Learning with Confidence

Oliver E. Richardson[1,2]

[1]Computer Science Dept., Université de Montréal, Montréal, Canada
[2]Mila – Quebec AI Institute

Abstract

We characterize a notion of confidence that arises in learning or updating beliefs: the amount of trust one has in incoming information and its impact on the belief state. This *learner's confidence* can be used alongside (and is easily mistaken for) probability or likelihood, but it is fundamentally a different concept—one that captures many familiar concepts in the literature, including learning rates and number of training epochs, Shafer's weight of evidence, and Kalman gain. We formally axiomatize what it means to learn with confidence, give two canonical ways of measuring confidence on a continuum, and prove that confidence can always be represented in this way. Under additional assumptions, we derive more compact representations of confidence-based learning in terms of vector fields and loss functions. These representations induce an extended language of compound "parallel" observations. We characterize *Bayesian* learning as the special case of an *optimizing learner* whose loss representation is a linear expectation.

1 INTRODUCTION

What does it mean to have a high degree of confidence in a statement ϕ? It is often taken to mean that ϕ is likely. We argue that there is also another conception of confidence that arises when learning—one that complements likelihood and, moreover, unifies several different concepts in the literature. This kind of confidence is a measure of *trust* in an observation ϕ, rather than its likelihood; it quantifies how seriously to take ϕ in updating our beliefs. So at one extreme, if we observe ϕ but have no confidence in it, we do not change our beliefs at all; at the other, if we have full confidence in ϕ, we fully (and irreversibly) incorporate it into our beliefs.

Example 1. Suppose our belief state is a probability measure P, and we observe an event ϕ. The standard way to learn ϕ is to condition on it (i.e., adopt belief state $P \mid \phi$). This is a full-confidence update; ϕ has probability 1 afterwards, and conditioning on it again has no further effect. Here is one obvious way to interpret intermediate degrees of confidence: starting with prior P and learning ϕ with confidence $\alpha \in [0,1]$, we end up with posterior $(1-\alpha)P + \alpha(P \mid \phi)$. Thus, having high confidence in ϕ leads to posterior beliefs that give ϕ high probability. The converse is false, however. If an untrusted source tells us ϕ which we already happen to believe, then our prior assigns ϕ high probability, we learn ϕ with low confidence, and our posterior beliefs still give ϕ high probability. □

Confidence allows us to be uncertain about observations, which is quite different in principle from making observations that are uncertain. *Jeffrey's rule* (1968) is a well-established approach to the latter. An important feature of the former, however, is that it enables learning without fully committing to new observations. Full-confidence updates, such as conditioning in Example 1, are irreversible: from ϕ and the posterior $P|\phi$, it is not possible to recover the prior belief P. The same is true of Jeffrey's rule, which, in our conception, also prescribes full-confidence updates. The concept we propose here is more similar to that behind of Shafer's *Theory of Evidence* (1976), although his account is specialized to a specific representation of uncertainty that has since fallen out of fashion.

Example 2. Suppose our beliefs are represented by a *(Dempster-Shafer) belief function*, which generalizes a probability measure over a finite set W of possible worlds. Like a probability, a belief function Bel assigns to each event $U \subseteq W$ a number $Bel(U) \in [0,1]$, with $Bel(\emptyset) = 0$ and $Bel(W) = 1$. It need not necessarily be that $Bel(U) + Bel(\overline{U}) = 1$, but Bel must satisfy certain axioms (whose details do not matter for our purposes) ensuring that $Bel(U) + Bel(\overline{U}) \leq 1$. Bel can be equivalently represented by its *plausibility function* $Plaus(U) := 1 - Bel(\overline{U})$. It

is easy to see that $Bel(U) \leq Plaus(U)$, and if Bel is a probability measure, then $Bel = Plaus$.

Suppose we come accross evidence that supports an event $\phi \subseteq W$ to a degree $\alpha \in [0, 1]$. Together, ϕ and our confidence α in it can be represented by the *simple support function*

$$Bel_{(\alpha,\phi)}(U) := \begin{cases} 1 & \text{if } U = W \\ \alpha & \text{if } \phi \subseteq U \subsetneq W \\ 0 & \text{otherwise.} \end{cases}$$

To combine belief functions, Shafer argues for Dempster's *rule of combination* ($\oplus$). If we use $\oplus$ to combine two simple support functions for ϕ with degrees of support α_1 and α_2, we get another simple support function for ϕ, with combined support $\alpha_1 + \alpha_2 - \alpha_1\alpha_2$. As we will see Section 3.1, confidence also has an additive form. In Shafer's theory, this is the *weight of evidence* $w = -k \log(1 - \alpha)$ for some $k > 0$ [Shafer, pg 78]. The additive form of confidence plays a fundamental role in Shafer's theory, as it does in ours.

Using $\oplus$ to combine our prior with our evidence leads to posterior belief $Bel' := Bel \oplus Bel_{(\alpha,\phi)}$, whose plausibility measure happens to be

$$Plaus'(U) = \frac{\alpha\ Plaus(U \cap \phi) + (1 - \alpha)\ Plaus(U)}{1 - \alpha + \alpha\ Plaus(\phi)}. \quad (1)$$

It is easy to verify that $Bel' = Bel$ when $\alpha = 0$, and it can also be shown that $Bel'(\phi) = Plaus'(\phi) = 1$ when $\alpha = 1$. So, as before, confidence $\alpha \in [0, 1]$ parametrizes a continuous path from ignoring ϕ to fully incorporating it. Yet the meaning of intermediate degrees of confidence can be subtle. In the special case where $Bel = Plaus$ is a probability measure, a full confidence update ($\alpha = 1$) yields the same conditioned probability $Plaus' = (Plaus|\phi)$ as in Example 1. Furthermore, the set of possible posteriors for intermediate $\alpha \in (0, 1)$ is the same in both cases. However, the two paths are parameterized differently; in fact, for all $\alpha \in (0, 1)$ the two updates disagree. It follows that the appropriate numerical value of α must depend on more than just an intuition of "fraction of the way to the update". □

Shafer's theory aims to address two seemingly problematic aspects of Bayesianism: it prescribes a belief representation that can better handle ignorance, and enables observations other than those that "establish a single proposition with certainty" [Shafer, 1976, Chapter 1: §7,§8]. Ironically, in solving the first problem, his solution to the second becomes inaccessible to those who do not work with Dempster-Shafer belief functions. Our notion of learner's confidence directly addresses Shafer's second concern, but applies far more broadly. A significant strength of our approach that we do not take a stand on how beliefs should be represented—the concept of trust applies whether you use probability measures, belief functions, graphical models, imprecise probabilities, or something entirely different. To illustrate, we now unpack the role of confidence in neural networks.

Example 3 (Training a NN)**.** The "belief state" of a neural network may viewed as a setting $\theta \in \Theta \subseteq \mathbb{R}^d$ of weight parameters. For definiteness, suppose we are talking about a classifier, so that there is a space X of inputs, a finite set Y of labels, and a parameterized family of functions $\{f_\theta : X \to \Delta Y\}_{\theta \in \Theta}$ mapping inputs $x \in X$ to distributions $f_\theta(x) \in \Delta Y$ over labels. In the supervised setting, an observation is a pair (x, y) consisting of an input x labeled with class y.

Suppose we now observe $\phi = (x, y)$ with some degree of confidence; how should we update the weights θ? In contrast with previous examples, it is not so obvious how to learn ϕ with full confidence. Instead, modern learning algorithms tend to be iterative procedures $\texttt{step} : (X \times Y) \times \Theta \to \Theta$ that make small adjustments $\theta \mapsto \texttt{step}(\phi, \theta)$ to the weights. Each step is essentially a low-confidence update. There is no guarantee, for example, that $f_{\texttt{step}(\phi,\theta)}(x)$ gives high probability to y—only that it is higher than $f_\theta(y|x)$. This lower level of confidence is arguably what makes these learning algorithms robust to noisy and contradictory inputs.

Higher confidence updates can be obtained by applying `step` more than once. From initial weights θ_0 and defining $\theta_{n+1} = \texttt{step}(\phi, \theta_n)$, we get a sequence $(\theta_0, \theta_1, \theta_2, ...)$ that converges to some $\theta_* \in \Theta$. These limiting weights fully incorporate ϕ in the sense that $\theta_* = \texttt{step}(\phi, \theta_*)$, and also that $f_{\theta_*}(x)(y) = 1$ (at least if the network is sufficiently over-parameterized), i.e., x is classified as y with probability 1. Correspondingly, adopting belief θ_* is appropriate only if we have complete trust in ϕ, meaning we find it critical that x be classified as y. (At the other extreme, if we have no confidence in ϕ, we should not update θ at all.) Thus, the number of training iterations n is a measure of confidence: it interpolates between no confidence (zero iterations of `step`) and full confidence (infinitely many iterations of `step`). Like Shafer's weight of evidence (Example 2), the number of training iterations is an additive measure of confidence.

In the simplest settings, training examples do not come with confidence annotations, in which case one effectively treats them all with the same default confidence (by selecting a learning rate). The number of times that $\phi = (x, y)$ appears in a dataset is then the de-facto measure of confidence in ϕ. Often, though, these are not our intended confidences, which is why it can be helpful to remove duplicates [Lee et al., 2021]. In richer settings, a more nuanced degree of confidence specific to each training example often arises, such as agreement between annotators [Artstein, 2017], or confidence scores in self-training [Zou et al., 2019].

