
From Bayes’ Rule to Bayes Rules: Optimal Information Processing and Axiomatic Foundations Beyond Probability

Jeremie Houssineau¹

Badr-Eddine Chérif-Abdellatif^{*2}

¹Nanyang Technological University, Singapore

²CNRS, LPSM, Sorbonne Université, France

Abstract

This paper develops principled updating rules for possibilistic inference, where uncertainty about a fixed parameter is represented by a possibility function, the maxitive analogue of a probability distribution, and comparisons are made pointwise via a partial order. From two complementary foundations, an information-conservation viewpoint and an axiomatic viewpoint, we derive the same canonical update: the posterior is the prior-likelihood product followed by supremum normalisation. The two derivations agree for an arbitrary loss, differing only in where the learning-rate parameter enters. This parameter controls epistemic strength and is not identifiable from the normalising evidence alone, clarifying the role of analogous learning-rate parameters in generalised Bayesian updating.

1 INTRODUCTION

Bayes’ rule occupies a unique position in statistical inference: it is at once an operational framework for updating beliefs and a benchmark against which alternative update rules are judged. Yet both classical statistics and modern machine learning routinely depart from standard Bayesian updating. Classical examples include inverse-probability methods and alternative postdata constructions, while modern examples include loss-based “generalised Bayes” updates, tempered posteriors, and evidence-fusion schemes in which uncertainty is not naturally additive. These departures raise a basic question that is easy to state but conceptually difficult: *what exactly makes Bayes’ rule the “right” way to process information, and how does that answer depend on the mathematical representation of information?*

One way to revisit this question is to treat inference as

information processing. Starting from two inputs, i) prior information about an unknown but fixed quantity of interest and ii) information supplied by data through a likelihood or loss, one asks for a principled information processing rule (IPR) that outputs a postdata representation of the state of knowledge. In the probabilistic setting, such a perspective is often associated with the work of Zellner [1988] on the information conservation principle (ICP), where Bayes’ rule arises as the unique IPR that conserves a chosen scalar measure of information. In parallel, the framework of Bissiri et al. [2016] shows that, when data enter through losses rather than a data-generating mechanism (DGM), a Gibbs (exponentiated-loss) posterior is forced by coherence requirements on how losses should accumulate and how updating should behave under restriction.

In this work, we argue that these “foundational” stories are not exclusive to probability, but translate naturally when uncertainty about the parameter is represented by *possibility functions*. Possibility theory [Zadeh, 1978] is designed to model epistemic uncertainty about a fixed state of nature. A prior state of knowledge about $\theta \in \Theta$ is represented by a possibility function $\pi : \Theta \rightarrow [0, 1]$ with $\sup_{\theta \in \Theta} \pi(\theta) = 1$, where larger values indicate greater compatibility with the available information. Crucially, the set of possibility functions $\mathcal{F}(\Theta)$ on a set Θ is naturally ordered pointwise: one possibility function can directly be said to be (weakly) more informative than another without first compressing them into a single number such as the (cross-)entropy. This order structure has two immediate consequences for information processing: i) it becomes meaningful to distinguish *information loss* from *information creation* at the level of information objects themselves, and ii) combining independent informational inputs naturally corresponds to multiplication followed by supremum normalisation, yielding a maxitive analogue of the familiar prior-likelihood product.

Our central message is that, under two conceptually distinct viewpoints, the same possibilistic update rule emerges as

^{*}Corresponding author: badr-eddine.cherief-abdellatif@cnrs.fr

inevitable:

$$\pi(\theta | y) = \frac{\pi(\theta)L(\theta | y)}{\sup_{\vartheta \in \Theta} \pi(\vartheta)L(\vartheta | y)}, \quad (1)$$

with $L(\theta | y)$ a likelihood and $\pi(\theta)$ a prior possibility function. When the likelihood $L(\theta | y)$ is a genuine sampling distribution $p_\theta(y)$, this update corresponds to a sup-normalised prior-likelihood product. When the input is a generic loss $\ell(\theta, y)$, the same structure appears with $L(\theta | y) \propto \exp(-w\ell(\theta, y))$ for a learning-rate parameter $w \geq 0$. In this sense, Bayes’ rule is not a single formula but a member of a family of *Bayes’ rules* indexed by the underlying calculus of information (additive vs. maxitive) and by the nature of the data-to-parameter link (probabilistic likelihood vs. loss-derived evidence).

Contributions. The paper is organised around two derivations of the possibilistic Bayes’ rule, each mirroring a classical probabilistic counterpart but revealing structural differences that are specific to ordered, sup-normalised information objects.

- **View 1 (optimal information processing).** We reformulate information conservation without appealing to scalar cross-entropies. By introducing a pointwise information score and enforcing additivity of information under independent conjunction, we obtain a canonical “ $-\log$ ” information map on possibility values. This yields an information imbalance *function* (rather than a scalar), making information loss/creation genuine phenomena. Imposing conservation in the induced partial order identifies the sup-normalised prior-likelihood product as the unique lossless and non-creative IPR.
- **View 2 (axiomatic approach).** We propose a possibilistic analogue of the coherence axioms of Bissiri et al. [2016], phrased for updates that map a prior possibility function and a loss into a posterior possibility function. These axioms force a sup-normalised Gibbs form $\pi(\theta | y) \propto \pi(\theta) \exp(-w\ell(\theta, y))$, hence deriving the same structural rule without presupposing it.
- **Unification and interpretation.** We show that the two views yield the *same* one-parameter family of updates for an arbitrary loss, with the learning rate w exogenous to both derivations (an encoding choice in View 1, an axiom-underdetermined constant in View 2). This clarifies the role of w as an epistemic-strength calibration parameter rather than a data-identifiable quantity, and explains why evidence-based learning of w is ill-posed in the possibilistic setting. The log-loss, $\ell(\theta, y) = -\log p_\theta(y)$, recovers the more familiar (tempered) Bayesian update as a special case.

2 RELATED WORK

Beyond Bissiri et al. [2016], foundational perspectives on generalised Bayes include divergence-/scoring-rule motivations [Jewson et al., 2018, Knoblauch et al., 2022], quasi-posteriors based on generic criterion functions [Chernozhukov and Hong, 2003], and work on calibrating the loss scale/learning rate for misspecification-robustness [Grünwald and van Ommen, 2017, Syring and Martin, 2019, Lyddon et al., 2019]. In the information-processing tradition related to Zellner [1988], Bayes updating is also closely related to information-theoretic projection principles (maximum entropy / minimum cross-entropy and minimum discrimination information) and their axiomatic characterisations [Jaynes, 1957, Shore and Johnson, 1980, Csiszár, 1975].

Bayes-like updates in possibility theory have long been studied through *conditional possibility* and notions of independence/noninteraction, leading to several competing conditioning rules and systematic treatments [Hisdal, 1978, de Cooman, 1997], with Walley and de Cooman [1999] providing a behavioural/coherence perspective on such conditioning rules. A likelihood-based approach for possibility measures is developed by Dubois et al. [1997]. Our contribution differs in that we derive a *sup-normalised prior-likelihood product* as the unique information-processing rule compatible with (loss-based) updating and information conservation, rather than positing a conditioning rule *a priori*. Our approach aligns with recent imprecise-probabilistic perspectives in which sup-normalisation emerges as the appropriate calibration step when combining a likelihood with partial prior information [Martin, 2026]. More broadly, this perspective aligns with the emerging Imprecise Probabilistic Machine Learning (IPML) literature on learning and decision-making under imprecise information such as Caprio et al. [2024], Singh et al. [2024] or Chau et al. [2025]. A survey of possibility theory and its relation to probability is given by Dubois and Prade [2015, Ch. 5], on which we draw to justify combining the two: our practical focus is the mixed setting of a probabilistic likelihood $p_\theta(y)$ with a possibilistic prior, i.e. the data treated as random while beliefs about θ remain possibilistic.

Among imprecise-probability models, we work with possibility measures because their *pointwise* representation by a possibility function is what enables the partial order on $\mathcal{F}(\Theta)$ underlying View 1, where loss and creation become pointwise phenomena rather than a scalar summary. Set-function models, e.g., coherent lower previsions [Walley, 1991], the imprecise Dirichlet model [Bernard, 2005], p-boxes, probability intervals, or distortion models, also support generalised Bayes updating and can be ordered by credal-set inclusion, but they admit no comparably direct pointwise order, so an information-conservation analysis there would require a different notion of imbalance.

Most closely related is the companion paper of Singh et al. [2025] which develops a maxitive analogue of the Donsker–Varadhan variational formula and applies it to *possibilistic variational inference*. The two are complementary: Singh et al. [2025] is computational and objective-driven (variational characterisations and tractable candidate families), whereas the present paper is foundational, axiomatising the *update rule itself* through the information-processing and axiomatic viewpoints and explaining why the sup-normalised prior-likelihood product is canonical. The viewpoints connect directly: our no-loss/no-creation characterisation of the posterior (Section 4.3) coincides with their dual consistency bounds, and we recover their variational uniqueness result as an information-conservation statement. This programme has already reached deep learning: Ni et al. [2026] instantiate the possibilistic posterior on the probability simplex, learning a Dirichlet possibility function that approximates a supremum-projected posterior by minimising the maxitive divergence of Singh et al. [2025], and obtain second-order uncertainty predictors competitive with the state of the art at the cost of a single forward pass, a downstream application of the very update rule characterised here.

