}^{\text{pred}}) \tag{2}$$

and

$$\mathrm{EU}(\mathcal{P}_{\text{pred}}) \approx \mathrm{EU}(\hat{\mathcal{P}}_{\text{pred}}) \in [\overline{H}(\hat{P}^{\text{pred}}) - \underline{H}(\hat{P}^{\text{pred}}), \overline{H}(\hat{P}^{\text{pred}}) - \underline{H}(\hat{P}^{\text{pred}}) + \log(K \times S)], \tag{3}$$

where (i) $\underline{H}(\hat{P}^{\text{pred}}) = \min_{k,s} H(\hat{P}^{\text{pred}}_{k,s})$; (ii) $\overline{H}(\hat{P}^{\text{pred}}) = \max_{k,s} H(\hat{P}^{\text{pred}}_{k,s})$; and (iii) $\mathcal{P}_{\text{pred}}$ is the "true" predictive FGCS, that is, the one we would have obtained had we been able to compute the exact posteriors in **Step 3** and the exact predictive distributions in **Step 4**. Let us point out a salient feature of CBDL. The AU and the (bounds for) the EU associated with the VI-approximated predictive FGCS $\hat{\mathcal{P}}_{\text{pred}}$ embed uncertainty comparable to an uncountably infinite ensemble of BNNs, i.e., an ensemble of BNNs of cardinality $\aleph_1$, despite the simple and intuitive mathematics over the finite set $\mathrm{ex}\hat{\mathcal{P}}_{\text{pred}}$. They are not merely pessimistic results on the uncertainties associated with a finite ensemble of BNNs.

We observe that we compute the AU and EU associated with the (VI-approximated) predictive credal set $\hat{\mathcal{P}}_{\text{pred}}$, rather than those related to the (VI-approximated) posterior credal set $\breve{\mathcal{P}}_{\text{post}}$. We do so because we are ultimately interested in reporting the uncertainty around the predicted outputs given a new input in the problem at hand, more than the uncertainty on the parameters of the NN.

Before commenting on the next step, let us pause here and add a remark. While a "bad choice" of priors and likelihoods may lead to maximal upper and lower entropy – $\overline{H}(\hat{P}^{\text{pred}})$ and $\underline{H}(\hat{P}^{\text{pred}})$, respectively – this is not a risk confined to our procedure. Poor modeling choices are an unavoidable risk in model-based techniques. This gave rise to the famous adage by George Box "essentially, all models are wrong but some are useful" (Box, 1976).[13] We maintain that our method is indeed useful, since it overcomes some of the shortcomings of traditional Bayesian techniques – as explained in Appendices A and B. As for "regular" Bayesian methods, though, for our approach too the designer will need to make "plausible" choices for priors and likelihoods.

Finally, in **Step 6**, we compute and return the $\alpha$-level Imprecise Highest Density Region $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$ for the VI-approximated predictive FGCS $\hat{\mathcal{P}}_{\text{pred}}$, which approximates the IHDR for the "true" predictive FGCS $\mathcal{P}_{\text{pred}}$. It is the smallest subset of $\mathcal{Y}$ such that $\hat{P}^{\text{pred}}[\{\tilde{y} \in IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})\}] \geq 1 - \alpha$, for all $\hat{P}^{\text{pred}} \in \hat{\mathcal{P}}_{\text{pred}}$, and for some $\alpha \in [0, 1]$. It can be interpreted as the smallest collection of outputs $\tilde{y}$ that have a high probability of being the correct ones for the new input $\tilde{x}$, according to all the distributions in $\hat{\mathcal{P}}_{\text{pred}}$. Or equivalently, we can say that the correct output for the new input $\tilde{x}$ belongs to IHDR $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$ with lower probability of at least $1 - \alpha$, a probabilistic guarantee for the set of outputs generated by our procedure. Notice also that the size of $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$ is an increasing function of both AU and EU. As a consequence, it is related, but it is not equal, to the AU the agent faces. If we want to avoid to perform the procedure only to discover that $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$ is "too large", then we can add an "AU check" after **Step 5**. This, together with computing $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$ in a classification setting, is explored in Appendices G and H.

As we have seen in section 2.2, we find $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$ by taking the union of the $K \times S$ HDRs $R_\alpha(\hat{P}^{\text{pred}}_{k,s})$ of the extreme elements $\mathrm{ex}\hat{\mathcal{P}}_{\text{pred}}$ of $\hat{\mathcal{P}}_{\text{pred}}$. The HDR of a well-known distribution (for instance, a Normal) can be routinely obtained in R, e.g. using package HDInterval (Juat et al., 2022).

We conclude this section with two remarks. First, we point out how CBDL does not depend on the method used to approximate the posterior and the predictive distributions: Variational Inference can be substituted by other approaches. CBDL can also be easily adapted to other TU, AU, and EU measures, as long as the measure chosen for the total uncertainty is bounded. Second, **Step 6** can be effortlessly modified so that CBDL produces the collection of outputs having the highest lower density. This is in line with the Naive Credal Classifier theory (Cozman, 2000a; Zaffalon, 2002). In this case, we forego control on the accuracy level $1 - \alpha$ of the output region produced by CBDL. It is implemented as follows. In the modified version of **Step 6**, we compute

$$\arg\max_{\tilde{y} \in \mathcal{Y}} \min_{k,s} \hat{p}^{\text{pred}}_{k,s}(\tilde{y}),$$

where $\hat{p}^{\text{pred}}_{k,s}$ is the pdf of the VI-approximated predictive distribution $\hat{P}^{\text{pred}}_{k,s}$. If such arg max is not a singleton, and the user is set on CBDL outputting a unique value $\tilde{y}$ for the new input $\tilde{x}$, they can select one element uniformly at random from the arg max.

[13] Curiously, a similar motivation was brought forward by Pseudo-Dionysius the Areopagite in favor of the use of sacred images in the Christian tradition (Migne, 1857). While they do not capture the essence of God, these defective approximations help the believer's thought to elevate.

### 3.2 Theoretical Properties of CBDL

Working with credal sets makes CBDL more robust to distribution shifts than single BNNs. To see this, we present the following general result, and then we apply it to our case.

Let $\Pi'$ be an FGCS as in Remark 2, and consider a probability measure $\Psi$ such that $\Psi \notin \Pi'$.

**Proposition 8.** *Call $d$ any metric and div any divergence on the space of probability measures of interest. Let $d(\Pi', \Psi) := \inf_{P' \in \Pi'} d(P', \Psi)$ and $div(\Pi' \| \Psi) := \inf_{P' \in \Pi'} div(\Pi' \| \Psi)$. Then, for all $P' \in \Pi'$, $d(\Pi', \Psi) \leq d(P', \Psi)$ and $div(\Pi', \Psi) \leq div(P' \| \Psi)$.*

Proposition 8 holds if $\Pi'$ is any set of probabilities, not just an FGCS.[14] In Appendix I, we show that the above result still holds if the elements of $\Pi'$ and $\Psi$ are defined on Euclidean spaces having different dimensions (Cai & Lim, 2022; Caprio, 2022).

Let us now apply Proposition 8 to CBDL. Suppose that, when designing a single BNN, an agent chooses likelihood $\mathbf{L}$, while when implementing CBDL, they specify in **Step 2** a finite set of likelihoods $\text{ex}\mathcal{P}_{\text{lik}} = \{L_s^{\text{ex}}\}_{s=1}^S$, and then let the induced credal set $\mathcal{P}_{\text{lik}} = \text{Conv}(\text{ex}\mathcal{P}_{\text{lik}})$ represent their uncertainty around the sampling model. Assume that $\mathbf{L} \in \text{ex}\mathcal{P}_{\text{lik}}$. This means that when designing the single BNN, the agent chooses arbitrarily which of the elements of $\text{ex}\mathcal{P}_{\text{lik}}$ to use. Suppose also that the "oracle" data generating process $L^o$ is different from $\mathbf{L}$, $L^o \neq \mathbf{L}$, so that we are actually in the presence of distribution shift. Then, we have two cases. (1) If the true sampling model $L^o$ belongs to $\mathcal{P}_{\text{lik}}$, then the distance – measured via a metric or a divergence – between $\mathcal{P}_{\text{lik}}$ and $L^o$ is 0, while that between $\mathbf{L}$ and $L^o$ is positive. (2) If $L^o \notin \mathcal{P}_{\text{lik}}$, then the distance between $\mathcal{P}_{\text{lik}}$ and $L^o$ is no larger than the distance between $\mathbf{L}$ and $L^o$, no matter (i) which metric or distance we use (Proposition 8), (ii) whether or not $L^o$ and the elements of $\mathcal{P}_{\text{lik}}$ are defined on the same Euclidean space (Appendix I, Lemma 17). A visual representation is given in Figure 5.

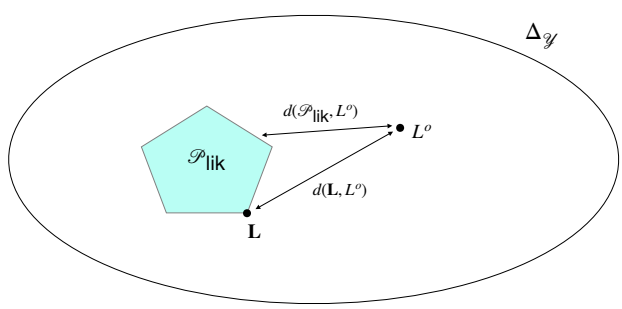



Figure 5: CBDL is more robust to distribution shifts than single BNNs. Here $\mathcal{P}_{\text{lik}}$ is the convex hull of five plausible likelihoods, and $d$ denotes a generic metric on the space $\Delta_{\mathcal{Y}}$ of probabilities on $\mathcal{Y}$. We see how $d(\mathcal{P}_{\text{lik}}, L^o) < d(\mathbf{L}, L^o)$; if we replace metric $d$ by a generic divergence div, the inequality would still hold.

Let us add a remark here. Assume that the "oracle" prior $P^o$ is in the prior credal set $\mathcal{P}_{\text{prior}}$ and that the "oracle" likelihood $L^o$ is in the likelihood credal set $\mathcal{P}_{\text{lik}}$. Then, it is immediate to see that the "oracle" posterior $P^o(\cdot \mid D)$ belongs to the posterior credal set $\mathcal{P}_{\text{post}}$. Naturally, this does not imply that the posterior credal set collapses to $P^o(\cdot \mid D)$. In general, it is unlikely that a finite amount of data is able to completely annihilate all the epistemic uncertainty faced by the agent. What may happen is that if the training set is large enough, $\mathcal{P}_{\text{post}}$ may be inscribed in a ball of small radius around $P^o(\cdot \mid D)$. This does not mean that we suffer from under-confidence due to larger-than-necessary epistemic uncertainty. Rather, the relative epistemic uncertainty (measured by the difference between prior and posterior uncertainty, divided by the prior uncertainty) drops significantly. In addition, working with sets of prior and likelihoods allows us to hedge against prior and likelihood misspecification, a consequence of Proposition 8.

[14] More in general, Proposition 8 holds for any type of function $f$, not just metrics or divergences, because Proposition 8 essentially only applies the definition of the infimum.

# 4 Experiments

From the previous sections, it is clear that CBDL improves on the uncertainty quantification capabilities of single BNNs.[15] CBDL allows a better quantification of AU because of its robustness to misspecification stemming from Proposition 8. In a sense, the AU quantified by a single BNN is a function of the choices of prior and likelihood made by the user. In addition, as we pointed out before, EU cannot be quantified by a single distribution (Hüllermeier & Waegeman, 2021), and hence it cannot be obtained in a theoretically principled manner from a single BNN.

In this section, we present our experimental findings. In section 4.1, we show that CBDL is better at quantifying and disentangling predictive AU and EU than ensemble of BNNs. In section 4.2, we analyze the downstream task performance of CBDL. We show that they perform better than an ensemble of BNNs (EBNN). To demonstrate the utility of our method, we study the behavior of certain safety-critical settings under distribution shifts and its ramifications. One, for motion prediction in autonomous driving scenarios, and two, to model blood glucose and insulin dynamics for artificial pancreas control.

## 4.1 Uncertainty Quantification

Distribution shifts can introduce uncertainties in a system, which in turn can render the predictions meaningless. This can be due to naturally occurring corruptions, as introduced in Hendrycks & Dietterich (2019) for image datasets. The authors introduced 18 different noise types, which can be varied across 5 different severity levels. The intuition is that in the current context, increasing the noise severity should generally result in higher uncertainty. We evaluate our CBDL method on 4 standard image datasets CIFAR-10 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), Fashion-MNIST (Xiao et al., 2017), and MNIST (Lecun et al., 1998). We use a slightly different set of perturbations than those introduced in Mu & Gilmer (2019) for gray-scale images like MNIST and Fashion-MNIST. Additionally, we perform cross domain testing for each dataset, where we expect the uncertainties to be higher. We implement and train a Resnet-20 Bayesian Neural Network model inside the library Bayesian-torch (Krishnan et al., 2019). For each dataset, we train 4 different networks initialized with different seeds on the prior and with the same architecture. This corresponds to eliciting a prior FGCS $\mathcal{P}_{\text{prior}}$ such that $\text{ex}\mathcal{P}_{\text{prior}} = \{P_1^{\text{ex}}, \dots, P_4^{\text{ex}}\}$, so that $K$ in **Step 1** of Algorithm 1 is equal to 4, and a likelihood FGCS that is a singleton, $\mathcal{P}_{\text{lik}} = \text{ex}\mathcal{P}_{\text{lik}} = \{L\}$, so that $S$ in **Step 2** of Algorithm 1 is equal to 1. We use a learning-rate of 0.001, batch-size of 128, and train the networks using Mean-Field Variational Inference for 200 epochs. The inference is carried out by performing multiple forward passes through parameters drawn from the posterior distribution. We used 20 Monte-Carlo samples in the experiments.

**Baselines.** In Egele et al. (2021), the authors pursue the same heuristic endeavor as we do, but take an ensemble route. They too consider different BNNs, but instead of keeping them separate and use them to build a predictive credal set, they average them out. Similar to theirs, we elicit the following procedure, that we call ensemble of BNNs (EBNN).

Consider $R \in \mathbb{N}_{\geq 2}$ different BNNs, and compute the posterior distribution on the parameters. They induce $R$ predictive distributions on the output space $\mathcal{Y}$, each having mean $\mu_r$ and variance $\sigma_r^2$, $r \in \{1, \dots, R\}$. We call *EBNN distribution* $P_{\text{ens}}$ a Normal having mean $\mu_{\text{ens}} = 1/R \sum_{r=1}^{R} \mu_r$ and covariance matrix $\sigma_{\text{ens}}^2 I$, where $\sigma_{\text{ens}}^2 = 1/R \sum_{r=1}^{R} \sigma_r^2 + 1/(R-1) \sum_{r=1}^{R} (\mu_r - \mu_{\text{ens}})^2$. We use the $\alpha$-level HDR $R_\alpha(P_{\text{ens}})$ associated with $P_{\text{ens}}$ as a baseline for the IHDR $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$ computed at **Step 6** of Algorithm 1.

**Results.** Following Egele et al. (2021), for EBNN we posit that $1/R \sum_{r=1}^{R} \sigma_r^2$ captures the aleatoric uncertainty associated with $P_{\text{ens}}$, and $1/(R-1) \sum_{r=1}^{R} (\mu_r - \mu_{\text{ens}})^2$ captures the epistemic uncertainty associated with $P_{\text{ens}}$; we use these values as baselines.[16] For CBDL, we look at the value of $\text{AU}(\hat{\mathcal{P}}_{\text{pred}}) = \underline{H}(\hat{P}^{\text{pred}})$ from equation 2, and at the lower bound $\overline{H}(\hat{P}^{\text{pred}}) - \underline{H}(\hat{P}^{\text{pred}})$ for $\text{EU}(\hat{\mathcal{P}}_{\text{pred}})$ from equation 3. It is enough to focus

[15] This includes empirical Bayes methodologies (Krishnan et al., 2020).

[16] Notice that in this case we retain the assumption that the EBNN distribution $P_{\text{ens}}$ on $\mathcal{Y}$ has mean $\mu_{\text{ens}}$ and covariance matrix $\sigma_{\text{ens}}^2 I$, but we do not require that it is a Normal. That is because in the four image datasets that we consider, the output space $\mathcal{Y}$ is finite. We assume, though, that the probability mass function of $P_{\text{ens}}$ is a regularly looking, symmetric, bell-shaped histogram.

on the lower bound because the upper bound is given by $\overline{H}(\hat{P}^{\text{pred}}) - \underline{H}(\hat{P}^{\text{pred}}) + \log(4)$, and $\log(4) \approx 0.6$ is a fixed value. Hence, the trend of both lower and upper bounds for $\text{EU}(\hat{\mathcal{P}}_{\text{pred}})$ as the severity of corruption changes is the same.

We discuss the results for CIFAR-10 presented in Table 2. We report the average uncertainties across all test samples as measured by the respective procedures. Note that we abstracted the severity levels to low (severity $= 1$), medium (severity $= 2, 3$), and high (severity $= 4, 5$). We observe that CBDL reports lower levels of aleatoric uncertainty when subjected to low levels of corruption, which gradually increases as we move to higher levels. The epistemic uncertainty grows as well, but the change in the aleatoric component is more pronounced. However, for EBNN this is not true. Even though the epistemic uncertainty increases, the aleatoric shows the reverse trend, contrary to our intuition of aleatoric uncertainty. Using CBDL, we achieve the highest uncertainties, both aleatoric and epistemic, when applied to entirely different test sets, namely MNIST, SVHN, and Fahion-MNIST. We report these numbers as well.