It is worth emphasizing that confidence is not always just a matter of accuracy. Suppose, for example, that the classifier is intended to screen job applications, and that we want to make hiring practices less discriminatory. In this case, we should have low confidence in training data based on prior hiring decisions—not because it is inaccurate, but because we do not trust it to inform our new hiring practice. □

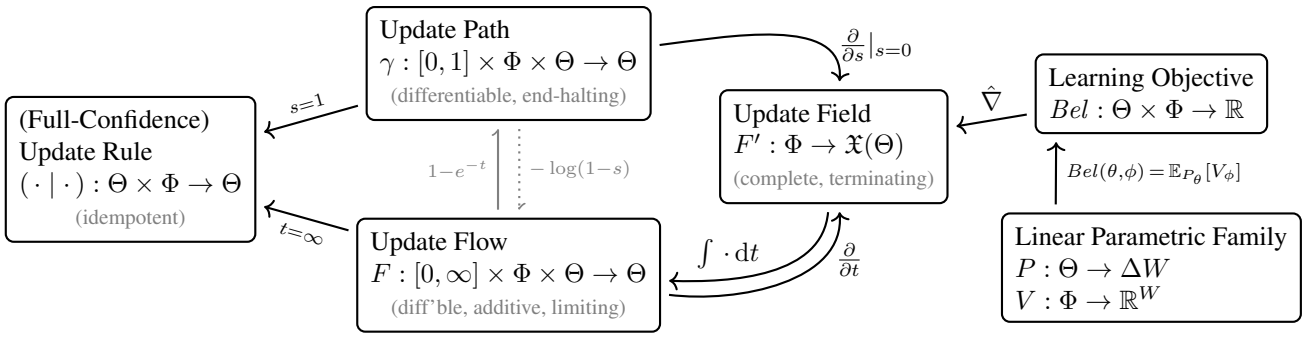

Figure 1: Relationships between different representations of confidence-based learners.

Perhaps the most important application of learner's confidence is in treating different sources of information with different degrees of trust. Sensor fusion, which aims to combine readings from multiple sensors of various reliabilities, is a clear example—and Kalman filtering [Kalman, 1960, Brown and Hwang, 1997], the standard approach to this problem, indeed comes with its own account of confidence.

Example 4 (1D Kalman Filter)**.** Suppose we are modeling a dynamical system whose state is a real number $x \in \mathbb{R}$, and we receive noisy measurements z of x. The Kalman Filter tells us how to track this information with belief state $(\hat{x}, \sigma^2)$, where $\hat{x} \in \mathbb{R}$ is our current estimate of x, and σ^2 is an uncertainty in that estimate, in the form of a variance. We now receive an observation $z \sim \mathcal{N}(x, r^2)$ from a sensor. How should we update our beliefs?

The answer ranges from ignoring z to replacing $\hat{x}$ with it, depending on how much we trust the sensor. The Kalman filter measures this trust with two (entangled) kinds of confidence: the precision r^{-2} of the sensor, and a quantity K called *Kalman gain*. The updated state $(\hat{x}', \sigma^{2\prime})$ is then:

$$\hat{x}' = \hat{x} + K(z - \hat{x}), \quad \sigma^{2\prime} = (1-K)^2\sigma^2 + (K)^2 r^2.$$

Like the other confidence measures we have seen, K interpolates (linearly) between our prior mean $\hat{x}$ and the new observation z, and ("quadratically") between our prior uncertainty σ^2 and the sensor variance r^2.

More than in previous examples, we can also say something prescriptive about how to select a degree of confidence. Assuming the goal is to maintain an unbiased estimate of x with minimal uncertainty (as measured by expected squared error of $\hat{x}$), and that z is indeed the result of adding independent noise to x, then the optimal Kalman gain is $K_{\text{opt}} = \sigma^2/(\sigma^2 + r^2)$ [Brown and Hwang, 1997, p. 146], and K is typically chosen this way in practice [Becker, 2003]. Let us now revisit the extremes. If $K = 0$, which is optimal when z has unbounded variance, the belief state remains unchanged: intuitively, there is so much noise in observations that we ignore them. At the other extreme, if no noise is added ($r^2 = 0$), then $K_{\text{opt}} = 1$ and we end up with a posterior $(z, 0)$ based solely on the new observation. □

Example 4 features three kinds of (un)certainty:

1. **Learner's Confidence:** a subjective trust in how seriously to take an observation for updating (e.g., K).
2. **Internal (Epistemic) Confidence:** the degree of uncertainty present in a given belief state, either overall (σ^2) or in a given statement (e.g., the density $\phi \mapsto \mathcal{N}(\phi|\hat{x}, \sigma^2)$). Internal confidences in our other examples include the probability $\Pr(\phi)$ in Example 1, the degree of belief $Bel(\phi)$ in Example 2, and the value of the loss function $\mathcal{L}(\theta, \phi)$ used to train the classifier in Example 3.
3. **Statistical (Aleatoric) Confidence:** an objective measure of the (un)reliablility of an observation, based on historical data and/or modeling assumptions about how observations arise (e.g., the noise level r^2).

The three senses of the word "confidence" are related, but different in nature. A great deal of work has already gone into understanding the differences between senses 2 and 3 [Der Kiureghian and Ditlevsen, 2009, Hüllermeier and Waegeman, 2021]. We (obviously) focus on sense 1, which we have tried to distinguish from more pervasive usage of the word (sense 2) to quantify subjective likelihood, degree of belief, or (un)certainty. Nevertheless, epistemic confidences (sense 2) may be thought of as aggregate reflections of learner's confidence (sense 1) in past observations; conversely, it is often possible to define learner's confidence by its effect on epistemic confidence (Section 3.2).

One should also distinguish learner's confidence (sense 1), at least in principle, from statistical confidences (sense 3) such as the variance in readings of a sensor (Example 4) or annotator agreement (Example 3). When available, the statistical reliability of an information source should absolutely play a role in determining how seriously we take it in updating our beliefs; learner's confidence informed exclusively by a probabilistic model can be seen as an important ("aleatoric") special case of our theory. Still, statistical confidence often presupposes that observations are drawn (independently) from a (fixed) distribution, while learners's confidence is meaningful even without such assumptions.

Contributions. We hope that these examples have given the reader an intuitive sense of what confidence is, how ubiquitously it arises, and why it is important. In the remainder of the paper, we study confidence more formally, making a series of successively stronger assumptions (all satisfied by Examples 1 to 4). Each set of assumptions enables a new more compact representation for a learning rule, summarized in Figure 1. In Section 2, we develop a formal framework laying out axioms for our notion of confidence. In Section 3, we focus on the properties of confidence in a continuum, developing vector-field and loss-based representations of learners. This can enable simultaneous orderless updates, even in settings where it was not previously possible. Finally, we analyze Bayesian updating in Section 4.

2 A FORMAL MODEL OF CONFIDENCE, LEARNING, AND BELIEF

Our formalism consists of three components: a domain $[\bot,\top]$ of confidence values, a space Θ of belief states, and a language Φ of possible observations. For instance:

- In Example 1, Θ is the set of probability measures on some measurable space $(\Omega, \mathcal{F})$, Φ is the σ-algebra $\mathcal{F}$, and the confidence domain is $[0,1]$.
- In Example 2, Θ is the set of belief functions over a finite set W, $\Phi = 2^W$ is the set of subsets of W, and confidence is a degree of support $\alpha \in [0,1]$ or a weight of evidence $w \in [0,\infty]$.
- In Example 3, $\Theta \subseteq \bar{\mathbb{R}}^d$ is the space of network parameters, $\Phi = X \times Y$ is the space of input-lablel pairs, and the confidence domain is the extended natural numbers $\{0,1,\dots,\infty\}$ under addition.
- In Example 4, $\Theta = \Phi = \mathbb{R}$, The domain of K is $[0,1]$, and the domain of σ^2 is $[\infty,0]$. Together, the pair (K,σ^2) acts as a confidence.

We call $(\Theta, \Phi, [\bot,\top])$ a *learning setting*. In this setting, a *learner* is a function $Lrn : \Phi \times [\bot,\top] \times \Theta \to \Theta$ that describes the belief update process. Explicitly: from a prior belief θ, and a statement ϕ observed with some degree of confidence χ, a learner produces a posterior belief state $Lrn(\phi,\chi,\theta) \in \Theta$. We use superscripts and subscripts to fix some arguments of Lrn and view it as a function of the others. So $Lrn(\phi,\chi,\theta)$ can equivalently be written as $Lrn_\phi(\chi,\theta) = Lrn^\chi_\phi(\theta) = Lrn^\chi(\phi,\theta) = Lrn_{(\theta,\phi)}(\chi)$. The rest of Section 2 develops axioms for Lrn and supporting concepts intended to capture intuitions about learning.

We proceed in three stages. After starting with an abstract theory of confidence domains $[\bot,\top]$ themselves (Section 2.1), we then axiomatize confidence-based updates to beliefs in Θ (Section 2.2). Finally, we bring in observations Φ and the function Lrn (Section 2.3).