3 POSSIBILITY THEORY

Let θ^* be an unknown parameter in a set Θ , which we model by an *uncertain variable* θ , the analogue of a random variable in possibility theory. The prior uncertainty about θ (equiv. θ^*) is represented by a possibility function $\pi(\theta)$ in $\mathcal{F}(\Theta) = \{f : \Theta \rightarrow [0, 1] : \sup_{\theta \in \Theta} f(\theta) = 1\}$, which directly models the available information about θ . The possibility function $\pi(\theta)$ only *describes* the information about θ as there is no canonical epistemic uncertainty (as opposed to the probabilistic case where an underlying probability measure, often denoted $\mathbb{P}$, defines the *true* probability of any suitable event). We consider that information about θ is available through an experiment which yields an observation y , with corresponding likelihood $L(\theta | y)$. We will consider multiple cases for the likelihood, which can be defined as i) a probability distribution $p_\theta(y)$ in y , ii) a possibility function $f_\theta(y)$ in y , and iii) a loss $\ell(\theta, y)$, via $\exp(-\ell(\theta, y))$. The last case overlaps with the first two cases by setting $\ell(\theta, y) = -\log p_\theta(y)$ or $\ell(\theta, y) = -\log f_\theta(y)$. Conversely, if a loss $\ell(\theta, y)$ is specified first, it can be transformed into a possibility function whenever it is bounded below. Writing $m(y) \doteq \inf_{\theta \in \Theta} \ell(\theta, y)$, the relative evidence

$$h_y(\theta) \doteq \exp(-(\ell(\theta, y) - m(y))) \in (0, 1]$$

is a possibility function. This transformation is meaningful since both losses and possibility functions capture information, whereas transforming a loss into a probability distribution (when possible) does not guarantee that the obtained probability will have any meaning as a DGM. Unlike

a genuine sampling density, however, a loss carries no intrinsic scale: for any strength $w \geq 0$ the tempered evidence $h_y^w \in \mathcal{F}(\Theta)$ is an equally valid encoding of the same information, so w is a modelling input rather than a property of the loss. We make this strength explicit from the outset and return to its role below.

Priors, regularisers, and provenance. The same scaling question applies to the prior when prior information comes from a regulariser. When the prior is specified directly, or obtained as a relative likelihood $\pi(\theta) = p_0(\theta) / \sup_{\vartheta} p_0(\vartheta)$ from a genuine model [Walley and Moral, 1999, Denoeux, 2014], its scale is fixed by its provenance, and tempering it to π^β with $\beta \neq 1$ departs from that model. When the prior instead encodes a penalty or regulariser $R : \Theta \rightarrow [0, \infty)$ through $\pi(\theta) = \exp(-\beta(R(\theta) - \inf_{\vartheta} R(\vartheta)))$, the specificity of $\beta = 1$ vanishes and $\beta \geq 0$ is a free strength parameter, exactly as w is for the loss. Nothing in the formalism singles out the prior as the scale-fixed factor: prior and evidence enter (1) symmetrically as multiplicative possibilistic terms, so freezing the prior’s scale is a modelling stance. Unless otherwise stated, we assume that π is specified directly and defer further discussion on this to Appendix A.

Independence. For two uncertain variables θ and θ' represented jointly by a possibility function $f_{\theta, \theta'}(\theta, \theta')$ on $\Theta \times \Theta'$, independence is characterised by the existence of two possibility functions $f_\theta(\theta)$ and $f_{\theta'}(\theta')$ such that $f_{\theta, \theta'}(\theta, \theta') = f_\theta(\theta)f_{\theta'}(\theta')$ for all $(\theta, \theta') \in \Theta \times \Theta'$. Although this characterisation of independence resembles the probabilistic one, the fact that possibility functions are not induced by the uncertain variable they describe means that independence is a property of possibility functions rather than a property of uncertain variables. In a Bayesian context, the prior and likelihood can now be said to be independent when they are both possibility functions. This can improve conceptual transparency in epistemic-uncertainty and evidence-fusion settings, where “independence” is fundamentally about sources rather than randomness. This advantage comes with the responsibility to assess overlap between sources, otherwise beliefs could be unintentionally over-concentrated.

Non-identifiability of strength. Whenever an information source is a possibility function rather than a genuine sampling density (e.g., a possibilistic likelihood, the loss-based evidence h_y , or a regulariser prior), its strength is a free parameter, since possibility functions are closed under power. Indeed, $f^w \in \mathcal{F}(\Theta)$ for every $w \geq 0$ and every $f \in \mathcal{F}(\Theta)$. It is tempting to place a prior on the strengths and learn them from data. However, the consistency

$$\begin{aligned} c(y) &\doteq \sup_{\theta \in \Theta} \pi(\theta)^\beta h_y(\theta)^w \\ &= \exp\left(-\inf_{\theta \in \Theta} [\beta R(\theta) + w(\ell(\theta, y) - m(y))]\right), \end{aligned}$$

with $R \doteq -\log \pi \geq 0$, is non-increasing in both β and w , so maximising the possibilistic evidence merely drives each admissible strength to its smallest admissible value. The strengths are therefore not identifiable from the evidence alone: scaling reflects epistemic strength, not data-generating variability, and, in short, one cannot learn how well one knows something. This is illustrated in Appendix B. The phenomenon is familiar in generalised Bayesian inference [Bissiri et al., 2016] for a loss-based likelihood; in the possibilistic setting it arises intrinsically and symmetrically, for every factor whose provenance does not fix its scale. Another example of this indeterminacy arises in decentralised possibilistic fusion [Houssineau et al., 2026], where a possibility function f may be split into independent shares f^w and f^{1-w} and fused back to f exactly for any $w \in [0, 1]$. In this setting, one need not *choose* the discounting weight, just as here one cannot *learn* it.

Thus the rest of the paper treats the prior π , the evidence possibility h_y^w , and the posterior $\pi(\cdot | y)$ as elements of $\mathcal{F}(\Theta)$, all compared through the same pointwise order.

Existing results. The update rule (1) already shares a number of properties with its probabilistic counterpart such as a Bernstein–von Mises theorem [Hieu et al., 2025]. We do not focus on these properties in this work as they are mainly *consequential* rather than *characterising*, i.e., they start from (1) instead of leading to it.

4 VIEW 1: OPTIMAL INFORMATION PROCESSING

Zellner [1988] shows that the standard Bayesian posterior can be recovered from an IPR which transforms input information into output information. The argument relies on the introduction of a generic *postdata* distribution as one of the components in the output information (along with the evidence), which is then compared to the input information, based on the prior (or *antedata*) distribution and the likelihood. A detailed review of the argument by Zellner [1988] is provided in Appendix C for completeness.

4.1 OUR VIEW ON ZELLNER’S ANALYSIS

We begin by emphasizing a point of fundamental importance: in Zellner’s ICP, what is being conserved is not information itself, but rather a *numerical measure* of information, that is, an *amount of information*. Zellner [1988] himself frames the efficiency of an IPR in terms of output information measured in a suitably chosen metric, so that comparing input and output is metric-dependent by construction. Statements about information being “lost”, “created”, or “conserved” only become meaningful once this information is mapped to a scalar quantity through the choice of a metric. From this perspective, Zellner’s ICP can be interpreted as a

principle concerning the conservation of a specific numerical representation of information, rather than information itself as an abstract object.

A natural candidate for measuring information is Shannon entropy. However, Shannon entropy is an absolute quantity: it is not defined relationally between two informational objects. As such, it does not provide a direct mechanism for comparing input and output information arising from different sources. In particular, entropy alone does not induce a meaningful conservation principle linking these distinct objects. To overcome this difficulty, Zellner adopts the Shannon cross-entropy, which is a relative quantity defined with respect to a reference distribution. This choice renders the ICP mathematically well-posed and leads to a unique postdata distribution satisfying the conservation condition, namely the Bayesian posterior. Equivalently, Bayes’ rule minimises the scalar functional $\Delta(\text{IPR})$, though this variational form is a reformulation of the conservation property rather than a primitive principle. Importantly, however, this uniqueness is a consequence of the chosen information measure rather than an intrinsic property of information processing itself.

Within Zellner’s framework, the information loss $\Delta(\text{IPR})$ induced by an IPR can be rewritten as:

$$\begin{aligned} \Delta(\text{IPR}) &\doteq \text{Input info}(\text{IPR}) - \text{Output info}(\text{IPR}) \\ &= \text{KL}(q(\cdot | y) \| \pi_{\text{add}}(\cdot | y)) \geq 0, \end{aligned}$$

where $q(\cdot | y)$ is the postdata distribution produced by the IPR and π_{add} is the Bayesian posterior, the subscript “add” referring to the additive nature of the posterior probability. This result entails two immediate and somewhat paradoxical consequences, corresponding to the sign of $\Delta(\text{IPR})$ and to its case of equality. First, since $\Delta(\text{IPR}) \geq 0$, **no IPR can create extraneous information**: the output can never exceed the input, so even a rule that adds structure to the belief gains nothing in these terms. This already rules out Zellner’s own suggestion that some IPRs might raise output above input. Second, since $\Delta(\text{IPR}) = 0$ only when $q(\cdot | y) = \pi_{\text{add}}(\cdot | y)$, **every IPR other than Bayes’ rule strictly loses information**. While mathematically consistent within the adopted framework, this conflicts with common intuitions about informativeness: an IPR whose postdata belief is more concentrated around the true parameter value than the Bayesian posterior would usually be judged more informative, not less. The notion of information loss arising here is therefore tightly linked to coherence with the prior and likelihood, rather than to alternative notions of informativeness such as concentration, precision, or proximity to the truth.