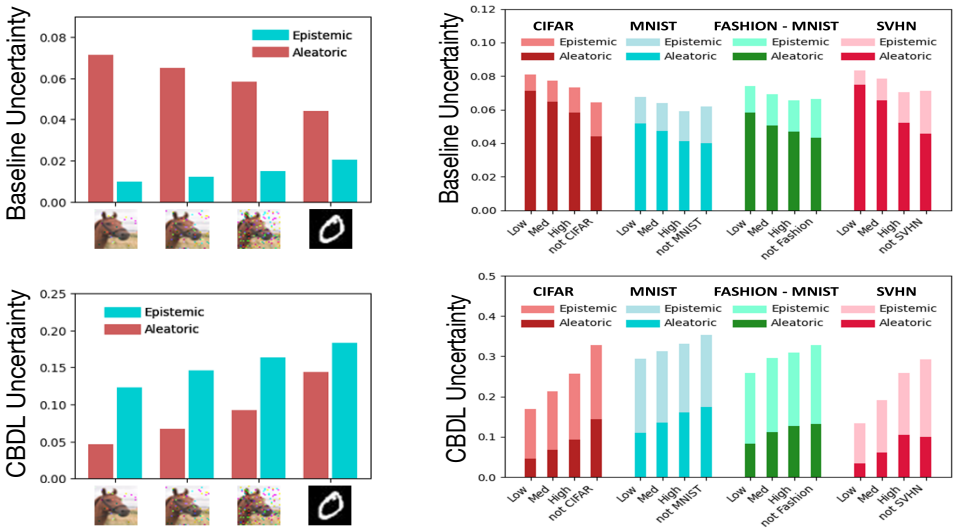
| | | CBDL | | | | | | Baseline | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Epistemic | | | Aleatoric | | | Epistemic | | | Aleatoric | | |
| | | Low | Med | High | Low | Med | High | Low | Med | High | Low | Med | High |
| CIFAR-10 Corruption | gaussian noise | 0.145 | 0.169 | **0.176** | 0.066 | 0.099 | **0.102** | 0.012 | **0.016** | **0.016** | **0.065** | 0.056 | 0.055 |
| | shot noise | 0.129 | 0.158 | **0.174** | 0.053 | 0.084 | **0.104** | 0.01 | 0.014 | **0.016** | **0.069** | 0.061 | 0.055 |
| | speckle noise | 0.128 | 0.158 | **0.174** | 0.053 | 0.081 | **0.102** | 0.01 | 0.014 | **0.016** | **0.069** | 0.061 | 0.055 |
| | impulse noise | 0.134 | 0.161 | **0.174** | 0.052 | 0.085 | **0.118** | 0.011 | 0.015 | **0.017** | **0.069** | 0.059 | 0.051 |
| | defocus blur | 0.105 | 0.129 | **0.176** | 0.032 | 0.048 | **0.095** | 0.008 | 0.01 | **0.016** | **0.076** | 0.071 | 0.057 |
| | gaussian blur | 0.105 | 0.154 | **0.186** | 0.032 | 0.069 | **0.11** | 0.008 | 0.013 | **0.017** | **0.076** | 0.064 | 0.053 |
| | motion blur | 0.137 | 0.166 | **0.175** | 0.055 | 0.085 | **0.1** | 0.011 | 0.015 | **0.016** | **0.068** | 0.06 | 0.056 |
| | zoom blur | 0.148 | 0.161 | **0.176** | 0.063 | 0.077 | **0.1** | 0.012 | 0.014 | **0.016** | **0.066** | 0.062 | 0.056 |
| | snow | 0.131 | 0.157 | **0.163** | 0.05 | 0.077 | **0.086** | 0.01 | 0.014 | **0.015** | **0.07** | 0.062 | 0.059 |
| | fog | 0.104 | 0.122 | **0.154** | 0.033 | 0.045 | **0.09** | 0.008 | 0.009 | **0.013** | **0.076** | 0.072 | 0.06 |
| | brightness | 0.103 | 0.108 | **0.124** | 0.031 | 0.034 | **0.043** | 0.008 | 0.008 | **0.01** | **0.076** | 0.075 | 0.072 |
| | contrast | 0.107 | 0.145 | **0.165** | 0.035 | 0.066 | **0.148** | 0.008 | 0.012 | **0.017** | **0.075** | 0.065 | 0.046 |
| | elastic | 0.134 | 0.144 | **0.168** | 0.053 | 0.059 | **0.089** | 0.011 | 0.012 | **0.015** | **0.069** | 0.067 | 0.058 |
| | pixelate | 0.116 | 0.135 | **0.162** | 0.042 | 0.065 | **0.096** | 0.009 | 0.011 | **0.015** | **0.073** | 0.066 | 0.058 |
| | jpeg | 0.13 | 0.147 | **0.156** | 0.049 | 0.064 | **0.075** | 0.01 | 0.012 | **0.013** | **0.07** | 0.066 | 0.063 |
| | spatter | 0.117 | 0.145 | **0.147** | 0.041 | **0.065** | 0.062 | 0.009 | **0.012** | **0.012** | **0.073** | 0.065 | 0.066 |
| | saturate | 0.112 | 0.113 | **0.131** | 0.041 | 0.039 | **0.046** | 0.008 | 0.009 | **0.011** | 0.073 | **0.074** | 0.07 |
| | frost | 0.124 | 0.153 | **0.166** | 0.047 | 0.075 | **0.099** | 0.01 | 0.013 | **0.015** | **0.071** | 0.063 | 0.056 |
| | | Epistemic | | | Aleatoric | | | Epistemic | | | Aleatoric | | |
| Dataset | MNIST | 0.184 | | | 0.142 | | | 0.021 | | | 0.044 | | |
| | Fashion MNIST | 0.183 | | | 0.147 | | | 0.022 | | | 0.042 | | |
| | SVHN | 0.183 | | | 0.141 | | | 0.019 | | | 0.046 | | |

Table 2: The 4 BNNs trained have the following accuracies : $90, 89, 90, 89$ in percentage terms and rounded to the nearest whole number. For different categories of corruptions, increasing severity leads to higher levels of aleatoric uncertainty for CBDL. When exposed to completely unseen datasets, this reaches its peak. In contrast, the baseline has a reverse trend.

**Summary.** We summarize our results for CIFAR-10 in Figure 6.(a), and for all four datasets in Figure 6.(b). With increasing corruption severity, the aleatoric uncertainty for CBDL becomes more pronounced. This is expected since for higher corruptions, it is hard to reconcile the performance gap by simple data-augmentation, without any additional knowledge. For EBNN, even though the epistemic uncertainty grows, the aleatoric part does not. This counterintuitive – and erroneous – behavior makes us conclude that CBDL is able to quantify and disentangle (predictive) AU and EU better than the baseline.

Note that the absolute uncertainty differs due to the quantities being fundamentally different: variance versus entropy. However, the relative trends are consistent for each dataset, demonstrating the utility of CBDL.

**Other datasets.** The results for MNIST, SVHN, and Fashion-MNIST are presented in Tables 3, 4, and 5, respectively. We observe trends similar to those seen for CIFAR-10: the aleatoric uncertainty increases with corruption severity, and is the highest for completely unknown datasets.




(a) Results for CIFAR-10.

(b) Results for all four datasets, CIFAR-10, MNIST, Fashion-MNIST, and SVHN.

Figure 6: The different trends with increasing corruption severity for EBNN (baseline) versus CBDL. CBDL better inform the degree of shift as compared to EBNN.

| | | CBDL | | | | | | Baseline | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Epistemic | | | Aleatoric | | | Epistemic | | | Aleatoric | | |
| | | Low | Med | High | Low | Med | High | Low | Med | High | Low | Med | High |
| MNIST Corruption | brightness | 0.21 | **0.246** | 0.195 | 0.128 | 0.087 | **0.158** | 0.023 | **0.026** | 0.024 | 0.045 | **0.05** | 0.045 |
| | canny edges | **0.162** | 0.158 | **0.162** | 0.174 | **0.176** | 0.174 | **0.016** | 0.015 | **0.016** | **0.042** | **0.042** | **0.042** |
| | dotted line | **0.18** | 0.173 | 0.176 | 0.097 | **0.102** | **0.102** | **0.014** | 0.013 | **0.014** | **0.055** | **0.055** | **0.055** |
| | fog | 0.185 | 0.187 | **0.19** | 0.163 | **0.167** | 0.166 | **0.023** | **0.023** | **0.023** | **0.043** | 0.042 | **0.043** |
| | glass blur | **0.161** | 0.145 | 0.155 | 0.181 | **0.206** | 0.2 | 0.016 | 0.016 | **0.02** | **0.038** | 0.033 | 0.033 |
| | impulse noise | **0.186** | 0.172 | 0.181 | 0.083 | 0.132 | **0.158** | 0.013 | 0.015 | **0.021** | **0.057** | 0.049 | 0.042 |
| | motion blur | **0.184** | 0.161 | 0.157 | 0.123 | 0.185 | **0.195** | 0.016 | 0.018 | **0.019** | **0.049** | 0.035 | 0.029 |
| | rotate | **0.189** | 0.167 | 0.142 | 0.072 | 0.134 | **0.196** | 0.013 | 0.014 | **0.015** | **0.059** | 0.048 | 0.033 |
| | scale | **0.196** | 0.169 | 0.121 | 0.08 | 0.152 | **0.232** | 0.014 | **0.015** | 0.013 | **0.057** | 0.044 | 0.025 |
| | shear | **0.188** | 0.177 | 0.153 | 0.065 | 0.102 | **0.184** | 0.012 | 0.014 | **0.017** | **0.061** | 0.055 | 0.036 |
| | shot noise | **0.188** | 0.179 | 0.18 | 0.061 | 0.08 | **0.113** | 0.012 | 0.012 | **0.014** | **0.062** | 0.059 | 0.052 |
| | spatter | **0.186** | 0.174 | 0.176 | 0.074 | **0.132** | 0.127 | 0.013 | **0.016** | **0.016** | **0.059** | 0.047 | 0.049 |
| | stripe | 0.18 | 0.182 | **0.184** | **0.165** | 0.163 | 0.161 | **0.021** | 0.02 | **0.021** | **0.04** | **0.04** | **0.04** |
| | translate | 0.191 | **0.192** | **0.192** | 0.061 | 0.086 | **0.128** | 0.013 | 0.015 | **0.019** | **0.061** | 0.056 | 0.046 |
| | zigzag | **0.18** | 0.176 | 0.179 | 0.119 | **0.123** | 0.122 | **0.016** | **0.016** | **0.016** | **0.051** | 0.05 | 0.05 |
| | | Epistemic | | | Aleatoric | | | Epistemic | | | Aleatoric | | |
| Dataset | CIFAR | 0.185 | | | 0.167 | | | 0.023 | | | 0.042 | | |
| | Fashion MNIST | 0.162 | | | 0.187 | | | 0.019 | | | 0.035 | | |
| | SVHN | 0.190 | | | 0.168 | | | 0.022 | | | 0.043 | | |

Table 3: The 4 BNNs trained have the following accuracies : 99, 99, 99, 98 in percentage terms and rounded to the nearest whole number. For different categories of corruptions, increasing severity leads to higher levels of aleatoric uncertainty for CBDL. When exposed to completely unseen datasets, this gets close to the highest aleatoric uncertainty. The same is not true for the baseline. The epistemic uncertainty for CBDL shows a less consistent trend in this case.

### 4.2 Downstream Tasks Performance

As we have shown in the previous section, CBDL is better than single BNNs and ensemble of BNNs at quantifying and disentangling AU and EU. In this section, we show with two applications – motion prediction in autonomous driving scenarios, and blood glucose and insulin dynamics for artificial pancreas control – that CBDL has better downstream tasks capability than EBNN. We do not compare CBDL against belief tracking techniques because these latter require extra assumptions that CBDL do not, see Appendix K.

| | | CBDL | | | | | | Baseline | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Epistemic | | | Aleatoric | | | Epistemic | | | Aleatoric | | |
| | | Low | Med | High | Low | Med | High | Low | Med | High | Low | Med | High |
| SVHN Corruption | brightness | 0.062 | 0.074 | **0.102** | 0.011 | 0.019 | **0.053** | 0.005 | 0.006 | **0.008** | **0.083** | 0.08 | 0.07 |
| | contrast | 0.077 | 0.099 | **0.143** | 0.022 | 0.048 | **0.174** | 0.006 | 0.008 | **0.018** | **0.08** | 0.072 | 0.04 |
| | defocus blur | 0.109 | **0.154** | 0.138 | 0.044 | 0.147 | **0.218** | 0.009 | 0.017 | **0.02** | **0.072** | 0.045 | 0.029 |
| | elastic | 0.182 | 0.177 | **0.192** | **0.14** | 0.122 | 0.102 | **0.023** | 0.021 | 0.022 | 0.04 | 0.046 | **0.05** |
| | fog | 0.157 | 0.171 | **0.181** | 0.095 | 0.124 | **0.135** | 0.016 | 0.02 | **0.023** | **0.056** | 0.047 | 0.042 |
| | frost | 0.089 | 0.118 | **0.131** | 0.026 | 0.052 | **0.063** | 0.007 | 0.01 | **0.012** | **0.078** | 0.069 | 0.066 |
| | gaussian blur | 0.068 | 0.131 | **0.14** | 0.017 | 0.087 | **0.208** | 0.005 | 0.012 | **0.019** | **0.081** | 0.06 | 0.032 |
| | gaussian noise | 0.116 | 0.163 | **0.182** | 0.031 | 0.061 | **0.103** | 0.01 | 0.018 | **0.025** | **0.074** | 0.061 | 0.044 |
| | impulse noise | 0.13 | 0.169 | **0.185** | 0.035 | 0.06 | **0.102** | 0.012 | 0.018 | **0.026** | **0.072** | 0.061 | 0.045 |
| | jpeg | 0.07 | 0.083 | **0.129** | 0.014 | 0.018 | **0.041** | 0.005 | 0.006 | **0.012** | **0.082** | 0.08 | 0.071 |
| | motion blur | 0.1 | 0.157 | **0.167** | 0.027 | 0.078 | **0.149** | 0.008 | 0.015 | **0.02** | **0.077** | 0.06 | 0.041 |
| | pixelate | 0.061 | 0.123 | **0.18** | 0.011 | 0.036 | **0.095** | 0.004 | 0.011 | **0.02** | **0.083** | 0.073 | 0.052 |
| | saturate | 0.063 | 0.065 | **0.141** | 0.011 | 0.012 | **0.075** | 0.005 | 0.005 | **0.014** | **0.083** | **0.083** | 0.062 |
| | shot noise | 0.119 | 0.17 | **0.183** | 0.031 | 0.067 | **0.106** | 0.011 | 0.019 | **0.026** | **0.074** | 0.059 | 0.044 |
| | snow | 0.131 | **0.16** | **0.16** | 0.043 | 0.076 | **0.096** | 0.012 | **0.017** | **0.017** | **0.07** | 0.059 | 0.055 |
| | spatter | 0.088 | 0.127 | **0.163** | 0.021 | 0.037 | **0.06** | 0.007 | 0.012 | **0.017** | **0.079** | 0.072 | 0.063 |
| | speckle noise | 0.1 | 0.143 | **0.181** | 0.023 | 0.046 | **0.085** | 0.008 | 0.015 | **0.023** | **0.077** | 0.067 | 0.052 |
| | zoom blur | 0.061 | 0.061 | **0.064** | 0.011 | 0.012 | **0.013** | 0.004 | 0.004 | **0.005** | **0.083** | **0.083** | **0.083** |
| | | Epistemic | | | Aleatoric | | | Epistemic | | | Aleatoric | | |
| Dataset | CIFAR | 0.181 | | | 0.121 | | | 0.026 | | | 0.040 | | |
| | MNIST | 0.198 | | | 0.063 | | | 0.024 | | | 0.056 | | |
| | Fashion MNIST | 0.199 | | | 0.113 | | | 0.026 | | | 0.041 | | |

Table 4: The 4 BNNs trained have the following accuracies : $95, 95, 96, 95$ in percentage terms and rounded to the nearest whole number. For different categories of corruptions, increasing severity leads to higher levels of aleatoric uncertainty for CBDL. When exposed to completely unseen datasets, this reaches its peak. The same is not true for the baseline. For the epistemic uncertainty as well, there is a clear trend with increasing corruption severity.