2.1 ABSTRACT CONFIDENCE DOMAINS

A *confidence domain* $(D, \leq, \bot, \top, *, \mathfrak{g})$ is a set D of confidence values equipped with a preorder $\leq$, a least element $\bot$ ("no confidence"), a greatest element $\top$ ("full confidence"), and an operation $*$ that combines two independent degrees of confidence. We often abbreviate a confidence domain as $D = [\bot,\top]$, leaving $\leq$ and $*$ implicit. We want to ignore independent information we have no confidence in, and, if already fully confident, remain so in the face of new independent information. Formally, this amounts to requiring, for all $\chi, \chi', \chi'' \in D$:

$$
\begin{aligned}
&(\chi * \chi') * \chi'' = \chi * (\chi' * \chi'') && \text{(associativity)},\\
&\bot * \chi = \chi && \text{(that } \bot \text{ is neutral)},\\
&\top * \chi = \top && \text{(and that } \top \text{ is absorbing)}.
\end{aligned}
$$

Finally, D comes with geometric information $\mathfrak{g}$, which may include topology or differentiable structure. We are especially interested in two continuous domains from our examples. The first is the *fractional domain* $[0,1]$, whose elements $s \in [0,1]$ represent the "proportion of the way towards complete trust". If you go proportion s towards fully trusting something, then s' of the remaining way, then overall you have gone $s * s' := s + s'(1-s) = s + s' - s \cdot s'$ of the way to complete trust. The other confidence domain of particular interest is the *additive domain* $[0,\infty]$, which is ideal for analogies of time and weight.

Proposition 1. *The fractional domain* $[0,1]$ *and the additive domain* $[0,\infty]$ *are isomorphic. Furthermore, the space of isomorphisms between them is in natural bijection with* $(0,\infty)$. *Specifically, for each* $\beta \in (0,\infty)$, *there is an isomorphism* $\varphi_\beta : [0,1] \to [0,\infty]$ *given by* $\varphi_\beta(s) = -\frac{1}{\beta}\log(1-s)$ *with inverse* $\varphi_\beta^{-1}(t) = 1 - e^{-\beta t}$.

The fact that these two domains are equivalent but only up to β—a "choice of units" in the additive domain, or "tempering" in the fractional domain—-implies that many standard ways of quantifying confidence are equivalent, yet also highlights the fundamental difficulty of doing so in absolute terms (as we began to see at the end of Example 2).

Keep in mind that there are confidence domains as well. The interval $[0,1]$ with $* = \max$ is an important one that is not isomorphic to the additive or fractional domains. Confidence domains can also be multi-dimensional or discrete—but our results in Sections 3 and 4 say little about these cases.

2.2 BELIEF STATES AND COMMITMENT FUNCTIONS

We now reintroduce belief states $\theta \in \Theta$ in order to describe the role of confidence in belief updating. Observations ϕ come later (Section 2.3); we find that the most essential aspects of confidence can already be understood through the

behavior of a function $F = Lrn_\phi : [\bot,\top] \times \Theta \to \Theta$ that describes the learning process for some fixed and abstract ϕ. We call such a function F a *commitment function* if it obeys the axioms in this subsection (L1–5) intended to ensure that F respects the structure of the confidence domain.

No Confidence. Having no confidence ($\chi = \bot$) in an observation ϕ should lead us to ignore it.

[L1] $\forall \phi, \theta. \quad Lrn_\phi^\bot(\theta) = Lrn_\phi(\bot, \theta) = \theta.$

Full-confidence. Since the purpose of $Lrn_\phi^\top$ is to *fully* incorporate ϕ into our beliefs, two successive full-confidence updates with the same information ought to have the same effect as a single one: having fully integrated ϕ into our beliefs, there is nothing to do upon observing ϕ again.

[FC] Full-confidence updates are idempotent. That is, for all $\phi \in \Phi$, $Lrn_\phi^\top \circ Lrn_\phi^\top = Lrn_\phi^\top$.

Once Θ, Φ, and any relevant relationships between them are specified, there is often a natural choice of full-confidence update rule. We illustrate with three examples. In each case, the possible belief states $\Theta := \Delta W$ be the set of all probability distributions over a finite set W of possible worlds.

(1) **Conditioning.** First, consider the case where observations are events, i.e., $\Phi := 2^W$. The overwhelmingly standard way to update is to condition: starting with $P \in \Delta W$, the conditional measure $P|A \in \Delta W$ is given by $(\mu|A)(B) = P(B \cap A)/P(A)$, provided $P(A) > 0$. Note that $(P|A)|A = P|A$, so the update is idempotent.

(2) **Imaging** [Lewis, 1976]**.** Suppose we already have a full-confidence update rule $f : \Phi \times W \to W$ that, given $\phi \in \Phi$ and $w \in W$, produces the world $f(\phi, w) \in W$ "most similar to w, in which ϕ is true" [Gardenfors, 1982]. Idempotence of $f_\phi : W \to W$ means the world most similar to $f(\phi, w)$ in which ϕ is true, is $f(\phi, w)$ itself. We can then lift f to a full confidence update rule for ΔW, by $F(\phi, P)(A) := P(\{w : f(w, \phi) \in A\})$, intuitively moving the mass of w to $f(\phi, w)$. Since f is idempotent, so is F.

(3) **Jeffrey's Rule.** The two previous approaches to updating establish an event with probability 1. Jeffrey's rule (J) addresses this limitation by allowing for uncertain (i.e., probabilistic) observations. Formally, let Φ be the set of pairs (X, π) where $X : W \to S$ is a random variable taking values in a set S, and $\pi \in \Delta S$ is a probability on S. Jeffrey's update rule is: $J((X, \pi), P) := \sum_{x \in S} \pi(X{=}x) P\big|(X{=}x)$. When π places all mass on some $x \in S$, J conditions on $X{=}x$. For this reason, J is thought to generalize conditioning to observations of "lower confidence". Yet even when π is not deterministic, J *fully* incorporates π into the posterior beliefs: the marginal of $J((X, \pi), P)$ on X is $\pi(X)$, and the prior belief $P(X)$ has been destroyed. Indeed, $J_{(X,\pi)}$ is idempotent. Therefore, J still establishes observations with full confidence—it's just that those observations are probabilities. Experience suggests that this point can be counter-intuitive; we submit that the confusion is clarified by a conception of confidence distinct from likelihood.

FC implies that full-confidence updates are not invertable: they destroy information in the prior, often making for a simpler posterior. This potential simplification of future calculations is a major benefit of fully trusting information. However, full-confidence updates are extreme. An agent that updates by conditioning, for instance, permanently commits to believing everything it ever learns (and thus gains nothing from making the same observation again later). Clearly humans are not like this; revisiting information helps us learn [Ausubel and Youssef, 1965]. Similarly, artificial neural networks are trained with many incremental updates, and benefit from seeing the training data many times. We would like an account that allows for less extreme belief alterations, in which information is only partially incorporated. This is the role of intermediate degrees of confidence.

Geometry. Learner's confidence interpolates between ignoring new information and fully defering to it, and we would like that interpolation to be continuous and differentiable.

[L2] If $[\bot,\top]$ and Θ are both topological spaces, then for all θ and ϕ, the map $Lrn_{(\theta,\phi)} = \chi \mapsto Lrn(\theta, \chi, \phi)$ is continuous. If $[\bot,\top]$ and Θ are both manifolds, then $Lrn_{(\theta,\phi)}$ is differentiable—and also Lrn_ϕ^χ is differentiable on a subset Θ_ϕ defined in Proposition 3 below.

Ideally the posterior would be continuous in our prior beliefs as well. This suggests a simpler strengthening of L2: that Lrn_ϕ also be continuous (and differentiable) as a function of (χ, θ)—yet this is often too much to ask for.

Proposition 2. *Take $\Theta = \Delta W$ and $\phi \subseteq W$. There exists no continuous function $Lrn_\phi : \Delta W \times [0, 1] \to \Delta W$ with the property that $Lrn_\phi(\mu, 1) = \mu|\phi$ when $\mu(\phi) > 0$.*

This result is yet another perspective on the familiar difficulties with conditioning on events of probability zero [], but intuitively this should be an edge case. Instead of imposing an axiom, we observe that it is possible to capture the phenomenon in a useful way even at this abstract level.

Proposition 3. *For all $\phi \in \Phi$, there is a maximal open set $\Theta_\phi \subseteq \Theta$ such that the restriction $Lrn_\phi|_{\Theta_\phi} : [\bot,\top) \times \Theta_\phi \to \Theta$ of Lrn_ϕ to Θ_ϕ is continuous.*

In our examples, Θ_ϕ consists of those belief states that do not flatly contradict ϕ. In Example 1, Propositions 2 and 3 imply that $\Theta_\phi = \{\mu \in \Delta W : \mu(\phi) > 0\}$ is the set of distributions for which conditioning on ϕ is defined. In Example 3, $\Theta_{(x,y)}$ is the set of parameters at which gradients $\nabla_\theta \ell(f_\theta(x), y)$ of the loss ℓ are finite.

Order. For a learner, the defining feature of the ordering $\chi < \chi'$ is that learning with higher confidence (χ') can done

by first making the more conservative, lower-confidence (χ) update, followed by a nontrivial residual update.

[L3] $\exists s : \{(\chi', \chi) : \chi' > \chi_1\} \to [\bot, \top]$ continous such that $Lrn_\phi(s(\chi', \chi), Lrn_\phi(\chi, \theta)) = Lrn_\phi(\chi', \theta)$.

Furthermore, learning is not cyclic: if learning with confidences χ_0 and χ_1 have the same effect, then the same is true of all confidences $\chi_0 \leq \chi \leq \chi_1$ between them.

[L4] If $\chi_0 \leq \chi \leq \chi_1$ and $Lrn_\phi(\chi_0, \theta) = Lrn_\phi(\chi_1, \theta)$, then $Lrn_\phi(\chi, \theta) = Lrn_\phi(\chi_0, \theta)$.

Independent Combination. Lrn should be used to incorporate information to the extent that it is novel, i.e., information that is not already accounted for in our prior beliefs. Thus, we would like a sequence of two independent observations in the same observation ϕ to be equivalent to a single observation of ϕ with their combined degree of confidence.