We argue that these paradoxes are epistemological in nature. Comparing input and output information implicitly assumes that informational objects can be meaningfully ordered or compared. When information is represented by probability distributions, this assumption is problematic: the set

$\mathcal{P}(\Theta)$ of probability distributions over Θ has no canonical pointwise informativeness order analogous to the possibility-function order. Any comparison between distributions must therefore be achieved through an externally chosen functional, such as entropy or cross-entropy. Consequently, the ICP does not compare information objects directly, but only compares scalar summaries derived from them. Therefore, the conclusions drawn from such comparisons are tied to the choice of information measure and cannot be regarded as intrinsic properties of information processing.

Our claim is that possibility theory provides a more appropriate framework for comparing informational objects directly. The set $\mathcal{F}(\Theta)$ of possibility functions over Θ is naturally partially ordered by

$$f \preceq g \iff \forall \theta \in \Theta, f(\theta) \leq g(\theta).$$

This structure admits a greatest element (the constant function equal to one) which represents total ignorance, a notion that has no direct analogue in probability theory. Within this ordered setting, information can be compared without recourse to a scalar summary: one informational object can be directly said to contain more or less information than another.

In the following subsection, we show that when information is represented by possibility functions and the ICP is formulated in terms of the order $\preceq$, the possibilistic Bayes' rule emerges as the unique postdata belief that conserves information. Moreover, this framework allows for a clear identification of IPRs that genuinely induce information loss or information gain, without relying on arbitrary numerical measures.

4.2 POSSIBILISTIC BAYES' RULE AS OPTIMAL INFORMATION PROCESSING

We now reformulate statistical inference as an *information processing* problem within a possibilistic framework. As in the previous section, an *information processing rule* (IPR) transforms given *input information* into *output information*. The crucial difference lies in the representation of uncertainty about the unknown parameter: uncertainty about the parameter θ is represented by *possibility functions*, while the likelihood is loss-based, with a strength parameter $w \geq 0$, as identified in Section 3 as an exogenous input rather than a feature of the loss.

The goal here is threefold. First, we show that within this framework, both information loss and information creation can occur. Second, we show that Bayes' rule emerges as the *unique* IPR that neither loses nor creates information. Finally, we contrast this result with the probabilistic setting, where such distinctions were impossible by construction.

We consider a postdata possibility function $g(\theta | y)$ with no assumption that it must follow Bayes' rule. Instead, it is

entirely determined by the chosen IPR, whose optimality is to be characterised. In contrast with the probabilistic case, information is no longer summarised by scalar expectations, i.e. the cross-entropy. Instead, information is represented directly as a *function over the parameter space*.

Assumptions (Pointwise information). Let $S : (0, 1] \rightarrow [0, \infty)$ be a scalar *information score* on possibility values and define, for any positive function $f : \Theta \rightarrow (0, 1]$, the pointwise information score $\mathcal{I}[f] : \Theta \rightarrow [0, \infty)$ by

$$\mathcal{I}[f](\theta) = S(f(\theta)).$$

Assume:

A1.1 (*Order-consistency*) For any $0 < a < b \leq 1$, it holds that $S(a) > S(b)$.

A1.2 (*Normalisation*) $S(1) = 0$.

A1.3 (*Additivity under independent conjunction*) For any $a, b \in (0, 1]$, $S(ab) = S(a) + S(b)$.

A1.4 (*Regularity*) S is continuous on $(0, 1]$.

Assumptions A1.1–A1.4 are the pointwise counterparts of the standard axioms characterising Shannon self-information, and lead to the same logarithmic map.

Proposition 1 (Canonical pointwise information score). *Under Assumptions A1.1–A1.4, there exists a constant $k > 0$ such that for all $a \in (0, 1]$,*

$$S(a) = -k \log a.$$

Consequently, for any positive $f : \Theta \rightarrow (0, 1]$,

$$\mathcal{I}[f](\theta) = -k \log f(\theta), \quad \theta \in \Theta.$$

Proof. Assumption A1.3 implies that S satisfies Cauchy's functional equation on $(0, 1]$ under multiplication. By Assumption A1.4, the only solutions are of the form $S(a) = c \log a$ for some $c \in \mathbb{R}$. Assumptions A1.1 and A1.2 yield $c < 0$ and $S(1) = 0$, hence $S(a) = k(-\log a)$ with $k = -c > 0$. $\square$

Accordingly, we define $\mathcal{I}[f] : \Theta \rightarrow \mathbb{R}$ for any positive function $f : \Theta \rightarrow (0, 1]$ as the function $\mathcal{I}[f](\theta) = -\log f(\theta)$ for all $\theta \in \Theta$ (so lower possibility corresponds to higher information content), by considering the convention $k = 1$ in Proposition 1. This convention fixes the *unit* in which information is measured and applies identically to every possibility function; it does not fix the *scale* of any particular input. A possibility function built from a loss carries its own strength inside the object being measured, so its information content is a multiple w of the (shifted) loss even though $\mathcal{I}$ itself uses $k = 1$ (see below, and Appendix A). Information is thus no longer compressed into a single numerical quantity. Instead, it is preserved as a structured object that can be compared pointwise over Θ . When possibility functions

take the value 0, we interpret $-\log 0 = +\infty$ and work with extended-real information scores, or equivalently restrict the pointwise arguments to the positive support and recover zero posterior values by continuity.

Remark 1 (On factorisation and conjunction). The factorisation $f_{\theta, \theta'}(\theta, \theta') = f_{\theta}(\theta)f_{\theta'}(\theta')$ and the additivity assumption (A1.3) should not be read as claiming that the product is the only possible conjunction in possibility theory. Rather, they fix the *information calculus* under which we operate: we work with the product t-norm as the conjunctive rule for combining independent information sources, which is precisely the regime where a log-additive information map is natural. Under an alternative conjunction $\otimes$, the corresponding additivity requirement would be formulated in terms of $\otimes$, and the induced information map and update rule would generally differ, yielding a different member of the family of “Bayes’ rules”. The present derivation is therefore a characterisation *conditional* on adopting product conjunction for independent information. The induced information map and update rule change accordingly (the minimum, for instance, yields minitive conditioning $\pi(\theta | y) \propto \min(\pi(\theta), L(\theta | y))$) and the product is the case where the two views collapse into one rule. We leave the full characterisation of such an alternative information calculus to future work.

Phrasing Bayes’ rule as combining information. The operator $\mathcal{I}[\cdot]$ takes possibility functions as arguments, but neither $L(\theta | y)$ nor $c(y)$ in (1) is necessarily a possibility function. We therefore rephrase (1) so that all components are possibilistic. Following Section 3, let $\ell(\theta, y)$ be any loss with $m(y) \doteq \inf_{\vartheta \in \Theta} \ell(\vartheta, y)$ finite, and let the evidence possibility at strength $w \geq 0$ be

$$h_y^w(\theta) = \exp(-w(\ell(\theta, y) - m(y))) \in (0, 1].$$

Measured with $\mathcal{I}$, its information content is $\mathcal{I}[h_y^w](\theta) = w(\ell(\theta, y) - m(y))$: the loss is the information carried by the evidence, up to the scale w . Unit strength $w = 1$ is the natural default, being the choice under which the shifted loss coincides exactly with the information content in nats; any other w rescales that information without changing its nature, which is precisely the non-identifiable strength of Section 3. When $\ell(\theta, y) = -\log p_{\theta}(y)$, unit strength recovers the relative likelihood $h_y^1(\theta) = p_{\theta}(y) / \sup_{\vartheta} p_{\vartheta}(y)$ whenever the supremum is finite; when $\sup_{\vartheta} p_{\vartheta}(y) = \infty$, the loss-based definition remains meaningful as soon as $m(y)$ is finite (e.g., after restricting Θ or regularising ℓ).

Introducing an uncertain variable θ' described by h_y^w alongside θ described by π , and assuming h_y^w and π independent,

$$f(\theta | \theta' = \theta) = \frac{h_y^w(\theta) \pi(\theta)}{\sup_{\vartheta \in \Theta} h_y^w(\vartheta) \pi(\vartheta)} = \pi(\theta | y). \quad (2)$$

The normalising constant $c(y) \doteq \sup_{\vartheta \in \Theta} h_y^w(\vartheta) \pi(\vartheta)$ lies in $(0, 1]$ and is itself a possibility, so $\mathcal{I}[\cdot]$ is now defined

for every element of (2). It also makes clear that $c(y)$, the degree of consistency, is an output information absent from the postdata possibility function, as it is what is conditioned on. Henceforth we take $L(\theta | y) \doteq h_y^w(\theta)$, so that (1) coincides with (2) and carries the strength w . Although (2) has the same multiplicative shape as the usual Bayesian product identity, its semantics are different: $c(y)$ is not a marginal data-generating probability, but an information term associated with the compatibility of prior and evidence.