| | | CBDL | | | | | | Baseline | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Epistemic | | | Aleatoric | | | Epistemic | | | Aleatoric | | |
| | | Low | Med | High | Low | Med | High | Low | Med | High | Low | Med | High |
| Fashion MNIST Corruption | brightness | 0.199 | **0.212** | 0.205 | **0.108** | 0.106 | 0.094 | 0.021 | **0.024** | 0.022 | **0.051** | 0.048 | **0.051** |
| | canny edges | **0.18** | 0.175 | 0.177 | 0.157 | **0.164** | 0.16 | **0.021** | 0.02 | 0.02 | **0.043** | 0.042 | **0.043** |
| | dotted line | 0.167 | **0.168** | **0.168** | **0.051** | **0.051** | **0.051** | 0.012 | **0.013** | **0.013** | **0.067** | 0.066 | 0.066 |
| | fog | **0.19** | 0.185 | 0.187 | 0.128 | **0.132** | 0.129 | **0.021** | 0.02 | **0.021** | **0.046** | 0.045 | **0.046** |
| | glass blur | 0.189 | 0.189 | **0.193** | 0.114 | 0.127 | **0.133** | 0.017 | 0.019 | **0.022** | **0.052** | 0.048 | 0.045 |
| | impulse noise | 0.169 | 0.196 | **0.202** | 0.052 | 0.09 | **0.111** | 0.013 | 0.019 | **0.02** | **0.066** | 0.055 | 0.051 |
| | motion blur | **0.182** | 0.171 | 0.163 | 0.106 | 0.157 | **0.174** | 0.017 | **0.02** | 0.019 | **0.053** | 0.041 | 0.038 |
| | rotate | **0.185** | 0.173 | 0.162 | 0.074 | 0.164 | **0.184** | 0.016 | **0.02** | **0.02** | **0.06** | 0.037 | 0.035 |
| | scale | 0.154 | **0.172** | 0.127 | 0.054 | 0.123 | **0.223** | 0.01 | **0.015** | 0.014 | **0.067** | 0.05 | 0.028 |
| | shear | 0.172 | 0.177 | **0.182** | 0.063 | 0.13 | **0.144** | 0.014 | 0.019 | **0.02** | **0.063** | 0.045 | 0.039 |
| | shot noise | 0.16 | 0.178 | **0.197** | 0.046 | 0.059 | **0.088** | 0.01 | 0.013 | **0.017** | **0.068** | 0.064 | 0.056 |
| | spatter | 0.15 | 0.189 | **0.191** | 0.044 | **0.082** | 0.074 | 0.01 | **0.018** | 0.017 | **0.07** | 0.057 | 0.059 |
| | stripe | **0.21** | 0.208 | 0.208 | 0.125 | **0.126** | 0.125 | **0.025** | **0.025** | **0.025** | **0.041** | **0.041** | **0.041** |
| | translate | 0.153 | **0.189** | 0.17 | 0.046 | 0.094 | **0.158** | 0.01 | 0.018 | **0.019** | **0.069** | 0.054 | 0.041 |
| | zigzag | 0.187 | **0.19** | 0.189 | 0.065 | 0.065 | **0.066** | **0.016** | **0.016** | **0.016** | **0.061** | **0.061** | **0.061** |
| | | Epistemic | | | Aleatoric | | | Epistemic | | | Aleatoric | | |
| Dataset | CIFAR | 0.205 | | | 0.104 | | | 0.022 | | | 0.050 | | |
| | MNIST | 0.165 | | | 0.182 | | | 0.021 | | | 0.033 | | |
| | SVHN | 0.218 | | | 0.109 | | | 0.026 | | | 0.046 | | |

Table 5: The 4 BNNs trained have the following accuracies : $93, 92, 92, 92$ in percentage terms and rounded to the nearest whole number. For different categories of corruptions, increasing severity leads to higher levels of aleatoric uncertainty for CBDL, which is high when exposed to completely unseen datasets as well. The same is not true for the baseline.

#### 4.2.1 Motion Prediction for Autonomous Racing

In this case study, we demonstrate the utility of CBDL for motion prediction in autonomous driving scenarios. An important challenge in autonomous driving is understanding the intent of other agents and predicting their future trajectories to allow for safety-aware planning. In autonomous racing, where control is pushed

to the dynamical limits, accurate and robust predictions are even more essential for outperforming opponent agents while assuring safety. CBDL provides a straightforward method for quantifying uncertainty and deriving robust prediction regions for anticipating an agent's behavior.

We use the problem settings in Tumu et al. (2023) to define the problem of obtaining prediction sets for future positions of an autonomous racing agent. Our results show that the prediction regions have improved coverage when compared to EBNN. These results hold in both in-distribution and out-of-distribution settings, which are described below.

**Problem.** Let $O^i(t,l) \equiv O^i = \{\pi^i_{t-l}, \ldots, \pi^i_t\}$ denote the $i$-th trajectory instance of an agent at time $t$, consisting of the observed positions from time $t-l$ up to time $t$. Let then $C^i$ be a time-invariant context variable. Let also $F^i(t,h) \equiv F^i = \{\pi^i_{t+1}, \ldots, \pi^i_{t+h}\}$ be the collection of the next $h$ future positions. We wish to obtain a model $M$ that predicts region $\mathcal{R}_\alpha$ with probabilistic guarantees. In particular, for EBNN $\mathcal{R}_\alpha$ is the $\alpha$-level HDR $R_\alpha(P_{\text{ens}})$ of $P_{\text{ens}}$, so that $P_{\text{ens}}[F^i \in R_\alpha(P_{\text{ens}})] \geq 1-\alpha$, while for CBDL $\mathcal{R}_\alpha = IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$, so that $\underline{\hat{P}}^{\text{pred}}[F^i \in IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})] \geq 1-\alpha$, or equivalently, $\hat{P}^{\text{pred}}[F^i \in IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})] \geq 1-\alpha$, for all $\hat{P}^{\text{pred}} \in \hat{\mathcal{P}}_{\text{pred}}$.

The dataset consists of instances of $(O^i, F^i)$ divided into a training set $D_{\text{train}}$ and a testing set $D_{\text{test}}$. We train an uncertainty-aware model on $D_{\text{train}}$ that computes the triplet $(F^i_l, F^i_m, F^i_u) = M(O^i, C^i)$ where $F^i_l$, $F^i_u$, $F^i_m$ are the lower, upper, and mean predictions of the future positions, respectively.

The dataset $D_{\text{all}}$ is created by collecting simulated trajectories of autonomous race cars in the F1Tenth-Gym (O'Kelly et al., 2020); for details, see Tumu et al. (2023). As shown in Figure 7, different racing lines were utilized including the center, right, left, and optimal racing line for the Spielberg track.

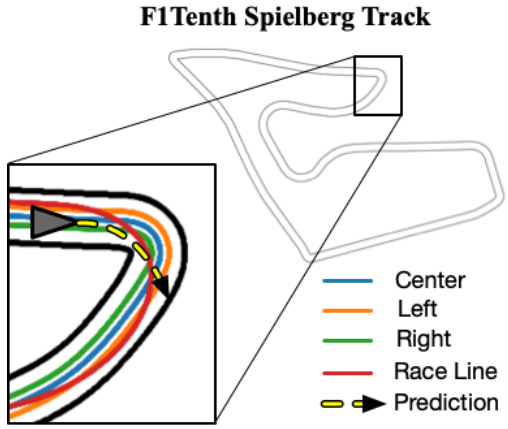



Figure 7: Motion Prediction for F1Tenth-Gym Environment (O'Kelly et al., 2020). Data is collected by simulating various racing lines on the Spielberg Track.

We denote these by $D_{\text{center}}$, $D_{\text{right}}$, $D_{\text{left}}$, and $D_{\text{race}}$, respectively. Position $\pi$ is a vector $\pi = (a, b, \vartheta, v)^\top$, where $a$ and $b$ are coordinates in a 2-dimensional Euclidean space, and $\vartheta$ and $v$ are the heading and speed, respectively. In total, the $D_{\text{all}}$ consists of 34686 train instances, 4336 validation instances, and 4336 test instances.

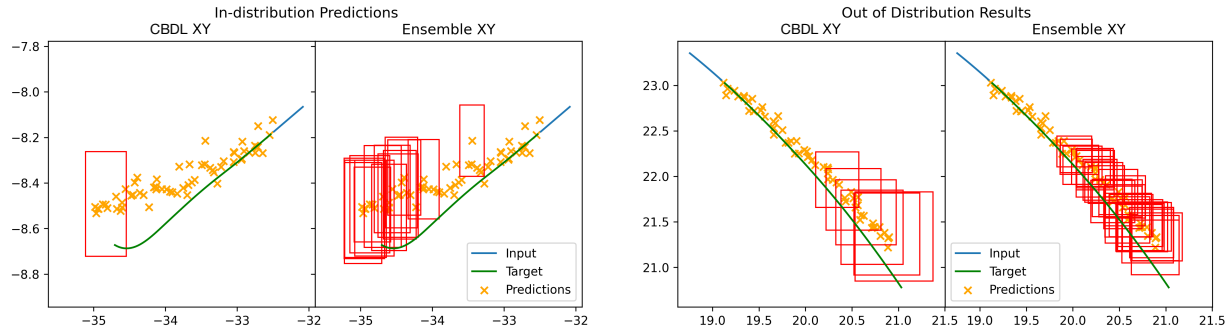
**In-distribution vs. Out-of-distribution.** We consider the prediction task to be in-distribution when $D_{\text{train}}, D_{\text{test}} \subset D_{\text{all}}$. It is out-of-distribution (OOD) when $D_{\text{train}} \subset D_{\text{center}} \cup D_{\text{right}} \cup D_{\text{left}}$ and $D_{\text{test}} \subset D_{\text{race}}$.

**Metrics.** We train the ensemble of BNNs and the CBDL models, $M_{\text{ens}}$ and $M_{\text{CBDL}}$ respectively, using the same architecture and different seeds. As for section 4.1, for CBDL this corresponds to having a non-singleton prior FGCS, and a singleton likelihood FGCS. We compare the performance with respect to the test set by computing the single-step coverage, where each prediction time-step is treated independently, and the multi-step coverage, which considers the entire $h$-step prediction.

Figure 8.(a) depicts a sample of the in-distribution evaluation for each of the models. For a given trajectory, the red boxes indicate when the prediction region did not cover the actual trajectory at that time-step. Qualitatively, $M_{\text{CBDL}}$ has less missed time steps when compared to $M_{\text{ens}}$. Table 6 shows that CBDL performs better in terms of both one-step and multi-step coverage. Similar results can be observed for the OOD scenario. There, all models were trained on racing lines which are predominantly parallel to the track curvature. As a consequence, when the test set consists of instances with higher curvatures, the overall coverage of all models degrades. This can be seen in Figure 8.(b), where the prediction of the models (orange) tends to be straight while the actual trajectory is more curved (green). Despite this, the figure and the coverage metrics in Table 6 show how CBDL exhibits a more robust behavior.

| In-Distribution Results | | | | | | |
|---|---|---|---|---|---|---|
| | Ensemble | | | CBDL | | |
| $1-\alpha$ | 0.9 | 0.95 | 0.99 | 0.9 | 0.95 | 0.99 |
| One-step | 0.962 | 0.980 | 0.992 | **0.992** | **0.995** | **0.997** |
| Multi-step | 0.638 | 0.826 | 0.937 | **0.914** | **0.948** | **0.979** |
| **Out-of-Distribution Results** | | | | | | |
| | Ensemble | | | CBDL | | |
| $1-\alpha$ | 0.9 | 0.95 | 0.99 | 0.9 | 0.95 | 0.99 |
| One-step | 0.919 | 0.950 | 0.980 | **0.979** | **0.988** | **0.995** |
| Multi-step | 0.532 | 0.703 | 0.860 | **0.825** | **0.884** | **0.943** |

Table 6: F1Tenth coverage results. We report one-step coverage and multi-step coverage across 3 different values of $\alpha$. CBDL exceed coverage of EBNNs in all settings.




(a) (b)

Figure 8: In both pictures, the red boxes indicate when the prediction region did not cover the actual trajectory at that time-step. **Left:** F1Tenth In-Distribution results. Given an input of past observations, CBDL exhibits better coverage of the future target trajectory. Predictions which do not cover the target within the desired $1-\alpha$ level are indicated in red. **Right:** F1Tenth Out-Of-Distribution (OOD) results. Robust performance is exhibited by CBDL when compared to EBNN in OOD settings.

#### 4.2.2 Artificial Pancreas Control

**Overall Setup.** In this next case study we consider the problem of data-driven control of human blood glucose-insulin dynamics, using an artificial pancreas system, see Figure 9.

External insulin delivery is accomplished by using an insulin pump controlled by the artificial pancreas software, which attempts to regulate the blood-glucose (BG) level of the patient within the euglycemic range of $[70, 180]$ mg/dl (Kushner et al., 2018). Levels below 70 mg/dl lead to hypoglycemia, which can

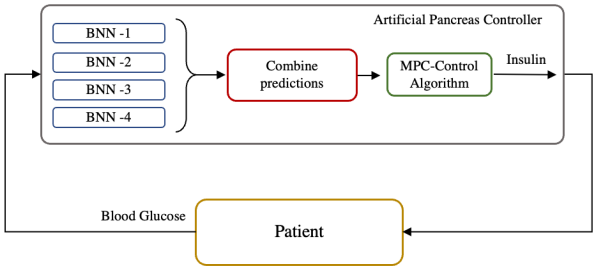



Figure 9: The Bayesian Neural Networks predict a future blood glucose value. These individual predictions are combined to get a robust estimate of the true value as an interval. This is used by the Model Predictive Control (MPC) algorithm to recommend insulin dosage for the patient. The patient block in our experiment is simulated using the virtual patient models from the UVa-Padova simulator.

lead to loss of consciousness, coma or even death. On the other hand, levels above 300 mg/dl lead to a condition called ketoacidosis, where the body can break down fat due to lack of insulin, and lead to build up of ketones. In order to treat this situation, patients receive external insulin delivery through insulin pumps. Artificial Pancreas (AP) systems can remedy this situation by measuring the blood glucose level, and automatically injecting insulin into the blood stream. Thus, we define the *unsafe regions* of the space as $G(t) \in (-\infty, 70) \cup (300, \infty)$, where $G(t)$ is the BG value at time $t$. This is the shaded region in Figure 10.

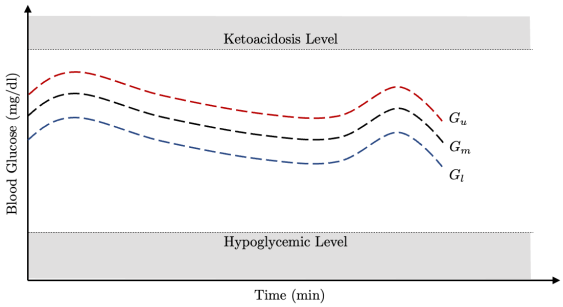



Figure 10: Starting from an initial glucose value, the task of the artificial pancreas controller is to maintain the blood glucose value within safe operating limits using insulin as a mode of control.

**Neural Network Models and Controller.** Deep Neural Networks are effective in capturing the BG-insulin dynamics for personalized medical devices (Kushner et al., 2018). This allows for improved device performance. Even though standard Feedforward Neural Networks can be used, Bayesian Neural Networks (BNNs), and especially a collection of multiple BNNs, offer a better alternative towards uncertainty aware predictions. Here, we test the ramifications of these prediction sets, when used inside an online receding horizon control scheme for insulin delivery. We use the standard MPC control scheme for this purpose, well-known in the literature (Dutta et al., 2018). More formally, let $G(t)$ and $I(t) \equiv I_t$ be the blood-glucose and insulin values at time $t$, respectively. We denote the finite length trajectory of length $H$ as $\overleftarrow{G}_H(t) := [G(t-H+1), \ldots, G(t)]$, and $\overleftarrow{I}_H(t) := [I(t-H+1), \ldots, I(t)]$. An uncertainty aware model $M$ computes the triplet $(G_l(t+l), G_m(t+l), G_u(t+l)) = M(\overleftarrow{G}_H(t), \overleftarrow{I}_H(t))$, where $G_m$ is the mean prediction output, and $G_l, G_u$ are the lower and upper predictions of the glucose value, respectively. By design, it is true that $G_l \leq G_m \leq G_u$. A model predictive control algorithm – whose cost function we denote by $J$ – solves $\arg\min_{I_0, I_1, \ldots, I_{k-1}} \sum_{i=0}^{k-1} J(M(\overleftarrow{G}_H(t+i), \overleftarrow{I}_H(t+i)))$.

After every time step, the control algorithm picks the first insulin input $I_0$ as the insulin bolus for the patient, and discards the rest. Cost function $J$ takes into account three factors, (i) Distance of the mean prediction level $G_m$ at each time step from a target value of 120 mg/dl, (ii) Distance of upper and lower predictions

($G_u$ and $G_l$) from the unsafe regions of the state space $G(t) > 300$ and $G(t) < 70$, and (iii) Total insulin injected $\sum_{t=0}^{k-1} I_t$.

Starting with some initial glucose value $G(0)$, we measure the performance of the artificial pancreas controller as the fraction of time it spends in the unsafe regions,

$$t_{\text{unsafe}} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1} \left\{ G(t) \in (-\infty, 70) \cup (300, \infty) \right\},$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. A lower value is more desirable. We compare EBNN and CBDL as different realizations of the model $M$.

**Distribution Shift using Meals.** A well known problem with learned models is distribution shift. Bayesian Neural Networks can address this issue by apprising the end user of the increased uncertainty. For regression models of the type described above, this appears as larger prediction intervals $[G_l, G_u]$. The artificial pancreas controller can run into this situation in the following way: the insulin-glucose time series data collected for training the data-driven model $M$ can be without meals, while at test time the patient can have meals. This creates a distribution shift between the training and test time data. Fortunately, the UVa-Padova simulator (Dalla Man et al., 2013) allows us to create datasets with and without meal inputs. In this case study, the training data was obtained by randomly initializing the BG value in the range $[120, 190]$, and simulating the patient for 720 minutes. The controller was executed at 5 minutes intervals. At test time the patient was supplied meals at specific time intervals (for details, see Appendix J). This creates a significant distribution shift since meals are effectively an unknown variable which can affect the system state. However, from the controller's perspective this is practical, since patients can have unannounced meals.