[L5] $\forall \phi, \chi, \chi'.\ Lrn_\phi(\chi, Lrn_\phi(\chi', \theta)) = Lrn_\phi(\chi * \chi', \theta)$

L5 appears to be a rather strong assumption. Since $\top$ is absorbing, for example, L5 implies FC. In the language of algebra, L1 and L5 (and L2) together require Lrn_ϕ to be a (smooth) action of the monoid $([\bot, \top], *, \bot)$ on Θ. However, if we are free to chose the confidence domain, L5 imposes no other restrictions on Lrn (see Proposition 12 in the appendix). It is also easy to verify that the confidences α and n of Examples 1 to 3 satisfy L5. Nevertheless, for the canonical domains $[0, 1]$ and $[0, \infty]$, L5 is indeed a strong assumption. In fact, of the confidences in Example 4, neither K alone nor σ^2 satisfy L5 out of the box—but sensor precision σ^{-2} does when $K = K_{\text{opt}}$ is the optimal gain, and the pair (K, σ^2) can be combined into a single domain satisfying L5, as we show in the appendix.

Our axioms so far have been conditions on the separate commitment functions $F : [\bot, \top] \times \Theta \to \Theta$, which we have called "$Lrn_\phi$", but we have not required that $F = Lrn_\phi$ have any relationship to observations ϕ. To address this, we must reintroduce the final pieces of our formalism.

2.3 OBSERVATIONS AND DEGREE OF BELIEF

Consider a function $Bel : \Theta \times \Phi \to [\bot, \top]$ that associates each belief state θ with a degree of belief in each statement ϕ. This usage of $[\bot, \top]$ represents "confidence" in the standard sense of likelihood, rather than of trust. Still, we can use Bel to articulate another key desideratum for the latter: learning ϕ with more confidence should lead to more belief in ϕ.

[LB1] $\forall \phi, \theta, \chi, \chi'. \quad \chi \geq \chi' \implies Bel(\phi, Lrn(\phi, \chi, \theta)) \geq Bel(\phi, Lrn(\phi, \chi', \theta))$.

We cannot ask for strict monotonicity, however: if we already fully believe ϕ (i.e., $Bel(\phi, \theta) = \top$), there is no way to attain a higher degree of belief, we cannot attain a higher degree of belief by learning ϕ. Instead, if we fully believe ϕ, learning ϕ should have no effect.

[LB2] If $Bel(\phi, \theta) = \top$, then $Lrn(\phi, \chi, \theta) = \theta$.

Perhaps even more importantly, if we learn something with full confidence, then we ought to fully believe it.

[LB3] $Bel(\phi, Lrn(\phi, \top, \theta)) = \top$.

While LB1–3 are serious constraints on Lrn if Bel is given, one can easily define Bel based on Lrn so as to ensure that LB1–3 hold trivially.

3 THE CONFIDENCE CONTINUUM

We now look deeper into the theory of learners whose confidence domain is a continuum (i.e., a connected, totally ordered, one-dimensional manifold with two endpoints).

With the domain $[0, \infty]$, L5 means Lrn is *additive*, making it amenable to analogies of weight (e.g., the weight of evidence w in Example 2) and time (e.g., the number of training iterations n in Example 3). Indeed, an additive learner can be implemented so that confidence really does coincide with time: imagine a machine with state space Θ, controlled by buttons labeled by Φ, that, while ϕ is pressed, evolves from initial state θ_0 according to $\theta(t) = Lrn(\phi, t, \theta)$. Conversely, this interpretation is coherent only if Lrn is additive—for otherwise there would exist t_1, t_2 such that the machine's state after pressing ϕ for t_1 seconds followed by t_2 additional seconds, would be different from the configuration after holding down ϕ for $t_1 + t_2$ seconds.

Temporal analogies may not always be appropriate (as they may clash with other, truer conceptions of "time"), yet they have such intuitive force that a function $f : [a, b] \times \Theta \to \Theta$ (with $0 \in [a, b] \subseteq \mathbb{R}$) satisfying L1, L2 and L5 is known generically as a *flow* [Lee, 2013]. Since L5 implies L3–4 for this domain, the only additional requirement of a commitment function is that $Lrn_{(\theta,\phi)}(\chi)$ have a well-defined limit as $\chi \to \infty$. This highlights the strength of the assumption that confidence lies in $[0, \infty]$ and combines additively, so one might understandably worry that this could limit applicability—but this is not the case. While the additive domain $([0, \infty], +)$ certainly restricts how confidence can be measured, it has little effect on what confidence can express.

Theorem 4. *If $[\bot, \top]$ is a continuum and $F : [\bot, \top] \times \Theta \to \Theta$ is a commitment function (i.e., satisfies L1–5), then there exists a continuous "translation" function $g : [\bot, \top] \times \Theta \to [0, \infty]$, and a commitment flow ${}^+F$ such that $\forall \theta, \chi.\ {}^+F(g(\chi, \theta), \theta) = F(\chi, \theta)$.*

Thus, updates performed with Lrn are equivalent to updates performed with ${}^+Lrn$ (its *additive form*), if confidences are translated (via g) appropriately. When the original domain $[\bot, \top]$ is isomorphic to the canonical domains $[0, \infty]$ and $[0, 1]$, the translation g need not depend on θ

and there is a unique such representation, up to a multiplicative constant in the output of g. However, by allowing for a belief-state-dependent translation of confidence, our construction provides in principle an additive representation even for very different confidence domains, such as when $*$ is not invertible (e.g., $* = \max$)—provided the points of non-differentiability can be handled appropriately, which is sometimes but not always possible.

The key to proving Theorem 4 is realizing that commitment flows can be equivalently represented by vector fields. This view, which we now unpack, confers other benefits as well.

3.1 ORDERLESS COMBINATION AND THE VECTOR FIELD REPRESENTATION

Is it the same to learn ϕ_1 and then ϕ_2 as it is to learn them in the opposite order? It is for belief functions (Example 2) and when conditioning. But, in general, the order of observations can have a significant impact on the result. Humans tend to have a recency bias: more recent observations have a stronger influence on beliefs. Examples 1 and 4 are not commutative either. But if the order matters for our update, what should we do if we receive two pieces of information simultaneously? There is a natural way to do this with the techniques used to prove Theorem 4.

Since Θ carries a differentiable structure, it makes sense to talk about its tangent space $T\Theta$, which consists of pairs $(\theta, \mathbf{v})$ where $\theta \in \Theta$, and $\mathbf{v}$, intuitively, is a direction that one can travel in Θ beginning at θ [Lee, 2013, §3]. A *vector field* $X \in \mathfrak{X}\Theta$ is a differentiable map $X : \Theta \to T\Theta$ assigning to each $\theta \in \Theta$ a vector $X(\theta) = (\theta, \mathbf{v}) \in T\Theta$ tangent to θ. L3 implies that the behavior of Lrn is generated by the way it handles updates of small confidence. So, in a sense, all we need to know about Lrn is how it handles infinitesimal confidences—which can be viewed as a vector field. More precisely, in most cases (such as when using either the fractional or additive confidence domains), a commitment function Lrn_ϕ can be represented by the vector field

$$Lrn'_\phi := \theta \mapsto \frac{\partial}{\partial \chi} Lrn(\theta, \chi, \phi)\Big|_{\chi=\bot} \qquad \in \mathfrak{X}\Theta. \tag{2}$$

(To handle edge cases involving the zero field, we may need a more complex but closely related definition; see the proof of Theorem 4 for details.) We can then recover ${}^+Lrn_\phi$ as the integral curves of Lrn'_ϕ [Lee, 2013, Thm 9.12]. It may seem counter-intuitive that the vector field Lrn'_ϕ, which does not mention confidence at all, alegedly captures confidence—but it does, intuitively, by specifying everything about the learning process except for the degree of confidence itself.

We now return to orderless combination of observations. One key property of vector fields is thier closure under linear combination—and since commitment flows and vector fields are equivalent, we can extend this linear structure to observations themselves. This gives us a natural way to combine observations in parallel. Concretely, given $\phi_1, \phi_2 \in \Phi$, we can form a new input $\phi_1 \oplus \phi_2$ and extend Lrn to handle it by taking $Lrn'_{\phi_1 \oplus \phi_2} := Lrn'_{\phi_1} + Lrn'_{\phi_2}$. Standard existence theorems (and uniqueness) theorems for ordinary differential equations then apply. Nevertheless, there are several wrinkles: in some cases, $Lrn_{\phi_1 \oplus \phi_2}$ may only continuously extend to a finite $\lim_{t\to\infty} Lrn^t_{\phi_1 \oplus \phi_2}$ may not exist, in which case we cannot continuously extend $Lrn_{\phi_1 \oplus \phi_2}$ to handle full confidence, and $Lrn_{\phi_1 \oplus \phi_2}$ might not satisfy L4. We leave $\phi_1 \oplus \phi_2$ undefined in such cases, but point out that having a loss representation for Lrn (the subject of Section 3.2) suffices to avoid both problems.

Observations ϕ_1 and ϕ_2 *commute* iff $Lrn^{\chi_1}_{\phi_1} \circ Lrn^{\chi_2}_{\phi_2} = Lrn^{\chi_2}_{\phi_2} \circ Lrn^{\chi_1}_{\phi_1}$ for all $\chi_1, \chi_2 \neq \top$. Clearly $\phi_1 \oplus \phi_2 = \phi_2 \oplus \phi_1$ when either is defined, so $\oplus$ provides a way of combining observations orderlessly, even in cases where ϕ_1 and ϕ_2 do not commute—and when they do, $\phi_1 \oplus \phi_2$ is equivalent to observing ϕ_1 and ϕ_2 in either order. The following proposition shows that $\phi_1 \oplus \phi_2$ is equivalent to an infinitely fine interleaving of ϕ_1 and ϕ_2 updates.