Information Conservation Principle (ICP). As before, we postulate that a good IPR should conserve information, i.e. “Input information = Output information”. An IPR satisfying this condition is said to be *optimal* or *100% informationally efficient*. In contrast with the probabilistic setting, this framework allows suboptimal IPRs to genuinely lose information or introduce extraneous information.

Information imbalance functional. To operationalise the ICP, we define the information imbalance induced by an IPR as the function:

$$\Delta(\text{IPR}) \doteq \underbrace{(\mathcal{I}[\pi] + \mathcal{I}[L(\cdot | y)])}_{\text{input information}} - \underbrace{(\mathcal{I}[g(\cdot | y)] + \mathcal{I}[c(y)])}_{\text{output information}}$$

that is, using $L(\cdot | y) = h_y^w$ from (2),

$$\begin{aligned} \Delta(\text{IPR})(\theta) &= -\log \pi(\theta) + w(\ell(\theta, y) - m(y)) \\ &\quad + \log g(\theta | y) + \log c(y), \end{aligned}$$

where $c(y)$ is constant in θ , so that $\mathcal{I}[c(y)] = -\log c(y)$. The strength enters only through the evidence term $\mathcal{I}[L] = w(\ell - m)$, while the global unit $k = 1$ in $\mathcal{I}$ is untouched. Unlike the probabilistic case, $\Delta(\text{IPR})$ is not a scalar but a function. Consequently, information loss and information creation are to be understood in terms of the pointwise (partial) order:

$$\Delta(\text{IPR}) \succeq 0 \iff \forall \theta \in \Theta, \Delta(\text{IPR})(\theta) \geq 0.$$

An IPR is said to conserve information exactly if it holds that $\Delta(\text{IPR}) = 0$.

Optimal information processing rule. For each fixed strength w , solving $\Delta(\text{IPR}) = 0$ pointwise yields a unique postdata possibility function, the strength- w update

$$\pi(\theta | y) = \frac{\exp(-w \ell(\theta, y)) \pi(\theta)}{\sup_{\vartheta \in \Theta} \exp(-w \ell(\vartheta, y)) \pi(\vartheta)},$$

which is (1) with $L(\cdot | y) = h_y^w$ and coincides with Bayes’ rule in possibilistic form. Conservation fixes the posterior once the loss scale is set but does not itself select w , which remains the exogenous strength of Section 3. Unit strength $w = 1$ recovers the sup-normalised prior-likelihood product that is the focus of this work.

Information creation and information loss. An IPR does not create extraneous information if and only if:

$$\Delta(\text{IPR}) \geq 0 \iff g(\theta | y) \geq \pi(\theta | y) \quad \forall \theta \in \Theta.$$

Similarly, an IPR does not lose information if and only if:

$$\Delta(\text{IPR}) \leq 0 \iff g(\theta | y) \leq \pi(\theta | y) \quad \forall \theta \in \Theta.$$

Thus, Bayes' rule at strength w is the only IPR that belongs simultaneously to both classes.

These conditions admit a natural interpretation in terms of possibility theory. If the information about an uncertain variable θ is described by a possibility function f on Θ , then it is also described by any possibility function g on Θ such that $g(\theta) \geq f(\theta)$ for all $\theta \in \Theta$. In this case, g is said to be *less informative* than f , as it rules out fewer values of θ . Accordingly, postdata possibility functions that dominate the Bayesian posterior everywhere do not introduce information beyond what is justified by the input, whereas postdata possibility functions that are everywhere dominated by the Bayesian posterior necessarily introduce additional information. These two classes do not, however, cover all of $\mathcal{F}(\Theta)$: since $\preceq$ is only partial, a generic postdata $g(\cdot | y)$ is *incomparable* to $\pi(\cdot | y)$, with $\Delta(\text{IPR})$ changing sign across Θ , creating information at some parameter values while losing it at others. The possibilistic setting therefore exhibits three qualitatively distinct suboptimal regimes, loss without creation ($g \succeq \pi(\cdot | y)$), creation without loss ($g \preceq \pi(\cdot | y)$), and mixed incomparability, with Bayes' rule the unique point of exact conservation. Far from a defect, this is the payoff of the pointwise order: it localises *where* a rule adds or discards information, which a scalar summary cannot express.

Summary. Within this possibilistic framework, at any fixed strength w (writing $\pi(\cdot | y)$ for the corresponding strength- w update): i) information loss and information creation are meaningful and observable phenomena that a rule may exhibit separately or, when its postdata belief is incomparable to $\pi(\cdot | y)$, together at different parameter values, ii) IPRs that upper-bound $\pi(\cdot | y)$ do not create information, iii) IPRs that lower-bound $\pi(\cdot | y)$ do not lose information, and iv) Bayes' rule at strength w is the unique IPR that conserves information exactly.

This stands in sharp contrast with the probabilistic framework where such distinctions are precluded by the scalar nature of the information measure.

4.3 CONNECTION TO MAXITIVE DONSKEK-VARADHAN DUALITY

To make the dependence of the imbalance $\Delta(\text{IPR})$ on the postdata possibility function $g(\cdot | y)$ more explicit, we write $\Delta_g(\theta)$ instead of $\Delta(\text{IPR})(\theta)$. At unit strength, the setting of the concurrent maxitive Donsker-Varadhan duality of

Singh et al. [2025], the imbalance and their consistency-bound integrand are two readings of the same function: writing $Z_{\max} \doteq \sup_{\vartheta \in \Theta} \pi(\vartheta) \exp(-\ell(\vartheta, y))$ for their maxitive marginal likelihood, so that $\log c(y) = m(y) + \log Z_{\max}$, it holds that

$$\Delta_g(\theta) = \log Z_{\max} - \left(-\ell(\theta, y) - \log \frac{g(\theta | y)}{\pi(\theta)} \right).$$

Their lower and upper consistency bounds are the infimum and supremum over θ of the bracketed integrand, so the correspondence holds at the level of solution sets: an IPR loses no information ($\Delta_g \preceq 0$, i.e. $g \preceq \pi(\cdot | y)$) precisely when its postdata belief maximises their lower bound, and creates no information ($\Delta_g \succeq 0$, i.e. $g \succeq \pi(\cdot | y)$) precisely when it minimises their upper bound. The two-sided slack is their max-relative entropy, $\sup_{\theta} \Delta_g = D_{\max}(g \| \pi(\cdot | y))$ and $-\inf_{\theta} \Delta_g = D_{\max}(\pi(\cdot | y) \| g)$, and their identification of the posterior as the unique common optimiser of the two bounds is the statement that Bayes' rule is the unique rule incurring neither loss nor creation. For a general strength, applying their formula to the rescaled loss $w\ell$ traces the one-parameter family of Theorem 1.

5 VIEW 2: AXIOMATIC APPROACH

In the probabilistic case, considering a function of the loss as a likelihood is a departure from the standard Bayesian paradigm whenever the loss is not based on the true DGM, hence the name *generalised* Bayesian inference. However, possibilistic inference does not require the likelihood to be the true DGM and any *information* about the relationship between an observation and the parameter can be encoded as a possibilistic likelihood whenever the loss is bounded below. For consistency, we follow Bissiri et al. [2016] and consider a loss $\ell(\theta, y)$ but highlight that a possibilistic likelihood $h_y(\theta) \propto \exp(-\ell(\theta, y))$ could be considered instead whenever $\inf_{\theta} \ell(\theta, y) > -\infty$.

In the following set of assumptions, we simply write $\ell(\theta)$ for the loss $\ell(\theta, y)$ at a data point y . Similarly, we write $\ell'(\theta)$ for the same loss evaluated at another data point y' , i.e., $\ell'(\theta) \doteq \ell(\theta, y')$.

Assumptions (Possibilistic coherence axioms). Let $\psi[\ell, \pi] : \Theta \rightarrow [0, 1]$ denote the posterior possibility function based on the loss $\ell : \Theta \rightarrow \mathbb{R}$ and the prior possibility function π . Assume:

A2.1 It holds that $\psi[\ell', \psi[\ell, \pi]] = \psi[\ell + \ell', \pi]$.

A2.2 For any $A \subseteq \Theta$ with $\sup_{\theta \in A} \pi(\theta) > 0$,

$$\frac{\psi[\ell, \pi](\theta)}{\sup_{\vartheta \in A} \psi[\ell, \pi](\vartheta)} = \psi[\ell, \pi_A](\theta), \quad \theta \in A,$$

where $\pi_A(\theta) = \mathbf{1}_A(\theta)\pi(\theta) / \sup_{\vartheta \in A} \pi(\vartheta)$.

A2.3 If $A \subseteq \Theta$ satisfies $\sup_{\theta \in A} \pi(\theta) > 0$, $\ell(\theta) > \ell'(\theta)$ for all $\theta \in A$, and $\ell(\theta) = \ell'(\theta)$ for all $\theta \notin A$, then

$$\sup_{\theta \in A} \psi[\ell, \pi](\theta) < \sup_{\theta \in A} \psi[\ell', \pi](\theta).$$

A2.4 If ℓ is constant, then $\psi[\ell, \pi] = \pi$.

