Decoupling epistemic and aleatoric uncertainties with possibility theory

Nong Minh Hieu
NTU & SMU, Singapore

Jeremie Houssineau
NTU, Singapore

Neil K. Chada
CityU, Hong Kong

Emmanuel Delande
CNES, France

Abstract

The special role of epistemic uncertainty in Machine Learning is now well recognised, and an increasing amount of research is focused on methods for dealing specifically with such a lack of knowledge. Yet, most often, a probabilistic representation is considered for both aleatoric and epistemic uncertainties, hence creating challenges in applications where decoupling these two types of uncertainty is necessary. In this work, we show that an alternative representation of epistemic uncertainty, based on possibility theory, maintains many of the convenient features of standard Bayesian inference while displaying specific behaviours and properties that closely match the ones of an intuitive notion of information. Our main contributions are i) a general framework for jointly representing epistemic and aleatoric uncertainties, ii) a Bernstein-von Mises theorem for the analogue of Bayes'rule in possibility theory, iii) a version of the law of large numbers and of the central limit theorem for the associated variables, and iv) an analysis of the properties of the possibilistic maximum a posteriori. These results highlight that a dedicated and principled representation of epistemic uncertainty, that is compatible with standard Bayesian inference and preserves many of its strengths, is attainable.

1 INTRODUCTION

Differentiating epistemic uncertainty, due to lack of knowledge, from aleatoric uncertainty, due to randomness, is key for a number of tasks within Machine Learning (Hüllermeier and Waegeman, 2021). Large language models (LLMs) provide a great example of the importance of such a distinction (Yadkori et al., 2024), where the inherent randomness of language, stemming from the existence of multiple potential answers to a question or multiple potential phrasings, should not be confused with the lack of knowledge about the correct answer. Successfully decoupling these two types of uncertainty in LLMs holds the promise of more trustworthy outputs with a better behaviour for out-of-distribution inputs. Yet, such an endeavour is made more difficult by the fact that both epistemic and aleatoric uncertainties are often modelled via the same representation of uncertainty: probability theory.

While probability theory and aleatoric uncertainty are perfectly matched, the same cannot be said of epistemic uncertainty and, historically, there has been a considerable effort dedicated to generalising probability theory, e.g., with Dempster-Shafer theory (DST) (Dempster, 1968) or imprecise probability (IP) (Augustin et al., 2014). Both DST and IP have been leveraged to further the ability of neural networks to capture epistemic uncertainty by modifying the last layer and the loss in classification tasks (Sensoy et al., 2018) or by modelling the uncertainty in labels (Lienen and Hüllermeier, 2021); yet, the sophistication of these methods mean that applying them more generally, e.g., to model the uncertainty in the high-dimensional parameters of deep neural networks, could be challenging in practice. In this work, we consider a particular formulation of possibility theory (Zadeh, 1978) which is a scalable and flexible framework dedicated to modelling epistemic uncertainty. The version of Bayesian inference stemming from the considered formulation of possibility theory has been leveraged by (Houssineau and Bishop, 2018) for state space models, by (Houssineau, 2021) for point processes, and by (Chen et al., 2021) for control problems.

The overall objective of this work is to show that the theories of possibility and probability can be combined

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

in a general way, without necessarily complexifying the corresponding inference problems, and retaining many of the desirable properties of Bayesian inference. After reviewing possibility theory in Section 2 and discussing the related work in Section 3, we introduce such a framework in Section 4 and highlight the main features of the approach to inference it underpins in Section 5. Our main theoretical contributions, which can be found in Section 6, are i) a Bernstein-von Mises theorem for the analogue of Bayes'rule in possibility theory, ii) a version of the law of large numbers (LLN) and of the central limit theorem (CLT) for the associated variables, and iii) an analysis of the properties of the possibilistic maximum a posteriori.

2 OVERVIEW OF POSSIBILITY THEORY

To understand the fundamentals of possibility theory, we consider inference problems that involve a deterministic yet uncertain parameter. We denote by Ω_{d} the sample space of deterministic phenomena and by $\omega^* \in \Omega_{\mathrm{d}}$ the true outcome. In this section, we provide a brief description of the core concepts in possibility theory and their equivalence in probability theory.

Random vs. deterministic uncertain variables A *deterministic uncertain variable* $\boldsymbol{\theta}$ is a mapping from the sample space Ω_{d} to a parameter space Θ such that $\boldsymbol{\theta}(\omega)$ would be the true value of the parameter if the true outcome were $\omega \in \Omega_{\mathrm{d}}$; in particular, $\boldsymbol{\theta}(\omega^*) = \theta^*$ is the true value of the parameter. Random variables will be assumed to be mappings on a probability space $(\Omega_{\mathrm{r}}, \mathcal{F}, \mathbb{P})$ where Ω_{r} is the sample space for random phenomena. We will not focus on measure-theoretic considerations and will instead implicitly assume subsets and functions to be suitably measurable.

Probability density function vs. possibility function A given random variable X in a space X gives rise to a probability measure P on X, which in many cases can be characterised by a corresponding probability density function (PDF) $p : \mathsf{X} \to [0, \infty)$. The probability of an event $X \in B$ is found by integrating the PDF over B, i.e., $\mathbb{P}(X \in B) = P(B) = \int_B p(x)\mathrm{d}x$. In possibility theory, a given deterministic uncertain variable $\boldsymbol{\theta}$ on Θ is *described* by a set function $\bar{P}$, which we assume to take the form $\bar{P}(A) = \sup_{\theta \in A} f_{\boldsymbol{\theta}}(\theta)$, with $f_{\boldsymbol{\theta}}$ a non-negative function with supremum equal to one that we refer to as a *possibility function*. For a subset A of Θ, the scalar $\bar{P}(A)$ defines the *credibility* of the event $\boldsymbol{\theta} \in A$. A credibility of 1 implies that there is no evidence against $\boldsymbol{\theta} \in A$, or equivalently $\theta^* \in A$. On the contrary, a credibility of 0 indicates that it is not possible that $\theta^* \in A$ given the current information.

The set function $\bar{P}$ is an outer measure, i.e., it satisfies i) $\bar{P}(\emptyset) = 0$, ii) if $A, B \subseteq \Theta$ with $A \subseteq B$, then $\bar{P}(A) \leq \bar{P}(B)$, and iii) for any subsets $A_1, A_2, \ldots$, $\bar{P}\big(\bigcup_{n=1}^{\infty} A_n\big) \leq \sum_{n=1}^{\infty} \bar{P}(A_n)$. We assume that the true parameter is in Θ, so the outer measure $\bar{P}$ also verifies $\bar{P}(\Theta) = 1$. We therefore refer to set functions such as $\bar{P}$ as *outer probability measures* (OPMs).

A crucial difference between probability measures such as P and OPMs such as $\bar{P}$ is that the former is characterised by the random variable X, whereas the latter simply expresses information about $\boldsymbol{\theta}$ and is not defined as the image of a more fundamental OPM on Ω_{d}. Such a behaviour is to be expected from a mathematical object modelling *information*, which can vary between modellers.

Change of variable Let $\boldsymbol{\psi}$ be an uncertain variable defined as $\boldsymbol{\psi} = T(\boldsymbol{\theta})$ for a given map $T : \Theta \to \Psi$. Then, the possibility function that describes $\boldsymbol{\psi}$ is defined by the change of variable formula (Baudrit et al., 2008)

$$f_{\boldsymbol{\psi}}(\psi) = \sup_{\theta \in T^{-1}[\psi]} f_{\boldsymbol{\theta}}(\theta),$$

with $T^{-1}[\psi] = \{\theta \in \Theta : T(\theta) = \psi\}$ the inverse image of the singleton $\{\psi\}$, and with the convention that $\sup \emptyset = 0$. If $f_{\boldsymbol{\theta},\boldsymbol{\psi}}$ models joint information about $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, then marginalisation, which can be seen as a change of variable through $(\theta, \psi) \mapsto \theta$, takes the form $f_{\boldsymbol{\theta}}(\theta) = \sup_{\psi \in \Psi} f_{\boldsymbol{\theta},\boldsymbol{\psi}}(\theta, \psi)$.

Expectation and variance Since the main operator behind possibility functions is the maximum, it makes sense to define the corresponding notion of expected value as the mode, i.e., as the point(s) at which a given possibility function is maximised. Specifically, for a deterministic uncertain variable $\boldsymbol{\theta}$ on Θ described by the possibility function $f_{\boldsymbol{\theta}} : \Theta \to [0, 1]$, the expected value is defined as

$$\mathbb{E}^*_{f_{\boldsymbol{\theta}}}[\boldsymbol{\theta}] \doteq \arg\max_{\theta \in \Theta} f_{\boldsymbol{\theta}}(\theta).$$

which we will write as $\mathbb{E}^*[\boldsymbol{\theta}]$ when there is no ambiguity. It corresponds to the most likely value(s) of θ^* when the available information is captured by $f_{\boldsymbol{\theta}}$, and satisfies similar properties as the maximum likelihood estimator (MLE). For instance, if T is any mapping on Θ then it holds that $\mathbb{E}^*[T(\boldsymbol{\theta})] = T(\mathbb{E}^*[\boldsymbol{\theta}])$. This property reinforces the interpretation of $\mathbb{E}^*[\boldsymbol{\theta}]$ as the most likely value(s) of the parameter, e.g., if s is the most likely value for a variance parameter s^* then so is $1/s$ for the corresponding precision parameter $\tau^* = 1/s^*$.When $\mathbb{E}^*[\boldsymbol{\theta}]$ is a singleton, say $\{\mu\}$, we do not distinguish between the set $\{\mu\}$ and the value μ.

Since there is no natural variability in deterministic uncertain variables, the associated notion of variance must be based on a different principle; instead, we aim to quantify the uncertainty related to $\mathbb{E}^*[\boldsymbol{\theta}]$ via the curvature at the mode, as is usual in the standard Gaussian approximation. This only makes sense when $\mathbb{E}^*[\boldsymbol{\theta}]$ is a singleton and when $f_{\boldsymbol{\theta}}$ is twice differentiable at $\mathbb{E}^*[\boldsymbol{\theta}]$. Because of the normalisation of possibility functions, it holds that $H(f_{\boldsymbol{\theta}}(\theta)) = H(\log f_{\boldsymbol{\theta}}(\theta))$ at $\theta = \mathbb{E}^*[\boldsymbol{\theta}]$, with $H(\cdot)$ denoting the Hessian matrix. We favour the latter expression and first define a possibilistic analogue of the Fisher information as

$$\bar{\mathcal{I}}_{f_{\boldsymbol{\theta}}}(\boldsymbol{\theta}) \doteq \mathbb{E}^*\big[- H\big(\log f_{\boldsymbol{\theta}}(\boldsymbol{\theta})\big)\big].$$

When $\bar{\mathcal{I}}_{f_{\boldsymbol{\theta}}}(\boldsymbol{\theta})$ is positive definite, the variance of $\boldsymbol{\theta}$ is defined as $\mathbb{V}^*_{f_{\boldsymbol{\theta}}}(\boldsymbol{\theta}) = \bar{\mathcal{I}}_{f_{\boldsymbol{\theta}}}(\boldsymbol{\theta})^{-1}$, which underscores the interpretation of the possibilistic variance as uncertainty rather than variability. As before, we will simply write $\bar{\mathcal{I}}(\cdot)$ and $\mathbb{V}^*(\cdot)$ when there is no ambiguity. We further motivate the definition of these notions of expected value and variance in Section 6 via the possibilistic version of the LLN and of the CLT, in which they appear naturally.

Gaussian possibility function: If $\Theta \subseteq \mathbb{R}^d$ then the Gaussian possibility function with the vector $\mu \in \Theta$ as expected value and with the positive semi-definite matrix Λ as precision matrix is defined as

$$\overline{\mathrm{N}}_d(\theta; \mu, \Lambda) = \exp\Big(- \frac{1}{2}(\theta - \mu)^\intercal \Lambda (\theta - \mu)\Big).$$

We will simply write $\overline{\mathrm{N}}_d(\mu, \Lambda)$ to refer to the entire function and $\overline{\mathrm{N}}(\mu, \Lambda)$ for the case where $d = 1$. Two important differences between the Gaussian possibility function and its probabilistic analogue are i) Θ can be a strict subset of $\mathbb{R}^d$ without requiring re-normalization ii) Λ does not need to be positive definite, e.g., it could be the case that Λ is the $d \times d$ zero matrix $\mathbf{0}_{d\times d}$, which motivates deviating from convention and parameterising the normal possibility function by the precision matrix. The Gaussian possibility function can be verified to satisfy most of the properties of its probabilistic counterpart; for instance Houssineau and Bishop (2018) proved that the Kalman filter remains the same in possibility theory. If $\boldsymbol{\theta}$ is described by $\overline{\mathrm{N}}_d(\mu, \Lambda)$, with Λ positive definite, then $\mathbb{E}^*[\boldsymbol{\theta}] = \mu$ and $\mathbb{V}^*(\boldsymbol{\theta}) = \Lambda^{-1}$, so that the standard Gaussian approximation corresponds to a moment-matching procedure in this context.