**Results and Discussion.** To capture the difference in performance between EBNN and CBDL, we compute $\text{Perf}_{\text{diff}} := (t_{\text{unsafe}}^{\text{EBNN}} - t_{\text{unsafe}}^{\text{CBDL}})/t_{\text{unsafe}}^{\text{EBNN}}$. Both $t_{\text{unsafe}}^{\text{EBNN}}$ and $t_{\text{unsafe}}^{\text{CBDL}}$ depend on interval $[G_l, G_u]$; for EBNN, this latter corresponds to the $\alpha$-level HDR $R_\alpha(P_{\text{ens}})$ associated with EBNN distribution $P_{\text{ens}}$, while for CBDL it corresponds to the IHDR $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$. We consider one case in which an CBDL is trained using a credal prior set and only one likelihood (we choose different seeds which initialize the prior distributions but we keep the same architecture for the BNNs), and another case in which we do the opposite (we use the same seed and different architectures).

We report $\text{Perf}_{\text{diff}}$, across different choices in Table 7. We observe that for lower values of $\alpha$ the gains of CBDL are more pronounced. This means that when larger significance levels need to be ensured, CBDL is to be preferred to ensemble of BNNs. As discussed before, the CBDL procedure considers all the infinitely many possible priors that can be expressed as a convex combination of the priors that the user specifies at the beginning of the analysis. The same holds for the likelihoods. While this results in a more conservative estimate as compared to the EBNN framework, CBDL produces controllers which respect the safety limits better. To see that CBDL is more conservative than EBNN, notice that when combining predictive distributions from multiple BNNs, CBDL combines the predictions via a finitely generated credal set (FGCS) whose extrema are the individual distributions. On the contrary, an EBNN takes an average of the individual distributions to compute the ensemble distribution $P_{\text{ens}}$. The union of the HDRs of the predictive distributions is more conservative (i.e., broader) than the HDR of the single ensemble distribution. While the approach by EBNN seems like a reasonable choice on the surface, it falls short in capturing the uncertainty necessary for the downstream task. For more details on this case study, see Appendix J.

| $1-\alpha$ | 0.9 | 0.95 | 0.99 |
|---|---|---|---|
| Varying Seeds | -5.8% | **5.5**% | **0.6**% |
| Varying Architectures | -6.6% | **0.9**% | **0.3**% |

Table 7: We report the performance improvements when using IBNNs as compared to EBNNs across 3 different values of $\alpha$. Row 1 corresponds to the case where the individual BNNs are trained with different seeds for the prior distribution; and Row 2 is the case when the BNNs have different architectures.

## 5 Related Work

In Corani et al. (2012), the authors introduce credal classifiers (CCs) as a generalization of classifiers based on Bayesian networks. Unlike CCs, CBDL does not require independence assumptions between non-descendant, non-parent variables. In addition, CBDL avoids NP-hard complexity issues of searching for optimal structure in the space of Bayesian networks (Chickering et al., 2004). In Manchingal & Cuzzolin (2022), an epistemic convolutional neural network (ECNN) is developed that explicitly models the epistemic uncertainty induced by training data of limited size and quality. A clear distinction is that ECNNs measure uncertainty in target-level representations whereas CBDL identifies the uncertainty measure on the output space $\mathcal{Y}$. Despite the merit of their work, we believe CBDL achieves greater generality, since it is able to quantify both aleatoric and epistemic uncertainties, and is applicable to problems beyond classification. For a review of the state of the art concerning the distinction between EU and AU we refer to Hüllermeier & Waegeman (2021) and to Manchingal & Cuzzolin (2022). We also point out how CBDL has been recently used to solve prior-likelihood conflicts in Bayesian statistics (Marquardt et al., 2023). Further references can be found in Appendix L.

We also point out how there exist other efficient methods which perform approximate Variational Inference via dropouts in deep neural networks (Kendall & Gal, 2017b; Gal & Ghahramani, 2016). As mentioned at the end of Section 3.1, we can easily adapt CBDL to use such dropout approximations in **Steps 3-4** of Algorithm 1. Since our contribution is centered around how different predictions can be combined via an FGCS, we used the de-facto standard for performing inference on BNNs, which is based on off-the-shelf VI techniques. In the future, we plan to study the effect on computational complexity and uncertainty quantification capability of a CBDL procedure that approximates posterior and predictive distributions via dropout.

We do not consider Bayesian Model Averaging (BMA) as a baseline for CBDL for two main reasons. First, BMA applied to deep learning needs to implement full batch Hamiltonian Monte Carlo in order to get to the true posterior (Izmailov et al., 2021b). Given the number of parameters in modern deep learning architectures – in the order of millions – this is realistically possible at an experimental level only to labs with access to industry scale computational resources. In order to be practically relevant, we limit our experiments to the more well-understood realm of Variational Inference on BNNs. In addition, Izmailov et al. (2021a) show the pitfalls of BMA in the context of Bayesian Neural Networks, a further reason not to use a Highest Density Region resulting from BMA as a baseline for CBDL.

## 6 Conclusion

We presented CBDL, a procedure that can be seen as a non-condensed, uncountably infinite ensemble of BNNs, carried out using only finitely many elements. It allows to distinguish between AU and EU, and to quantify them. We showed how it can be used to specify a set of outputs – the IHDR – that enjoys probabilistic guarantees. We showed empirically that it improves on the Bayesian state of the art at gauging AU and EU, and we also demonstrated its downstream tasks capabilities.

We point out how a region that improves on $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$, meaning that it would be tighter, is

$$IR'_\alpha(\hat{\mathcal{P}}_{\text{pred}}) = \{y \in \mathcal{Y} : \underline{\hat{p}}^{\text{pred}}(y) \geq \underline{\hat{p}}^{\alpha}\},$$

where $\underline{\hat{p}}^{\text{pred}} := \min_{k,s} \hat{p}^{\text{pred}}_{k,s}$, and $\underline{\hat{p}}^{\alpha}$ is the largest constant such that $\underline{\hat{P}}^{\text{pred}}[y \in IR'_\alpha(\hat{\mathcal{P}}_{\text{pred}})] \geq 1 - \alpha$. The problem with $IR'_\alpha$ is that, while the highest density regions $R_\alpha(\hat{P}^{\text{pred}}_{k,s})$ associated with the predictive distributions in $\text{ex}\hat{\mathcal{P}}_{\text{pred}}$ can be computed using off-the-shelf tools, calculating $\underline{\hat{p}}^{\text{pred}}$ and $\underline{\hat{p}}^{\alpha}$ would have been much more computationally expensive. In addition, it would have required to come up with a new technique to find $\underline{\hat{p}}^{\text{pred}}$ and $\underline{\hat{p}}^{\alpha}$. We defer studying this to future work.

We also plan to apply CBDL to continual learning (CL) to overcome the curse of dimensionality and to capture an agent's preference over the tasks to perform, similarly to Lu et al. (2023).

Furthermore, we intend to relate CBDL to Bayesian Model Selection (BMS) (Ghosh et al., 2019). This latter suffers from the same problem as "regular" Bayesian inference. That is, while it tries to come up with a

sophisticate prior that induces shrinkage, it still relies on the "correctness" of that prior, i.e. on correctly specifying the prior's parameters. In the future, an interesting way of combining CBDL with BMS will be to use a finite number of regularized horseshoe priors, as suggested by Ghosh et al. (2019, Section 3.2), as extreme elements of the prior credal set.

We also call attention to the fact that CBDL is a model-based approach. The relationship with model-free approaches such as conformal prediction (Shafer & Vovk, 2008) will be the object of future studies. In particular, we are interested in finding in which cases IHDRs are narrower than conformal regions, and vice versa, and whether IHDRs enjoy the same probabilistic guarantees as conformal regions.

Finally, we point out how one possible way of easing the burden of the combinatorial task in **Step 3** of Algorithm 1 is to specify a prior credal set whose size strikes the perfect balance between being "vague enough" so that we do not underestimate the EU, and being "small enough" so that CBDL is actually implementable. We suspect conjugacy of the priors may play a key role in this endeavor. Because of its centrality, we defer the study of "optimal prior credal sets" to future work.

## References


Joaquín Abellán and Serafín Moral. A non-specificity measure for convex sets of probability distributions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8(3):357–367, 2000.

Joaquín Abellán, George Jiří Klir, and Serafín Moral. Disaggregated total uncertainty measure for credal sets. *International Journal of General Systems*, 1(35):29–44, 2006.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14927–14937. Curran Associates, Inc., 2020.

Thomas Augustin, Frank P.A. Coolen, Gert De Cooman, and Matthias C.M. Troffaes (eds.). *Introduction to imprecise probabilities*. Wiley Series in Probability and Statistics. John Wiley and Sons, 2014.

David Avis and Komei Fukuda. Reverse search for enumeration. *Discrete Applied Mathematics*, 65:21–46, 1996.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816 – 845, 2023. doi: 10.1214/23-AOS2276. URL `https://doi.org/10.1214/23-AOS2276`.

James O. Berger. The robust Bayesian viewpoint. In Joseph B. Kadane (ed.), *Robustness of Bayesian Analyses*. Amsterdam : North-Holland, 1984.

Jose M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B*, 41(2):113–128, 1979.

George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

Yuhang Cai and Lek-Heng Lim. Distances between probability distributions of different dimensions. *IEEE Transactions on Information Theory*, 68(6):4020–4031, 2022.

Michele Caprio. Refined Pinsker's and reverse Pinsker's inequalities for probability distributions of different dimensions. *IEEE Access*, 10:116425–116431, 2022.

Michele Caprio and Ruobin Gong. Dynamic precise and imprecise probability kinematics. In Enrique Miranda, Ignacio Montes, Erik Quaeghebeur, and Barbara Vantaggi (eds.), *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*, volume 215 of *Proceedings of Machine Learning Research*, pp. 72–83. PMLR, 11–14 Jul 2023.

Michele Caprio and Sayan Mukherjee. Ergodic theorems for dynamic imprecise probability kinematics. *International Journal of Approximate Reasoning*, 152:325–343, 2023.

Michele Caprio, Kuk Jin Jang, Souradeep Dutta, Shireen Manchingal, Fabio Cuzzolin, Oleg Sokolsky, and insup Lee. Credal and interval deep evidential classification. *Technical Reports of the PRECISE Center*, 2024a.

Michele Caprio, Yusuf Sale, Eyke Hüllermeier, and Insup Lee. A Novel Bayes' Theorem for Upper Probabilities. In Fabio Cuzzolin and Maryam Sultana (eds.), *Epistemic Uncertainty in Artificial Intelligence*, pp. 1–12, Cham, 2024b. Springer Nature Switzerland.

Simone Cerreia-Vioglio, Fabio Maccheroni, and Massimo Marinacci. Ergodic theorems for lower probabilities. *Proceedings of the American Mathematical Society*, 144:3381–3396, 2015.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1356–1367. Curran Associates, Inc., 2020.

David M. Chickering, David Heckerman, and Christopher Meek. Large-sample learning of Bayesian networks is NP-hard. *The Journal of Machine Learning Research*, 5:1287–1330, 2004.

Frank P. A. Coolen. Imprecise highest density regions related to intervals of measures. *Memorandum COSOR*, 9254, 1992.

Giorgio Corani, Alessandro Antonucci, and Marco Zaffalon. *Bayesian Networks with Imprecise Probabilities: Theory and Application to Classification*, chapter 4 of Data Mining: Foundations and Intelligent Paradigms: Volume 1: Clustering, Association and Classification, pp. 49–93. Berlin, Germany : Springer, 2012.

Fabio G. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000a. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(00)00029-1. URL `https://www.sciencedirect.com/science/article/pii/S0004370200000291`.

Fabio Gagliardi Cozman. Credal networks. *Artificial Intelligence*, 120:199–233, 2000b.

Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc D. Breton, Boris Kovatchev, and Claudio Cobelli. The UVA/PADOVA type 1 diabetes simulator: New features. *Journal of Diabetes Science and Technology*, 8(1):26–34, 2013.

George B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1963. URL `http://www.jstor.org/stable/j.ctt1cx3tvg`.

Bruno de Finetti. *Theory of Probability*, volume 1. New York : Wiley, 1974.

Bruno de Finetti. *Theory of Probability*, volume 2. New York : Wiley, 1975.

Thierry Denœux. An evidential neural network model for regression based on random fuzzy numbers. In Sylvie Le Hégarat-Mascle, Isabelle Bloch, and Emanuel Aldea (eds.), *Belief Functions: Theory and Applications*, pp. 57–66, Cham, 2022. Springer International Publishing.

Thierry Denœux. Quantifying prediction uncertainty in regression using random fuzzy sets: the ENNreg model. *IEEE Transactions on Fuzzy Systems*, pp. 1–10, 2023.

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 *International Conference on Machine Learning*, pp. 1184–1193. PMLR, 2018.

Didier Dubois and Eyke Hüllermeier. Comparing probability measures using possibility theory: A notion of relative peakedness. *International Journal of Approximate Reasoning*, 45(2):364–385, 2007. Eighth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2005).

Souradeep Dutta, Taisa Kushner, and Sriram Sankaranarayanan. Robust data-driven control of artificial pancreas systems using neural networks. In Milan Češka and David Šafránek (eds.), *Computational Methods in Systems Biology*, pp. 183–202, Cham, 2018. Springer International Publishing. ISBN 978-3-319-99429-1.

Romain Egele, Romit Maulik, Krishnan Raghavan, Prasanna Balaprakash, and Bethany Lusch. Autodeuq: Automated deep ensemble with uncertainty quantification. *CoRR*, abs/2110.13511, 2021.

Daniel Ellsberg. Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4): 643–669, 1961.

Emlyn Flint, Florence Chikurunhe, and Anthony Seymour. Regime-based tactical allocation for equity factors and balanced portfolios. *SSRN Electronic Journal*, 01 2017.

Benoît Fortin, Samir Hachour, and François Delmotte. Multi-target PHD tracking and classification using imprecise likelihoods. *International Journal of Approximate Reasoning*, 90:17–36, 2017.

Vincent Fortuin, Adrià Garriga-Alonso, Mark van der Wilk, and Laurence Aitchison. BNNpriors: A library for bayesian neural network inference with different prior distributions. *Software Impacts*, 9:100079, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

Soumya Ghosh, Jiayu Yao, and Finale Doshi-Velez. Model selection in Bayesian neural networks via horseshoe priors. *Journal of Machine Learning Research*, 20(182):1–46, 2019. URL `http://jmlr.org/papers/v20/19-236.html`.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Available at arXiv:2305.12616*, 2023.

Itzhak Gilboa and Massimo Marinacci. Ambiguity and the Bayesian paradigm. In Daron Acemoglu, Manuel Arellano, and Eddie Dekel (eds.), *Advances in Economics and Econometrics, Tenth World Congress*, volume 1. Cambridge : Cambridge University Press, 2013.

Ethan Goan and Clinton Fookes. *Bayesian Neural Networks: An Introduction and Survey*, pp. 45–87. Cham, Switzerland : Springer International Publishing, 2020.

Ruobin Gong and Xiao-Li Meng. Judicious judgment meets unsettling updating: dilation, sure loss, and Simpson's paradox. *Statistical Science*, 36(2):169–190, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *CoRR*, abs/1903.12261, 2019.

Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. Hoboken, New Jersey : Wiley, 2nd edition, 2009.

Rob J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2): 120–126, 1996.

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

Pavel Izmailov, Patrick Nicholson, Sanae Lotfi, and Andrew G Wilson. Dangers of bayesian model averaging under covariate shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3309–3322. Curran Associates, Inc., 2021a.

Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. What are bayesian neural network posteriors really like? *CoRR*, abs/2104.14421, 2021b. URL https://arxiv.org/abs/2104.14421.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are Bayesian neural network posteriors really like? In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4629–4640. PMLR, 2021c.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society A*, 186(1007):453–461, 1946.

Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on Bayesian neural networks – A tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

Ngumbang Juat, Mike Meredith, and John Kruschke. Package 'hdinterval', 2022. URL https://cran.r-project.org/web/packages/HDInterval/HDInterval.pdf. Accessed on May 9, 2023.

Sanyam Kapoor, Wesley J Maddox, Pavel Izmailov, and Andrew Gordon Wilson. On uncertainty, tempering, and data augmentation in bayesian classification. *arXiv preprint arXiv:2203.16481*, 2022.

Ramneet Kaur, Xiayan Ji, Souradeep Dutta, Michele Caprio, Yahan Yang, Elena Bernardis, Oleg Sokolsky, and Insup Lee. Using semantic information for defining and detecting OOD inputs. *Available at arXiv:2302.11019*, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017a.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5580–5590, Red Hook, NY, USA, 2017b. Curran Associates Inc. ISBN 9781510860964.

Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5275–5285. PMLR, 13–18 Jul 2020.

John Klein, Christèle Lecomte, and Pierre Miché. Hierarchical and conditional combination of belief functions induced by visual tracking. *International Journal of Approximate Reasoning*, 51(4):410–428, 2010.

Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. MOPED: efficient priors for scalable variational inference in bayesian deep neural networks. *CoRR*, abs/1906.05323, 2019.

Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Specifying weight priors in Bayesian deep neural networks with empirical Bayes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (04):4477–4484, April 2020. doi: 10.1609/aaai.v34i04.5875. URL https://ojs.aaai.org/index.php/AAAI/article/view/5875.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 (Canadian Institute for Advanced Research). *Available at http://www.cs.toronto.edu/ kriz/cifar.html*, 2009.