A2.5 If $\tilde{\ell}(\theta) = \ell(\theta) + c$ for some constant $c \in \mathbb{R}$, then $\psi[\tilde{\ell}, \pi] = \psi[\ell, \pi]$.

The axioms respectively impose sequential coherence, restriction compatibility, monotonicity in loss, invariance to vacuous losses, and invariance to additive loss shifts. An axiom-by-axiom discussion is given in Appendix D. Although these axioms closely mirror the coherence logic of Bissiri et al. [2016], the point here is that this argument is not special to probability. It is not, however, a mechanical integral-to-supremum substitution. The axioms are the possibilistic counterparts of Bissiri’s coherence conditions, not a strengthening of them; the extra work lies in the proof. Because a possibility function is fixed by its pairwise odds only up to the supremum constraint, the probabilistic normalisation cannot simply be reused, and one must instead establish, from the same axioms, that maxitive restriction (A2.2) preserves pairwise odds, that binary possibility functions are recovered from those odds, and that the binary updates cohere on triples (Appendix E).

Theorem 1. *Under Assumptions A2.1–A2.5, and assuming that Θ contains at least three elements, then there exists $w > 0$ such that for every loss $\ell(\cdot, y)$ satisfying $0 < \sup_{\vartheta \in \Theta} \exp(-w\ell(\vartheta, y))\pi(\vartheta) < \infty$, and prior possibility function π , it holds that*

$$\psi[\ell, \pi](\theta) = \frac{\exp(-w\ell(\theta, y))\pi(\theta)}{\sup_{\vartheta \in \Theta} \exp(-w\ell(\vartheta, y))\pi(\vartheta)}.$$

The proof is deferred to Appendix E. The theorem leaves w undetermined since rescaling $\ell \mapsto a\ell$ reproduces the family with aw . This is consistent with its status as the exogenous strength of Section 3. When the prior itself is constructed from a regulariser, its strength is a second modelling input rather than a quantity determined by the formalism. In that case the relative prior-loss weight controls the posterior mode, while the overall strength controls the contraction of α -cuts. Neither is selected by possibilistic evidence alone. Details are given in Appendix A.

Convergence of the two views. The two derivations do not merely share a functional form; they yield the *same* one-parameter family of posteriors. For any loss $\ell(\theta, y)$ and any strength $w \geq 0$, View 1 encodes the loss as the evidence possibility $h_y^w(\theta) = \exp(-w(\ell(\theta, y) - m(y)))$ and, through information conservation, returns the strength- w update $\pi(\theta | y) \propto \pi(\theta) \exp(-w\ell(\theta, y))$ of (2), while View 2 derives the identical family from the coherence axioms in

Theorem 1. The agreement is structural rather than incidental: in both views the exponential form traces to the same fact, i.e., independent information accumulates additively on the log scale, which enters once as additivity of the pointwise information score (A1.3) and once as sequential coherence of updating (A2.1). Both are instances of Cauchy’s functional equation. Its monotone solution is the logarithm when it fixes the information score in View 1 and the exponential when it fixes the loss-to-possibility map in View 2. These are mutually inverse, which is precisely why the two derivations compose into a single rule.

What the two formalisms make explicit is that the strength w is exogenous to *both* derivations, though it enters differently. In View 1, conservation is transparent to scale: it fixes the posterior completely once w is chosen, but w is set upstream, in the encoding of the loss as h_y^w , and conservation itself never selects it. In View 2, w is the constant left undetermined by the coherence axioms. Neither principle pins it down, consistent with its status as the non-identifiable epistemic-strength parameter of Section 3 and Appendix A.

Continuous parameter spaces. In standard Bayesian inference, extending consistency results from discrete to continuous spaces often requires measure-theoretic tools such as Radon–Nikodym derivatives because individual points in a continuous space typically have probability mass zero. Since possibility functions are defined pointwise as degrees of possibility rather than probability densities, the pairwise consistency arguments used in the proofs extend directly to uncountably infinite Θ without measure-theoretic ambiguity. In addition, while the derivation relies on ratios $\pi(a)/\pi(b)$, which technically requires $\pi(b) > 0$, the resulting product form $\psi[\ell, \pi](\theta) \propto \exp(-w\ell(\theta))\pi(\theta)$ remains valid on the entire domain. If $\pi(\theta) = 0$ (prior impossibility), the posterior correctly vanishes.

To see the correspondence with the familiar Bayesian naming convention, consider the log-loss $\ell(\theta, y) = -\log p_\theta(y)$ (whenever $p_\theta(y) > 0$). Then Theorem 1 yields the tempered Bayes-like update

$$\pi(\theta | y) = \frac{p_\theta(y)^w \pi(\theta)}{\sup_{\vartheta \in \Theta} p_\vartheta(y)^w \pi(\vartheta)}.$$

For $w = 1$ this reduces exactly to the sup-normalised prior-likelihood product which is the focus of this work.

Remark 2 (Connection with Martin’s sup-normalised prior-likelihood baseline). Theorem 1 identifies a broad class of coherent possibilistic updating rules and shows that, under Assumptions A2.1–A2.5, the posterior possibility function must take a *sup-normalised product* form. This agrees with the observation of Martin [2022] (in the context of *validation* via outer consonant approximation) that supremum normalisation is the appropriate way to convert a non-negative prior-likelihood score into a posterior *possibility function*: unlike integral- or Choquet-type normal-

isations, dividing by the supremum enforces the defining constraint $\sup_{\theta \in \Theta} \pi(\theta | y) = 1$ for each fixed dataset y . Theorem 1 therefore extends this observation: it shows that sup-normalisation is not merely a convenient choice tied to a specific construction, but is forced by general coherence requirements on possibilistic updating.

6 DISCUSSION

This paper characterises a canonical Bayes-type update for possibilistic inference. Under both an information-conservation viewpoint and an axiomatic viewpoint, the posterior is the prior-likelihood product with supremum normalisation, and the two derivations yield the same one-parameter family for an arbitrary loss. The result also highlights that the learning rate is exogenous to both derivations: it governs epistemic strength and cannot, in general, be learned from the evidence alone.

Connection to imprecise probability. Although the posterior is a possibility function, genuine upper probabilities surface in prediction under a probabilistic likelihood p_θ : the posterior predictive

$$\bar{P}(B | y) = \sup_{\theta} \pi(\theta | y) \int \mathbf{1}_B(y') p_\theta(y') dy'$$

is an upper probability, whose use in decisions invokes standard imprecise-probability tools such as Choquet integrals or credal sets.

While our focus is conceptual, the update $\pi(\theta | y) \propto \pi(\theta) L(\theta | y)$ can be used directly as a postdata *compatibility score* for inference and prediction. Its α -cuts $\{\theta : \pi(\theta | y) \geq \alpha\}$ define nested regions that can be pushed through a forward map to obtain predictive possibility bands, and summarised either by level-set estimators (e.g., MAP/profiles) or by decision rules that are robust on cuts (optimising worst-case loss over a chosen α -cut). Computationally, working on the log scale turns the update into the addition of a prior penalty and a data-fit term, and sup-normalisation replaces intractable integrals by maximisation. Finally, the epistemic-strength parameter (e.g. w in a Gibbs form) is naturally selected by *calibration* rather than by marginal evidence, e.g., by enforcing coverage/validity constraints in the spirit of plausibility or inferential-model constructions [Martin, 2026], or via conformal-style calibration of prediction sets derived from $\pi(\cdot | y)$.

Information deletion. Another natural direction concerns the inverse problem of information processing. Whereas the present work asks how new information should be incorporated into an existing state of knowledge, one may equally ask how previously acquired information should be removed. In the probabilistic setting, this question has recently attracted considerable attention under the name

of Bayesian unlearning. Beyond its practical motivations, recent work on *optimal information deletion* by Montcho and Rue [2026] has shown that Bayesian unlearning itself admits a principled information-processing interpretation, providing a conceptual counterpart to the role played by Bayes' theorem for information acquisition.

This perspective suggests a natural extension of the present work. Rather than viewing possibilistic unlearning as merely an algorithmic question, it would be interesting to investigate whether it can itself be characterised from first principles. In particular, one may ask whether a canonical possibilistic unlearning rule can be derived from an optimal-information-deletion principle mirroring the information-conservation viewpoint developed in Section 4, or equivalently from an axiomatic characterisation paralleling the coherence arguments of Section 5. Such a result would provide a possibilistic counterpart to recent developments in probabilistic unlearning, while further testing the scope of the foundational principles developed here.

Future avenues. This paper does not aim to develop a full inferential or decision-making pipeline, nor to provide empirical comparisons; our contribution is to clarify the informational principles and coherence requirements that single out particular update rules within a broader family of “Bayes’ rules.” Important next steps include i) systematic calibration strategies for the epistemic-strength parameter (e.g. w) based on coverage/validity or prequential criteria, ii) the study of predictive set constructions obtained by propagating α -cuts through complex models, and iii) characterising how alternative conjunction operators (beyond the product t-norm) induce different update rules and different operational guarantees. These directions would connect the present foundations to concrete algorithms and benchmarks.

Acknowledgements

This research is supported by the Singapore Ministry of Digital Development and Information under the AI Visiting Professorship Programme (AIVP-2024-004). BECA acknowledges funding from the ANR grant project BACKUP ANR-23-CE40-0018-01.