Non-uniqueness of possibility functions One key distinction between possibility theory and probability theory is that deterministic uncertain variables do not characterizes possibility functions. Specifically, if the information about an uncertain variable $\boldsymbol{\theta}$ is described by the possibility function $f : \Theta \to [0, 1]$, then it is also described by any $g : \Theta \to [0, 1]$ such that $g(\theta) \geq f(\theta), \forall \theta \in \Theta$. Although g is less informative than f, there might be reasons to forgo information, e.g., to obtain certain statistical properties. For instance, g can be set to the normal possibility function $\overline{\mathrm{N}}_d(\mu, \Lambda)$ possessing some desired traits. This can be achieved in a principled manner using possibility theory by finding the tightest upper bounding normal possibility function whose precision matrix has the targeted properties. For example, in Kimchaiwong et al. (2024), conditional independence is enforced by finding an upper-bounding normal possibility function such that some of the off-diagonal entries in the precision matrix are 0.

Probability dilution When both epistemic and aleatoric uncertainties are present, a potential issue with fully probabilistic models is that an increase in epistemic uncertainty can translate into probabilities of (extreme) events being reduced. This can have serious consequences for decision making, see e.g. Balch et al. (2019), which considers collision risks caused by debris orbiting Earth. Indeed, the uncertainty about the state of debris is typically overwhelmingly epistemic and, in fully probabilistic models, more uncertainty about these debris often leads paradoxically to lower probabilities of collisions. The same ambiguity is likely to arise for other extreme events with seemingly low probabilities: either the probability of the event is truly low, or the epistemic uncertainty is too substantial to allow for a faithful assessment. The proposed framework, combining possibility and probability theories, allows to distinguish between these two cases.

3 RELATED WORK

Random (fuzzy) sets Similarly to the frequentist approach, specifically the notion of confidence interval, a natural approach to make statements about uncertain quantities is to use random sets. DST (Dempster, 1968) provides a framework to handle the corresponding beliefs. Stating that a parameter θ^* is within a set A is equivalent to describing $\boldsymbol{\theta}$ by the possibility function $\mathbf{1}_A$, defined as the indicator of the set A. To generalise this statement to any possibility function $f_{\boldsymbol{\theta}}$, one has to consider the fuzzy version (Yen, 1990) of DST, in which $f_{\boldsymbol{\theta}}$ is related to the so-called membership function of the corresponding fuzzy set. In Section 4, we will generalise such an approach by allowing different combinations of epistemic and aleatoric uncertainties, and we will prove in Section 6 that the associated approach to inference has a suitable asymptotic behaviour. Subjective logic Jøsang (2016), which underpins evidential neural networks Sensoy et al. (2018), is also based on DST.

Imprecise probability IP (Augustin et al., 2014; Walley, 1991), another generalisation of probability theory, also shares some of our motivation. Indeed, within IP, one considers *gambles*, *lower previsions*, and *lower probabilities*. Bayesian inference can be carried out in this context by considering a suitable class of probability distributions for the likelihood and/or for the prior Wasserman and Kadane (1990); Caprio et al. (2023). For instance, IP has been used in Lienen and Hüllermeier (2021) for expressing the uncertainty about labels in self-supervised learning. The relationship between our approach and IP is stated more formally in Section 5.4.

Energy-based models Although energy-based models (LeCun et al., 2006) have become almost synonymous with the associated probabilistic models via the corresponding Gibbs distribution, see e.g. Lee et al. (2023) and Xu et al. (2024) for recent work, their original interpretation and behaviour closely match the ones seen above for possibility functions. Specifically, for an energy function $E(z, y, x)$ with latent variable z, we consider the possibility function $f(z, y \,|\, x) \propto \exp(-\beta E(z, y, x))$ for some given $\beta > 0$; the corresponding marginal possibility function $\sup_z f(z, y \,|\, x)$ indeed yields the energy function $E(y, x) = \min_z E(z, y, x)$, and the conditional expected value $\mathbb{E}^*[\boldsymbol{y} \,|\, \boldsymbol{x} = x]$ is the outcome y that minimises the energy. Also, the Gibbs possibility function $f(y \,|\, x) \propto \exp(-\beta E(y, x))$ does not display the same limitations as its probabilistic analogue: the normalising constant is well defined as long as $E(y, x) > -\infty$ and computing it is generally possible since minimising $E(y, x)$ for a given x is required for predicting new labels.

4 GENERAL OUTER PROBABILITY MEASURES

In general, some of the sources of uncertainty in a model will be epistemic while others might be aleatoric, so that a general framework is needed to model both as faithfully as possible.

Example 1. Let $\Phi_{\text{bb}} : \mathsf{X} \to \mathsf{Y}$ be a black-box neural network taking as input a realisation $x \in \mathsf{X}$ of a random variable X drawn from some given data distribution D. We do not have access to the realisation x but we assume that we know D, and we have some information about the output $y = \Phi_{\text{bb}}(x) \in \mathsf{Y}$, for a given x, modelled as a conditional possibility function $f_{\text{bb}}(y \,|\, x)$. This situation arises in practice, e.g., when a sensor applies some post-processing to a raw data point x but only provides the processed data y. We can model the outcome y in this situation by an OPM of the form

$$\bar{P}_{\text{bb}}(B) = \int \left[\sup_{y \in B} f_{\text{bb}}(y \,|\, x) \right] D(x) \mathrm{d}x,$$

for any $B \subseteq \mathsf{Y}$. If nothing is known about the black-box neural network Φ_{bb}, then $f_{\text{bb}}(y \,|\, x) = 1$ for all $y \in \mathsf{Y}$ and $\bar{P}_{\text{bb}}(B) = 1$ for any subset B; conversely, if Φ_{bb} is known, then $f_{\text{bb}}(y \,|\, x) = \mathbf{1}_{\Phi_{\text{bb}}(x)}(y)$ and $\bar{P}_{\text{bb}}(\cdot)$ is the probability measure defined as the image of D via Φ_{bb}, as required. The uncertainty modelled by OPMs of the same form as $\bar{P}_{\text{bb}}(\cdot)$ is the one considered in fuzzy Dempster-Shafer theory (Yen, 1990).

To accommodate for the type of uncertainty discussed in Example 1, we consider the following notion combining random variables and deterministic uncertain variables.

Definition 1. An uncertain variable is a mapping $\mathcal{X}$ on $\Omega_{\text{d}} \times \Omega_{\text{r}}$ such that $\mathcal{X}(\omega, \cdot)$ is a random variable for any $\omega \in \Omega_{\text{d}}$.

To better differentiate the mathematical objects related to the two considered types of uncertainty, the notational conventions summarised in Table 1 will be used consistently throughout this work. To gain intuition on the generality of this notion, the two simplest yet non-trivial types of uncertain variables are considered next.

Statistical model Let $\boldsymbol{\theta}$ be a deterministic uncertain variable in Θ defined regardless on the random outcomes and, for a given value θ of $\boldsymbol{\theta}$, let X be a random variable on X defined based on θ (formally, we would need to define a parameterised family $\{X_\theta\}_{\theta \in \Theta}$ of random variables). This case corresponds to statistical inference with a family $\{p_X(\cdot \,|\, \theta) : \theta \in \Theta\}$ of possible distributions for X. If $f_{\boldsymbol{\theta}}$ is the possibility function describing $\boldsymbol{\theta}$, then the credibility of events of the form $(\boldsymbol{\theta}, X) \in C \subseteq \Theta \times \mathsf{X}$ is

$$\bar{P}_{\boldsymbol{\theta}, X}(C) = \sup_{\theta \in \Theta} f_{\boldsymbol{\theta}}(\theta) \int \mathbf{1}_C(\theta, x) p_X(x \,|\, \theta) \mathrm{d}x. \qquad (1)$$

Black-box model Let X be a random variable in X with PDF p_X and, for a given realisation x of X, let $\boldsymbol{y}$ be a deterministic uncertain variable in Y defined based on x and described by a possibility function of the form $f_{\boldsymbol{y}}(\cdot \,|\, x)$. This case corresponds to Example 1, and indeed the associated OPM is of the form

$$\bar{P}_{X, \boldsymbol{y}}(C) = \int \left[\sup_{y \in \mathsf{Y}} \mathbf{1}_C(x, y) f_{\boldsymbol{y}}(y \,|\, x) \right] p_X(x) \mathrm{d}x,$$

for any $C \subseteq \mathsf{X} \times \mathsf{Y}$. Despite the dependence of $\boldsymbol{y}$ on X, the possibility function $f_{\boldsymbol{y}}(\cdot \,|\, x)$ might not depend on x if information about such dependence is unavailable.

	Probability	Possibility	Mixed
Variable	Random variable (e.g., X)	Deterministic uncertain variable ($\boldsymbol{\theta}$)	Uncertain variable ($\mathcal{X}$)
Uncertainty	Probability distribution (p)	Possibility function ($f_{\boldsymbol{\theta}}$)	OPM ($\bar{P}$)
Expected value	$\mathbb{E}[X]$	$\mathbb{E}^*[\boldsymbol{\theta}]$	-

Table 1: Summary of notational conventions.

More sophisticated forms of OPMs can be useful in practice, with one example being detailed in Section 5.2.

5 INFERENCE WITH OUTER PROBABILITY MEASURES

5.1 Bayes' rule with a possibilistic prior

We first focus on OPMs of the same form as (1). By analogy to Bayes'rule, we consider the possibility function describing $\boldsymbol{\theta}$ given that $X = x$, characterised by

$$f_{\boldsymbol{\theta}|X}(\theta \,|\, x) = \frac{p_X(x \,|\, \theta) f_{\boldsymbol{\theta}}(\theta)}{\sup_{\theta' \in \Theta} p_X(x \,|\, \theta') f_{\boldsymbol{\theta}}(\theta')}, \tag{2}$$

for any $\theta \in \Theta$, where $\theta \mapsto p_X(x \,|\, \theta) f_{\boldsymbol{\theta}}(\theta)$ is assumed to be bounded. Equation 2 comes from the following axiom: for any OPM $\bar{P}$, it holds that $\bar{P}(C \mid D) = \bar{P}(C \cap D)/\bar{P}(D)$ for any subsets C and D such that $\bar{P}(D) > 0$. Using this form of conditioning with $\bar{P} = \bar{P}_{\boldsymbol{\theta},X}$ as defined in (1), with $C = A \times \mathsf{X}$ for some subset A of Θ, and with $D = \Theta \times \{x\}$, and, for the sake of simplicity, considering the case where X is discrete, we obtain

$$\bar{P}_{\boldsymbol{\theta}|X}(A \,|\, x) = \frac{\bar{P}_{\boldsymbol{\theta},X}(A \times \{x\})}{\bar{P}_{\boldsymbol{\theta},X}(\Theta \times \{x\})} = \sup_{\theta \in A} f_{\boldsymbol{\theta}|X}(\theta \,|\, x),$$

so that $f_{\boldsymbol{\theta}|X}(\cdot \,|\, x)$ indeed characterises the information about $\boldsymbol{\theta}$ given $X = x$. We will refer to $f_{\boldsymbol{\theta}}$ and $f_{\boldsymbol{\theta}|X}(\cdot \,|\, x)$ as the prior and posterior respectively. Despite its similarities with the standard Bayes' rule, the version (2) has distinct properties. The main difference is the fact that the denominator, often referred to as the evidence, has turned from an integration problem to an optimisation problem. There are many instances in which the latter will be more tractable than the former, especially in high dimension. Another difference is that the evidence no longer measures the fitness of the model, so that model selection is less straightforward than in standard Bayesian inference. Instead, the possibilistic evidence quantifies the consistency between the prior and the likelihood, which could be useful in practice, e.g., to detect outliers or model misspecification. Once again, this behaviour should be expected when modelling information, as the consistency between information sources is a natural notion.

The properties of Bayesian inference that do not depend on the expression of the evidence generally hold for (2). For instance, possibilistic conjugate priors can be introduced by simply shifting the considered range of hyper-parameters to include the cases which are not integrable and exclude the ones which are unbounded. For instance, the Gamma possibility function can be defined as

$$\overline{\mathrm{Ga}}(\theta; \alpha, \beta) = \big(\beta\theta/\alpha\big)^{\alpha} \exp(\alpha - \beta\theta),$$

with $\alpha > 0$ and $\beta > 0$ or $\alpha = \beta = 0$, in which case $\overline{\mathrm{Ga}}(\theta; \alpha, \beta) = 1$ for all $\theta > 0$.

The posterior expected value matches with the usual maximum a posteriori (MAP) estimator, defined as

$$\hat{\theta}_{\mathrm{MAP}}(x) = \mathbb{E}^*[\boldsymbol{\theta} \,|\, X = x] = \arg\max_{\theta \in \Theta} f_{\boldsymbol{\theta}|X}(\theta \,|\, x),$$

which inherits the general properties of the expected value $\mathbb{E}^*[\cdot]$. In particular, the MAP for $T(\boldsymbol{\theta})$ will be $T\big(\hat{\theta}_{\mathrm{MAP}}(x)\big)$ for any mapping T on Θ, a property that does not hold in general for the standard MAP, as opposed to the MLE, despite being intuitively appealing when dealing with epistemic uncertainty.