Meelis Kull and Peter A. Flach. Reliability Maps: A Tool to Enhance Probability Estimates and Improve Classification Accuracy. In Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo (eds.), *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pp. 18–33, Berlin, Heidelberg, 2014. Springer. ISBN 978-3-662-44851-9.

Taisa Kushner, David Bortz, David M. Maahs, and Sriram Sankaranarayanan. A data-driven approach to artificial pancreas verification and synthesis. In *Proceedings of the 9th ACM/IEEE International Conference on Cyber-Physical Systems*, ICCPS '18, pp. 242–252. IEEE Press, 2018. ISBN 9781538653012.

Jouko Lampinen and Aki Vehtari. Bayesian approach for neural networks - review and case studies. *Neural Networks*, 4:257–274, 2001.

Daniel Lassiter. Representing credal imprecision: from sets of measures to hierarchical Bayesian models. *Philosophical Studies*, 177(6):1463–1485, 2020.

Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Peter Lenk and Bryan Orme. The value of informative priors in Bayesian inference with sparse data. *Journal of Marketing Research*, 46(6):832–845, 2009. ISSN 00222437.

Isaac Levi. *The Enterprise of Knowledge*. London, UK : MIT Press, 1980.

Vivian Lin, Kuk Jin Jang, Souradeep Dutta, Michele Caprio, Oleg Sokolsky, and Insup Lee. Reversing distribution shifts using reinforcement learning. *Available at arXiv:2302.10341*, 2023.

Pengyuan Lu, Michele Caprio, Eric Eaton, and Insup Lee. Zero-shot task preference addressing enabled by imprecise Bayesian continual learning. *Available at arXiv:2305.14782*, 2023.

Shireen Kudukkil Manchingal and Fabio Cuzzolin. Epistemic deep learning. *Available at arxiv:2206.07609*, 2022.

Massimo Marinacci and Luigi Montrucchio. Introduction to the mathematics of ambiguity. In Itzhak Gilboa (ed.), *Uncertainty in economic theory: a collection of essays in honor of David Schmeidler's 65th birthday*. London : Routledge, 2004.

Alexander Marquardt, Julian Rodemann, and Thomas Augustin. An empirical study of prior-data conflicts in Bayesian neural networks. *Poster presented at 13th International Symposium on Imprecise Probabilities: Theories and Applications (ISIPTA) 2023*, 2023.

Jacques Paul Migne. De coelesti hierarchia. In *Patrologia Graeca*, volume 3. Paris : Imprimerie Catholique, 1857.

Norman Mu and Justin Gilmer. MNIST-C: A robustness benchmark for computer vision. *CoRR*, abs/1906.02337, 2019.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Available at arXiv:2402.19460*, 2024.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

Matthew O'Kelly, Hongrui Zheng, Dhruv Karthik, and Rahul Mangharam. F1TENTH: An Open-source Evaluation Environment for Continuous Control and Reinforcement Learning. In *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, pp. 77–89. PMLR, 2020.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. PAC confidence sets for deep neural networks via calibrated prediction. In *8th International Conference on Learning Representations*, 2020.

Luis Raul Pericchi. Sets of prior probabilites and bayesian robustness. *Documentation Section on the website of the Society for Imprecise Probability Theory and Applications (SIPTA)*, 1998.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018.

Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. Second-Order Uncertainty Quantification: A Distance-Based Approach. *Available at arxiv:2312.00995*, 2023a.

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 (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1795–1804. PMLR, 31 Jul–04 Aug 2023b. URL `https://proceedings.mlr.press/v216/sale23a.html`.

Robin Senge, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, January 2014. ISSN 0020-0255.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Marek Smieja and Jacek Tabor. Entropy of the mixture of sources and entropy dimension. *IEEE Transactions on Information Theory*, 58(5):2719–2728, 2012.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

Bjørnar Tessem. Interval probability propagation. *International Journal of Approximate Reasoning*, 7(3): 95–120, 1992.

D. Michael Titterington. Bayesian methods for neural networks and related models. *Statistical Science*, 19 (1):128–139, 2004.

Krasymyr Tretiak, Georg Schollmeyer, and Scott Ferson. Neural network model for imprecise regression with interval dependent variables. *Available at arXiv:2206.02467*, 2022.

Matthias C.M. Troffaes and Gert de Cooman. *Lower Previsions*. Chichester, United Kingdom : John Wiley and Sons, 2014.

Renukanandan Tumu, Lars Lindemann, Truong Nghiem, and Rahul Mangharam. Physics Constrained Motion Prediction with Uncertainty Quantification, 2023. arXiv:2302.01060.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Cham : Springer, second edition, 2022.

Peter Walley. Coherent lower (and upper) probabilities. Technical report, University of Warwick, Coventry, 1981.

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

Hao Wang and Dit-Yan Yeung. A survey on Bayesian deep learning. *ACM Computing Surveys*, 53(5):1–37, 2021.

Larry A. Wasserman and Joseph B. Kadane. Bayes' theorem for Choquet capacities. *The Annals of Statistics*, 18(3):1328–1339, 1990.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Marco Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1): 5–21, 2002. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(01)00201-4. URL `https://www.sciencedirect.com/science/article/pii/S0378375801002014`. Imprecise Probability Models and their Applications.

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180 – 2207, 2020.

## A Why do we need IPs?

The main motivations for working with credal sets are two. Let $(\Omega, \mathcal{F})$ be the measurable space of interest.

(i) A single probability distribution does not suffice to represent ignorance in the sense of lack of knowledge; this is well documented in the literature, see e.g. Hüllermeier & Waegeman (2021) and references therein. Consider the example of complete ignorance (CI) in the case of a finite state space $\Omega$ (Hüllermeier & Waegeman, 2021, Section 3.3). In standard Bayesian analysis, CI is modeled in terms of the uniform distribution $\text{Unif}(\Omega)$; this is justified by Laplace's "principle of indifference". Then, however, it is not possible to distinguish between precise probabilistic knowledge about a random event – called *prior indifference*; think of the tossing of a fair coin – and a complete lack of knowledge due to an incomplete description of the experiment – called *prior ignorance*. Another problem is given by the additive nature of probability distributions. Consider again the example of a uniform distribution. First, let us observe that it is not invariant under reparametrization. In addition, if we model the ignorance about the length $x$ of the side of a cube in $\mathbb{R}^3$ via a uniform measure on the interval $[l, u] \subset \mathbb{R}$, then this does not yield a uniform distribution of $x^3$ on $[l^3, u^3]$, which suggests some degree of informedness about the cube's volume. Finally, as pointed out in Walley (1991), if we ask a subject – even an expert – about their opinion regarding some events, it is much more likely that they will report interval of probabilities rather than single values.

(ii) Working with credal sets allows to achieve prior and likelihood *robustness*: realistically large sets $\mathcal{P}_{\text{prior}}$ of priors and $\mathcal{P}_{\text{lik}}$ of likelihoods are elicited. Using credal sets, the agent recognizes that prior beliefs and knowledge about the sampling model are limited and imprecise. Combining each pair of functions in $\mathcal{P}_{\text{prior}}$ and $\mathcal{P}_{\text{lik}}$ using Bayes' rule, a class of posterior distributions – reflecting the updated state of uncertainty – is formed. If the available information is not sufficient to identify a unique posterior distribution, or a set of posteriors whose diameter is small, credal sets allow to represent *indecision*, thus leading to a less informative but more robust conclusions.[17]

## B On the use of credal sets

Let us address a critique raised against the use of credal sets. Lassiter (2020) argues against the use of sets of probabilities to model an agent's prior beliefs and their knowledge of the sampling model, while debating in favor of using hierarchical Bayesian models. As reported in Hüllermeier & Waegeman (2021, Secton 4.6.2), the argument against credal sets that is more cogent for the machine learning literature is that modeling a lack of knowledge in a set-based manner may hamper the possibility of inductive inference, up to a point where learning from empirical data is not possible any more. With this, we mean the following. As Pericchi (1998) points out, the natural candidate for a class of priors to represent complete ignorance is the class $\mathcal{P}_{\text{all}}$ of all distributions. When this class leads to non-vacuous and useful conclusions, these are quite compelling and uncontroversial. It turns out that the posterior probabilities obtained from this class are vacuous, that is, their lower and upper bounds are 0 and 1: no finite sample is enough to annihilate a sufficiently extreme prior belief. There is then a compromise to be made, and this is the compromise of *near-ignorance*. The near-ignorance class should be vacuous a priori in some respects, typically the ones that are the most important for the analysis at hand. This way of proceeding is labeled as arbitrary by Lassiter (2020), who instead advocates for the use of hierarchical Bayesian procedures. We find this critique not compelling, as during the analysis the job of the agent is to model reality: as pointed out in Hüllermeier & Waegeman (2021, Secton 5), statistical inference is not possible without underlying assumptions, and conclusions drawn from data are always conditional on those assumptions. If we were to work every time with the maximum level of generality, we would hardly be able to reach any conclusions. For example, in a statistical analysis we never consider the state $\Omega$ of *apparently possible states* (Walley, 1991, section 2.1.2), that is, the one that contains all the states $\omega$ that are logically consistent with the available information. If we consider a coin toss, we let the state space be $\Omega = \{\text{heads, tails}\}$, certainly not $\Omega = \{\text{heads, tails, coin landing on its edge, coin braking into pieces on landing, coin disappearing down a crack in the floor}\}$. The same holds for sets

[17] Here "diameter" has to be understood as the distance between upper and lower probability of event $A$, for all $A \in \mathcal{F}$.

of probabilities: it makes much more sense to work with near-ignorance credal sets than to work with $\mathcal{P}_{\text{all}}$. A final reason to rebut the point in Lassiter (2020) is that the problems indicated in (i) in section A that make the use of the uniform prior distribution – often interpreted as representing epistemic uncertainty in standard Bayesian inference – at least debatable are inherited by hierarchical Bayesian modeling, as specified in Bernardo (1979); Hüllermeier & Waegeman (2021); Jeffreys (1946). Furthermore, a single distribution that is the result of a hierarchical Bayesian procedure is unable to gauge epistemic uncertainty (Hüllermeier & Waegeman, 2021).

### B.1 A note on how to specify credal sets

As pointed out in Corani et al. (2012, Section 3.3), there is a way of obtaining credal sets starting from sets of probability intervals; in addition, standard algorithms can compute the extreme elements of a credal set for which a probability interval has been provided (Avis & Fukuda, 1996). However, the resulting number of extrema is exponential in the size of the possibility space (Tessem, 1992).[18] For this reason we prefer to specify prior and likelihood finitely generated credal sets instead.

## C A further IP concept: the core

Let again $(\Omega, \mathcal{F})$ be the measurable space of interest. Because of the conjugacy property of upper and lower probabilities, let us focus on upper probabilities only. We say that upper probability $\overline{P}$ is *concave* if $\overline{P}(A \cup B) \leq \overline{P}(A) + \overline{P}(B) - \overline{P}(A \cap B)$, for all $A, B \in \mathcal{F}$. Recall that $\Delta(\Omega, \mathcal{F})$ denotes the set of all probability measures on $(\Omega, \mathcal{F})$. Upper probability $\overline{P}$ is *compatible* with the convex set (Gong & Meng, 2021)

$$\begin{aligned}\text{core}(\overline{P}) :&= \{P \in \Delta(\Omega, \mathcal{F}) : P(A) \leq \overline{P}(A), \forall A \in \mathcal{F}\} \\ &= \{P \in \Delta(\Omega, \mathcal{F}) : \underline{P}(A) \leq P(A) \leq \overline{P}(A), \forall A \in \mathcal{F}\}\end{aligned}$$

where the second equality is a characterization (Cerreia-Vioglio et al., 2015, Page 3389). Notice that the core is convex (Marinacci & Montrucchio, 2004, Section 2.2). We assume it is nonempty and weak$^\star$-closed.[19] Then, it is weak$^\star$-compact as a result of Marinacci & Montrucchio (2004, Proposition 3).

Since the core is convex, the set $\text{ex}[\text{core}(\overline{P})]$ of extreme points of the core is well defined. It contains all the elements of the core that cannot be written as a convex combination of one another. The following important result is a consequence of Walley (1991, Theorem 3.6.2).

**Theorem 9.** *Suppose core$(\overline{P})$ is nonempty and weak$^\star$-closed. Then, the following holds.*

*(a) ex$[\text{core}(\overline{P})] \neq \emptyset$.*

*(b) core$(\overline{P})$ is the closure in the weak$^\star$ topology of the convex hull of ex$[\text{core}(\overline{P})]$.*

*(c) If $\overline{P}(A) = \sup_{P \in \text{core}(\overline{P})} P(A)$, for all $A \in \mathcal{F}$, then $\overline{P}(A) = \sup_{P \in \text{ex}[\text{core}(\overline{P})]} P(A)$, for all $A \in \mathcal{F}$.*

So in order to define an upper probability $\overline{P}$ that setwise dominates the elements of $\text{core}(\overline{P})$ it is enough to specify the extreme points of the core.

## D A new Bayes' theorem for IPs

We present Theorem 10, a result that – although appealing – does not lend itself well to be applied to the CBDL procedure. An extension of Theorem 10 is given in Caprio et al. (2024b).

Call $\Theta$ the parameter space of interest and assume it is Polish, that is, the topology for $\Theta$ is complete, separable, and metrizable. This ensures that the set $\Delta(\Theta, \mathcal{B})$ of probability measures on $\Theta$ is Polish as well, where $\mathcal{B}$ denotes the Borel $\sigma$-algebra for $\Theta$. Let $\mathscr{X}$ be the set of all bounded, non-negative, $\mathcal{B}$-measurable

[18] Recall that the possibility space of a random variable is the space of the values it can take on.

[19] Recall that in the weak$^\star$ topology, a net $(P_\alpha)_{\alpha \in I}$ converges to $P$ if and only if $P_\alpha(A) \to P(A)$, for all $A \in \mathcal{F}$.

functionals on $\Theta$. Call $\mathscr{D} = \mathcal{X} \times \mathcal{Y}$ the sample space endowed with the product $\sigma$-algebra $\mathcal{A} = \mathcal{A}_{\mathbf{x}} \times \mathcal{A}_{\mathbf{y}}$, where $\mathcal{A}_{\mathbf{x}}$ is the $\sigma$-algebra endowed to $\mathcal{X}$ and $\mathcal{A}_{\mathbf{y}}$ is the $\sigma$-algebra endowed to $\mathcal{Y}$. Let the agent elicit $\mathcal{L}_\theta := \{P_\theta \in \Delta(\mathscr{D}, \mathcal{A}) : \theta \in \Theta\}$. Assume that each $P_\theta \in \mathcal{L}_\theta$ has density $L(\theta) = p(D \mid \theta)$ with respect to some $\sigma$-finite dominating measure $\nu$ on $(\mathscr{D}, \mathcal{A})$; this represents the likelihood function for $\theta$ having observed data $D \subset \mathscr{D}$. We assume for now that $L \in \mathscr{X}$, for all $D \subset \mathscr{D}$.

Let the agent specify a set $\mathcal{P}$ of probabilities on $(\Theta, \mathcal{B})$. Then, compute $\overline{P}$, and consider $\mathcal{P}^{\text{co}} := \text{core}(\overline{P})$; it represents the agent's initial beliefs.[20] We assume that every $P \in \mathcal{P}^{\text{co}}$ has density $p$ with respect to some $\sigma$-finite dominating measure $\mu$ on $(\Theta, \mathcal{B})$, that is, $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$. We require the agent's beliefs to be represented by the core for two main reasons. The first, mathematical, one is to ensure that the belief set is compatible with the upper probability. The second, philosophical, one is the following (Caprio & Gong, 2023; Caprio & Mukherjee, 2023). A criticism brought forward by Walley (1991, Section 2.10.4.(c)) is that, given an upper probability $\overline{P}$, there is no cogent reason for which the agent should choose a specific $P_T$ that is dominated by $\overline{P}$, or – for that matter – a collection of "plausible" probabilities. Because the core considers all (countably additive) probability measures that are dominated by $\overline{P}$, it is the perfect instrument to reconcile Walley's behavioral and sensitivity analysis interpretations.

Let the agent compute $\overline{P}_\theta$, and consider $\mathcal{L}_\theta^{\text{co}} := \text{core}(\overline{P}_\theta)$; it represents the set of plausible likelihoods. Let

$$\mathscr{L} := \left\{ L = \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu},\, P_\theta \in \mathcal{L}_\theta^{\text{co}} \right\}, \tag{4}$$

and denote by $\overline{L}(\theta) := \sup_{L \in \mathscr{L}} L(\theta)$ and by $\underline{L}(\theta) := \inf_{L \in \mathscr{L}} L(\theta)$, for all $\theta \in \Theta$. Call

$$\mathcal{P}_D^{\text{co}} := \left\{ P_D \in \Delta(\Theta, \mathcal{B}) : \frac{\mathrm{d}P_D}{\mathrm{d}\mu} = p(\theta \mid D) = \frac{L(\theta)p(\theta)}{\int_\Theta L(\theta)p(\theta)\mathrm{d}\theta}, \right.$$
$$\left. p = \frac{\mathrm{d}P}{\mathrm{d}\mu},\, P \in \mathcal{P}^{\text{co}},\, L = \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu},\, P_\theta \in \mathcal{L}_\theta^{\text{co}} \right\}$$

the class of posterior probabilities when the prior is in $\mathcal{P}^{\text{co}}$ and the likelihood is in $\mathcal{L}_\theta^{\text{co}}$, and let $\overline{P}_D(A) = \sup_{P_D \in \mathcal{P}_D^{\text{co}}} P_D(A)$, for all $A \in \mathcal{B}$. Then, the following is a generalization of Bayes' theorem in Wasserman & Kadane (1990).