References

- Jean-Marc Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2-3):123–150, 2005.
- Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103–1130, 2016.
- Michele Caprio, Maryam Sultana, Eleni G Elia, and Fabio

- Cuzzolin. Credal learning theory. *Advances in Neural Information Processing Systems*, 37:38665–38694, 2024.
- Siu Lun Chau, Michele Caprio, and Krikamol Muandet. Integral imprecise probability metrics. *Advances in Neural Information Processing Systems*, 2025.
- Victor Chernozhukov and Han Hong. An MCMC approach to classical estimation. *Journal of econometrics*, 115(2): 293–346, 2003.
- Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.
- Gert de Cooman. Possibility theory II: Conditional possibility. *International Journal Of General System*, 25(4): 325–351, 1997.
- Thierry Denoeux. Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7): 1535–1547, 2014.
- Didier Dubois and Henry Prade. Possibility theory and its applications: Where do we stand? In *Springer handbook of computational intelligence*, pages 31–60. Springer, 2015.
- Didier Dubois, Serafin Moral, and Henri Prade. A semantics for possibility theory based on likelihoods. *Journal of Mathematical analysis and applications*, 205(2):359–380, 1997.
- Peter Grünwald and Thijs van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069 – 1103, 2017.
- Nong Minh Hieu, Jeremie Houssineau, Neil K. Chada, and Emmanuel Delande. Decoupling epistemic and aleatoric uncertainties with possibility theory. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2899–2907. PMLR, 2025.
- Ellen Hisdal. Conditional possibilities independence and noninteraction. *Fuzzy Sets and Systems*, 1(4):283–297, 1978.
- Jeremie Houssineau, Han Cai, Murat Uney, and Emmanuel Delande. Decentralised possibilistic inference with applications to target tracking. *Signal Processing*, page 110628, 2026.
- Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.
- Jack Jewson, Jim Q Smith, and Chris Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.
- Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on Bayes’ rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.
- Simon P Lyddon, Chris C Holmes, and Stephen G Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.
- Ryan Martin. Valid and efficient imprecise-probabilistic inference with partial priors, II. General framework. *arXiv preprint arXiv:2211.14567*, 2022.
- Ryan Martin. Possibilistic inferential models: a review. *Journal of the American Statistical Association*, 121(553): 807–826, 2026.
- Hans Montcho and Håvard Rue. Optimal information deletion and bayes’ theorem. *arXiv preprint arXiv:2602.09061*, 2026.
- Yao Ni, Jeremie Houssineau, Yew Soon Ong, and Piotr Koniusz. Possibilistic predictive uncertainty for deep learning. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026.
- John Shore and Rodney Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory*, 26(1):26–37, 1980.
- Anurag Singh, Siu Lun Chau, Shahine Bouabid, and Krikamol Muandet. Domain generalisation via imprecise learning. In *International Conference on Machine Learning*, 2024.
- Jasraj Singh, Shelvia Wongso, Jeremie Houssineau, and Badr-Eddine Chérif-Abdellatif. Maxitive Donsker-Varadhan formulation for possibilistic variational inference. *arXiv preprint arXiv:2511.21223*, 2025.
- Nicholas Syring and Ryan Martin. Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486, 2019.
- Peter Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, 1991.
- Peter Walley and Gert de Cooman. Coherence of rules for defining conditional possibility. *International Journal of Approximate Reasoning*, 21(1):63–107, 1999.

Peter Walley and Serafin Moral. Upper probabilities based only on the likelihood function. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):831–847, 1999.

Lotfi Asker Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 1(1):3–28, 1978.

Arnold Zellner. Optimal information processing and Bayes' theorem. *The American Statistician*, 42(4):278–280, 1988.

From Bayes' Rule to Bayes Rules: Optimal Information Processing and Axiomatic Foundations Beyond Probability (Supplementary Material)

Jeremie Houssineau¹

Badr-Eddine Chérif-Abdellatif^{*2}

¹Nanyang Technological University , Singapore

²CNRS, LPSM, Sorbonne Université , France

A LOSSES, REGULARISERS, AND POSTERIOR CONCENTRATION

Suppose the prior encodes a regulariser $R : \Theta \rightarrow [0, \infty)$ (with $\inf_{\theta} R = 0$) and the evidence a loss, so that, writing $\tilde{\ell}(\theta) \doteq \ell(\theta, y) - m(y)$, Theorem 1 gives the sup-normalised posterior $\pi(\theta | y) \propto \exp(-[\beta R(\theta) + w\tilde{\ell}(\theta)])$ at strengths $\beta, w \geq 0$. Reparametrise by the relative weight $\lambda \doteq \beta/w$ and the overall strength $s \doteq w$, and let $Q_{\lambda} \doteq \lambda R + \tilde{\ell}$. Then

$$\pi(\theta | y) = \exp(-s [Q_{\lambda}(\theta) - \inf_{\vartheta} Q_{\lambda}(\vartheta)]) = [\pi_{\lambda}^{(1)}(\theta | y)]^s,$$

the s -th power of the unit-strength posterior determined by λ alone. This highlights two distinct degrees of freedom. The *location* of the posterior, in particular its mode $\arg \min_{\theta} Q_{\lambda}$, depends only on the ratio λ , since scaling by $s > 0$ does not move the $\arg \min$. This is the regime of penalised optimisation, where a loss and a regulariser are combined and only their relative weight is meaningful. The *concentration*, however, depends on s : for $\alpha \in (0, 1)$ the α -cut $\{\theta : \pi(\theta | y) \geq \alpha\} = \{\theta : Q_{\lambda}(\theta) - \inf_{\vartheta} Q_{\lambda} \leq -s^{-1} \log \alpha\}$ contracts as s increases. The possibilistic posterior therefore contains information that is not relevant for penalised optimisation: optimisation focuses on one aspect (λ), possibilistic updating on two (λ and s). A special case is $\beta = w \neq 1$, i.e. $\lambda = 1$ with $s \neq 1$: the penalised objective $R + \tilde{\ell}$ and its minimiser are unaffected, yet the posterior is tempered.

This also answers whether attaching a strength to each factor is the same as choosing a different constant k in the pointwise information score $\mathcal{I}[\cdot] = -k \log(\cdot)$ of View 1. The answer is positive for the inputs: measuring the prior with $k_{\pi} = \beta$ and the loss-evidence with $k_L = w$ reproduces the factors π^{β} and h_y^w . However, a *common* k appears as a multiplicative factor in every term of $\Delta(\text{IPR})$ and cancels in the conservation identity $\Delta(\text{IPR}) = 0$, which is why View 1 fixes the update without fixing any scale. *Per-factor* constants differ across that identity and do not cancel. The obtained differences are exactly the parameters (λ, s) above.

B EXAMPLE: SQUARED-LOSS REGRESSION UNDER MISSPECIFICATION

Let $D_n = \{(x_i, y_i)\}_{i=1}^n$ with fixed design $x_i \in \mathbb{R}$ and data generated from a misspecified model

$$y_i = f_0(x_i) + \varepsilon_i, \quad f_0(x) = \beta x + \gamma x^2, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2),$$

but we fit the linear working model $m_{\theta}(x) = \theta x$ using the squared loss

$$\ell_n(\theta) \doteq \sum_{i=1}^n (y_i - \theta x_i)^2.$$

Write the ‘‘Gibbs evidence’’ $L_w(\theta | D_n) = \exp(-w\ell_n(\theta))$, where $w > 0$ controls the epistemic strength of the data relative to the prior.

^{*}Corresponding author: badr-eddine.cherief-abdellatif@cnrs.fr

^{*}Corresponding author: badr-eddine.cherief-abdellatif@cnrs.fr

Generalised Bayes (integral normalisation). Take a Gaussian prior $p_0(\theta) \propto \exp\{-(\theta - \mu_0)^2/(2\tau^2)\}$. The generalised posterior is

$$p_w(\theta | D_n) \propto p_0(\theta) \exp\{-w\ell_n(\theta)\}.$$

Since $\ell_n(\theta)$ is quadratic in θ , $p_w(\theta | D_n)$ is Gaussian with precision

$$\Lambda_w = \frac{1}{\tau^2} + 2w \sum_{i=1}^n x_i^2, \quad v_w = \Lambda_w^{-1}, \quad m_w = v_w \left(\frac{\mu_0}{\tau^2} + 2w \sum_{i=1}^n x_i y_i \right).$$

Under misspecification, m_w still targets the best linear approximation to f_0 in the design-weighted L^2 sense, while the spread v_w depends on w and does not correspond to a correct noise variance unless w is calibrated.

Possibilistic Bayes (sup normalisation). Take the corresponding Gaussian-shaped *prior possibility* $\pi_0(\theta) = \exp\{-(\theta - \mu_0)^2/(2\tau^2)\}$. The possibilistic update is

$$\pi_w(\theta | D_n) = \frac{\pi_0(\theta) \exp\{-w\ell_n(\theta)\}}{\sup_{\vartheta} \pi_0(\vartheta) \exp\{-w\ell_n(\vartheta)\}} = \exp\left(-[Q_w(\theta) - \inf_{\vartheta} Q_w(\vartheta)]\right),$$

where $Q_w(\theta) \doteq (\theta - \mu_0)^2/(2\tau^2) + w\ell_n(\theta)$. Hence $\pi_w(\cdot | D_n)$ is a log-quadratic possibility function with the same mode $\hat{\theta}_w = \arg \min_{\theta} Q_w(\theta)$ as the generalised Bayes posterior, but it is normalised by a supremum rather than an integral.