5.2 The uninformative possibilistic prior

The assumptions on possibility functions allow for considering a constant function, regardless of the nature of Θ. As a possibility function its supremum must be equal to 1, such that it is constant and equal to 1 everywhere on Θ. We denote such a function as $\mathbf{1}_{\Theta}$, or simply by $\mathbf{1}$ when there is no ambiguity. Rewriting (2) with the prior $f_{\boldsymbol{\theta}}$ equal to $\mathbf{1}$ yields

$$f_{\boldsymbol{\theta}|X}(\theta \,|\, x) = \frac{p_X(x \,|\, \theta)}{\sup_{\theta' \in \Theta} p_X(x \,|\, \theta')},$$

which is simply a likelihood ratio. This connection between the proposed possibilistic version of Bayesian inference with frequentist inference can be pushed further: the possibility $\sup_{\theta \in A} f_{\boldsymbol{\theta}|X}(\theta \,|\, x)$ that $\boldsymbol{\theta}$ is in a subset A of Θ is related to the corresponding likelihood ratio test, the MAP estimator $\hat{\theta}_{\mathrm{MAP}}(x)$ is the MLE $\hat{\theta}_{\mathrm{MLE}}(x)$ and, denoting by $H_{\theta}(\cdot)$ the Hessian matrix in the variable θ, the posterior precision is

$$\begin{aligned}\mathbb{V}^*(\boldsymbol{\theta} \,|\, X = x)^{-1} &= \mathbb{E}^*\big(- H_{\theta}\big(\log p(x \,|\, \boldsymbol{\theta})\big) \,\big|\, X = x\big) \\ &= -H_{\theta}\big(\log p\big(x \,|\, \hat{\theta}_{\mathrm{MLE}}(x)\big)\big),\end{aligned}$$

which is the *observed information*, a finite sample version of the Fisher information. Despite these connections, there is a fundamental difference between frequentist inference and possibilistic Bayesian inference with $\mathbf{1}$ as a prior: In the latter, we do not target coverage properties based on i.i.d. replications of the observation and focus instead on directly characterising the posterior information as in standard Bayesian inference.

For a given inference problem, i.e., when the parameter set Θ is fixed, the possibility function $f_{\boldsymbol{\theta}} = \mathbf{1}$ can be seen as *the* uninformative prior. Indeed, it trivially satisfies a number of desirable properties:

- P.1 It is proper (as a possibility function) with no assumption on Θ.
- P.2 It satisfies the likelihood principle.
- P.3 It is invariant under reparametrisation: if T is a mapping from Θ to a set Ψ, then $\boldsymbol{\psi} = T(\boldsymbol{\theta})$ is described by

$$\begin{aligned} f_{\boldsymbol{\psi}}(\psi) &= \sup\left\{f_{\boldsymbol{\theta}}(\theta) : \exists \theta \in \Theta, \psi = T(\theta)\right\} \\ &= \mathbf{1}_{T(\Theta)}(\psi). \end{aligned}$$

  Indeed, the only information about $\boldsymbol{\psi}$ is that it must be in the image of T.

The absence of assumption on the mapping T in P.3 means that $\mathbf{1}$ also transforms coherently under partitioning of Θ since this is equivalent to assuming that T is surjective, with each element of Θ mapping to the element of the partition it belongs to. A lot of work has been devoted to finding uninformative probabilistic priors (Berger, 2006; Casella and Moreno, 2006; Consonni et al., 2018), with Jeffrey's priors (Jeffreys, 1946) being the most well-known; yet, they typically only verify a restricted version of P.3, where T is assumed to be bijective, at the expense of both P.1 and P.2. For instance, Jeffrey's priors are derived based on the likelihood, so that P.2 does not hold for them, and they can be improper even for simple inference problems such as when X is normally distributed with mean θ and known variance, so that P.1 does not hold in general either.

All the possibilistic conjugate prior families include $\mathbf{1}$ as a special case, e.g., $\overline{\mathrm{Ga}}(0,0) = \mathbf{1}$ or $\overline{\mathrm{N}}_d(\mu, \mathbf{0}_{d\times d}) = \mathbf{1}$ as previously noted. This means that, whenever conjugacy applies, one can consider the prior $\mathbf{1}$ and still remain within the conjugate prior family. In general, the availability of the uninformative prior $\mathbf{1}$ removes one of the difficulties with standard Bayesian inference in complex problems, where expressing priors for hyper-parameters with little physical meaning is often challenging. Similarly, since possibility functions are not densities, there is no technical difficulty with defining them, $\mathbf{1}$ included, on infinite-dimensional spaces.

The flexibility of the considered framework also allows for the prior to be a general OPM, with some components of the parameters being characterised by a probability distribution and other components being described by a possibility function, which could be the uninformative possibility function $\mathbf{1}$. This is illustrated in the following example which is based on a well-known difficulty of improper probabilistic priors.

Example 2 (Marginalisation paradox, from Dawid et al. (1973))**.** Consider independent random variables $X_1, \dots, X_n$ and M, such that M is a random integer in $\{1, \dots, n-1\}$ with distribution p_M, and X_i is distributed according to $\mathrm{Exp}(\theta)$, the exponential distribution with parameter θ, when $i \leq M$ and according to $\mathrm{Exp}(c\theta)$ when $i > M$, with $c \neq 1$ a known constant. The parameter $\theta > 0$ is unknown and, together with M, is the object of the inference problem. The likelihood for a realisation $x = (x_1, \dots, x_n)$ of $(X_1, \dots, X_n)$ is

$$p(x \,|\, \theta, m) = c^{n-m}\theta^n \exp\left(-\theta\left(\sum_{i=1}^{m} x_i + c\sum_{i=m+1}^{n} x_i\right)\right).$$

We model the unknown parameter θ as a deterministic uncertain variable $\boldsymbol{\theta}$, about which we know nothing. One possible prior OPM $\bar{P}$ is characterised by

$$\bar{P}_{M,\boldsymbol{\theta}}(\varphi) = \sum_{m=1}^{n-1} p_M(m) \sup_{\theta>0} \varphi(m, \theta),$$

for any bounded function φ on $\{1, \dots, n-1\} \times \Theta$. Another potential prior would have the sum and supremum swapped, however it is easier to learn about $\boldsymbol{\theta}$ for a given value of M rather than the other way around, so that the considered prior is more convenient. The corresponding posterior OPM describing $\boldsymbol{\theta}$ is characterised by

$$\bar{P}_{\boldsymbol{\theta}|X}(\varphi \,|\, x) = \sum_{m=1}^{n-1} p_M(m \,|\, x) \sup_{\theta>0} \varphi(\theta)\overline{\mathrm{Ga}}(\theta; n, \beta(m)),$$

for any bounded function φ on Θ, with $\beta(m) = \sum_{i=1}^{m} z_i + c\sum_{i=m+1}^{n} z_i$, where $z_i = x_i/x_1$ for any $i \in \{1, \dots, n\}$, and $p_M(m \,|\, x) \propto p_M(m) c^{-m} \beta(m)^{-n}$. The posterior distribution for M only depends on $z = (z_1, \dots, z_n)$ and, as pointed out by Dawid et al. (1973), the likelihood for z can in fact be expressed as a function of m only as $p(z \,|\, m) \propto c^{-m}\beta(m)^{-n}$. In the fully probabilistic case with a uniform improper prior on the random variable associated with θ, there is an inconsistency between the posterior obtained with the likelihood $p(z \,|\, m)$ and the one obtained with the original likelihood $p(x \,|\, \theta, m)$, yielding the claimed paradox; yet, in the proposed approach, no such inconsistency arises since $p(m \,|\, x)$ is proportional to $p(z \,|\, m)p(m)$.

Although this paradox can easily be avoided by considering the recommended improper prior distribution proportional to $1/\theta$, following such recommendations does not work in all the examples identified in Dawid et al. (1973). In the proposed approach, there is no need to select the uninformative prior based on considerations related to the likelihood, $\mathbf{1}$ is *the* uninformative prior.

Improper prior distributions can also be ill-behaved for large-dimensional inference problems where ensuring that the posterior is proper can be non-trivial. For instance, the improper log-uniform prior distribution proposed in Kingma et al. (2015) for variational dropout was shown to yield improper posterior distributions in Hron et al. (2017). Because of the lack of interpretability of parameters of deep neural networks and of the difficulties with improper priors, it is common to use arbitrary Gaussian priors in deep learning, despite their known adverse effects (de G. Matthews et al., 2018). In contrast, by expressing the lack of information via the proper possibility function $\mathbf{1}$, a proper posterior possibility function can be obtained whenever the likelihood is bounded.

5.3 Role of the likelihood

As can be seen in (2), the posterior will be a possibility function when the prior is, even if the likelihood is based on a probability distribution. We could also consider a likelihood that is based either on a possibility function of the form $f_{\boldsymbol{x}}(x \,|\, \theta)$ or even a more general conditional OPM. A possibilistic likelihood corresponds naturally to the case where the underlying objective is a loss function, as in energy-based model or generalised Bayesian inference. There are crucial differences between a possibilistic and probabilistic likelihood, in particular, it is not generally possible to learn scale parameters of a possibilistic likelihood; this is because the amount of uncertainty in a possibilistic likelihood is inherently subjective as opposed to the variability of a random process. For instance, if we consider the possibilistic likelihood $\theta \to \overline{\mathrm{N}}(x; 0, \theta)$ for a given observation x with the uninformative prior $\mathbf{1}$, then $\mathbb{E}^*[\boldsymbol{\theta} \,|\, x] = 0$, i.e., the MAP corresponds to the case where the likelihood is uninformative. In the context of energy-based models, this type of likelihood corresponds to the generalised perceptron loss which is known to produce "flat [...] energy surfaces if the architecture allows it" (LeCun et al., 2006, Section 2.2.2).

5.4 Relationship with IP

To better illustrate the relationship between IP and the proposed approach, we first describe a large class of problem where the two would be in agreement: If we consider a random variable X in a set X, then the most general parametrisation we can consider in the proposed framework is to define Θ as the set of probability distributions on X, in which case we have $P(B \,|\, \theta) = \theta(B)$, for any (measurable) subset B of X. Since possibility functions are not densities, there is no issue when considering a possibility function $f_{\boldsymbol{\theta}}$ describing the unknown probability distribution $\boldsymbol{\theta}$, regardless of the nature of X. We could then consider the possibility function $f_{\boldsymbol{\theta}}$ defined as the indicator of $\mathcal{M} = \{\theta \in \Theta : \mathbb{E}[X \,|\, \theta] \geq m\}$, where $\mathbb{E}[X \,|\, \theta]$ is the expected value of $X \sim \theta$ and where m is a scalar, referred to as a *lower prevision* in the IP literature. In general, since $\mathbb{E}[X \,|\, \boldsymbol{\theta}]$ is a deterministic uncertain variable, we can define the expected value of X as the set

$$\mathbb{E}^*[\mathbb{E}[X \,|\, \boldsymbol{\theta}]] = \big\{\mathbb{E}[X \,|\, \theta] : \theta \in \mathbb{E}^*[\boldsymbol{\theta}]\big\},$$

which matches with the standard plug-in estimate when $\mathbb{E}^*[\boldsymbol{\theta}]$ is a singleton. In the considered context, $\mathbb{E}^*[\boldsymbol{\theta}] = \mathcal{M}$ so that $\mathbb{E}^*[\mathbb{E}[X \,|\, \boldsymbol{\theta}]] = [m, \infty)$, hence agreeing with IP.

IP extends the construction above by considering gambles as real-valued functions of uncertain outcomes and the corresponding lower (and upper) previsions. To model gambles, we could consider an uncertain variable $\mathcal{X}$ rather than the random variable X. Yet, in this case, the structure of $\mathcal{X}$ needs to be made explicit before writing its expected value since deterministic outcomes could depend on realisations of random variables, e.g., the next move of a poker player after a new card is revealed, hence inducing a form of causality. By working directly with lower previsions, IP bypasses the need for identifying this type of structure, allowing for a high-level assessment of the desirability of gambles whereas our framework takes a fully model-based approach.

This highlights that IP and the considered framework take different paths, the former focusing on the value and the desirability of given gambles, and the latter explicitly modelling all the sources of uncertainty. Both approaches have their own merit, as is usual when contrasting methodologies that are fully model-based against those that are partially so.

6 ASYMPTOTIC ANALYSIS

In this section, we present our main results on the asymptotic properties of inference with outer probability measures. We show that most of our results can be translated naturally from probability theory.

6.1 Bernstein-von Mises theorem

In the context of Bayesian inference, the Bernstein-von Mises (BvM) theorem describes the behaviour of the posterior distribution as the sample size increases. The most important implication of the probabilistic BvM theorem is: with enough data, the inference made using the posterior distribution is asymptotically correct from the perspective of a frequentist. Furthermore, the information contained in the prior is forgotten as the number of observations increases. In the result that follows, we provide a BvM theorem in the case where the prior, and hence the posterior, is a possibility function. This version of the BvM theorem carries over the same implications from the probabilistic settings. We consider the following assumptions:

- A.1 The parameter space Θ is a compact and convex subset of $\mathbb{R}^d$
- A.2 The parameter θ^* is an element of Θ and the MLE is consistent
- A.3 The prior possibility function $f_{\boldsymbol{\theta}}$ is continuous and positive in a neighbourhood of θ^*

These assumptions are (the analogues of) the standard ones for the BvM theorem, but could be weakened using standard techniques. Yet, the focus of this work is to demonstrate the viability of the considered inferential approach and a more in-depth analysis is kept for future work.