**Theorem 10.** *Suppose $\mathcal{P}^{co}, \mathcal{L}_\theta^{co}$ are nonempty and weak$^\star$-closed. Then for all $A \in \mathcal{B}$,*

$$\overline{P}_D(A) \leq \frac{\sup_{P \in \mathcal{P}^{co}} \int_\Theta \overline{L}(\theta) \mathbb{1}_A(\theta) P(d\theta)}{\mathbf{c}}, \tag{5}$$

*provided that the ratio is well defined. Here, $\mathbf{c} := \sup_{P \in \mathcal{P}^{co}} \int_\Theta \overline{L}(\theta) \mathbb{1}_A(\theta) P(d\theta) + \inf_{P \in \mathcal{P}^{co}} \int_\Theta \underline{L}(\theta) \mathbb{1}_{A^c}(\theta) P(d\theta)$, and $\mathbb{1}_A$ denotes the indicator function for $A \in \mathcal{B}$. In addition, if $\overline{P}$ is concave, then the inequality in equation 5 is an equality for all $A \in \mathcal{B}$.*

This result is particularly appealing because, given some assumptions, it allows to perform a (generalized) Bayesian update of a prior upper probability (PUP) by carrying out only one operation, even when the likelihood is ill specified so that a set of likelihoods is needed. We also have the following.

**Lemma 11.** *Suppose $\mathcal{P}^{co}, \mathcal{L}_\theta^{co}$ are nonempty and weak$^\star$-closed. Then, if $\overline{P}$ is concave, we have that $\overline{P}_D$ is concave as well.*

This lemma is important because it tells us that the generalized Bayesian update of Theorem 10 preserves concavity, and so it can be applied to successive iterations. If at time $t$ the PUP is concave, then the PUP at time $t+1$ – that is, the posterior upper probability at time $t$ – will be concave too. Necessary and sufficient conditions for a generic upper probability to be concave are given in Marinacci & Montrucchio (2004, Section 5).

[20] Superscript "co" stands for convex and core.

In the future, these results can be generalized to the case in which the elements of $\mathscr{X}$ are unbounded using techniques in Troffaes & de Cooman (2014), and to the case in which the elements of $\mathscr{X}$ are $\mathbb{R}^d$-valued, for some $d \in \mathbb{N}$, since we never used specific properties of $\mathbb{R}$ in our proofs.

Despite being attractive, the generalized Bayesian update of Theorem 10 hinges upon three assumptions, namely that $\mathcal{P}^{\text{co}}$ and $\mathcal{L}_\theta^{\text{co}}$ are both cores of an upper probability, that they are nonempty and weak$^\star$-closed, and that the prior upper probability $\overline{P}$ is concave. As the proverb goes, there is no free lunch. Having to check these assumptions, together with computing a supremum, an infimum, and the integrals in equation 5, makes Theorem 10 inadequate to be applied in the context of CBDL.

## E Bounds on upper and lower entropy

In this section we find an upper bound for $\overline{H}(P)$ and a lower bound for $\underline{H}(P)$ that are extremely interesting. We first need to introduce three new concepts.

**Definition 12.** *Consider a set $\mathscr{P}$ of probabilities on a generic measurable space $(\Omega, \mathcal{F})$. We say that lower probability $\underline{P}$ is convex if $\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B) - \underline{P}(A \cap B)$, for all $A, B \in \mathcal{F}$.*

*Then, let $\mathsf{P}$ be either an upper or a lower probability, and consider a generic bounded function $f$ on $(\Omega, \mathcal{F})$, that is, $f \in B(\Omega)$. We define the Choquet integral of $f$ with respect to $\mathsf{P}$ as follows*

$$\int_\Omega f(\omega)\mathsf{P}(d\omega) := \int_0^\infty \mathsf{P}\left(\{\omega \in \Omega : f(\omega) \geq t\}\right) dt + \int_{-\infty}^0 \left[\mathsf{P}\left(\{\omega \in \Omega : f(\omega) \geq t\}\right) - \mathsf{P}(\Omega)\right] dt,$$

*where the right hand side integrals are (improper) Riemann integrals. If $\mathsf{P}$ is additive, then the Choquet integral reduces to the standard additive integral.*

*Finally, if $\Omega$ is uncountable, define for all $\omega \in \Omega$*

$$\underline{\pi}(\omega) := \inf_{P \in \mathscr{P}} \frac{dP}{d\mu}(\omega) \quad \textit{and} \quad \overline{\pi}(\omega) := \sup_{P \in \mathscr{P}} \frac{dP}{d\mu}(\omega).$$

*We call them lower and upper densities, respectively.*

The following theorem gives the desired bounds.

**Theorem 13.** *Consider a set $\mathscr{P}$ of probabilities on a generic measurable space $(\Omega, \mathcal{F})$. If $\Omega$ is uncountable, assume that every $P \in \mathscr{P}$ is dominated by a $\sigma$-finite measure $\mu$ and that the Radon-Nikodym derivatives $\frac{dP}{d\mu}$ are continuous and bounded, for all $P \in \mathscr{P}$. Define*

$$H(\underline{P}) := \begin{cases} -\int_\Omega \log\left[\overline{\pi}(\omega)\right] \underline{P}(d\omega) & \textit{if } \Omega \textit{ is uncountable} \\ -\sum_{\omega \in \Omega} \underline{P}(\{\omega\}) \log[\overline{P}(\{\omega\})] & \textit{if } \Omega \textit{ is at most countable} \end{cases}$$

*and similarly*

$$H(\overline{P}) := \begin{cases} -\int_\Omega \log\left[\underline{\pi}(\omega)\right] \overline{P}(d\omega) & \textit{if } \Omega \textit{ is uncountable} \\ -\sum_{\omega \in \Omega} \overline{P}(\{\omega\}) \log[\underline{P}(\{\omega\})] & \textit{if } \Omega \textit{ is at most countable} \end{cases},$$

*Then, $\overline{H}(P) \leq H(\overline{P})$ and $\underline{H}(P) \geq H(\underline{P})$. In addition, if $\overline{P}$ is concave, the first bound is tighter, and if $\underline{P}$ is convex, the second bound is tighter.*

**Remark 14.** *In Hüllermeier & Waegeman (2021, Section 4.6.1), the authors point out that Abellán & Moral (2000) presents a generalization of the Hartley measure, called generalized Hartley measure $GH(P)$, that can be used to disaggregate the total uncertainty captured by $\overline{H}(P)$ into aleatoric and epistemic uncertainties. We prefer not to introduce it in the present work because $GH(P)$ is defined based on the mass function of a belief function (Gong & Meng, 2021, Definition 2.4).*[21] *This entails that the authors assume that lower*

[21] A belief function is a mathematical concept that should not be confused with the term "belief" we used throughout the paper to address the agent's knowledge.

*probability $\underline{P}$ associated with the set of probabilities of interest is a belief function, and so that for every collection $\{A, A_1, \dots, A_k\}$ such that $A \subset A_i$, the following holds*

$$\underline{P}(A) \geq \sum_{\emptyset \neq I \subset \{1,\dots,k\}} (-1)^{\#I-1} \underline{P}(\cap_{i \in I} A_i),$$

*for all $k \in \mathbb{N}$. As it is immediate to see, this is assumption is not needed in the context of CBDL. Other ways of disentangling and quantifying AU and EU can be found in Sale et al. (2023a;b)*

## F How to derive a predictive distribution

Suppose we performed a Bayesian updating procedure so to obtain posterior pdf $p(\theta \mid x_1, y_1, \dots, x_n, y_n)$. Recall that $\{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ denotes the training set. We obtain the predictive distribution $p(\tilde{y} \mid \tilde{x}, x_1, y_1, \dots, x_n, y_n)$ on $\mathcal{Y}$ as follows

$$\begin{aligned} p(\tilde{y} \mid \tilde{x}, x_1, y_1, \dots, x_n, y_n) &= \int_\Theta p(\tilde{y}, \theta \mid \tilde{x}, x_1, y_1, \dots, x_n, y_n) \mathrm{d}\theta \\ &= \int_\Theta p(\tilde{y} \mid \theta, \tilde{x}, x_1, y_1, \dots, x_n, y_n) \cdot p(\theta \mid \tilde{x}, x_1, y_1, \dots, x_n, y_n) \mathrm{d}\theta \\ &= \int_\Theta p(\tilde{y} \mid \theta, \tilde{x}) \cdot p(\theta \mid x_1, y_1, \dots, x_n, y_n) \mathrm{d}\theta, \end{aligned}$$

where $p(\tilde{y} \mid \theta, \tilde{x})$ is the likelihood used to derive the posterior. Notice that the last equality comes from output $\tilde{y}$ only depending on input $\tilde{x}$ and parameter $\theta$, and from having assumed $D_{\mathbf{x}} \perp\!\!\!\perp \theta$ (see section 2.1). From an applied point of view, a sample from $p(\tilde{y} \mid \tilde{x}, x_1, y_1, \dots, x_n, y_n)$ is obtained as follows:

1. specify input $\tilde{x}$;
2. sample a parameter $\tilde{\theta}$ from the posterior, $\tilde{\theta} \sim p(\theta \mid x_1, y_1, \dots, x_n, y_n)$;
3. plug $\tilde{\theta}$ in the likelihood and sample $\tilde{y} \sim p(y \mid \tilde{\theta}, \tilde{x})$.

## G Aleatoric uncertainty check for $\alpha$-level IHDR

The diameter of $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$ is a function of the aleatoric uncertainty (AU) faced by the agent.[22] If we want to avoid to perform the computation in **Step 5** of Algorithm 1 only to discover that $IR_\alpha(\hat{\mathcal{P}}_{\text{pred}})$ is "too large", then we can add an "AU check".

At the beginning of the analysis, compute the lower entropy $\underline{H}(\hat{P}^{\text{pred}}) = \min_{k,s} H(\hat{P}^{\text{pred}}_{k,s})$ associated with the set $\text{ex}\hat{\mathcal{P}}_{\text{pred}}$ of extreme elements of the VI-approximated predictive credal set $\hat{\mathcal{P}}_{\text{pred}}$. By equation 2, it is equal to the aleatoric uncertainty encoded in $\hat{\mathcal{P}}_{\text{pred}}$. We then verify whether the lower entropy $\underline{H}(\hat{P}^{\text{pred}})$ is "too high". That is, if $\underline{H}(\hat{P}^{\text{pred}}) > \varphi$, for some $\varphi > 0$, we want our procedure to abstain. This means that if the aleatoric uncertainty in set $\hat{\mathcal{P}}_{\text{pred}}$ is too high, then our procedure does not return any output set for input $\tilde{x}$. The value of $\varphi$ can be set equal to the entropy of the probability measures that are typically used in the context the agent works in. For example, in medical applications the agent may consider the entropy of a Normal distribution, while in financial applications the entropy of a distribution with fatter tails, such as a $t$-distribution or a Cauchy. We call these *reference $\varphi$ values*.

If we add this "AU check", the 2-tuple that the agent needs to specify at inference time in Algorithm 1 is $\langle \varphi, \alpha \rangle$.

[22] Since the diameter is a metric concept, we assume that we can find a well-defined metric $d_{\mathbf{y}}$ on $\mathcal{Y}$. If that is not the case, we substitute the diameter with the notion of cardinality.

## H $\alpha$-level IHDR in a Classification setting

In classification problems, BNNs compute the probability vector

$$\varpi := \frac{1}{\#\Theta} \sum_{\theta \in \Theta} \Phi_{\theta|D}(x),$$

where we write $\Phi_{\theta|D}$ to highlight the fact that $\theta$ is sampled from posterior $p(\theta \mid D)$, and then select the most likely class $\hat{y} := \arg\max_j \varpi_j$, where the $\varpi_j$'s are the elements of $\varpi$.

When applied to a classification setting, the general procedure introduced in Algorithm 1 becomes the following. Recall that we denote by $K$ the cardinality of $\text{ex}\mathcal{P}_{\text{prior}}$, and by $S$ the cardinality of $\text{ex}\mathcal{P}_{\text{lik}}$. Assume that $\mathcal{Y} = \{y_1, \dots, y_J\}$, that is, there are $J \in \mathbb{N}_{\geq 2}$ possible labels. Then, a VI-approximated predictive distribution $\hat{P}^{\text{pred}}_{k,s}$ in $\text{ex}\hat{\mathcal{P}}_{\text{pred}}$ can be seen as $J$-dimensional probability vector $\hat{\mathbf{p}}^{\text{pred}}_{k,s} = (\hat{p}^{\text{pred}}_{k,s,1}, \dots, \hat{p}^{\text{pred}}_{k,s,J})^\top$, where $\hat{p}^{\text{pred}}_{k,s,j} = \hat{P}^{\text{pred}}_{k,s}(\{y_j\})$, for all $k \in \{1, \dots, K\}$, $s \in \{1, \dots, S\}$, and $j \in \{1, \dots, J\}$.

Now fix any $k$ and any $s$, and define the partial order $\preceq_{k,s}$ on $\mathcal{Y}$ as $y_l \preceq_{k,s} y_i \iff \hat{p}^{\text{pred}}_{k,s,l} \geq \hat{p}^{\text{pred}}_{k,s,i}$ and $y_l \prec_{k,s} y_i \iff \hat{p}^{\text{pred}}_{k,s,l} > \hat{p}^{\text{pred}}_{k,s,i}$, where $i, l \in \{1, \dots, J\}$, $i \neq l$. This means that we can order the labels according to the probability that $\hat{P}^{\text{pred}}_{k,s}$ assigns to them: the first label will be the one having highest probability according to $\hat{P}^{\text{pred}}_{k,s}$, the second label will have the second-highest probability according to $\hat{P}^{\text{pred}}_{k,s}$, and so on.

Now order the label space $\mathcal{Y}$ according to $\preceq_{k,s}$ so to obtain

$$\mathcal{Y}^{k,s} := \{y^{k,s}_1, \dots, y^{k,s}_J\}.$$

This means that $y^{k,s}_1 \preceq_{k,s} y^{k,s}_j$, for all $j \in \{2, \dots, J\}$, $y^{k,s}_2 \preceq_{k,s} y^{k,s}_j$, for all $j \in \{3, \dots, J\}$ (but $y^{k,s}_1 \preceq_{k,s} y^{k,s}_2$), and so on. That is, we order the labels from the most to the least likely according to $\hat{P}^{\text{pred}}_{k,s}$.

Then, we call *$\alpha$-level credible set* according to $\hat{P}^{\text{pred}}_{k,s}$, $\alpha \in [0, 1]$, the set

$$\begin{aligned} CS_\alpha(\hat{P}^{\text{pred}}_{k,s}) := \Bigg\{ y^{k,s}_1, \dots, y^{k,s}_j : \sum_{i=1}^{j} \hat{P}^{\text{pred}}_{k,s}(\{y^{k,s}_i\}) \in [1-\alpha, 1-\alpha+\varepsilon],\ j \leq J, \\ \text{and } \nexists j' < j : \sum_{i=1}^{j'} \hat{P}^{\text{pred}}_{k,s}(\{y^{k,s}_i\}) \in [1-\alpha, 1-\alpha+\varepsilon] \Bigg\}, \end{aligned} \tag{6}$$

for some $\varepsilon > 0$. It corresponds to the $\alpha$-level HDR $R_\alpha(\hat{P}^{\text{pred}}_{k,s})$. Notice that we require $\sum_{i=1}^{j} \hat{P}^{\text{pred}}_{k,s}(\{y^{k,s}_i\}) \in [1-\alpha, 1-\alpha+\varepsilon]$ because we may need to go slightly above level $1-\alpha$. Just as a toy example, we may have 7 labels, 3 of which would give a 0.945 coverage, while 4 would give a coverage of 0.953. If we are interested in the $\alpha = 0.05$-level credible set, we ought to include the fourth label, thus yielding a coverage slightly higher than $1 - \alpha = 0.95$. The interpretation to $CS_\alpha(\hat{P}^{\text{pred}}_{k,s})$ is the following: it consists of the smallest collection of labels to which $\hat{P}^{\text{pred}}_{k,s}$ assigns probability of at least $1-\alpha$ (that is, those having the highest probability of being the correct one for the new input $\tilde{x}$).

Finally, we call *$\alpha$-level imprecise credible set*, $\alpha \in [0, 1]$, the set

$$ICS_\alpha(\hat{\mathcal{P}}_{\text{pred}}) := \bigcup_{k,s} CS_\alpha(\hat{P}^{\text{pred}}_{k,s}).$$

In turn, we have that $\underline{\hat{P}}^{\text{pred}}[\{\tilde{y} \in ICS_\alpha(\hat{\mathcal{P}}_{\text{pred}})\}] \geq 1 - \alpha$.