Interpretable uncertainty via α -cuts. For any $\alpha \in (0, 1]$, the α -cut is the interval

$$\{\theta : \pi_w(\theta | D_n) \geq \alpha\} = \{\theta : Q_w(\theta) \leq Q_w(\hat{\theta}_w) - \log \alpha\} = \left[\hat{\theta}_w \pm \sqrt{\frac{-\log \alpha}{a_w}} \right],$$

where $a_w = \frac{1}{2} \frac{d^2}{d\theta^2} Q_w(\theta) = \frac{1}{2\tau^2} + w \sum_{i=1}^n x_i^2$ (constant in θ). These level-set regions are likelihood-ratio-like and avoid any appeal to a correct sampling likelihood.

Role of misspecification and calibration of w . In both approaches w controls concentration: larger w yields narrower $p_w(\theta | D_n)$ and tighter α -cuts. However, in the possibilistic update the normalising constant

$$c_w(D_n) \doteq \sup_{\theta} \pi_0(\theta) \exp\{-w\ell_n(\theta)\} = \exp\{-\inf_{\theta} Q_w(\theta)\}$$

is non-increasing in w , so “evidence maximisation” pushes w to the smallest admissible value. This highlights that w should be chosen by *calibration* (for example, selecting w so that predictive α -cuts achieve a desired coverage level on held-out data or via a conformal-style procedure), rather than by possibilistic evidence alone.

C BAYES’ RULE AS OPTIMAL INFORMATION PROCESSING (PROBABILITY)

Zellner [1988] formulates statistical inference as an *information processing* problem. An *information processing rule* (IPR) transforms given *input information* (prior distribution and likelihood) into *output information* (postdata distribution and evidence). Different IPRs may lose information or introduce extraneous information.

SETUP

Let Θ be a parameter space and let $\theta \in \Theta$ be an unknown parameter. Let y denote the observed data, with likelihood

$$L(\theta | y) = p_{\theta}(y),$$

and let $\pi(\theta)$ be the prior (antedata) density. An IPR outputs a postdata density $q(\theta | y)$ and the evidence; we do not assume a priori that $q(\cdot | y)$ is given by Bayes’ rule. The evidence is

$$p(y) = \int_{\Theta} \pi(\theta) p_{\theta}(y) d\theta.$$

INFORMATION MEASURES

Information is quantified via Shannon's cross-entropy, i.e. as expectations of minus log-densities with respect to the postdata pdf $q(\theta | y)$:

$$\begin{aligned}\mathcal{I}[\pi] &= \mathbb{E}_{\theta \sim q(\theta | y)} [-\log \pi(\theta)] , \\ \mathcal{I}[L(\cdot | y)] &= \mathbb{E}_{\theta \sim q(\theta | y)} [-\log L(\theta | y)] , \\ \mathcal{I}[q(\cdot | y)] &= \mathbb{E}_{\theta \sim q(\theta | y)} [-\log q(\theta | y)] , \\ \mathcal{I}[p(y)] &= -\log p(y) .\end{aligned}$$

Input information consists of $\mathcal{I}[\pi] + \mathcal{I}[L(\cdot | y)]$, and output information consists of $\mathcal{I}[q(\cdot | y)] + \mathcal{I}[p(y)]$.

INFORMATION CONSERVATION PRINCIPLE (ICP)

A good IPR should conserve information:

$$\text{Input information} = \text{Output information}.$$

An IPR satisfying this condition is said to be either *optimal* or *100% informationally efficient*. Loss of information or creation of extraneous information is undesirable.

CRITERION FUNCTIONAL

To operationalise the ICP, Zellner defines the *information loss* criterion

$$\begin{aligned}\Delta(\text{IPR}) &= \text{Input information(IPR)} - \text{Output information(IPR)} \\ &= (\mathcal{I}[\pi] + \mathcal{I}[L(\cdot | y)]) - (\mathcal{I}[q(\cdot | y)] + \mathcal{I}[p(y)]) \\ &= \mathbb{E}_{\theta \sim q(\cdot | y)} [-\log \pi(\theta) - \log L(\theta | y) + \log q(\theta | y) + \log p(y)] \\ &= \mathbb{E}_{\theta \sim q(\cdot | y)} [-\log L(\theta | y)] + \text{KL}(q(\cdot | y) \| \pi) + \log p(y) .\end{aligned}$$

This criterion measures the difference between input and output information. An optimal IPR is a zero of $\Delta(\cdot)$.

OPTIMAL INFORMATION PROCESSING RULE

The unique optimal IPR is

$$\pi_{\text{add}}(\theta | y) = \frac{\pi(\theta) L(\theta | y)}{\int \pi(\theta) L(\theta | y) d\theta} ,$$

which is exactly obtained via *Bayes' rule*.

Bayes' rule is derived as the *optimal information processing rule* for the chosen information measures. The analysis links Bayesian updating to entropy and relative entropy concepts, showing that Bayesian inference neither loses nor creates information relative to the specified inputs.

SKETCH OF THE PROOF

The proof is in fact very simple. We just need to notice that

$$\Delta(\text{IPR}) = \text{KL}(q(\cdot | y) \| \pi_{\text{add}}(\cdot | y)) ,$$

which is equal to 0 only when $q(\cdot | y) = \pi_{\text{add}}(\cdot | y)$, which is given by Bayes' rule.

STATISTICAL INFERENCE AS OPTIMISATION

When adopting an information processing perspective on statistical inference, we remark that the optimal IPR (Bayes' rule) can be defined as *the single information loss minimiser* (since no IPR creates information under Zellner's information measures). This provides the standard variational formulation of (additive) Bayesian inference:

$$\{\pi_{\text{add}}(\cdot | y)\} = \arg \min_{q(\cdot | y)} \left\{ \mathbb{E}_{\theta \sim q(\cdot | y)} [-\log L(\theta | y)] + \mathbb{E}_{\theta \sim q(\cdot | y)} \left[\log \frac{q(\theta | y)}{\pi(\theta)} \right] \right\},$$

which directly follows from rewriting the information loss quantity as:

$$\begin{aligned} \Delta(\text{IPR}) &= \text{Input information(IPR)} - \text{Output information(IPR)} \\ &= \mathbb{E}_{\theta \sim q(\cdot | y)} [-\log L(\theta | y)] + \text{KL}(q(\cdot | y) \parallel \pi) + \log p(y). \end{aligned}$$

Note that in the possibilistic framework considered in this work, the optimal IPR is no longer defined via a single optimisation problem but as the intersection of two complementary optimisation problems:

$$\{\pi(\cdot | y)\} = \arg \max_{g(\cdot | y) \in \mathcal{F}(\Theta)} \inf_{\theta \in \Theta} \left\{ -\ell(\theta, y) - \log \frac{g(\theta | y)}{\pi(\theta)} \right\} \cap \arg \min_{g(\cdot | y) \in \mathcal{F}(\Theta)} \sup_{\theta \in \Theta} \left\{ -\ell(\theta, y) - \log \frac{g(\theta | y)}{\pi(\theta)} \right\},$$

where each problem is solved on its own by an entire class of postdata possibility functions, those below and those above $\pi(\cdot | y)$ respectively, and only their intersection is a singleton.

D AXIOM AUDIT

We briefly discuss which coherence requirements are responsible for each structural feature of the update and what may fail if an axiom is relaxed.

A2.1 (*Sequential coherence*) This axiom enforces that sequential incorporation of two losses is equivalent to incorporating their sum. It is the key ingredient behind the additive structure in the loss scale, and hence the multiplicative structure after exponentiation. Without (A2.1), the update may become path-dependent (order effects) and the representation $\psi[\ell, \pi] \propto \pi \exp(-w\ell)$ need not hold.

A2.2 (*Locality under restriction*) This locality condition is the main driver of the “odds-separability” property used in the characterisation. It rules out updates whose action on θ depends on losses outside the restricted set (e.g., through global normalisations other than sup or non-local aggregation). If (A2.2) is dropped, one can construct coherent-but-nonlocal updates where the relative update between two points depends on the loss landscape elsewhere, and the exponential form may fail.

A2.3 (*Monotonicity in loss*) This ensures that larger loss leads to (weakly) smaller posterior possibility, yielding $w \geq 0$ in the representation. Without (A2.3), one could admit updates that perversely increase plausibility under worse loss, and the learning-rate sign (and even existence) is no longer controlled.

A2.4 (*Vacuous loss leaves beliefs unchanged*) This pins down the identity element of updating and prevents spurious changes when the loss carries no discriminatory information. Relaxing (A2.4) allows arbitrary distortions even when ℓ is constant in θ , breaking the intended interpretation of ℓ as information.

A2.5 (*Shift invariance of loss*) This guarantees that only relative losses matter: adding a constant to ℓ has no effect. It is responsible for the appearance of a normalising constant that depends on y only, and for the natural sup-normalisation of the posterior possibility function. Dropping (A2.5) permits updates that depend on arbitrary baselines of ℓ , undermining comparability across datasets and destroying the canonical normalisation.