Theorem 1 (Bernstein-von Mises)**.** *Let $x_1, x_2, \dots$ be i.i.d. observations sampled from a distribution $p_X(\cdot \,|\, \theta^*)$, and let $p_n(\cdot \,|\, \theta)$ denote the distribution of $x_{1:n} = (x_1, \dots, x_n)$ for any parameter $\theta \in \Theta$. Under Assumptions A.1-A.3 and for large values of n, it holds that*

$$f_{\boldsymbol{\theta}}(\theta \,|\, x_{1:n}) \approx \overline{\mathrm{N}}_d\left(\theta; \theta^* + \frac{\Delta_n}{\sqrt{n}}, \mathcal{J}_n\right), \tag{3}$$

where $\mathcal{J}_n = -H_\theta\big(\log p_n\big(x_{1:n} \,\big|\, \hat{\theta}_{\mathrm{MLE}}(x_{1:n})\big)\big)$ is the observed information at the MLE, and where

$$\Delta_n = \sqrt{n}\mathcal{J}_n^{-1} \nabla_\theta \log p_n(x_{1:n} \,|\, \theta^*).$$

The proof of Theorem 1, which can be found in the supplementary material, borrows heavily from the proof of the standard BvM theorem; yet, some of the usual steps are particularly straightforward with a possibilistic prior, e.g., the MLE appears naturally in the denominator of (2) as soon as the prior is neglected away. To emphasise this, we show how the asymptotic moments in the possibilistic BvM can be readily obtained from a simple calculations: The log-posterior possibility function is characterised by $\log f_{\boldsymbol{\theta}}(\theta|x_{1:n}) \propto \sum_{i=1}^n \log p_X(x_i|\theta) + \log f_{\boldsymbol{\theta}}(\theta) + c$ for some constant c not depending on θ; from this, the posterior expected value, or the MAP estimator, is computed as $\hat{\theta}_{\mathrm{MAP}}(x_{1:n}) = \arg\max_{\theta\in\Theta} \big[\sum_{i=1}^n \log p_X(x_i|\theta) + \log f_{\boldsymbol{\theta}}(\theta)\big]$, which tends to the MLE as $n \to \infty$, as in the probabilistic case. However, for the precision matrix, we also have that

$$\bar{\mathcal{I}}(\boldsymbol{\theta}|x_{1:n}) = -\sum_{i=1}^n H_\theta\big(\log p_X\big(x_i \,\big|\, \hat{\theta}_{\mathrm{MAP}}(x_{1:n})\big)\big) - H_\theta\big(\log f_{\boldsymbol{\theta}}(\hat{\theta}_{\mathrm{MAP}}(x_{1:n}))\big),$$

with the first term on the right hand side, corresponding to the observed information, dominating as $n \to \infty$.

6.2 Law of large numbers

To allow for further asymptotic analysis and to support the considered definition of $\mathbb{E}^*[\cdot]$, we introduce a possibilistic version of the LLN in the following theorem. Related results have been derived by Marinacci (1999); Maccheroni and Marinacci (2005); De Cooman and Miranda (2008); Cozman (2010) and Terán (2014), but with different representations of uncertainty or different notions of expected value.

Theorem 2. *If $\boldsymbol{x}, \boldsymbol{x}_1, \boldsymbol{x}_2, \dots$ is a sequence of independent deterministic uncertain variables on $\mathbb{R}^d$ with possibility function $f_{\boldsymbol{x}}$ such that*

(i) $f_{\boldsymbol{x}}$ is continuous on $\mathbb{R}^d$,
(ii) $f_{\boldsymbol{x}}$ is a twice continuously differentiable function on an open neighbourhood of each point in $\mathbb{E}^*[\boldsymbol{x}]$,
(iii) $\lim_{\|x\|\to\infty} f_{\boldsymbol{x}}(x) = 0$,

then the possibility function $f_{\boldsymbol{s}_n}$ describing $\boldsymbol{s}_n = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i$ verifies

$$\lim_{n\to\infty} f_{\boldsymbol{s}_n} = \mathbf{1}_{\mathrm{Conv}(\mathbb{E}^*[\boldsymbol{x}])},$$

where the convergence is point-wise, and where $\mathrm{Conv}(S)$ is the convex hull of a set $S \subseteq \mathbb{R}^d$.

The proof of Theorem 2, which can be found the supplementary material, differs significantly from the one of the standard LLN and is one of the main contributions in this work.

Although the limiting possibility function in Theorem 2 is not equal to $\mathbf{1}_{\mathbb{E}^*[\boldsymbol{x}]}$ unless $\mathbb{E}^*[\boldsymbol{x}]$ is convex, this version of the LLN still motivates the considered notion of expected value: since $\boldsymbol{s}_n$ is an average, it makes sense that any convex combination of points in $\mathbb{E}^*[\boldsymbol{x}]$ also appear as possible in the limit. Theorem 2 can be interpreted as follows: if for each $i \in \mathbb{N}$, we have an independent source of information modelling the uncertainty about $\boldsymbol{x}_i$ by $f_{\boldsymbol{x}}$, then only a finite number of $\boldsymbol{x}_i$'s can take value outside of $\mathbb{E}^*[\boldsymbol{x}]$. Otherwise, the available information against values outside of $\mathbb{E}^*[\boldsymbol{x}]$

would compound and the possibility of the corresponding value of $\boldsymbol{s}_n$ would tend to 0. In the limit, $\boldsymbol{s}_n$ will not be influenced by these (finitely many) $\boldsymbol{x}_i$'s outside of $\mathbb{E}^*[\boldsymbol{x}]$, but only by the (infinitely many) elements within it, whose average value sweeps the whole convex hull of $\mathbb{E}^*[\boldsymbol{x}]$. Appendix C provides further illustrations on the intuition behind the possibilistic LLN and CLT.

6.3 Central limit theorem

The existence of the possibilistic LLN and the unusual limiting quantity that appears in it raise questions about the way the empirical average of deterministic uncertain variables converge to the associated expected value. We answer this question in the following theorem.

Theorem 3. *If $\boldsymbol{x}, \boldsymbol{x}_1, \boldsymbol{x}_2, \dots$ is a sequence of deterministic uncertain variables on $\mathbb{R}$ independently and identically described by a possibility function $f_{\boldsymbol{x}}$ verifying*

(i) $f_{\boldsymbol{x}}$ is strictly log-concave and
(ii) $f_{\boldsymbol{x}}$ is twice differentiable,

then $\mathbb{E}^*[\boldsymbol{x}]$ is a singleton, and the possibility function $f_{\boldsymbol{t}_n}$ describing $\boldsymbol{t}_n = n^{-1/2}\sum_{i=1}^n(\boldsymbol{x}_i - \mathbb{E}^*[\boldsymbol{x}])$ verifies

$$\lim_{n\to\infty} f_{\boldsymbol{t}_n} = \overline{\mathrm{N}}\big(0, \bar{\mathcal{I}}(\boldsymbol{x})\big),$$

where the convergence is uniform.

As is usual for this type of result, this possibilistic CLT can be used to find approximate characterisations of the uncertainty for averages over a large number n of elements. The proof of Theorem 3 can be found in the supplementary material. The use of the possibilistic Fisher information $\bar{\mathcal{I}}(\boldsymbol{x})$ in the statement of Theorem 3 helps to handle the case where $\bar{\mathcal{I}}(\boldsymbol{x}) = 0$, in which the variance is undefined. The limiting possibility function, $\overline{\mathrm{N}}(0, \bar{\mathcal{I}}(\boldsymbol{x}))$, motivates both the considered notion of variance/information and the definition of the normal possibility function.

Despite the differences between the possibilistic and probabilistic LLNs, the associated CLTs are very similar. We leverage this similarity and analyse the asymptotic behaviour of the possibilistic MAP, when the likelihood is a possibility function, by following the standard argument for MLEs. For this result, we need a notion of boundedness in possibility, which we define as follows.

Definition 2. Let $\boldsymbol{z}_1, \boldsymbol{z}_2, \dots$ be a sequence of deterministic uncertain variable in $\mathbb{R}$, let $f_{\boldsymbol{z}_{1:n}}$ be the possibility function describing $(\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$, and let $a_1, a_2, \dots$ be a sequence in $\mathbb{R}$, then $\boldsymbol{z}_n = \bar{O}_p(a_n)$ means that, for any $\epsilon > 0$, there exists $M \in (0, \infty)$ and $N \in \mathbb{N}$ such that, for all $n > N$,

$$\sup\big\{f_{\boldsymbol{z}_{1:n}}(z_1, \dots, z_n) : z_{1:n} \in \mathbb{R}^n, \big|z_n/a_n\big| > M\big\} < \epsilon.$$

Corollary 1. *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots$ be observations independently and identically described by a possibility function $f_{\boldsymbol{x}}(\cdot \,|\, \theta^*)$, and let $f_n(\cdot \,|\, \theta)$ be the possibility function describing $\boldsymbol{x}_{1:n} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ for any parameter $\theta \in \Theta$. Define the corresponding MLE as $\arg\max_\theta f_n(\boldsymbol{x}_{1:n} \,|\, \theta)$. Under Assumptions A.1-A.3 with $d = 1$, and additionally assuming that*

(i) $f_{\boldsymbol{x}}(\cdot \,|\, \theta)$ is strictly log-concave for any $\theta \in \Theta$
(ii) the map $(\theta, x) \mapsto f_{\boldsymbol{x}}(x \,|\, \theta)$ is twice-differentiable in x as well as thrice-differentiable in θ
(iii) $\partial_\theta^3 \log f_n(\boldsymbol{x}_{1:n} \,|\, \theta) = \bar{O}_p(n)$,

the possibility function $f_{\boldsymbol{t}_n}$ describing the deterministic uncertain variable $\boldsymbol{t}_n = \sqrt{n}(\theta_{\mathrm{MAP}}(\boldsymbol{x}_{1:n}) - \theta^*)$ verifies

$$\lim_{n\to\infty} f_{\boldsymbol{t}_n} \to \overline{\mathrm{N}}\big(0, \bar{\mathcal{I}}(\theta^*)^2\tau_{\mathrm{s}}\big),$$

with $\tau_{\mathrm{s}} = \bar{\mathcal{I}}(\partial_\theta \log f_{\boldsymbol{x}}(\boldsymbol{x} \,|\, \theta^*))$ the precision of the score.

This result shows that the asymptotic behaviour can be analysed, even when the likelihood is based on a possibility function. This possibility function could be itself defined as the exponential of a negative loss, with no randomness involved.

To better understand the precision in the limit appearing in Corollary 1, we consider the following additional assumptions: the observations are in $\mathbb{R}$, the score $x \mapsto \partial_\theta \log f_{\boldsymbol{x}}(x \,|\, \theta^*)$ is invertible, and both the score and its inverse are twice differentiable. In this situation, simple calculations lead to

$$\tau_{\mathrm{s}} = \bar{\mathcal{I}}(\boldsymbol{x})\big(\partial_x\partial_\theta \log f_{\boldsymbol{x}}(\boldsymbol{x} \,|\, \theta^*)\big)^{-2},$$

which highlights the interplay between derivatives of the log-likelihood in x and θ in the limiting precision found in Corollary 1.

7 CONCLUSION

By considering a particular version of possibility theory that is close to probability theory, and by introducing a general framework including both, we have shown that Bayesian inference can be extended to encompass both epistemic and aleatoric uncertainties. We have illustrated how the proposed approach can address long-standing issues with standard Bayesian inference in contexts with significant epistemic uncertainty while preserving key results such as the Bernstein-von Mises theorem. Future work will aim to leverage the scalability of possibility theory, as an optimisation-based inference framework, for complex problems in machine learning where decoupling epistemic and aleatoric uncertainties is key, such as deep neural networks in general and LLMs in particular.

Acknowledgements

The authors would like to thank the 4 anonymous reviewers who all gave constructive feedback and helped improved the clarity and quality of the work. JH is supported by the Singapore Ministry of Digital Development and Information under the AI Visiting Professorship Programme, award number AIVP-2024-004. NKC is supported by a City University of Hong Kong Start-up Grant, project number 7200809.

References

Augustin, T., Coolen, F. P., De Cooman, G., and Troffaes, M. C. (2014). *Introduction to imprecise probabilities*, volume 591. John Wiley & Sons.

Balch, M. S., Martin, R., and Ferson, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society A*, 475(2227):20180565.

Baudrit, C., Dubois, D., and Perrot, N. (2008). Representing parametric probabilistic models tainted with imprecision. *Fuzzy sets and systems*, 159(15):1913–1928.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385 – 402.

Caprio, M., Sale, Y., Hüllermeier, E., and Lee, I. (2023). A novel Bayes' theorem for upper probabilities. In *International Workshop on Epistemic Uncertainty in Artificial Intelligence*, pages 1–12. Springer.