Proposition 5. *Suppose Lrn_{ϕ_1} and Lrn_{ϕ_2} are commitment flows. For $t \in [0, \infty]$, let $L_t := Lrn^t_{\phi_2} \circ Lrn^t_{\phi_1}$ denote learning ϕ_1 followed by ϕ_2 (both with confidence t), and for $n \in \mathbb{N}$, let $L^{(n)}_t(\theta) := L_t \circ \cdots \circ L_t(\theta)$ denote n repeated applications of L_t. Then $Lrn^\chi_{\phi_1 \oplus \phi_2}(\theta) = \lim_{n\to\infty} L^{(n)}_{\chi/n}(\theta)$.*[1]

3.2 OPTIMIZING LEARNERS

We have now seen how learners satisfying certain axioms can be represented as vector fields (Section 3.1). A particularly important way of specifying a vector field is via the gradient of a potential. This is especially true in modern machine learning, where training is idealized as loss-minimizing gradient flow [Arora et al., 2018], and where the substantial advances of the last two decades have repeatedly demonstrated value of casting learning as optimization [Sra et al., 2011]. Our framework allows us to express this idea as a simple relationship between Lrn and Bel:

[LB4] $\quad \frac{\partial}{\partial \chi} Lrn(\phi, \chi, \theta) = \nabla_\theta \, Bel(\phi, \theta)$

LB4 says that learning occurs by gradient ascent (i.e., using some measure of disbelief in observations as a loss): that learning is fundamentally (just) about locally increasing degree of belief—no more, and no less. It also gives us a way of turning Bel (whose output is an epistemic confidence) into a commitment flow Lrn (which takes a learner's confidence as input), which may have contributed to any ambient confusion about the distinction between the two readings of

[1] For completeness, note that Proposition 5 is closely related to the *Lie-Trotter product formula* [Trotter, 1959, Cohen et al., 1982], and can be viewed as an interpreted instantiation of it.

the word "confidence". Unlike LB1–3, LB4 imposes serious constraints on Lrn even if we are free to select Bel. We say Lrn is *optimizing* if there exists some Bel such that the pair satisfy LB4. This way of constructing a learner has another benefit: the flows formed from such vector fields are guaranteed to have limits and satisfy L4, meaning that orderless combination $\oplus$ is always well-defined.

Technically, to view the derivative of a function $\ell : \Theta \to \mathbb{R}$ as a vector field $\nabla\ell \in \mathfrak{X}\Theta$ (rather than a co-vector field), one needs more than a manifold structure on Θ; we will assume that Θ comes with what is called a *Riemannian Metric*. The details are unimportant; what matters is that we can always fall back on the Euclidean metric for subsets of $\mathbb{R}^n$, and that some other spaces (such as parametric families of distributions), have a different natural metric.

Optimizing Commitment for Probabilistic Beliefs. In many learning settings of interest, beliefs $\theta \in \Theta$ are associated with probability distributions $P_\theta \in \Delta\Omega$ over some measurable space Ω. Fortunately, this gives us a natural Riemannian metric on Θ—which, as explained above, is precisely what we need in order to make sense of gradients on a manifold. Specifically, the *Fisher Information Metric* (FIM) induced by the parameterization $\theta \mapsto P_\theta$ turns out to be the unique metric (up to scalar multiple) that is invariant under sufficient statistics [Chentsov, 1982][2]—a finding that has lead many to use the term *natural gradient* for gradients in this geometry, and formed the basis Information Geometry [Amari, 1998, Amari and Nagaoka, 2000].

To be rather technical for a paragraph, a *Riemannian metric* consists of an inner product $\langle\cdot,\cdot\rangle_\theta : T_\theta\Theta \times T_\theta\Theta \to \mathbb{R}$ on tangent vectors at each point $\theta \in \Theta$; it can therefore be viewed as a matrix $G(\theta)$ with components $G(\theta)_{i,j} = \langle e_i, e_j\rangle_\theta$, where $\{e_i\}$ are basis vectors of the tangent space $T_\theta\Theta$. The gradient of a function $f : \Theta \to \mathbb{R}$ in this geometry is then given by $\nabla_\theta f(\theta) := G(\theta)^\dagger \frac{\partial f}{\partial \theta}^\mathsf{T}(\theta)$ where $\mathcal{I}(\theta)^\dagger$ denotes the Moore-Penrose psuedoinverse of the matrix $\mathcal{I}(\theta)$ and $\frac{\partial f}{\partial \theta} = [\frac{\partial f}{\partial \theta_1}, \ldots, \frac{\partial f}{\partial \theta_n}]$ is the (co)-vector of partials (i.e., the transpose of the gradient of f in the Euclidean metric, which is sensitive to the choice of coordinates). In the special case where $\Theta = \Delta W$ is itself the set of probability distributions over a finite set $W = \{1, \ldots, n\}$ and $P_\theta = \theta$, the simplex representation $\theta = P = (p_1, \ldots, p_n) \in \Theta$ (in which $\sum_i p_i = 1$ and $p_i \geq 0$), yields $\mathcal{I}(P) = \mathrm{diag}(\frac{1}{p_1}, \ldots, \frac{1}{p_n})$. For readers who did not follow the details: we now have a representation-invariant way of calculating gradients.

Let us now revisit the examples from Section 1.

[2]For instance, if X and Y take values in Ω, and $p(Y|X)$ and $q(X|Y)$ are such that $P_\theta(X) = q \circ p \circ P_\theta(X)$ for all θ, then clearly the family $P_\theta(Y) := p \circ P_\theta(X)$ carries the same information about the parameters (and how to update them) as does $P_\theta(X)$. Chentsov's theorem (1982) tells us that the FIM is the only Riemannian metric on Θ (as a function of the parameterization $\theta \mapsto P_\theta$), that is the same whether derived from $P_\theta(X)$ or $P_\theta(Y)$.

- The update process of Example 1 can be shown to be the optimizing for log probability $Bel(P, \phi) = \log P(\phi)$. In other words, it is about minimizing surprisal.
- In Example 2, $Lrn(Bel, \alpha, \phi) = Bel \oplus Bel_{(\alpha,\phi)}$ is not optimizing; assuming that it is leads to a contradiction of Clairaut's theorem in the general case. However, in special case where the belief state $Bel = Plaus \in \Theta$ is restricted to probability measures, Lrn is optimizing with objective $Bel(Bel, \phi) = Bel(\phi)$, perhaps atoning for the clash of symbols. This differs from Example 1 only by a strictly increasing monotone function, which is why the two update rules differ only by reparameterization. This is also the *Bayesian* objective, as we will see in Section 4.
- The learner in Example 3 is, by definition, an optimizing learner for $Bel(\theta, (x, y)) = -\ell(\theta, x, y)$ to minimize loss.
- In Example 4, the field generated at $K = 0$ is the gradient of $Bel((\hat{x}, \sigma^2), z) = \frac{1}{2}(\hat{x} - z)^2 + \sigma^4$.

Expected-Value Optimizing Learners. Having fixed the geometry on Θ, there is a 1-1 correspondence between optimizing commitment flows (those that satisfy LB4) and loss functions $\mathcal{L} = -Bel_\phi : \Theta \to \mathbb{R}$. One class of such functions stands out as a natural starting point for our investigations: the linear ones $P \mapsto \mathbb{E}_P[V]$, that is, expectations of of random variables $V : W \to \mathbb{R}$. When $W = \{1, \ldots, n\}$, these functions are parameterized by vectors $\phi = V \in \mathbb{R}^n$. So, what learning procedure is induced by linear beliefs?

Proposition 6. *Suppose $\Theta = \Delta W$ and Φ consists of random variables $V : W \to \mathbb{R}$. The flow form of the optimizing learner that has $\mathcal{L} = -Bel(P, V) = \mathbb{E}_P[V]$ is*

$$\mathrm{Boltz}(P, \beta, V)(w) :\propto P(w)\exp(-\beta V(w)).$$

This is also known as the *softmax* distribution (relative to the base measure P) with logits V and temperature $1/\beta$. Intuitively, larger confidence β reflects increasingly certainty in states w that have low potential $V(w)$. Indeed, using Boltz_V to update a distribution P with high confidence ($\beta \to \infty$) conditions P on the minimizer(s) of V. So for this learner, confidence "tempers" the distribution and coincides with the concept of thermodynamic coldness.

Proposition 7. *(a)* Boltz *satisfies L1–5.*
(b) Boltz *updates commute and are invertible iff $\beta < \infty$.*
(c) $\mathrm{Boltz}_{U\oplus V} = \mathrm{Boltz}_{U+V}$.
(d) $\mathrm{Boltz}^{\beta_1}_{V_1} \circ \cdots \circ \mathrm{Boltz}^{\beta_n}_{V_n}(P) = \mathrm{Boltz}(\sum_{i=1}^n \beta_i V_i, 1, P)$.

Observe how well-behaved these learners are: any sequence of observations in any order is equivalent to a single observation of their weighted sum. This property may come at a significant cost, however: learning in brains and artificial neural networks exhibits a recency bias, an effect which is arguably optimal for bounded agents [Wilson, 2014, Fudenberg et al., 2014], or in changing environments.