E PROOF OF THEOREM 1

To avoid division by zero in odds ratios, we first prove the result for strictly positive priors. The extension to priors with zeros follows by applying the result on the positive support and assigning posterior possibility zero wherever $\pi(\theta) = 0$. We start with a simple lemma showing that applying Assumption A2.2 on a pair of points in Θ implies a form of ratio consistency.

Lemma 1. Under Assumption A2.2, for any $a, b \in \Theta$ such that $a \neq b$, and for any loss ℓ , it holds that

$$\frac{\psi[\ell, \pi](a)}{\psi[\ell, \pi](b)} = \frac{\psi[\ell, \pi_{\{a,b\}}](a)}{\psi[\ell, \pi_{\{a,b\}}](b)}.$$

Proof. Apply Assumption A2.2 with $A = \{a, b\}$. Dividing the two equalities for $\theta = a$ and $\theta = b$ cancels the common factor $\sup_{\vartheta \in A} \psi[\ell, \pi](\vartheta)$, from which we can conclude. $\square$

The next lemma shows how to recover a possibility function from odds in the binary case.

Lemma 2. Let $\Theta = \{0, 1\}$ and f be a possibility function and define $t = f(0)/f(1) \in (0, \infty)$, then

$$f(0) = t \wedge 1, \quad f(1) = t^{-1} \wedge 1.$$

Proof. If $t \geq 1$ then $f(0) \geq f(1)$ so $f(0) = 1$ and $f(1) = 1/t$, hence $f(0) = t \wedge 1$ and $f(1) = t^{-1} \wedge 1$. If $t \leq 1$ then $f(1) = 1$ and $f(0) = t$, giving the same formulas. $\square$

The next lemma shows how the binary odds update reduces to a two-argument map.

Lemma 3. Under Assumptions A2.1, A2.4, and A2.5, let $\Theta = \{a, b\}$, let $\Delta = \ell(a) - \ell(b)$ and let $t = \pi(a)/\pi(b)$. Then there exists a function $\Phi : \mathbb{R} \times (0, \infty) \rightarrow (0, \infty)$ such that

$$\frac{\psi[\ell, \pi](a)}{\psi[\ell, \pi](b)} = \Phi(\Delta, t),$$

and Φ satisfies $\Phi(0, t) = t$ and, for all $\Delta_1, \Delta_2 \in \mathbb{R}$ and all $t > 0$,

$$\Phi(\Delta_2, \Phi(\Delta_1, t)) = \Phi(\Delta_1 + \Delta_2, t).$$

Moreover, under Assumption A2.3 the map $\Delta \mapsto \Phi(\Delta, t)$ is strictly decreasing for a given t .

Proof. On $\Theta = \{a, b\}$, Assumption A2.5 implies the posterior ratio can only depend on the loss difference $\Delta = \ell(a) - \ell(b)$, not on common shifts. By Lemma 2, the prior is determined (up to normalisation) by $t = \pi(a)/\pi(b)$, so the posterior ratio is a function of (Δ, t) , which we define as $\Phi(\Delta, t)$. If $\Delta = 0$ then ℓ is constant, hence by Assumption A2.4, $\psi[\ell, \pi] = \pi$ and $\Phi(0, t) = \pi(a)/\pi(b) = t$. For coherence, we apply Assumption A2.1 on the 2-point space: updating first by a loss difference Δ_1 and then by Δ_2 yields the same posterior as updating once by $\Delta_1 + \Delta_2$. Indeed, Assumption A2.1 yields

$$\frac{\psi[\ell', \psi[\ell, \pi]](a)}{\psi[\ell', \psi[\ell, \pi]](b)} = \frac{\psi[\ell + \ell', \pi](a)}{\psi[\ell + \ell', \pi](b)} \iff \Phi\left(\Delta_2, \frac{\psi[\ell, \pi](a)}{\psi[\ell, \pi](b)}\right) = \Phi(\Delta_1 + \Delta_2, t),$$

which implies the stated associativity equation for Φ . Finally, Assumption A2.3 implies that increasing $\ell(a)$ relative to $\ell(b)$ lowers the posterior of a relative to b , hence $\Delta \mapsto \Phi(\Delta, t)$ is strictly decreasing. $\square$

For regions where the prior assigns total impossibility ($\pi(\theta) = 0$), the posterior is defined as 0 by taking the limit $t \rightarrow 0$ in the odds ratio formulation (Lemma 3), preserving the validity of the product form on the entire domain without violating the coherence axioms.

Lemma 4. Under Assumption A2.2 and considering the setup of Lemma 3, for $\Theta = \{0, 1, 2\}$, for all $\Delta_1, \Delta_2 \in \mathbb{R}$ and $t_1, t_2 > 0$,

$$\Phi(\Delta_1 + \Delta_2, t_1 t_2) = \Phi(\Delta_1, t_1) \Phi(\Delta_2, t_2).$$

Therefore, it holds that $\Phi(\Delta, t) = g(\Delta)t$ with $g(\Delta) \doteq \Phi(\Delta, 1)$.

Proof. Let $\Delta_{ab} \doteq \ell(a) - \ell(b)$ and $t_{ab} \doteq \pi(a)/\pi(b)$. Note that $\Delta_{02} = \Delta_{01} + \Delta_{12}$ and $t_{02} = t_{01} t_{12}$. By Lemma 1, for each pair (a, b) the posterior ratio on $\{0, 1, 2\}$ equals the posterior ratio in the restricted 2-point problem $\{a, b\}$, hence

$$\frac{\psi[\ell, \pi](a)}{\psi[\ell, \pi](b)} = \Phi(\Delta_{ab}, t_{ab}).$$

Since it holds that

$$\frac{\psi[\ell, \pi](0)}{\psi[\ell, \pi](2)} = \frac{\psi[\ell, \pi](0)}{\psi[\ell, \pi](1)} \cdot \frac{\psi[\ell, \pi](1)}{\psi[\ell, \pi](2)},$$

substituting the Φ -representations and using the identities for Δ_{ab} and t_{ab} yields

$$\Phi(\Delta_{01} + \Delta_{12}, t_{01}t_{12}) = \Phi(\Delta_{01}, t_{01})\Phi(\Delta_{12}, t_{12}).$$

Since $(\Delta_{01}, \Delta_{12}, t_{01}, t_{12})$ can be chosen arbitrarily by varying ℓ and π on the three points, this proves the first result of the lemma. Setting $t_2 = 1$ and using $\Phi(0, t) = t$ from Lemma 3, we obtain that

$$\Phi(\Delta, t) = \Phi(\Delta, 1)\Phi(0, t) = g(\Delta)t,$$

as required. $\square$

Lemma 5. *Under Assumptions A2.1–A2.5, there exists $w > 0$ such that $g(\Delta) = \exp(-w\Delta)$, so that $\Phi(\Delta, t) = \exp(-w\Delta)t$.*

Proof. From Lemma 4 and Lemma 3,

$$g(\Delta_1 + \Delta_2) = g(\Delta_1)g(\Delta_2),$$

with $g(0) = 1$. Therefore, $h(\Delta) \doteq \log g(\Delta)$ satisfies Cauchy's equation $h(\Delta_1 + \Delta_2) = h(\Delta_1) + h(\Delta_2)$. The monotonicity of g implies that $h(\Delta) = -w\Delta$ for some $w \in \mathbb{R}$. Since g is decreasing, $w > 0$. Therefore $g(\Delta) = \exp(-w\Delta)$ and $\Phi(\Delta, t) = g(\Delta)t = \exp(-w\Delta)t$ as claimed. $\square$

We are now in a position to prove the main result, extending the previous results to arbitrary sets Θ via pairwise odds, giving rise to a form of maxitive Gibbs posterior.

Proof of Theorem 1. Let $a, b \in \Theta$ such that $a \neq b$. By Lemma 1, the posterior ratio between a and b equals the posterior ratio in the restricted 2-point problem $\{a, b\}$. Hence, by Lemmas 3, 4 and 5,

$$\frac{\psi[\ell, \pi](a)}{\psi[\ell, \pi](b)} = \exp(-w(\ell(a) - \ell(b))) \frac{\pi(a)}{\pi(b)}.$$

Pick an anchor $\theta_0 \in \Theta$. Rearranging gives, for all $\theta \in \Theta$,

$$\psi[\ell, \pi](\theta) = \psi[\ell, \pi](\theta_0) \exp(-w(\ell(\theta) - \ell(\theta_0))) \frac{\pi(\theta)}{\pi(\theta_0)} = C \exp(-w\ell(\theta))\pi(\theta),$$

where $C \doteq \psi[\ell, \pi](\theta_0) \exp(w\ell(\theta_0))/\pi(\theta_0)$ does not depend on θ . Finally, $\psi[\ell, \pi]$ is a possibility function, so $\sup_{\theta} \psi[\ell, \pi](\theta) = 1$. Therefore

$$1 = \sup_{\theta \in \Theta} \psi[\ell, \pi](\theta) = C \sup_{\theta \in \Theta} \exp(-w\ell(\theta))\pi(\theta),$$

which yields

$$C^{-1} = \sup_{\theta \in \Theta} \exp(-w\ell(\theta))\pi(\theta)$$

and the claimed sup-normalised Gibbs form follows. $\square$