Casella, G. and Moreno, E. (2006). Objective bayesian variable selection. *Journal of the American Statistical Association*, 101(473):157–167.

Chen, Z., Ristic, B., Houssineau, J., and Kim, D. Y. (2021). Observer control for bearings-only tracking using possibility functions. *Automatica*, 133:109888.

Consonni, G., Fouskakis, D., Liseo, B., and Ntzoufras, I. (2018). Prior Distributions for Objective Bayesian Analysis. *Bayesian Analysis*, 13(2):627 – 679.

Cozman, F. G. (2010). Concentration inequalities and laws of large numbers under epistemic and regular irrelevance. *International journal of approximate reasoning*, 51(9):1069–1084.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 35(2):189–213.

De Cooman, G. and Miranda, E. (2008). Weak and strong laws of large numbers for coherent lower previsions. *Journal of Statistical Planning and Inference*, 138(8):2409–2432.

de G. Matthews, A. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*.

Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232.

Houssineau, J. (2021). A linear algorithm for multitarget tracking in the context of possibility theory. *IEEE Transactions on Signal Processing*, 69:2740–2751.

Houssineau, J. and Bishop, A. N. (2018). Smoothing and filtering with a class of outer measures. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):845–866.

Hron, J., Matthews, A. G. d. G., and Ghahramani, Z. (2017). Variational Gaussian dropout is not Bayesian. *arXiv preprint arXiv:1711.02989*.

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

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Jøsang, A. (2016). *Subjective logic*, volume 3. Springer.

Kimchaiwong, C., Houssineau, J., and Johansen, A. M. (2024). Redesigning the ensemble kalman filter with a dedicated model of epistemic uncertainty.

Kingma, D. P., Salimans, T., and Welling, M. (2015). Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F., et al. (2006). A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Lee, H., Jeong, J., Park, S., and Shin, J. (2023). Guiding energy-based models via contrastive latent variables. In *The Eleventh International Conference on Learning Representations*.

Lienen, J. and Hüllermeier, E. (2021). Credal self-supervised learning. *Advances in Neural Information Processing Systems*, 34:14370–14382.

Maccheroni, F. and Marinacci, M. (2005). A strong law of large numbers for capacities. *The Annals of Probability*, 33(3):1171–1178.

Marinacci, M. (1999). Limit laws for non-additive probabilities and their frequentist interpretation. *Journal of Economic Theory*, 84(2):145–195.

Sensoy, M., Kaplan, L., and Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Terán, P. (2014). Law of large numbers for the possibilistic mean value. *Fuzzy Sets and Systems*, 245:116–124.

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman & Hall.

Wasserman, L. A. and Kadane, J. B. (1990). Bayes' theorem for Choquet capacities. *The Annals of Statistics*, pages 1328–1339.

Xu, X., Qin, Y., Mi, L., Wang, H., and Li, X. (2024). Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In *The Twelfth International Conference on Learning Representations*.

Yadkori, Y. A., Kuzborskij, I., György, A., and Szepesvári, C. (2024). To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*.

Yen, J. (1990). Generalizing the Dempster-Schafer theory to fuzzy sets. *IEEE Transactions on Systems, man, and Cybernetics*, 20(3):559–570.

Zadeh, L. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*.

Reproducibility Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] The assumptions are all listed in Section 6
   (b) Complete proofs of all theoretical results. [Yes] The proofs of all results are provided in the supplementary material.
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Decoupling epistemic and aleatoric uncertainties with possibility theory: Supplementary Material

A PROOF OF BERNSTEIN-VON MISES THEOREM

We present two proofs for the BvM theorem, the first one is simpler as it assumes that the parameter is a scalar, i.e., $d = 1$, whereas the second proof handles the general case.

Proof of Theorem 1 when $d = 1$. Suppose $\Theta \subset \mathbb{R}$. Let $\boldsymbol{\theta}$ be the variable that represents the uncertainty about the true parameter θ^*. Define the uncertain variable $\boldsymbol{\psi} = \sqrt{n}(\boldsymbol{\theta} - \theta^*)$ and let $L_n(.; x_{1:n})$ be the likelihood function based on the observations $x_1, \ldots, x_n$, $f_{\boldsymbol{\theta}}$ be the prior possibility function, $\zeta : \Theta \to \Theta$ be the mapping defined as $\zeta(\theta) = \sqrt{n}(\theta - \theta^*)$. By the change of variable formula, we have the following possibility function that describes $\boldsymbol{\psi}$:

$$\begin{aligned} f_{\boldsymbol{\psi}}(\psi|x_{1:n}) &= \sup_{\theta \in \zeta^{-1}[\varphi]} \frac{L_n(\theta; x_{1:n}) f_{\boldsymbol{\theta}}(\theta)}{\sup_{\xi \in \Theta} L_n(\xi; x_{1:n}) f_{\boldsymbol{\theta}}(\xi)} \\ &= \frac{L_n(\theta^* + \psi/\sqrt{n}; x_{1:n}) f_{\boldsymbol{\theta}}(\theta^* + \psi/\sqrt{n})}{\sup_{\varphi \in \Theta} L_n(\theta^* + \varphi/\sqrt{n}; x_{1:n}) f_{\boldsymbol{\theta}}(\theta^* + \varphi/\sqrt{n})} \\ &\approx \frac{L_n(\theta^* + \psi/\sqrt{n}; x_{1:n})}{\sup_{\varphi \in \Theta} L_n(\theta^* + \varphi/\sqrt{n}; x_{1:n})}. \end{aligned}$$

In the last step, we consider the approximation with large values of n in the argument of the prior possibility function. Note that $L_n(\theta; x_{1:n})$ achieves its maximum when $\theta = \hat{\theta}_{\mathrm{MLE}}(x_{1:n})$. Therefore, we have:

$$f_{\boldsymbol{\psi}}(\psi|x_{1:n}) \approx \frac{L_n(\theta^* + \varphi/\sqrt{n}; x_{1:n})}{L_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}); x_{1:n})} \tag{4}$$

To relate $\hat{\theta}_{\mathrm{MLE}}(x_{1:n})$ to the true parameter θ^*, we perform Taylor expansion on the first derivative of the log-likelihood $\ell_n(\theta) = \ln L_n(\theta; x_{1:n})$ at θ^* around $\hat{\theta}_{\mathrm{MLE}}(x_{1:n})$:

$$\begin{aligned} \partial_\theta \ell_n(\theta^*) &= \underbrace{\partial_\theta \ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))}_{0} + (\theta^* - \hat{\theta}_{\mathrm{MLE}}(x_{1:n})) \partial_\theta^2 \ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n})) + \frac{1}{2}(\theta^* - \hat{\theta}_{\mathrm{MLE}}(x_{1:n}))^2 \partial_\theta^3 \ell_n(\psi_n) \\ &= (\theta^* - \hat{\theta}_{\mathrm{MLE}}(x_{1:n})) \partial_\theta^2 \ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n})) + \frac{1}{2}(\theta^* - \hat{\theta}_{\mathrm{MLE}}(x_{1:n}))^2 \partial_\theta^3 \ell_n(\psi_n), \end{aligned}$$

where ψ_n lies in the interval formed by θ^* and $\hat{\theta}_{\mathrm{MLE}}(x_{1:n})$. From the above, we have:

$$\sqrt{n}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}) - \theta^*) = \frac{\frac{1}{\sqrt{n}} \partial_\theta \ell_n(\theta^*)}{-\frac{1}{n} \partial_\theta^2 \ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n})) + \frac{1}{2n}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}) - \theta^*) \partial_\theta^3 \ell_n(\psi_n)} \tag{5}$$

Since the MLE is consistent, we can approximate the second term of the denominator with a small value. Hence, we have:

$$\sqrt{n}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}) - \theta^*) \approx -\sqrt{n} \frac{\partial_\theta \ell_n(\theta^*)}{\partial_\theta^2 \ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))} = \sqrt{n} \frac{\partial_\theta \ell_n(\theta^*)}{\mathcal{J}_n} \doteq \Delta_n. \tag{6}$$

From (4), we have:

$$f_{\boldsymbol{\psi}}(\psi|x_{1:n}) \approx \frac{L_n(\theta^* + \varphi/\sqrt{n})}{L_n(\theta^* + \Delta_n/\sqrt{n})} = \exp\Big(\ell_n(\theta^* + \varphi/\sqrt{n}) - \ell_n(\theta^* + \Delta_n/\sqrt{n})\Big).$$

Performing Taylor expansion on $\ell_n(\theta)$ around $\hat{\theta}_{\mathrm{MLE}}(x_{1:n})$, we have:

$$\begin{aligned}\ell_n(\theta) &= \ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n})) + \frac{1}{2}(\theta - \hat{\theta}_{\mathrm{MLE}}(x_{1:n}))^2\partial_\theta^2\ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n})) + \frac{1}{6}(\theta - \hat{\theta}_{\mathrm{MLE}}(x_{1:n}))^3\partial_\theta^3\ell_n(\psi_n)\\ &= \ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n})) - \frac{\mathcal{J}_n}{2}(\theta - \hat{\theta}_{\mathrm{MLE}}(x_{1:n}))^2 + \frac{1}{6}(\theta - \hat{\theta}_{\mathrm{MLE}}(x_{1:n}))^3\partial_\theta^3\ell_n(\psi_n),\end{aligned}$$

where ψ_n lies in the interval formed by θ and θ^*. Use the above expansion for $\theta = \theta^* + \varphi/\sqrt{n}$ and $\theta = \theta^* + \Delta_n/\sqrt{n}$, neglecting the third order term that is of order $O(n^{-1/2})$, we obtain

$$\begin{aligned}&\ell_n(\theta^* + \varphi/\sqrt{n}) - \ell_n(\theta^* + \Delta_n/\sqrt{n})\\ &\quad\approx \frac{\partial_\theta^2\ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))}{2n}(\varphi^2 + 2\varphi\sqrt{n}(\theta^* - \hat{\theta}_{\mathrm{MLE}}(x_{1:n})) - \Delta_n^2 - 2\Delta_n\sqrt{n}(\theta^* - \hat{\theta}_{\mathrm{MLE}}(x_{1:n})))\\ &\quad= \frac{\partial_\theta^2\ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))}{2n}(\varphi^2 - 2\varphi\Delta_n - \Delta_n^2 + 2\Delta_n^2)\\ &\quad= \frac{\partial_\theta^2\ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))}{2n}(\varphi - \Delta_n)^2.\end{aligned}$$

Making use of the consistency of MLE once more, we have

$$\ell_n(\theta^* + \varphi/\sqrt{n}) - \ell_n(\theta^* + \Delta_n/\sqrt{n}) \approx -\frac{\mathcal{J}_n}{2n}(\varphi - \Delta_n)^2.$$

Therefore, $f_{\boldsymbol{\psi}}(\varphi|x_{1:n}) \approx \overline{\mathrm{N}}(\varphi; \Delta_n, \mathcal{J}_n/n)$. Using the relation $\boldsymbol{\psi} = \sqrt{n}(\boldsymbol{\theta} - \theta^*)$, we obtain the posterior that describes $\boldsymbol{\theta}$ as $f_{\boldsymbol{\theta}}(\theta|x_{1:n}) \approx \overline{\mathrm{N}}(\theta; \theta^* + \Delta_n/\sqrt{n}, \mathcal{J}_n)$ as desired. □

Proof of Theorem 1 when $d > 1$. Suppose that $\Theta \subset \mathbb{R}^d$ for $d \geq 2$. For all $\theta \in \Theta$, we denote θ_i, $1 \leq i \leq d$ as the i^{th} component of θ. Let the uncertain variable $\boldsymbol{\psi}$ be $\boldsymbol{\psi} = \sqrt{n}(\boldsymbol{\theta} - \theta^*)$. Similar to the univariate case, our aim is to prove that for large sample size n:

$$f_{\boldsymbol{\psi}}(\varphi|x_{1:n}) \approx \overline{\mathrm{N}}\left(\varphi; \Delta_n, \frac{\mathcal{J}_n}{n}\right),$$

where $\mathcal{J}_n$ denotes the Hessian matrix of the log-likelihood evaluated at the MLE based on the observations $x_1, \ldots, x_n$ and $\Delta_n \doteq \sqrt{n}\mathcal{J}_n^{-1}\nabla_\theta\ell_n(\theta^*)$. Denote $\mathbf{M}$ as the $d \times d \times d$ tensor defined as follows:

$$\mathbf{M}_i = \begin{bmatrix} \frac{\partial \mathbf{H}_{11}}{\partial\theta_i} & \frac{\partial \mathbf{H}_{12}}{\partial\theta_i} & \cdots & \frac{\partial \mathbf{H}_{1d}}{\partial\theta_i} \\ \frac{\partial \mathbf{H}_{21}}{\partial\theta_i} & \frac{\partial \mathbf{H}_{22}}{\partial\theta_i} & \cdots & \frac{\partial \mathbf{H}_{2d}}{\partial\theta_i} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{H}_{d1}}{\partial\theta_i} & \frac{\partial \mathbf{H}_{d2}}{\partial\theta_i} & \cdots & \frac{\partial \mathbf{H}_{dd}}{\partial\theta_i} \end{bmatrix}, \quad \forall 1 \leq i \leq d, \tag{7}$$

where $\mathbf{H}$, $\mathbf{H}_{ij}$ refers to the Hessian matrix of the log-likelihood function (based on the observations $x_1, \ldots, x_n$) and its (i, j)-entry, respectively. Then, we denote $\mathbf{v} = \hat{\theta}_{\mathrm{MLE}}(x_{1:n}) - \theta^*$. Using Taylor expansion on $\nabla_\theta\ell_n(\theta^*)$ around the MLE, we have:

$$\nabla_\theta\ell_n(\theta^*) = -\mathbf{H}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))\mathbf{v} + \frac{1}{2}\mathbf{M}(\psi_n)\mathbf{v}\mathbf{v}^\top,$$

where $\psi_n = \alpha\theta^* + (1 - \alpha)\hat{\theta}_{\mathrm{MLE}}(x_{1:n})$ for some $\alpha \in (0, 1)$. Making use of the consistency in MLE, we have:

$$\begin{aligned}\sqrt{n}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}) - \theta^*) &= -\sqrt{n}\mathbf{H}^{-1}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))\left(\nabla_\theta\ell_n(\theta^*) - \mathbf{M}(\psi_n)\mathbf{v}\mathbf{v}^\top\right)\\ &= -n\sqrt{n}\mathbf{H}^{-1}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))\left(\frac{1}{n}\nabla_\theta\ell_n(\theta^*) - \frac{1}{2n}\mathbf{M}(\psi_n)\mathbf{v}\mathbf{v}^\top\right).\end{aligned}$$