4 BOLTZMANN AND BAYES

Many believe that "correctly" accounting for confidence in updating (probabilistic) beliefs is a matter of properly applying *Bayes' Rule (BR)*. To some, this simply means that belief updates are given by conditioning (i.e., $Lrn(\mu, \top, \phi) = \mu|\phi$ with the trivial confidence domain $\{\bot, \top\}$), in which case BR is a helpful theorem. Others reject that learning necessarily establishes a proposition in the posterior with certainty (at least as far as one's belief state is concerned); for these people, BR describes the update itself. We now analyze these accounts of Bayesianism within our framework.

Definition 1. Lrn is *Bayesian* iff

(a) belief states correspond to distinct probability distributions over a measurable space $\mathcal{H}$ of hypotheses (i.e., there is an injection $\theta \mapsto P_\theta : \Theta \to \Delta\mathcal{H}$).
(b) there is a measurable space $(\mathcal{X}, \mathcal{A})$ in which every observation ϕ can be viewed as event (i.e., $\mathcal{A} \supseteq \Phi$);
(c) there is a conditional probability (i.e., a Markov kernel) $P(X \mid H) : \mathcal{H} \to \Delta\mathcal{X}$, associating each hypothesis h with a probability measure over $\mathcal{X}$;
(d) there exists $\star \in [\bot, \top]$ such that, for all ϕ and θ, $P_{Lrn^\star_\phi(\theta)}(h) = P_\theta(h)P(\phi|h) / \sum_{h'} P_\theta(h')P(\phi|h')$. □

Item (d) is Bayes' rule, and prescribes posterior the posterior belief "$P(H|\phi)$". Note that ϕ is not an event in the sample space $\mathcal{H}$, but in the space $\mathcal{X}$; we regard it as event in $\mathcal{X} \times \mathcal{H}$ for the purposes of conditioning the joint measure $P(X, H) := P(X|H)P_\theta(H)$. To obtain a new belief state of the same type as the original (i.e., a distribution over $\mathcal{H}$), however, we must also marginalize out $\mathcal{X}$. Thus, apart from its effect on the hypotheses, ϕ is forgotten after the update.

In the special case where $P(X|H)$ is deterministic (i.e., theories are *complete* enough to determine observations), the extended sample space $\mathcal{H} \times \mathcal{X}$ is not meaningfully different from $\mathcal{H}$, and we simply update by conditioning (as in Example 1 with full confidence). At the other end of the spectrum, when $P(X|H)$ has full support, Bayesian updates are characterized by optimizing learners with linear beliefs.

Proposition 8. *Lrn is a Boltzmann learner for a potential $v \geq 0$ if and only if it is Bayesian with $P(\cdot \mid \cdot) > 0$.*

This result may not be surprising to experienced readers, although one direction of the correspondence is more subtle than it might first appear. It also has a significant implication: Bayesian updating corresponds to a very special kind of optimizing learning where degree of belief can be viewed as the expectation of a fixed random variable. This induces significant limitations on how a given belief representation can be used—for example, high confidence updates always lead to the boundary of the probability simplex. This rules out situations like Jeffrey's rule, for which this is not the case. This raises some interesting questions. Is there a generic way to capture all learners with Bayesian updates (with a necessarily much larger belief space)? Alternatively, are some natural learning procedures provably incompatible with the Bayesian frame?

We point out that the use of relative entropy (KL divergence) as the target of optimization (instead of linear expectation) appears to be far more useful in practice (e.g., in Example 3). This starting point leads Richardson and Bao [2024] to an alternate natural derivation of *probabilistic dependency graphs* [Richardson and Halpern, 2021], leading well beyond ordinary probabilistic modeling to capture inconsistency and much of machine learning.

5 CONCLUSION

Metaphorically: if certainty is black and white, then probability allows for shades of gray, and learner's confidence is about *transparency*. The idea is an old one, having been deployed many times before in various contexts; this paper unifies the approaches, providing axiomatic grounding for the concept writ large (L1–5). We have identified the critical aspects of confidence in a very general setting, and related it to probabilistic notions of confidence (e.g., via LB4 and Proposition 8). The resulting framework connects many seemingly different representations of confidence and learning, for an overview of which we invite the reader to revisit Figure 1. We contend that this framework clarifies common points of confusion in literature (see Section 2.3).

There are many examples and applications of this framework. An obvious continuation point—a deeper analysis of which learning functions correspond to which loss functions when Θ is a parametric family of distributions—has already born fruit that we were not able to cover here.

A key question remains open: how should we decide how much confidence to place in an observation? With enough modeling assumptions, there can be a clear answer—such as in Example 4, where the optimal Kalman gain is related to the current uncertainty and the variance of the sensor. However, as illustrated by the discussion in Example 3, one's willingness to be influenced by an observation may not be merely a matter of probabilistic modeling. This makes the question surprisingly profound; we suspect that the search for a good answer will take us far beyond the present scope. Having laid the formal and conceptual foundations, we are eager to report back on these projects in the future.

Acknowledgements

I would like to sincerely thank Joe Halpern, my PhD advisor, who contributed significantly to the introduction of this paper, reading dozens of earlier drafts. The work was supported in part by AFOSR grant FA23862114029, MURI grant W911NF-19-1-0217, ARO grant W911NF-22-1-0061, and NSF grant FMitF-2319186.

References

Ralph Abraham, Jerrold E Marsden, and Tudor Ratiu. *Manifolds, tensor analysis, and applications*, volume 75. Springer Science & Business Media, 2012.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2000.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

Ron Artstein. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313, 2017.

David P Ausubel and Mohamed Youssef. The effect of spaced repetition on meaningful retention. *The Journal of General Psychology*, 73(1):147–150, 1965.

Alex Becker. Tutorial on the kalman filter, 2003. URL https://www.kalmanfilter.net/.

Robert Grover Brown and Patrick YC Hwang. Introduction to random signals and applied kalman filtering: with matlab exercises and solutions. *Introduction to random signals and applied Kalman filtering: with MATLAB exercises and solutions*, 1997.

Nikolai Nikolaevich Chentsov. *Statistical Decision Rules and Optimal Inference*, volume 53. American Mathematical Society, 1982. ISBN 0-8218-4502-0.

Joel E Cohen, Shmuel Friedland, Tosio Kato, and Frank P Kelly. Eigenvalue inequalities for products of matrix exponentials. *Linear Algebra and its Applications*, 45, 1982. Equation (17).

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

Drew Fudenberg, David K Levine, et al. Learning with recency bias. *Proceedings of the National Academy of Sciences*, 111:10826–10829, 2014.

Peter Gardenfors. Imaging and conditionalization. *The Journal of Philosophy*, 79(12):747–760, 1982. ISSN 0022362X. URL http://www.jstor.org/stable/2026039.

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

R. C. Jeffrey. Probable knowledge. In I. Lakatos, editor, *International Colloquium in the Philosophy of Science: The Problem of Inductive Logic*, pages 157–185. North-Holland, Amsterdam, 1968.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

John M Lee. *Smooth Manifolds*. Springer, 2013.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

David Lewis. Probabilities of conditionals and conditional probabilities. In *Ifs*, pages 129–147. Springer, 1976.

Oliver E Richardson and Jialu Bao. Mixture languages, 2024. Principles of Programming Languages (POPL) Workshop: Languages for Inference (LAFI).

Oliver E Richardson and Joseph Y Halpern. Probabilistic dependency graphs. *AAAI '21*, 2021.

Glenn Shafer. *A Mathematical Theory of Evidence*, volume 42. Princeton university press, 1976.

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

Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. *Optimization for machine learning*. MIT press, 2011.

Hale F Trotter. On the product of semi-groups of operators. *Proceedings of the American Mathematical Society*, 10 (4):545–551, 1959.

Andrea Wilson. Bounded memory and biases in information processing. *Econometrica*, 82(6):2257–2294, 2014.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019.

Learning with Confidence (Supplementary Material)

Oliver E. Richardson[1,2]

[1]Computer Science Dept., Université de Montréal, Montréal, Canada
[2]Mila – Quebec AI Institute

A PROOFS OF MAIN RESULTS

We begin with the claims of the main (i.e. numbered) results. For convenience, we repeat the statements of the propositions before proving them.

Proposition 1. *The fractional domain* $[0,1]$ *and the additive domain* $[0,\infty]$ *are isomorphic. Furthermore, the space of isomorphisms between them is in natural bijection with* $(0,\infty)$*. Specifically, for each* $\beta \in (0,\infty)$*, there is an isomorphism* $\varphi_\beta : [0,1] \to [0,\infty]$ *given by* $\varphi_\beta(s) = -\frac{1}{\beta}\log(1-s)$ *with inverse* $\varphi_\beta^{-1}(t) = 1 - e^{-\beta t}$*.*

Proof. Clearly φ_β and φ_β^{-1} are continuously differentiable, and one can verify with a few steps of simple algebra that the two are inverses. In both cases, the only possible wrinkle is the at the point of high confidence, but there are no problems there either, because:

$$\lim_{s\to 1}\varphi_\beta(s) = \frac{1}{\beta}\lim_{s\to 1}\log\left(\frac{1}{1-s}\right) = \infty \qquad \text{and} \qquad \lim_{t\to\infty}\varphi_\beta^{-1}(t) = \lim_{t\to\infty} 1 - e^{-\beta t} = 1.$$

Next, we show that φ_β preserves the structure of the confidence domain. We just saw that φ_β and φ_β^{-1} preserve the top element $\top$ of both confidence domains. It is even more immediate that it preserves the bottom element. It is also easy to see that both functions preserve the order (i.e., are monotonic). For example, $\frac{\mathrm{d}}{\mathrm{d}s}\varphi_\beta(s) = \frac{1}{\beta(1-s)} \geq 0$.