**Remark 15.** *Notice that if a credible set of level $\approx \alpha$ is enough, then we can replace the left endpoint of the interval in equation 6 with $1 - (\alpha + \varepsilon_{k,s})$, for some $\varepsilon_{k,s} > 0$. Strictly speaking, in this case we obtain an $(\alpha + \varepsilon_{k,s})$-level credible set, which we denote by $CS_{\alpha_{k,s}}(\hat{P}^{pred}_{k,s})$, where $\alpha_{k,s} := \alpha + \varepsilon_{k,s}$. Going back to our toy example, we will have a credible set with 3 labels that yields a coverage of $1-(\alpha+\varepsilon_{k,s}) = 0.945 \approx 0.95 = 1-\alpha$, so $\varepsilon_{k,s} = 0.005$. In turn, this implies that the imprecise credible set will have a coverage of $1-(\alpha+\max_{k,s} \varepsilon_{k,s})$, that is, it will have level $\alpha + \max_{k,s} \varepsilon_{k,s}$. We denote it by $ICS_{\tilde{\alpha}}(\hat{\mathcal{P}}_{pred})$, where $\tilde{\alpha} := \alpha + \max_{k,s} \varepsilon_{k,s}$.*

# I Distance between a set of distributions $\mathcal{P}$ and a single distribution $P'$ having different dimensions

Many concepts in this section are derived from Cai & Lim (2022). Let $m, n \in \mathbb{N}$ such that $m \leq n$, and let $p \in [1, \infty]$. Call $M^p(\mathbb{R}^j)$ the set of probability measures on $\mathbb{R}^j$ having finite $p$-th moment, and $M_d(\mathbb{R}^j)$ the set of probability measures on $\mathbb{R}^j$ having density with respect to some $\sigma$-finite dominating measure $\mu$, $j \in \{m, n\}$. Let $O(m, n) := \{V \in \mathbb{R}^{m\times n} : VV^\top = I_m\}$, where $I_m$ denotes the $m$-dimensional identity matrix, and for any $V \in O(m, n)$ and any $b \in \mathbb{R}^m$, define the following function

$$\varphi_{V,b} : \mathbb{R}^n \to \mathbb{R}^m, \quad x \mapsto \varphi_{V,b}(x) := Vx + b.$$

Let $\mathcal{B}(\mathbb{R}^n)$ be the Borel $\sigma$-algebra on $\mathbb{R}^n$, and for any $Q \in \Delta(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, define $\varphi_{V,b}(Q) := Q \circ \varphi_{V,b}^{-1}$, the pushforward measure. Consider then two generic probability measures $Q, S$ such that $Q \in M^p(\mathbb{R}^m)$ and $S \in M^p(\mathbb{R}^n)$, and call

$$\begin{aligned}
\Phi_p^+(Q, n) &:= \{\alpha \in M^p(\mathbb{R}^n) : \varphi_{V,b}(\alpha) = Q, \text{ for some } V \in O(m, n), b \in \mathbb{R}^m\},\\
\Phi_d^+(Q, n) &:= \{\alpha \in M_d(\mathbb{R}^n) : \varphi_{V,b}(\alpha) = Q, \text{ for some } V \in O(m, n), b \in \mathbb{R}^m\},\\
\Phi^-(S, m) &:= \{\beta \in M(\mathbb{R}^m) : \varphi_{V,b}(S) = \beta, \text{ for some } V \in O(m, n), b \in \mathbb{R}^m\}.
\end{aligned}$$

Recall now the definition of $p$-Wasserstein metric between two generic distributions defined on the *same* Euclidean space. Let $P_1, P_2 \in M^p(\mathbb{R}^n)$, for some $n \in \mathbb{N}$ and some $p \in [1, \infty]$. Then, the $p$-Wasserstein distance between them is defined as

$$W_p(P_1, P_2) := \left[\inf_{\gamma\in\Gamma(P_1,P_2)} \int_{\mathbb{R}^{2n}} \|x - y\|_2^p \gamma(\mathrm{d}(x, y))\right]^{1/p},$$

where $\|\cdot\|_2$ denotes the Euclidean distance, $p = \infty$ is interpreted as the essential supremum, and $\Gamma(P_1, P_2) := \{\gamma \in \Delta(\mathbb{R}^{2n}, \mathcal{B}(\mathbb{R}^{2n})) : \mathrm{proj}_1^n(\gamma) = P_2, \mathrm{proj}_2^n(\gamma) = P_1\}$ is the set of couplings between $P_1$ and $P_2$, where $\mathrm{proj}_1^n$ is the projection onto the first $n$ coordinates, and $\mathrm{proj}_2^n$ is the projection to the last $n$ coordinates.

Recall then the definition of $f$-divergence between two generic distributions defined on the *same* Euclidean space. Let $P_1, P_2 \in \Delta(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, for some $n \in \mathbb{N}$, and assume $P_1 \ll P_2$. Then, for any convex functional $f$ on $\mathbb{R}$ such that $f(1) = 0$, the $f$-divergence between $P_1$ and $P_2$ is defined as

$$\mathrm{div}_f(P_1\|P_2) := \int_{\mathbb{R}^n} f\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}(x)\right) P_2(\mathrm{d}x).$$

Aside from the Rényi divergence, the $f$-divergence includes just about every known divergences as special case (Cai & Lim, 2022). The following are the main results of this section.

**Lemma 16.** *Let $m, n \in \mathbb{N}$ such that $m \leq n$, and let $p \in [1, \infty]$ and $f$ be any convex functional on $\mathbb{R}$ such that $f(0) = 1$. Consider a generic $\mathscr{P} \subset M^p(\mathbb{R}^m)$ and $P' \in M^p(\mathbb{R}^n)$. Let $\Phi_p^+(\mathscr{P}, n) := \cup_{P\in\mathscr{P}}\Phi_p^+(P, n)$ and $\Phi_d^+(\mathscr{P}, n) := \cup_{P\in\mathscr{P}}\Phi_d^+(P, n)$. Define*

- *$W_p^+(P, P') := \inf_{\alpha\in\Phi_p^+(P,n)} W_p(\alpha, P')$, for all $P \in \mathscr{P}$;*
- *$div_f^+(P\|P') := \inf_{\alpha\in\Phi_d^+(P,n)} div_f(P\|P')$, for all $P \in \mathscr{P}$;*
- *$W_p^+(\mathscr{P}, P') := \inf_{\alpha\in\Phi_p^+(\mathscr{P},n)} W_p(\alpha, P')$;*
- *$div_f^+(\mathscr{P}\|P') := \inf_{\alpha\in\Phi_d^+(\mathscr{P},n)} div_f(P\|P')$;*

*Then, for all $P \in \mathscr{P}$ the following holds*

$$W_p^+(\mathscr{P}, P') \leq W_p^+(P, P') \quad \textit{and} \quad div_f^+(\mathscr{P}\|P') \leq div_f^+(P\|P').$$

**Lemma 17.** *Let $m, n \in \mathbb{N}$ such that $m \leq n$, and let $p \in [1, \infty]$ and $f$ be any convex functional on $\mathbb{R}$ such that $f(0) = 1$. Consider a generic $\mathscr{P} \subset M^p(\mathbb{R}^n)$ and $P' \in M^p(\mathbb{R}^m)$. Let $\Phi^-(\mathscr{P}, m) := \cup_{P \in \mathscr{P}} \Phi^-(P, m)$. Define*

- $W_p^-(P, P') := \inf_{\alpha \in \Phi^-(P,m)} W_p(\alpha, P')$, *for all* $P \in \mathscr{P}$;
- $div_f^-(P \| P') := \inf_{\alpha \in \Phi^-(P,m)} div_f(P \| P')$, *for all* $P \in \mathscr{P}$;
- $W_p^-(\mathscr{P}, P') := \inf_{\alpha \in \Phi^-(\mathscr{P},m)} W_p(\alpha, P')$;
- $div_f^-(\mathscr{P} \| P') := \inf_{\alpha \in \Phi^-(\mathscr{P},m)} div_f(P \| P')$;

*Then, for all $P \in \mathscr{P}$ the following holds*

$$W_p^-(\mathscr{P}, P') \leq W_p^-(P, P') \quad \text{and} \quad div_f^-(\mathscr{P} \| P') \leq div_f^-(P \| P').$$

A visual representation of the application of Lemma 17 in the context of CBDL is given in Figure 11.

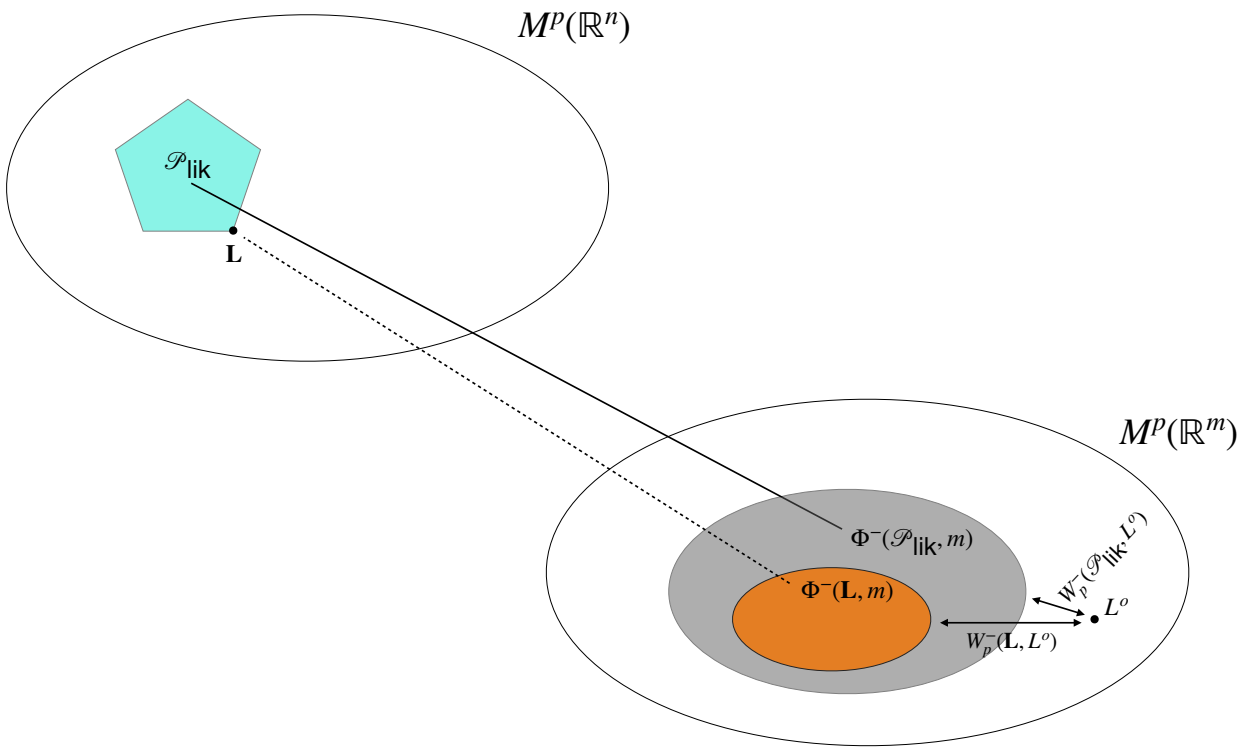



Figure 11: We assume that $n > m$ and that the oracle distribution $L^o$ belongs to $M^p(\mathbb{R}^m)$, while likelihood FGCS $\mathcal{P}_{\text{lik}}$ is a subset of $M^p(\mathbb{R}^n)$, for some finite $p \geq 1$. We also assume that $\mathbf{L}$ is one of the extreme elements of $\mathcal{P}_{\text{lik}}$. We see how $W_p^-(\mathcal{P}_{\text{lik}}, L^o) < W_p^-(\mathbf{L}, L^o)$; if we replace metric $W_p^-$ by a generic $f$-divergence $\text{div}_f$, the inequality would still hold thanks to Lemma 17.

## J Details on Artificial Pancreas Example

**Artificial Pancreas Model.** An important factor when designing the controller for an artificial pancreas is to adapt the insulin delivery algorithm to the particular details of the patient. This is because patients display a wide range of variability in their response to insulin, depending on age, Body Mass Index (BMI),

and other physiological parameters. The Bayesian Neural Network models have 2 hidden layers, with 10 neurons each, for the case study with 4 different seeds. This choice was informed by the experiments in Dutta et al. (2018). For the case study with different architectures, we trained BNNs with 4 different widths: 10, 20, 30, and 40. The horizon length is $H = 10$ time steps, and the prediction horizon is 5 steps into the future. The neural networks were trained for 200 time steps, with a learning rate of 0.001, and batch size of 128 using Mean-Field Variational Inference. The training dataset consisted of 28400 training samples, recorded without meals.

**Controller.** We implemented a simple model predictive controller, using an off-the-shelf implementation of the covariance matrix adaptation evolution strategy (CMA-ES). The model predictive control planning horizon was $k = 5$, with a fixed seed for the randomized solver.

## K Other possible baselines for CBDL

Our experiments may seem like a type of belief tracking (Fortin et al., 2017; Klein et al., 2010). Two comments are in order. First, this line of literature does not use deep learning techniques. Second, in these works the authors rely on Dempster-Shafer theory, a field in imprecise probability theory where lower probabilities are assumed to be belief functions, see Remark 14. We do not rely on this assumption in our work.

## L Further related work

Modeling uncertainty has been a longstanding goal of ML/AI research and a variety of approaches have been developed for doing so (Guo et al., 2017; Park et al., 2020; Jospin et al., 2022). Recently, emphasis has been placed on discerning between aleatoric and epistemic uncertainties (Senge et al., 2014; Kull & Flach, 2014; Kendall & Gal, 2017a). In Tretiak et al. (2022), the authors present an IP-based neural network which uses a regression technique based on probability intervals. Contrary to CBDL, their NN is rooted in the frequentist approach to imprecise probabilities (Huber & Ronchetti, 2009).

In Manchingal & Cuzzolin (2022, Sections 2.1, 2.3), the authors focus on belief-functions-based classification methods. CBDL cannot be directly compared with these methodologies because (i) they do not require that the user expresses their knowledge via a belief function, but rather through a credal set; (ii) they can be used for regression and classification; (iii) they are rooted in Bayesian theory, as opposed to Dempster-Shafer theory. Other works in ML using a belief function approach are those from the field of evidential machine learning (EML), see e.g. Denœux (2022; 2023) and references therein. Existing models mainly address clustering, classification, and regression problems. The reasons why CBDL cannot be directly compared with methods from the EML literature are (i) and (iii) above.

We also point out how there exists a field within ML called evidential deep learning (EDL, not to be confused with EML). It aims at quantifying and disentangling EU and AU, and its methodologies can be applied to classification and regression, see e.g. Amini et al. (2020); Sensoy et al. (2018) and references therein. CBDL is not immediately comparable to models from EDL because these latter are non-Bayesian. In addition, their definitions of AU and EU are slightly different from the canonical ones in ensemble deep learning.

## M Proofs

*Proof of Proposition 3.* If $\overline{P}(A) = \sup_{P' \in \Pi'} P'(A)$, for all $A \in \mathcal{F}$, then it is immediate to see that $\overline{P}(A) = \sup_{P \in \Pi} P(A) = \sup_{P' \in \text{ex}\Pi'} P'(A) = \sup_{P' \in \Pi'} P'(A)$, for all $A \in \mathcal{F}$, since $\Pi \subset \Pi'$.

Suppose now that $\overline{P}(A) = \sup_{P \in \Pi} P(A)$, for all $A \in \mathcal{F}$. Then, we have that for all $P \in \Pi$ and all $A \in \mathcal{F}$, $P(A) \leq \overline{P}(A)$. Pick now any $P' \in \Pi'$. We can write it as $P' = \sum_{j=1}^{k} \beta_j P_j$, where $\beta_j \in [0, 1]$, for all

$j \in \{1, \dots, k\}$, $\sum_{j=1}^{k} \beta_j = 1$, and $\{P_j\}_{j=1}^{k} = \Pi$. Pick then any $A \in \mathcal{F}$; we have

$$P'(A) = \sum_{j=1}^{k} \beta_j P_j(A) \leq \sum_{j=1}^{k} \beta_j \overline{P}(A) = \overline{P}(A).$$

So $\overline{P}(A) \geq P'(A)$. Because this holds for all $P' \in \Pi'$ and all $A \in \mathcal{F}$, the claim is proven. □

*Proof of Proposition 6.* The lower bound for the upper entropy of $\Pi'$ comes immediately from $\Pi'$ being a superset of $\Pi$.