Denote $\mathbf{K} = \frac{1}{2n}\mathbf{M}(\psi_n)\mathbf{v}\mathbf{v}^\top$. Then, $\mathbf{K}$ is a vector where the i^{th} entry admits the following bound:

$$\begin{aligned}
|\mathbf{K}_i| &= \left| \frac{1}{2n} \sum_{j,k=1}^{d} \mathbf{M}(\psi_n)_{i,j,k}\mathbf{v}_j\mathbf{v}_k \right| \\
&\leq \frac{1}{2n} \left(\sum_{j,k=1}^{d} |\mathbf{M}(\psi_n)_{i,j,k}|^2 \right)^{1/2} \left(\sum_{j,k=1}^{d} |\mathbf{v}_j\mathbf{v}_k|^2 \right)^{1/2} \qquad \text{(Cauchy-Schwarz)}.
\end{aligned}$$

Using the Cauchy-Schwarz inequality for double sums (Lemma 1), we have $\left(\sum_{j,k=1}^{d} |\mathbf{v}_j\mathbf{v}_k|^2 \right)^{1/2} \leq \|\mathbf{v}\|_2^2\sqrt{d}$. Therefore, $|\mathbf{K}_i|$ admits the following upper bound:

$$|\mathbf{K}_i| \leq \frac{\sqrt{d}\|\mathbf{v}\|_2^2}{2n} \left(\sum_{j,k=1}^{d} |\mathbf{M}(\psi_n)_{i,j,k}|^2 \right)^{1/2} = \frac{\sqrt{d}\|\mathbf{v}\|_2^2 \cdot \|\mathbf{M}(\psi_n)_i\|_{\mathrm{Fr}}}{2n},$$

where $\|\mathbf{M}(\psi_n)_i\|_{\mathrm{Fr}}$ is the Frobenius norm of the i^{th} matrix in the tensor of third order derivatives. We have $\|\mathbf{M}(\psi_n)_i\|_{\mathrm{Fr}} \in O(n\sqrt{d})$. Additionally, due to consistency of the MLE, we have $\|\mathbf{v}\|_2 \xrightarrow{p} 0$. Therefore, we can approximate $|\mathbf{K}_i|$ with a small value when n is large. Hence, we have $\mathbf{K} \approx \vec{\mathbf{0}}$. As a result, we can approximate $\sqrt{n}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}) - \theta^*)$ as follows:

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}) - \theta^*) &\approx -\sqrt{n}\mathbf{H}^{-1}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))\nabla_\theta \ell_n(\theta^*) \\
&= \sqrt{n}\mathcal{J}_n^{-1}\nabla_\theta \ell_n(\theta^*) \qquad (\mathcal{J}_n = -\mathbf{H}^{-1}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))) \\
&\doteq \Delta_n.
\end{aligned}$$

In line with the proof strategy from the univariate case, denote $\mathbf{u} = \theta - \hat{\theta}_{\mathrm{MLE}}(x_{1:n})$, we have the following Taylor expansion for $\ell_n(\theta)$ around $\hat{\theta}_{\mathrm{MLE}}(x_{1:n})$:

$$\begin{aligned}
\ell_n(\theta) &= \ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n})) + \underbrace{\nabla_\theta \ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))^\top}_{=\vec{\mathbf{0}}}\mathbf{u} + \frac{1}{2}\mathbf{u}^\top\mathbf{H}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))\mathbf{u} + \frac{1}{6}\sum_{i,j,k=1}^{d}\mathbf{M}(\psi_n)_{i,j,k}\mathbf{u}_i\mathbf{u}_j\mathbf{u}_k \\
&= \ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n})) + \frac{1}{2}\mathbf{u}^\top\mathbf{H}(\hat{\theta}_{\mathrm{MLE}}(x_{1:n}))\mathbf{u} + \frac{1}{6}\sum_{i,j,k=1}^{d}\mathbf{M}(\psi_n)_{i,j,k}\mathbf{u}_i\mathbf{u}_j\mathbf{u}_k \qquad (8) \\
&= \ell_n(\hat{\theta}_{\mathrm{MLE}}(x_{1:n})) - \frac{1}{2}\mathbf{u}^\top\mathcal{J}_n\mathbf{u} + \frac{1}{6}\sum_{i,j,k=1}^{d}\mathbf{M}(\psi_n)_{i,j,k}\mathbf{u}_i\mathbf{u}_j\mathbf{u}_k,
\end{aligned}$$

where $\psi_n = \alpha\theta + (1-\alpha)\theta^*$ for some $\alpha \in (0,1)$. Next, we apply the expansion in equation 8 with $\theta = \theta^* + \varphi/\sqrt{n}$ and $\theta = \theta^* + \Delta_n/\sqrt{n}$. For both expansions, we have $\mathbf{u} \in O(n^{-1/2})$. Therefore, we have:

$$\begin{aligned}
\left| \sum_{i,j,k=1}^{d} \mathbf{M}(\psi_n)_{i,j,k}\mathbf{u}_i\mathbf{u}_j\mathbf{u}_k \right| &\leq \sum_{i,j,k=1}^{d} |\mathbf{M}(\psi_n)_{i,j,k}\mathbf{u}_i\mathbf{u}_j\mathbf{u}_k| \qquad \text{(Triangle inequality)} \\
&\leq \left(\underbrace{\sum_{i,j,k=1}^{d} |\mathbf{M}(\psi_n)_{i,j,k}|^2}_{O(d^3n^2)} \right)^{1/2} \left(\underbrace{\sum_{i,j,k=1}^{d} |\mathbf{u}_i\mathbf{u}_j\mathbf{u}_k|^2}_{O(d^3n^{-3})} \right)^{1/2} \qquad \text{(Cauchy-Schwarz)} \\
&\in O(d^3n^{-1/2}).
\end{aligned}$$

Hence, the final third-order terms for both expansions are in the order of at most $O(n^{-1/2})$, which are negligible

for large sample size n. As a result, we have the following approximation:

$$\begin{aligned}
&\ell_n(\theta^* + \varphi/\sqrt{n}) - \ell_n(\theta^* + \Delta_n/\sqrt{n}) \\
&\qquad \approx -\frac{1}{2}\left((\varphi/\sqrt{n} - \mathbf{v})^\top \mathcal{J}_n(\varphi/\sqrt{n} - \mathbf{v}) - (\Delta_n/\sqrt{n} - \mathbf{v})^\top \mathcal{J}_n(\Delta_n/\sqrt{n} - \mathbf{v})\right) \\
&\qquad = -\frac{1}{2}((\varphi - \Delta_n)/\sqrt{n})^\top \mathcal{J}_n((\varphi + \Delta_n)/\sqrt{n} - 2\mathbf{v}) \\
&\qquad = -\frac{1}{2}(\varphi - \Delta_n)^\top \frac{\mathcal{J}_n}{n}(\varphi + \Delta_n - \underbrace{2\sqrt{n}\mathbf{v}}_{2\Delta_n}) \\
&\qquad = -\frac{1}{2}(\varphi - \Delta_n)^\top \frac{\mathcal{J}_n}{n}(\varphi - \Delta_n).
\end{aligned}$$

Finally, we have:

$$\begin{aligned}
f_{\boldsymbol{\psi}}(\psi|x_{1:n}) &\approx \frac{L_n(\theta^* + \varphi/\sqrt{n})}{L_n(\theta^* + \Delta_n/\sqrt{n})} \\
&= \exp\left(\ell_n(\theta^* + \varphi/\sqrt{n}) - \ell_n(\theta^* + \Delta_n/\sqrt{n})\right) \\
&\approx \exp\left(-\frac{1}{2}(\varphi - \Delta_n)^\top \frac{\mathcal{J}_n}{n}(\varphi - \Delta_n)\right) \\
&= \overline{\mathrm{N}}\left(\varphi; \Delta_n, \frac{\mathcal{J}_n}{n}\right),
\end{aligned}$$

as desired. □

Remark. *Note that the proofs for both univariate and multivariate cases have no bearing on the likelihood function. Therefore, we have the freedom to define the likelihood either as a probability density function or a possibility function. This means that the BvM theorem (both univariate and multivariate cases) works for sequences of uncertain variables as well.*

For the sake of completeness, we provide the proof for Cauchy-Schwarz inequality for double sums of finite sequences below:

Lemma 1 (Cauchy-Schwarz for double sum)**.** *Let $\{a_i\}_{i=1}^d$ and $\{b_i\}_{i=1}^d$ be two finite sequences of real numbers where $d \in \mathbb{N}$. Then, we have:*

$$\sum_{i=1}^d a_i b_i + \sqrt{\sum_{i=1}^d a_i^2 \sum_{i=1}^d b_i^2} \geq \frac{2}{d} \sum_{i,j=1}^d a_i b_j. \tag{9}$$

Proof. Define two sequences $\{x_i\}_{i=1}^d, \{y_i\}_{i=1}^d$ as follows:

$$\begin{cases} x_i &= a_i \Big/ \sqrt{\sum_{i=1}^d a_i^2} \\ \\ y_i &= b_i \Big/ \sqrt{\sum_{i=1}^d b_i^2} \end{cases}, \qquad 1 \leq i \leq d.$$

Since $\{x_i\}_{i=1}^d, \{y_i\}_{i=1}^d$ are the normalized sequences of $\{a_i\}_{i=1}^d, \{b_i\}_{i=1}^d$, we have $\sum_{i=1}^d x_i^2 = \sum_{i=1}^d y_i^2 = 1$. Dividing both sides of equation 9 by $\sqrt{\sum_{i=1}^d a_i^2 \sum_{i=1}^d b_i^2}$, proving the original inequality is equivalent to proving the following inequality:

$$\begin{aligned}
&\sum_{i=1}^d x_i y_i + 1 \geq \frac{2}{d} \sum_{i,j=1}^d x_i y_j, \\
\text{or } &\sum_{i=1}^d (x_i + y_i)^2 \geq \frac{4}{d} \sum_{i,j=1}^d x_i y_j.
\end{aligned}$$

Using Cauchy-Schwarz inequality and the fact that $(x+y)^2 \geq 4xy$, we have:

$$\begin{aligned}\sum_{i=1}^{d}(x_i+y_i)^2 &= \frac{1}{d}\left(\sum_{i=1}^{d}1^2\right)\cdot\left(\sum_{i=1}^{d}(x_i+y_i)^2\right)\\ &\geq \frac{1}{d}\left(\sum_{i=1}^{d}(x_i+y_i)\right)^2\\ &\geq \frac{4}{d}\sum_{i=1}^{d}x_i\sum_{j=1}^{d}y_j = \frac{4}{d}\sum_{i,j=1}^{d}x_iy_j.\end{aligned}$$

Hence, we obtained the desired bound. □

B CENTRAL LIMIT THEOREM & LAW OF LARGE NUMBERS

Definition 3 (Convergence in OPM)**.** consider a sequence of uncertain variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ on some state space $\mathcal{X}$ described by the sequence of possibility functions $f_{\boldsymbol{x}}^{(1)}, f_{\boldsymbol{x}}^{(2)}, \ldots$. We say that this sequence *converge in outer probability measure* to an uncertain variable $\boldsymbol{x}$ described by $f_{\boldsymbol{x}}$, denoted $\boldsymbol{x_n} \to \boldsymbol{x}$, if:

$$\lim_{n\to\infty} f_{\boldsymbol{x}}^{(n)}(x) = f_{\boldsymbol{x}}(x), \quad \forall x \in \mathcal{X}. \tag{10}$$

Lemma 2 (Slutsky's Lemma)**.** *Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ and $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ be two sequences of uncertain variables on $\mathbb{R}$. If it holds that $f_{\boldsymbol{x}_n}$ is continuous for all $n \geq 1$, that $(f_{\boldsymbol{x}_n})_{n\geq 1}$ converges uniformly to $f_{\boldsymbol{x}}$, and that for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \bar{\mathbb{P}}\big(|\boldsymbol{y}_n - c| \geq \epsilon\big) = 0, \tag{11}$$

for some constant $c \neq 0$, then $\boldsymbol{x}_n/\boldsymbol{y}_n \to \boldsymbol{x}/\alpha$.