Next we show that φ_β and its inverse preserve independent combination $(*)$. For $a, b \in [0,1]$, we have

$$\begin{aligned}
\varphi_\beta(a * b) &= \varphi_\beta(a + b - ab) \\
&= -\frac{1}{\beta}\log(1 - a - b + ab) \\
&= -\frac{1}{\beta}\log((1-a)(1-b)) \\
&= -\frac{1}{\beta}\log(1-a) - \frac{1}{\beta}\log(1-b) \\
&= \varphi_\beta(a) + \varphi_\beta(b).
\end{aligned}$$

A similar calculation shows, for all $t, u \in [0,\infty]$, that

$$\begin{aligned}
\varphi_\beta^{-1}(t) * \varphi_\beta^{-1}(u) &= 1 - e^{-\beta t} + 1 - e^{-\beta u} - (1 - e^{-\beta t})(1 - e^{-\beta u}) \\
&= 2 - e^{-\beta t} - e^{-\beta u} - 1 + e^{-\beta t} + e^{-\beta u} - e^{-\beta(u+t)} \\
&= 1 - e^{-\beta(u+t)} \\
&= \varphi_\beta^{-1}(u+t).
\end{aligned}$$

Finally, we must show that these are the only isomorphisms between the two confidence domains. For this, we refer to a standard argument that is most directly seen as the solution to Cauchy's exponential functional equation $g(x+y) = g(x)g(y)$ after the change of variables $g = 1 - f$.

A similar argument is provided by Shannon [1948] in defense of entropy, and a much more direct analogue appears in the form we need by Shafer [1976], who shows directly that every continuous mappings of $[0, 1]$ to $[0, \infty]$ for which multiplication becomes addition in this way, must be of the form $s \mapsto -k \log(1 - s)$, for some $k > 0$. □

Proposition 2. *Take $\Theta = \Delta W$ and $\phi \subseteq W$. There exists no continuous function $Lrn_\phi : \Delta W \times [0, 1] \to \Delta W$ with the property that $Lrn_\phi(\mu, 1) = \mu|\phi$ when $\mu(\phi) > 0$.*

Proof. Fix a non-empty subset $\phi \subseteq W$ and consider a function $F : \Delta W \times [0, 1] \to \Delta W$ such that $F(\mu, 0) = \mu$ and $F(\mu, 1) = \mu|\phi$ whenever $\mu(\phi) > 0$. Our aim is to show that F cannot be continuous.

Fix distribution $\mu_0 \in \Delta W$ with the property that $\mu_0(\phi) = 0$. For each $\delta > 0$, consider the set

$$B_\delta(\mu_0) := \{\mu \in \Delta W : \mathrm{TV}(\mu, \mu_0) < \delta\} = \{(1 - \delta)\mu_0 + \delta P\}_{P \in \Delta W}$$

of distributions within δ total variation distance of μ_0. By assumption, $F(-, 1)$ updates by conditioning on ϕ, which means all mass not on ϕ is removed, and the rest is renormalized. More precisely, this means $F((1 - \delta)\mu_0 + \delta P, 1) = P$ for all $\delta \in (0, 1)$, and thus the image of $B_\delta(\mu_0)$ under F is all of ΔW. Therefore, for every $\epsilon \in (0, 1)$, there cannot be $\delta > 0$ such that $\mu \in B_\delta(\mu_0)$ implies $F(\mu, 1) \in B_\epsilon(F(\mu_0, 1))$. Thus F cannot be continuous. □

Proposition 3. *For all $\phi \in \Phi$, there is a maximal open set $\Theta_\phi \subseteq \Theta$ such that the restriction $Lrn_\phi|_{\Theta_\phi} : [\bot, \top) \times \Theta_\phi \to \Theta$ of Lrn_ϕ to Θ_ϕ is continuous.*

Proof. As noted in the main text, the observation ϕ is not mathematically relevant to the argument; to simplify notation, we work with the commitment function $F := Lrn_\phi : \Theta \times [\bot, \top] \to \Theta$. In this context, the belief space Θ and confidence domain $[\bot, \top]$ both implicitly have topologies. Let $\tau \subseteq 2^\Theta$ denote the topology associated with Θ (i.e., the collection of all open subsets of Θ). Given $U \subseteq \Theta$, we use the standard notation $F|_U$ to denote the restriction of the function F to domain $U \times [\bot, \top]$.

By assumption (L2), for each fixed $\theta \in \Theta$, the function $F_\theta : [\bot, \top] \to \Theta$ is continuous. Let

$$\mathcal{U} := \left\{ U \in \tau \;\middle|\; F|_U : U \times [\bot, \top] \to \Theta \text{ is continuous} \right\}$$

be the set of all open subsets of Θ on which the restriction of F is continuous. Since unions of open sets are open, we know that $\Theta_\phi := \bigcup \mathcal{U} \subseteq \Theta$ is open. We now show that it is the maximal open set on which F is continuous, as promised by the theorem.

Recall that a function $f : X \to Y$ is continuous iff the preimage $f^{-1}(V) = \{x \in X : f(x) \in V\}$ of an open set $V \subseteq Y$ is itself an open set. Given $V \subseteq \Theta$, observe that

$$\begin{aligned}
(\theta, \chi) \in (F|_{\Theta_\phi})^{-1}(V) &\iff \exists U \in \mathcal{U}.\ \theta \in U \text{ and } F(\theta, \chi) \in V \\
&\iff \exists U \in \mathcal{U}.\ (\theta, \chi) \in (F|_U)^{-1}(V) \\
&\iff (\theta, \chi) \in \bigcup_{U \in \mathcal{U}} (F|_U)^{-1}(V).
\end{aligned}$$

In other words, we have shown that $(F|_{\Theta_\phi})^{-1}(V) = \bigcup_{U \in \mathcal{U}} (F|_U)^{-1}(V)$.

It follows that the preimage $(F|_{\Theta_\phi})^{-1}(V)$ of an open set $V \subseteq \Theta$ is a union of open sets (since each F_U was assumed to be continuous), and hence itself open. Therefore $F|_{\Theta_\phi}$ is continuous, and since Θ_ϕ contains every other open set satisfying that property, it is the maximal such open set. □

We will return to Theorem 4 in Appendix A.1. Previously, the following result was in the main text, but we no longer believe it important to state formally; we give it again here for completeness, as it still supports the discussion in Section 3.1.

Proposition 9. *If Lrn is a commitment flow and $\phi_1, \phi_2 \in \Phi$, then there is at most one commitment flow $Lrn_{\phi_1 \oplus \phi_2} : [0, \infty] \times \Theta \to \Theta$ such that $Lrn'_{\phi_1 \oplus \phi_2} = Lrn'_{\phi_1} + Lrn'_{\phi_2}$.*

Proof. Most of the work is done by an important result in differential geometry:

Fact 10 (The Fundemental Theorem on Flows)**.** *If $X \in \mathfrak{X}(\Theta)$ is a somoth vector field, then there is a unique function $f : \mathcal{D} \to \Theta$ where $\mathcal{D} \subseteq \mathbb{R} \times \Theta$ is maximal, satisfying $f(a, f(b, \theta)) = f(a + b, \theta)$ whenever $(a + b, \theta) \in \mathcal{D}$, and $\frac{\partial}{\partial t} f(t, \theta)|_{t=0} = X(\theta)$ for all $(t, \theta) \in \mathcal{D}$.*

The statement above is a gloss and selective restatement of the statement of the result as presented by Lee [2013, Theorem 9.12], which inlines the definition of a flow (Equations 9.6 and 9.7). A further alteration: we are interested in a minor variant in which the vector field X and the function of interest are not necessarily smooth (i.e., infinitely differentiable), but rather merely twice differentiable (C^2). As discussed in Appendix C of Lee and more directly treated by Abraham et al. [2012, §4.1], precisely the same techniques suffice to establish the analogous result without assuming smoothness.

Applying the C^k analogue of Fact 10 to the vector field $X = Lrn_{\phi_1 \oplus \phi_2} = Lrn_{\phi_1} + Lrn_{\phi_2}$, we find that there is a unique flow $F : \mathcal{D} \to \Theta$ whose derivative is X and whose domain $\mathcal{D} \subseteq \mathbb{R} \times \Theta$ is maximal. Thus, there is at most one function satisfying L1, L2 and L5, and hence at most one commitment flow. The primary missing piece is that the resulting flow may no longer be *complete*—following the sum of the two fields may "leave" the manifold Θ in finite time, and, even if it stays within the manifold, it may exhibit cyclic behavior, violating L4 or standing in the way of a well-defined continuous completion at the limit $t \to \infty$. □

Proposition 5. *Suppose Lrn_{ϕ_1} and Lrn_{ϕ_2} are commitment flows. For $t \in [0, \infty]$, let $L_t := Lrn^t_{\phi_2} \circ Lrn^t_{\phi_1}$ denote learning ϕ_1 followed by ϕ_2 (both with confidence t), and for $n \in \mathbb{N}$, let $L^{(n)}_t(\theta) := L_t \circ \cdots \circ L_t(\theta)$ denote n repeated applications of L_t. Then $Lrn^\chi_{\phi_1 \oplus \phi_2}(\theta) = \lim_{n \to \infty} L^{(n)}_{\chi/n}(\theta)$.*

Proof. $Lrn^\chi_{\phi_1 \oplus \phi_2}(\theta)$ is, by definition, the result of integrating a vector field from $t = 0$ to $t = \chi$. That integration can be thought of as taking a process of taking (infinitely) many (infinitesimal) sequential steps in the direction of that field.