Let us now prove the lower entropy equality. Let $\Pi = \{P_1, \dots, P_k\}$ and $\Pi' = \text{Conv}\Pi$. Pick any $P' \in \Pi'$. By the definition of $\Pi'$, there exists a collection of non-negative reals $\{\beta_j\}_{j=1}^{k}$ such that $\sum_{j=1}^{k} \beta_j = 1$ and $\sum_{j=1}^{k} \beta_j P_j = P'$. By the concavity of the entropy, we have that

$$H(P') = H\left(\sum_{j=1}^{k} \beta_j P_j\right) \geq \sum_{j=1}^{k} \beta_j H(P_j) \geq \sum_{j=1}^{k} \beta_j \underline{H}(P) = \underline{H}(P) := \inf_{P \in \Pi} H(P).$$

Since $P'$ was chosen arbitrarily and $\Pi$ is finite, this implies that

$$\inf_{P' \in \Pi'} H(P') =: \underline{H}(P') \geq \underline{H}(P).$$

In addition, we have that, since $\Pi \subset \Pi'$, $\underline{H}(P') \leq \underline{H}(P)$. Combining this with the above result, we obtain

$$\underline{H}(P') = \underline{H}(P).$$

□

*Proof of Proposition 8.* Pick any metric $d$ on the space $\Delta_\Omega \equiv \Delta(\Omega, \mathcal{F})$ of probabilities on $(\Omega, \mathcal{F})$, and any $\mathbf{P}' \in \Pi'$. Because $\mathbf{P}'$ belongs to $\Pi'$, $\inf_{P' \in \Pi'} d(P', \Psi)$ can only be either equal to or smaller than $d(\mathbf{P}', \Psi)$. By the definition of $d(\Pi', \Psi)$, if $\inf_{P' \in \Pi'} d(P', \Psi) = d(\mathbf{P}', \Psi)$, then $d(\Pi', \Psi) = d(\mathbf{P}', \Psi)$. If instead $\inf_{P' \in \Pi'} d(P', \Psi) < d(\mathbf{P}', \Psi)$, then $d(\Pi', \Psi) < d(\mathbf{P}', \Psi)$. The proof is similar for a generic divergence div on $\Delta(\Omega, \mathcal{F})$. □

*Proof of Theorem 10.* Assume $\mathcal{P}^{\text{co}}, \mathcal{L}_\theta^{\text{co}}$ are nonempty and weak$^\star$-closed. Then, by Marinacci & Montrucchio (2004, Proposition 3), they are also weak$^\star$-compact. Pick any $A \in \mathcal{B}$. Recall that we can rewrite the usual Bayes' updating rule as

$$\begin{aligned} P_D(A) &= \frac{\int_\Theta L(\theta) \mathbb{1}_A(\theta) P(\mathrm{d}\theta)}{\int_\Theta L(\theta) \mathbb{1}_A(\theta) P(\mathrm{d}\theta) + \int_\Theta L(\theta) \mathbb{1}_{A^c}(\theta) P(\mathrm{d}\theta)} \\ &= \frac{1}{1 + \frac{\int_\Theta L(\theta) \mathbb{1}_{A^c}(\theta) P(\mathrm{d}\theta)}{\int_\Theta L(\theta) \mathbb{1}_A(\theta) P(\mathrm{d}\theta)}}, \end{aligned}$$

which is maximized when

$$\frac{\int_\Theta L(\theta) \mathbb{1}_{A^c}(\theta) P(\mathrm{d}\theta)}{\int_\Theta L(\theta) \mathbb{1}_A(\theta) P(\mathrm{d}\theta)}$$

is minimized. But

$$\frac{\int_\Theta L(\theta) \mathbb{1}_{A^c}(\theta) P(\mathrm{d}\theta)}{\int_\Theta L(\theta) \mathbb{1}_A(\theta) P(\mathrm{d}\theta)} \geq \frac{\inf_{P \in \mathcal{P}^{\text{co}}} \int_\Theta \underline{L}(\theta) \mathbb{1}_{A^c}(\theta) P(\mathrm{d}\theta)}{\sup_{P \in \mathcal{P}^{\text{co}}} \int_\Theta \overline{L}(\theta) \mathbb{1}_A(\theta) P(\mathrm{d}\theta)},$$

which proves the inequality in equation 5. Assume now that $\overline{P}$ is concave. By Wasserman & Kadane (1990, Lemma 1), we have that there exists $\mathbf{P} \in \mathcal{P}^{\text{co}}$ such that

$$\sup_{P \in \mathcal{P}^{\text{co}}} \int_\Theta L(\theta) \mathbb{1}_A(\theta) P(\mathrm{d}\theta) = \int_\Theta L(\theta) \mathbb{1}_A(\theta) \mathbf{P}(\mathrm{d}\theta), \tag{7}$$

for all $L \in \mathscr{L}$. In addition, by Wasserman & Kadane (1990, Lemma 4), we have that for all $X \in \mathscr{X}$ and all $\epsilon > 0$, there exists a non-negative, upper semi-continuous function $h \leq X$ such that

$$\begin{aligned}\left[\sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta X(\theta)P(\mathrm{d}\theta)\right] - \epsilon &< \sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta h(\theta)P(\mathrm{d}\theta)\\ &\leq \sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta X(\theta)P(\mathrm{d}\theta).\end{aligned} \tag{8}$$

Let now $X = \overline{L}\mathbb{1}_A$. Notice that since $\mathcal{L}_\theta^{\text{co}}$ is weak$^\star$-compact, by equation 4 so is $\mathscr{L}$. This implies that $\underline{L}, \overline{L} \in \mathscr{L}$, since a compact set always contains its boundary, so $\overline{L} \in \mathscr{X}$ as well, and in turn $\overline{L}\mathbb{1}_A \in \mathscr{X}$. Fix then any $L \in \mathscr{L}$ and put $h = L\mathbb{1}_A$. It is immediate to see that $h$ is non-negative and upper semi-continuous. Then, by equation 8, we have that for all $\epsilon > 0$

$$\begin{aligned}&\left[\sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta)\right] - \epsilon <\\ &\sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta L(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta) \leq \sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta).\end{aligned} \tag{9}$$

Combining equation 7 andequation 9, we obtain

$$\begin{aligned}&\left[\sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta)\right] - \epsilon\\ &< \int_\Theta L(\theta)\mathbb{1}_A(\theta)\mathbf{P}(\mathrm{d}\theta) \leq \sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta),\end{aligned} \tag{10}$$

for all $L \in \mathscr{L}$.

Pick now any $\epsilon > 0$ and put

$$\begin{aligned}k := &\sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta)\\ &+ \inf_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \underline{L}(\theta)\mathbb{1}_{A^c}(\theta)P(\mathrm{d}\theta) > 0.\end{aligned}$$

Choose any $L \in \mathscr{L}$ and $\delta \in (0, \epsilon k)$. By equation 10 we have that $[\sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta)] - \delta < \int_\Theta L(\theta)\mathbb{1}_A(\theta)\mathbf{P}(\mathrm{d}\theta)$ and that $[\inf_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \underline{L}(\theta)\mathbb{1}_{A^c}(\theta)P(\mathrm{d}\theta)] + \delta > \int_\Theta L(\theta)\mathbb{1}_{A^c}(\theta)\mathbf{P}(\mathrm{d}\theta)$. Recall that $\mathbf{c} := \sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta) + \inf_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \underline{L}(\theta)\mathbb{1}_{A^c}(\theta)P(\mathrm{d}\theta)$, and define $\mathbf{d} := \int_\Theta L(\theta)\mathbb{1}_A(\theta)\mathbf{P}(\mathrm{d}\theta) + \int_\Theta L(\theta)\mathbb{1}_{A^c}(\theta)\mathbf{P}(\mathrm{d}\theta)$. Then,

$$\begin{aligned}\mathbf{P}_D(A) &= \frac{\int_\Theta L(\theta)\mathbb{1}_A(\theta)\mathbf{P}(\mathrm{d}\theta)}{\mathbf{d}}\\ &\geq \frac{\left[\sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta)\right] - \delta}{\mathbf{c} + \delta - \delta}\\ &= \frac{\sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta)}{\mathbf{c}} - \frac{\delta}{k}\\ &> \frac{\sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta)}{\mathbf{c}} - \epsilon.\end{aligned}$$

Since this holds for all $\epsilon > 0$, we have that

$$\sup_{P_D\in\mathcal{P}_D^{\text{co}}} P_D(A) = \frac{\sup_{P\in\mathcal{P}^{\text{co}}}\int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta)}{\mathbf{c}},$$

concluding the proof. □

*Proof of Lemma 11.* Walley (1981); Wasserman & Kadane (1990) show that concave upper probabilities are closed with respect to the generalized Bayes' rule. In particular, this means that, if we let $\mathbf{b} := \sup_{P\in\mathcal{P}^{\text{co}}} \int_\Theta L(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta) + \inf_{P\in\mathcal{P}^{\text{co}}} \int_\Theta L(\theta)\mathbb{1}_{A^c}(\theta)P(\mathrm{d}\theta)$, for any fixed $A \in \mathcal{B}$, if $\overline{P}$ is concave, then for all $L \in \mathscr{L}$

$$\overline{P}_D(A) = \frac{\sup_{P\in\mathcal{P}^{\text{co}}} \int_\Theta L(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta)}{\mathbf{b}} \tag{11}$$

is concave. But since $\mathcal{L}_\theta^{\text{co}}$ is weak$^\star$-compact (by our assumption and Marinacci & Montrucchio (2004, Proposition 3)), by equation 4 so is $\mathscr{L}$. This implies that $\underline{L}, \overline{L} \in \mathscr{L}$, since a compact set always contains its boundary. Call then $L' = \overline{L}\mathbb{1}_A + \underline{L}\mathbb{1}_{A^c}$. It is immediate to see that $L' \in \mathscr{L}$. Then, by equation 11 we have that if we call $\mathbf{b}' := \sup_{P\in\mathcal{P}^{\text{co}}} \int_\Theta L'(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta) + \inf_{P\in\mathcal{P}^{\text{co}}} \int_\Theta L'(\theta)\mathbb{1}_{A^c}(\theta)P(\mathrm{d}\theta)$, it follows that

$$\begin{aligned}\overline{P}_D(A) &= \frac{\sup_{P\in\mathcal{P}^{\text{co}}} \int_\Theta L'(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta)}{\mathbf{b}'} \\ &= \frac{\sup_{P\in\mathcal{P}^{\text{co}}} \int_\Theta \overline{L}(\theta)\mathbb{1}_A(\theta)P(\mathrm{d}\theta)}{\mathbf{c}}\end{aligned}$$

is concave, concluding the proof. □

*Proof of Theorem 13.* Suppose $\Omega$ is uncountable. First notice that, since we assumed $\frac{\mathrm{d}P}{\mathrm{d}\mu}$ to be continuous and bounded for all $P \in \mathscr{P}$, then so is $\log \circ \frac{\mathrm{d}P}{\mathrm{d}\mu}$, for all $P \in \mathscr{P}$, since composing a continuous function with a continuous and bounded one gives us a continuous and bounded function. This entails that the Choquet integrals of $\log \circ \frac{\mathrm{d}P}{\mathrm{d}\mu}$ with respect to $\underline{P}$ and $\overline{P}$ are both well defined. In addition, being continuous and bounded, both $\frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $\log \circ \frac{\mathrm{d}P}{\mathrm{d}\mu}$ attain their infima and suprema thanks to Weierstrass' extreme value theorem, for all $P \in \mathscr{P}$. Hence, all the Choquet integrals used in this proof are well defined.

Then, we have the following

$$\begin{aligned}\overline{H}(P) :&= \sup_{P\in\mathscr{P}} H(P) = \sup_{P\in\mathscr{P}} \left( -\int_\Omega \log\left[\frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right] P(\mathrm{d}\omega)\right) \\ &= \sup_{P\in\mathscr{P}} \int_\Omega (-\log)\left[\frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right] P(\mathrm{d}\omega) \\ &\leq \sup_{P\in\mathscr{P}} \int_\Omega \sup_{P\in\mathscr{P}} \left\{(-\log)\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right)\right\} P(\mathrm{d}\omega) && (12)\\ &\leq \int_\Omega \sup_{P\in\mathscr{P}} \left\{(-\log)\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right)\right\} \overline{P}(\mathrm{d}\omega) && (13)\\ &= -\int_\Omega \inf_{P\in\mathscr{P}} \left\{\log\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right)\right\} \overline{P}(\mathrm{d}\omega) && (14)\\ &= -\int_\Omega \log\left(\inf_{P\in\mathscr{P}} \frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right) \overline{P}(\mathrm{d}\omega) && (15)\\ &= -\int_\Omega \log\left[\underline{\pi}(\omega)\right] \overline{P}(\mathrm{d}\omega) = H(\overline{P}).\end{aligned}$$

The inequality in equation 12 is true because for all $\omega \in \Omega$,

$$\sup_{P\in\mathscr{P}} \left\{(-\log)\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right)\right\} \geq (-\log)\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right).$$

The inequality in equation 13 is a property of Choquet integrals taken with respect to upper probabilities (Marinacci & Montrucchio, 2004). The equality in equation 14 is true because for a generic function $f$, we have that $\sup -f = -\inf f$. Finally, the equality in equation 15 is true because the logarithm is a strictly increasing function. By Marinacci & Montrucchio (2004, Theorem 38), if $\overline{P}$ is concave, then inequality equation 13 holds with an equality, and so the bound is tighter.

The proof for $\underline{H}(P) \geq H(\underline{P})$ is similar; we use the facts that

- $\inf_{P\in\mathscr{P}}\left\{(-\log)\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right)\right\} \leq (-\log)\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right)$, for all $\omega \in \Omega$;
- by Marinacci & Montrucchio (2004),

$$\begin{aligned}&\inf_{P\in\mathscr{P}}\int_\Omega \inf_{P\in\mathscr{P}}\left\{(-\log)\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right)\right\}P(\mathrm{d}\omega)\\ &\geq \int_\Omega \inf_{P\in\mathscr{P}}\left\{(-\log)\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}(\omega)\right)\right\}\underline{P}(\mathrm{d}\omega);\end{aligned} \tag{16}$$

- for a generic function $f$, $\inf -f = -\sup f$;
- by Marinacci & Montrucchio (2004, Theorem 38), if $\underline{P}$ is convex, then equation 16 holds with an equality.

Suppose now $\Omega$ is at most countable; in this case, we do not need any assumptions to make the Choquet integrals well defined, since we will not deal with density functions. The following holds

$$\begin{aligned}\overline{H}(P) :&= \sup_{P\in\mathscr{P}} H(P) = \sup_{P\in\mathscr{P}}\left(-\sum_{\omega\in\Omega}P(\{\omega\})\log\left[P(\{\omega\})\right]\right)\\ &= \sup_{P\in\mathscr{P}}\sum_{\omega\in\Omega}P(\{\omega\})(-\log)\left[P(\{\omega\})\right]\\ &\leq \sum_{\omega\in\Omega}\sup_{P\in\mathscr{P}}\left\{P(\{\omega\})(-\log)\left[P(\{\omega\})\right]\right\} &(17)\\ &\leq \sum_{\omega\in\Omega}\overline{P}(\{\omega\})\sup_{P\in\mathscr{P}}(-\log)\left[P(\{\omega\})\right] &(18)\\ &= -\sum_{\omega\in\Omega}\overline{P}(\{\omega\})\inf_{P\in\mathscr{P}}\log\left[P(\{\omega\})\right] &(19)\\ &= -\sum_{\omega\in\Omega}\overline{P}(\{\omega\})\log\left[\inf_{P\in\mathscr{P}}P(\{\omega\})\right] &(20)\\ &= -\sum_{\omega\in\Omega}\overline{P}(\{\omega\})\log\left[\underline{P}(\{\omega\})\right] = H(\overline{P}).\end{aligned}$$

The inequality in equation 17 comes from the well known fact that the sum of the suprema is at least equal to the supremum of the sum. The inequality in equation 18 comes from the fact that for differentiable functions, the product of the suprema is at least equal to the supremum of the product. The equality in equation 19 is true because for a generic function $f$, we have that $\sup -f = -\inf f$. Finally, the equality in equation 20 is true because the logarithm is a strictly increasing function. By Marinacci & Montrucchio (2004, Theorem 38), if $\overline{P}$ is concave, then inequality equation 17 holds with an equality, and so the bound is tighter.

The proof for $\underline{H}(P) \geq H(\underline{P})$ is similar; we use the facts that

- the sum of the infima is at most equal to the infimum of the sum;
- for differentiable functions, the product of the infima is at most equal to the infimum of the product;
- for a generic function $f$, $\inf -f = -\sup f$;
- by Marinacci & Montrucchio (2004, Theorem 38), if $\underline{P}$ is convex, then $\inf_{P\in\mathscr{P}}\sum_{\omega\in\Omega}P(\{\omega\})(-\log)\left[P(\{\omega\})\right] = \sum_{\omega\in\Omega}\inf_{P\in\mathscr{P}}\left\{P(\{\omega\})(-\log)\left[P(\{\omega\})\right]\right\}$.

□

*Proof of Lemma 16.* Fix any $p \in [1, \infty]$ and pick any $\mathbf{P} \in \mathscr{P}$. Because $\Phi_p^+(\mathbf{P}, n) \subset \Phi_p^+(\mathscr{P}, n)$, then $\inf_{\alpha \in \Phi_p^+(\mathscr{P},n)} W_p(\alpha, P')$ can only be either equal or smaller than $\inf_{\alpha \in \Phi_p^+(\mathbf{P},n)} W_p(\alpha, P')$. Now, if $\inf_{\alpha \in \Phi_p^+(\mathscr{P},n)} W_p(\alpha, P') = \inf_{\alpha \in \Phi_p^+(\mathbf{P},n)} W_p(\alpha, P')$, then $W_p^+(\mathscr{P}, P') = W_p^+(\mathbf{P}, P')$. If instead $\inf_{\alpha \in \Phi_p^+(\mathscr{P},n)} W_p(\alpha, P') < \inf_{\alpha \in \Phi_p^+(\mathbf{P},n)} W_p(\alpha, P')$, then $W_p^+(\mathscr{P}, P') < W_p^+(\mathbf{P}, P')$. This concludes the first part of the proof. Fix then any convex functional $f$ on $\mathbb{R}$ such that $f(0) = 1$; the proof is similar for $f$-divergences. □

*Proof of Lemma 17.* The proof is very similar to that of Lemma 16. □