Proof. We consider the possibility $\bar{p}_n \doteq \bar{\mathbb{P}}(\boldsymbol{x}_n/\boldsymbol{y}_n = z)$ for some fixed z, and bound it above and below as follows

$$\bar{\mathbb{P}}(\boldsymbol{x}_n/\boldsymbol{y}_n = z, \boldsymbol{y}_n = c) \leq \bar{p}_n \leq \bar{\mathbb{P}}(\boldsymbol{x}_n/\boldsymbol{y}_n = z, |\boldsymbol{y} - c| \leq \epsilon) + \bar{\mathbb{P}}(\boldsymbol{x}_n/\boldsymbol{y}_n = z, |\boldsymbol{y}_n - c| > \epsilon).$$

The term second term in the upper bound is itself upper bounded by $\bar{\mathbb{P}}(|\boldsymbol{y}_n - c| > \epsilon)$, so it holds that

$$\begin{aligned}\bar{p}_n &\leq \bar{\mathbb{P}}(\boldsymbol{x}_n/\boldsymbol{y}_n = z, |\boldsymbol{y} - c| \leq \epsilon) + \bar{\mathbb{P}}(|\boldsymbol{y}_n - c| > \epsilon)\\ &\leq \bar{\mathbb{P}}(\boldsymbol{x}_n/(c+\epsilon) \leq z \leq \boldsymbol{x}_n/(c-\epsilon)) + \bar{\mathbb{P}}(|\boldsymbol{y}_n - c| > \epsilon)\end{aligned}$$

Taking the limit $n \to \infty$ and using the convergence of $\boldsymbol{y}_n$ to c, we find that

$$\begin{aligned}\lim_{n\to\infty}\bar{p}_n &\leq \lim_{n\to\infty}\bar{\mathbb{P}}(\boldsymbol{x}_n/(c+\epsilon) \leq z \leq \boldsymbol{x}_n/(c-\epsilon))\\ &= \lim_{n\to\infty}\sup\big\{f_{\boldsymbol{x}_n}(x) : x/(c+\epsilon) \leq z \leq x/(c-\epsilon)\big\}.\end{aligned}$$

Based on the fact that the convergence of $f_{\boldsymbol{x}_n}$ to $f_{\boldsymbol{x}}$ is uniform, we can swap the limit and the supremum to find that

$$\lim_{n\to\infty}\bar{p}_n \leq \sup\big\{f_{\boldsymbol{x}}(x) : x \in [z(c-\epsilon), z(c+\epsilon)]\big\}.$$

Since this statement is valid for any $\epsilon > 0$, it holds that

$$\lim_{n\to\infty}\bar{p}_n \leq f_{\boldsymbol{x}}(cz) = \bar{\mathbb{P}}(\boldsymbol{x}/c = z).$$

It remains to prove that $\lim_{n\to\infty}\bar{p}_n$ can also be lower bounded by the same quantity. For this purpose, we rewrite the lower bound stated earlier as

$$\bar{p}_n \geq \bar{\mathbb{P}}(\boldsymbol{x}_n = zc, \boldsymbol{y}_n = c) = \bar{\mathbb{P}}(\boldsymbol{y}_n = c \mid \boldsymbol{x}_n = zc)\bar{\mathbb{P}}(\boldsymbol{x}_n = zc).$$

However, it holds that $\lim_{n\to\infty}\bar{\mathbb{P}}(\boldsymbol{y}_n = c \mid \boldsymbol{x}_n = zc) = 1$, so that

$$\lim_{n\to\infty}\bar{p}_n \geq \lim_{n\to\infty}\bar{\mathbb{P}}(\boldsymbol{x}_n = zc).$$

We conclude that $\lim_{n\to\infty}\bar{\mathbb{P}}(\boldsymbol{x}_n/\boldsymbol{y}_n = z) = \bar{\mathbb{P}}(\boldsymbol{x}/c = z)$ as required. □

B.1 PROOF OF THEOREM 2 (LLN)

Definition 4 (Convex hull)**.** We denote by $\mathcal{C}_f = \mathrm{Conv}(\arg\max f)$ the convex hull of $\arg\max f$, and by $d_{\mathcal{C}_f}$ the distance to the set $\mathcal{C}_f$, i.e. the function

$$d_{\mathcal{C}_f}(x) = \inf_{y \in \mathcal{C}_f} \|x - y\|, \quad x \in \mathcal{X}.$$

Since $\mathcal{C}_f$ is convex, for any $x \in \mathcal{X}$, there exists a unique point $x^{\mathcal{C}_f} \in \mathcal{C}_f$ such that $d_{\mathcal{C}_f}(x) = \left\|x - x^{\mathcal{C}_f}\right\|$. We also recall, from the definition of $\boldsymbol{s}_n$, that

$$f_{\boldsymbol{s}_n}(y) = \sup \left\{ \prod_{i=1}^{n} f(x_i) \,\middle|\, n^{-1} \sum_{i=1}^{n} x_i = y \right\}, \quad n \in \mathbb{N}. \tag{12}$$

Theorem 2. We aim to prove that

$$\lim_{n\to\infty} f_{\boldsymbol{s}_n}(y) = \begin{cases} 1, & \text{if } y \in \mathcal{C}_f, \\ 0, & \text{otherwise.} \end{cases}$$

We will consider the two cases above separately.

Case 1 - $y \in \mathcal{C}_f$: The result being evident on $\arg\max f$, let $y \in \mathcal{C}_f \setminus \arg\max f$ be an arbitrary point on the convex hull $\mathcal{C}_f$ that does not belong to $\arg\max f$.

Using Carathéodory's theorem, y can be written as the convex combination of at most $d+1$ points of $\arg\max f$, i.e., there exists $2 \leq p \leq d+1$ such that

$$y = \sum_{i=1}^{p} c_i a_i,$$

where $c_i > 0$ and $a_i \in \arg\max f$, $1 \leq i \leq p$, with $\sum_1^p c_i = 1$.

For any $n \geq \max_i\{c_i^{-1}\} + 1$, we consider the sequence of points $(x_{n,i})_{i=1}^n \in \mathcal{X}^n$ defined as

$$x_{n,i} = \begin{cases} \dfrac{c_i n}{\lfloor c_i n \rfloor} a_i, & \text{if} \quad \displaystyle\sum_{j=1}^{p'-1} \lfloor c_j n \rfloor + 1 \leq i \leq \sum_{j=1}^{p'} \lfloor c_j n \rfloor, \quad 1 \leq p' \leq p-1, \\ \dfrac{c_p n}{n - \sum_{j=1}^{p-1} \lfloor c_j n \rfloor} a_p, & \text{if} \quad \displaystyle\sum_{j=1}^{p-1} \lfloor c_j n \rfloor + 1 \leq i \leq n, \end{cases}$$

and one can easily verify that $y = n^{-1} \sum_{i=1}^n x_{n,i}$. We can also write

$$\prod_{i=1}^{n} f(x_{n,i}) = \prod_{i=1}^{p-1} f\left(\frac{c_i n}{\lfloor c_i n \rfloor} a_i\right)^{\lfloor c_i n \rfloor} f\left(\frac{c_p n}{n - \sum_{j=1}^{p-1} \lfloor c_j n \rfloor} a_p\right)^{n - \sum_{j=1}^{p-1} \lfloor c_j n \rfloor} \tag{13a}$$

$$= \prod_{i=1}^{p} f\left(\left(1 + \frac{\alpha_{n,i}}{\beta_{n,i}}\right) a_i\right)^{\beta_{n,i}}, \tag{13b}$$

where

$$\alpha_{n,i} = c_i n - \lfloor c_i n \rfloor \in [0, 1) \quad \text{and} \quad \beta_{n,i} = \lfloor c_i n \rfloor \geq c_i n - 1, \quad 1 \leq i \leq p-1,$$

and

$$\alpha_{n,p} = \sum_{j=1}^{p-1} (\lfloor c_j n \rfloor - c_j n) \in (1-p, 0] \quad \text{and} \quad \beta_{n,p} = n - \sum_{j=1}^{p-1} \lfloor c_j n \rfloor \geq c_p n,$$

so that $\lim_n \alpha_{n,i}/\beta_{n,i} = 0$ for any $1 \leq i \leq p$.

For any $1 \leq i \leq p$, then, since f attains its supremum value 1 in a_i and f is $\mathcal{C}^2$ in some open neighbourhood of a_i, Taylor's theorem yields

$$f\left(\left(1 + \frac{\alpha_{n,i}}{\beta_{n,i}}\right) a_i\right) = 1 + \frac{1}{2} \frac{\alpha_{n,i}^2}{\beta_{n,i}^2} a_i^{\mathrm{t}} H_f(a_i) a_i + o\left(\frac{1}{2} \frac{\alpha_{n,i}^2}{\beta_{n,i}^2} \|a_i\|^2\right),$$

where $H_f(a_i)$ is the Hessian matrix of f in a_i. That is,

$$f\left(\left(1+\frac{\alpha_{n,i}}{\beta_{n,i}}\right)a_i\right)^{\beta_{n,i}} = \exp\left[\beta_{n,i}\log\left(1+\frac{1}{2}\frac{\alpha_{n,i}^2}{\beta_{n,i}^2}a_i^{\mathrm{t}}H_f(a_i)a_i + o\left(\frac{1}{2}\frac{\alpha_{n,i}^2}{\beta_{n,i}^2}\|a_i\|^2\right)\right)\right]$$
$$\sim_n \exp\left[\frac{1}{2}\frac{\alpha_{n,i}^2}{\beta_{n,i}}a_i^{\mathrm{t}}H_f(a_i)a_i + o\left(\frac{1}{2}\frac{\alpha_{n,i}^2}{\beta_{n,i}}\|a_i\|^2\right)\right],$$

so that $\lim_n f\left(\left(1+\frac{\alpha_{n,i}}{\beta_{n,i}}\right)a_i\right)^{\beta_{n,i}} = 1$.

From (13b) it holds that $\lim_n \prod_{i=1}^n f(x_{n,i}) = 1$, and from (12) if follows that $\lim_n f_{\boldsymbol{s}_n}(y) = 1$.

Case 2 - $y \notin \mathcal{C}_f$: Let $y \in \mathcal{X} \setminus \mathcal{C}_f$ be an arbitrary point outside the convex hull $\mathcal{C}_f$, and let us denote by $\delta = d_{\mathcal{C}_f}(y) > 0$ its distance to $\mathcal{C}_f$. We define the open set

$$B_0 = \{x \in \mathcal{X} \,|\, d_{\mathcal{C}_f}(x) < \delta/2\},$$

and the sequence of increasing closed sets $\{B_n\}_{n\in\mathbb{N}^*}$ as

$$B_n = \{x \in \mathcal{X} \,|\, \delta/2 \leq d_{\mathcal{C}_f}(x) \leq \delta(1+\sqrt{n})\}, \quad n \in \mathbb{N}^*.$$

We define $b_n = \sup_{x\in B_n} f(x)$ and $\bar{b}_n = \sup_{x\in\mathcal{X}\setminus(B_n\cup B_0)} f(x)$, $n \in \mathbb{N}$, and we note that

- Since f is bounded and continuous and the sets $\{B_n\}_{n\in\mathbb{N}^*}$ are closed, the supremums b_n are all reached and we have $0 \leq b_n < 1$ for $n \geq 1$. We define $b_* = \sup_{n\geq 1} b_n$ and we have $0 \leq b_* < 1$.
- Since $\lim_{\|x\|\to\infty} f(x) = 0$, $\lim_{n\to\infty} \bar{b}_n = 0$.

Recall from (12) that, for any $n \in \mathbb{N}$, we need to consider the sequences of points $x_{1:n}$ satisfying $n^{-1}\sum_{i=1}^n x_i = y$. We will focus first on the set

$$Y_n = \{x_{1:n} \in \mathcal{X}^n \,|\, x_1, \ldots, x_n \in B_n \cup B_0, n^{-1}\sum_{i=1}^n x_i = y\},$$

i.e., the admissible sequences whose points are all contained within $B_n \cup B_0$, and then on the set

$$\bar{Y}_n = \{x_{1:n} \in \mathcal{X}^n \,|\, n^{-1}\sum_{i=1}^n x_i = y\} \setminus Y_n,$$

i.e, those with a least one point in the remaining space $\mathcal{X} \setminus B_n \cup B_0$.

Denote by $\hat{n} = \min_{x_{1:n}\in Y_n} \sum_{i=1}^n \mathbf{1}_{B_n}(x_i)$ the minimum number of points in B_n across every sequence in Y_n, and consider a sequence $\hat{x}_{1:n} \in Y_n$ with $\hat{n}$ points in B_n, indexed from 1 to $\hat{n}$. Since $\hat{x}_i^{\mathcal{C}_f} \in \mathcal{C}_f$ for any $1 \leq i \leq n$ and $\mathcal{C}_f$ is convex, we have $\frac{1}{n}\sum_{i=1}^n \hat{x}_i^{\mathcal{C}_f} \in \mathcal{C}_f$. We may then write

$$\begin{aligned}
\delta &= d_{\mathcal{C}_f}(y) \\
&\leq \left\| y - \frac{1}{n}\sum_{i=1}^n \hat{x}_i^{\mathcal{C}_f} \right\| \\
&= \left\| \frac{1}{n}\sum_{i=1}^n \hat{x}_i - \frac{1}{n}\sum_{i=1}^n \hat{x}_i^{\mathcal{C}_f} \right\| \\
&\leq \frac{1}{n}\sum_{i=1}^{\hat{n}} \left\| \hat{x}_i - \hat{x}_i^{\mathcal{C}_f} \right\| + \frac{1}{n}\sum_{i=\hat{n}+1}^{n} \left\| \hat{x}_i - \hat{x}_i^{\mathcal{C}_f} \right\| \\
&\leq \frac{\hat{n}}{n}\delta(1+\sqrt{n}) + \frac{n-\hat{n}}{n}\delta/2.
\end{aligned}$$

It follows that $\frac{n}{1+2\sqrt{n}} \leq \hat{n}$, and thus $\lim_n \hat{n} = \infty$. However, since

$$\sup_{x_{1:n} \in Y_n} \prod_{i=1}^{n} f(x_i) \leq b_n^{\hat{n}} \leq b_*^{\hat{n}},$$

and $b_* < 1$, it follows that $\lim_n \sup_{x_{1:n} \in Y_n} \prod_{i=1}^{n} f(x_i) = 0$.

Any sequence in $\bar{Y}_n$ has at least one point in $\mathcal{X} \setminus (B_n \cup B_0)$, and thus

$$\sup_{x_{1:n} \in \bar{Y}_n} \prod_{i=1}^{n} f(x_i) \leq \bar{b}_n.$$

Since $\lim_n \bar{b}_n = 0$, it follows that $\lim_n \sup_{x_{1:n} \in \bar{Y}_n} \prod_{i=1}^{n} f(x_i) = 0$.

We can then write

$$\begin{aligned}\lim_{n\to\infty} f_{\boldsymbol{s}_n}(y) &= \lim_{n\to\infty} \sup_{x_{1:n} \in Y_n \cup \bar{Y}_n} \prod_{i=1}^{n} f(x_i) \\ &= \lim_{n\to\infty} \max\left\{ \sup_{x_{1:n} \in Y_n} \prod_{i=1}^{n} f(x_i), \sup_{x_{1:n} \in \bar{Y}_n} \prod_{i=1}^{n} f(x_i) \right\},\end{aligned}$$

and since for two convergent sequences $\{u_n\}_n$ and $\{v_n\}_n$ it holds that

$$\lim_{n\to\infty} \max\{u_n, v_n\} = \max\left\{ \lim_{n\to\infty} u_n, \lim_{n\to\infty} v_n \right\},$$

it follows that $\lim_{n\to\infty} f_{\boldsymbol{s}_n}(y) = 0$. □

B.2 PROOF OF THEOREM 3 (CLT)

Theorem 3. One of the basic properties of strictly log-concave functions is that they are maximised at a single point which we denote by μ. We assume without loss of generality that $\mu = 0$ and We write f instead of $f_{\boldsymbol{x}}$ for the sake of simplicity. To find the supremum of $\prod_{i=1}^{n} f(x_i)$ over the set of x_i's verifying $n^{-1/2} \sum_{i=1}^{n} x_i = x$, we first use Lagrange multipliers to find that

$$f'(x_i) f(x_j) = f'(x_j) f(x_i),$$

for any $i, j \in \{1, \ldots, n\}$ so that a solution is $x_i = n^{-1/2} x$. In order to show that this solution is local maximizer, we consider the bordered Hessian corresponding to our constrained optimisation problem, defined as

$$H = \begin{bmatrix} 0 & 1/\sqrt{n} & & \ldots & & 1/\sqrt{n} \\ 1/\sqrt{n} & a & b & \ldots & & b \\ & b & & & & \\ \vdots & \vdots & & \ddots & & \vdots \\ & & & & & b \\ 1/\sqrt{n} & b & & \ldots & b & a \end{bmatrix}, \tag{14}$$

where $a = f''(y) f(y)^{n-1}$ and $b = f'(y)^2 f(y)^{n-2}$ with $y = x/\sqrt{n}$. For the solution $x_i = y$, $i \in \{1, \ldots, n\}$, to be a local maximum, the sign of the principal minors $M_3, \ldots, M_n$ of H has to be alternating, starting with M_3 positive. Basic matrix manipulations for the determinant yield

$$M_k = -\frac{k-1}{n} (a - b)^{k-2}, \tag{15}$$

which is alternating in sign. For M_3 to be positive, it has to hold that

$$f''(y) f(y) < f'(y)^2.$$

This condition can be recognized as a necessary and sufficient condition for a function to be strictly log-concave. It also follows from the assumption of log-concavity that the condition $f'(x_i)f(x_j) = f'(x_j)f(x_i)$, which can be expressed as $(\log f(x_i))' = (\log f(x_j))'$ can only be satisfied at $x_i = x_j$ so that this solution is a global maximum. We therefore study the behaviour of the function $f(\frac{x}{\sqrt{n}})^n$ as $n \to \infty$ and obtain

$$f\left(\frac{x}{\sqrt{n}}\right)^n = \exp\left(f'(0)\sqrt{n}x + \frac{1}{2}\big(f''(0) - f'(0)^2\big)x^2 + O\big(n^{-1/2}\big)\right).$$

The result of theorem 3 follows easily by taking the limit and by noting that $f'(0) = 0$ and that $f''(0)$ is non-positive since f decreases in the neighbourhood of its arg max. □

B.3 PROOF OF COROLLARY 1 (ASYMPTOTIC NORMALITY OF MAP)

Corollary 1. As usual, denote $\boldsymbol{\theta}$ as the variable that describes the uncertainty in θ^*. Let $\boldsymbol{\psi} = \sqrt{n}(\boldsymbol{\theta} - \theta^*)$. By theorem 1 (BvM for uncertain variables), we have the approximation $f_{\boldsymbol{\psi}}(\varphi|\boldsymbol{x}_{1:n}) \approx \overline{\mathrm{N}}(\varphi; \Delta_n, \mathcal{J}_n/n)$ for the posterior possibility function of $\boldsymbol{\psi}$. Therefore:

$$\sqrt{n}(\theta_{\mathrm{MAP}}(\boldsymbol{x}_{1:n}) - \theta^*) = \sup_{\varphi \in \Theta} f_{\boldsymbol{\psi}}(\varphi|\boldsymbol{x}_{1:n}) \approx \sup_{\varphi \in \Theta} \overline{\mathrm{N}}(\varphi; \Delta_n, \mathcal{J}_n/n) = \Delta_n.$$

We make the dependency on the observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ explicit by writing $\Delta_n(\boldsymbol{x}_{1:n})$ and $\mathcal{J}_n(\boldsymbol{x}_{1:n})$ instead. From the above, we have:

$$\begin{aligned}
\sqrt{n}(\theta_{\mathrm{MAP}}(\boldsymbol{x}_{1:n}) - \theta^*) &\approx \Delta_n(\boldsymbol{x}_{1:n}) \\
&= \frac{n}{\mathcal{J}_n(\boldsymbol{x}_{1:n})} \times \frac{\partial_\theta \log f_n(\boldsymbol{x}_{1:n}|\theta^*)}{\sqrt{n}} \\
&= \underbrace{\frac{n}{\mathcal{J}_n(\boldsymbol{x}_{1:n})}}_{\mathrm{A}} \times \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial_\theta \log f_{\boldsymbol{x}}(\boldsymbol{x}_i|\theta^*)}_{\mathrm{B}}.
\end{aligned}$$

Now, the strategy is to analyse the convergence in OPM of both A and B then apply Slutsky's lemma 2. By LLN for uncertain variable and the consistency of the MLE, we have:

$$\frac{\mathcal{J}_n(\boldsymbol{x}_{1:n})}{n} = -\frac{1}{n} \sum_{i=1}^{n} \partial_\theta^2 \log f_{\boldsymbol{x}}(\boldsymbol{x}_i|\hat{\theta}_{\mathrm{MLE}}(\boldsymbol{x}_{1:n})) \to \mathbb{E}^*[-\partial_\theta^2 \log f_{\boldsymbol{x}}(\boldsymbol{x}|\theta^*)] = \bar{\mathcal{I}}(\theta^*). \tag{16}$$

Considering the convergence of B, the CLT for uncertain variable yields:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \partial_\theta \log f_{\boldsymbol{x}}(\boldsymbol{x}_i|\theta^*) \to \overline{\mathrm{N}}(t; 0, \bar{\mathcal{I}}(\partial_\theta \log f_{\boldsymbol{x}}(\boldsymbol{x}|\theta^*))) = \overline{\mathrm{N}}(t; 0, \tau_s). \tag{17}$$

Then, by Slutsky's lemma for uncertain variables, we have:

$$\sqrt{n}(\theta_{\mathrm{MAP}}(\boldsymbol{x}_{1:n}) - \theta^*) \to \overline{\mathrm{N}}(t; 0, \tau_s \bar{\mathcal{I}}(\theta^*)^2),$$

as desired. □

C Intuitions for Main Results

In this section, we provide further illustrations and motivating examples for the possibilistic LLN and CLT.

C.1 Law of Large Numbers (Theorem 2)

Suppose that an object moves in the 2-dimensional Euclidean place $\mathbb{R}^2$, and that a sensor located at the origin measures its distance with the object every second. We denote by $\boldsymbol{x}_i$ the unknown position of the object after i seconds. No information about the motion of the object is known, and we wish to estimate its average position $\boldsymbol{s}_n = n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i$ through the first n seconds of the experiment.

Suppose first that the posterior information about the object's state, following the sensor observation at every second, is the possibility function $f_{\boldsymbol{x}}(x) = \overline{\mathrm{N}}(||x||; r^*, \sigma^2)$. That is, at every second, there is some evidence against the object being anywhere *but* on the circle $C_{r^*}(0)$ with radius r^* centred on the origin. When the number of such observations grows to infinity, the possibilistic LLN tells us that the information about the average position of the object is $\mathbf{1}_{B_{r^*}(0)}$, the indicator function of the ball $B_{r^*}(0)$ with radius r^* centred on the origin. In other words, if we keep getting evidence against the object being outside of the circle $C_{r^*}(0)$, then in the limit case we know that the true average position must be somewhere on the corresponding ball $B_{r^*}(0)$, but *nothing else*.

Suppose instead that the posterior information about the object's state, following the sensor observation at every second, is the probability distribution $p(x) \propto \mathrm{N}(||x||; r^*, \sigma^2)$. That is, at every second, it is most likely that the object lies somewhere on the circle $C_{r^*}(0)$. Then, the sample average S_n of i.i.d. samples drawn from p will tend to 0 by symmetry of the problem, i.e., in the limit case, the averaged position of the object will be *almost surely* 0.

The probabilistic assessment is overly optimistic in the general case. Indeed, we can imagine a wealth of behaviour patterns that are fully consistent with the sensor measurements – that is, the object's true position is on $C_{r^*}(0)$ at every second – and yet the object's average position in the limit case is not the origin: the object could be motionless and stay somewhere on $C_{r^*}(0)$, or could move back on forth (in a deterministic or random manner) on a quadrant of the circle, etc. On the other hand, *any* such behaviour pattern will lead to an average position somewhere on $B_{r^*}(0)$ in the limit case, as posited by the possibilistic assessment.

C.2 Central Limit Theorem (Theorem 3)

Continuing with the example of Section C.1, the CLT tells us that, under assumptions, the possibility function describing $\boldsymbol{s}_n$ for large n can be approximated as $f_{\boldsymbol{s}_n}(s) \approx \overline{\mathrm{N}}(s; \mathbb{E}^*[\boldsymbol{x}], \sqrt{n}\bar{\mathcal{I}}(\boldsymbol{x}))$. We can consider a variant of the example used for the LLN where the state of the object is in $[0, \infty)$. In this case, $\mathbb{E}^*[\boldsymbol{x}]$ is a singleton, and the CLT can be applied with additional assumptions on $f_{\boldsymbol{x}}$ to obtain a normal approximation of $f_{\boldsymbol{s}_n}$ for large n. The assumption that the underlying possibility function is strictly log-concave, which ensures that $\mathbb{E}^*[\boldsymbol{x}]$ is a singleton, could potentially be relaxed. This would likely result in $\boldsymbol{s}_n$ being asymptotically described by

$$\exp\left(-\frac{1}{2}\bar{\mathcal{I}}(\boldsymbol{x})d\big(s, \mathbb{E}^*[\boldsymbol{x}]\big)\right),$$

with $d(x, S)$ the distance between the point x and the set S. Although speculative, this limiting distribution highlights the flexibility of possibility functions, obtained from the simpler requirement of having a supremum equal to 1.