In the limit as $\epsilon \to 0$,

$$Lrn^\epsilon_{\phi_1 \oplus \phi_2}(\theta_0) = \theta_0 + \epsilon Lrn'_{\phi_1 \oplus \phi_2} = \theta_0 + \epsilon Lrn'_{\phi_1} + \epsilon Lrn'_{\phi_2}$$

can be viewed as a small linear addition to the original position (in any choice of local coordinates). Yet by the same approximation, this is also what results from as an infinitesimal update of Lrn_{ϕ_1} followed by Lrn_{ϕ_2}, which equals $L_\epsilon(\theta)$! As $\epsilon \to 0$, the Euler integration method of the field $Lrn'_{\phi_1 \oplus \phi_2}$ starting at θ from $t = 0$ to $t = \chi$ with step size ϵ, which equals $Lrn^\chi_{\phi_1 \oplus \phi_2}(\theta)$, is actually calculating $\lim_{n \to \infty} L^{(n)}_{\chi/n}(\theta)$. Therefore the two quantities are equal. □

Proposition 6. *Suppose $\Theta = \Delta W$ and Φ consists of random variables $V : W \to \mathbb{R}$. The flow form of the optimizing learner that has $\mathcal{L} = -Bel(P, V) = \mathbb{E}_P[V]$ is*

$$\mathrm{Boltz}(P, \beta, V)(w) :\propto P(w) \exp(-\beta V(w)).$$

Proof. First, we calculate the vector field given by the gradient of $Bel(\mu, V) = \mathbb{E}_\mu[V]$ in the natural (Fisher) geometry for $\Theta = \Delta X$.

$$\begin{aligned} \hat{\nabla}_\mu Bel(\mu, V) &= \hat{\nabla}_\mu \mathop{\mathbb{E}}_\mu[V] \\ &= \mathcal{I}(\mu)^{-1}(\nabla_\mu \mathbb{E}_\mu[V] - \lambda \mathbf{1}) \end{aligned}$$

where λ is the Lagrange multiplier associated with the constraint $g(\mu) = \sum_x \mu(x) - 1 = 0$, which has gradient $\nabla_\mu g(\mu) = \mathbf{1}$. The field is therefore given by

$$\begin{aligned} &= \left[\mu(x) \frac{\partial}{\partial \mu(x)} \mathop{\mathbb{E}}_\mu[V] - \lambda \mu(x) \right]_{x \in X} \\ &= x \mapsto \mu(x)(V(x) - \lambda) \end{aligned}$$

for some constant λ. We can solve for λ with the observation that the result must yield a vector tangent to the probability simplex, i.e., the sum across all components must equal zero; thus $\sum_{x\in X}\mu(x)(V(x)-\lambda)=\mathbb{E}_\mu[V]-\lambda=0$, and so we must have $\lambda=\mathbb{E}_\mu[V]$. Therefore,

$$\begin{aligned}\hat{\nabla}_\mu Bel(\mu,V) &= x\mapsto \mu(x)(V(x)-\mathbb{E}_\mu[V])\\ &= \mu\odot(V-\mathbb{E}_\mu[V]),\end{aligned}$$

where $\odot$ is used to emphasize that it is an element-wise product between vectors.

At the same time, we can calculate the path velocity of the Boltzman update rule. Letting $Z:=\mathbb{E}_\mu[\exp(-\beta V)]$ be the normalization constant, $\frac{\partial Z}{\partial\beta}=\mathbb{E}_\mu\left[\frac{\partial}{\partial\beta}\exp(-\beta V)\right]=\mathbb{E}_\mu[-V\exp(-\beta V)]$. Keeping that in mind, we can calculate:

$$\begin{aligned}\frac{\partial}{\partial\beta}\mathrm{Boltz}[V](\mu,\beta)\Big|_{\beta=0} &= x\mapsto \frac{\partial}{\partial\beta}\Big[\frac{\mu(x)\exp(-\beta V(x))}{\mathbb{E}_\mu[\exp(-\beta V)]}\Big]\\ &= x\mapsto \mu(x)\frac{\partial}{\partial\beta}\Big[\exp(-\beta V(x))\Big]_{\beta=0}+\mu(x)\exp(-\beta V(x))\frac{\partial}{\partial\beta}\Big[\frac{1}{Z}\Big]_{\beta=0}\\ &= x\mapsto \mu(x)\exp(-\beta V(x))\Big(-V(x)+\frac{\partial}{\partial\beta}\Big[\frac{1}{Z}\Big]_{\beta=0}\Big)\Big|_{\beta=0}\\ &= x\mapsto \mu(x)\Big(-V(x)-\frac{1}{Z^2}\frac{\partial Z}{\partial\beta}\Big)\\ &= x\mapsto \mu(x)\Big(-V(x)-\frac{\mathbb{E}_\mu[-V\exp(-\beta V)]}{\mathbb{E}_\mu[\exp(-\beta V)]^2}\Big|_{\beta=0}\Big)\\ &= x\mapsto \mu(x)(-V(x)-\mathbb{E}_\mu[-V])\\ &= \mu\odot(\mathbb{E}_\mu[V]-V).\end{aligned}$$

Since this is the same field as before, Proposition 9 tells us that Boltz_V is the unique flow representation of the optimizing learner with potential $\mathbb{E}_\mu[V]$. □

Proposition 7.

(a) Boltz *satisfies L1–5.*
(b) Boltz *updates commute and are invertible iff* $\beta<\infty$.
(c) $\mathrm{Boltz}_{U\oplus V}=\mathrm{Boltz}_{U+V}$.
(d) $\mathrm{Boltz}^{\beta_1}_{V_1}\circ\cdots\circ\mathrm{Boltz}^{\beta_n}_{V_n}(P)=\mathrm{Boltz}(\sum_{i=1}^n\beta_iV_i,1,P)$.

Proof. (a) L1 and L2 are obvious. L4 follows from the fact that (as shown in Proposition 6), the field is the gradient of a potential, and so it cannot have closed integral curves. L5 is actually part (c), and L3 follows from L5 and the fact that adding numbers makes them larger.

(b) Boltzmann updates commute because

$$\mathrm{Boltz}^{\beta_1}_u\circ\mathrm{Boltz}^{\beta_2}_v(\mu)\propto\mu\exp(-\beta_1u)\exp(-\beta_2v)=\mu\exp(-\beta_2v)\exp(-\beta_1u)\propto\mathrm{Boltz}^{\beta_2}_v\circ\mathrm{Boltz}^{\beta_1}_u(\mu).$$

If $\beta<\infty$, the update Boltz^β_u can be inverted by $\mathrm{Boltz}^\beta k-u$ where k is any constant. If $\beta=\infty$, then it amounts to conditioning, and hence is not invertible.

(c) Adding the vector fields discovered in the proof of Proposition 6,

$$\begin{aligned}\mathrm{Boltz}'_{u\oplus v} &= \mathrm{Boltz}'_u+\mathrm{Boltz}'_v\\ &= \mu\odot(\mathbb{E}_\mu[u]-u)+\mu\odot(\mathbb{E}_\mu[v]-v)\\ &= \mu\odot(\mathbb{E}_\mu[u+v]-(u+v))\\ &= \mathrm{Boltz}'_{u+v}.\end{aligned}$$

□

(d) Slightly generalizing the calculation of part (b):

$$\begin{aligned}\mathrm{Boltz}_{v_1}^{\beta_1} \circ \cdots \circ \mathrm{Boltz}_{v_n}^{\beta_n}(\mu) &\propto \mu \prod_{i=1}^{n} \exp(-\beta_i v_i) \\ &\propto \mu \exp\Big(-\sum_{i=1}^{n} \beta_i v_i\Big) \\ &\propto \mathrm{Boltz}_{\sum_{i=1}^{n} \beta_i v_i}\end{aligned}$$

Proposition 8. *Lrn is a Boltzmann learner for a potential $v \geq 0$ if and only if it is Bayesian with $P(\cdot \mid \cdot) > 0$.*

Proof. One direction is easy: if Lrn is Bayesian with likelihood $P(\cdot \mid \cdot) > 0$, then belief states are probability distributions, and so for $\star := \beta = 1$, a Bayesian update with likelihood $P(X \mid H)$ can be written as

$$\begin{aligned}P_{Lrn(\theta,\star,\phi)}(h) &\propto P_\theta(h) \cdot P(\phi \mid h) \\ &\propto P_\theta(h) \cdot \exp(\log P(\phi \mid h)),\end{aligned}$$

and so coincides with the Boltzmann update with confidence 1 and potential $-\log P(\phi \mid h)$. This simple well-known fact is largely responsible for the prevalence of "tempering" and exponential families in the Bayesian literature. In effect, it just converts between the additive and multiplicative domains.

The opposite direction is less well-known, and considerably less intuitive. We cannot simply invert the construction above, because, owing to the fact that probabilities are constrained to sum to one, not every potential can be obtained by the logarithm of a conditional probability in this way. However, we can circumvent this by choosing a new measurable space $\mathcal{X}$.

Concretely, suppose we are given a potential $u : \Phi \times \mathcal{H} \to [0, \infty)$. In this case, define X to be a variable whose can take on values 2^Φ, and define the likelihood $P(X|h)$ according to:

$$P(X{=}A \mid h) := \prod_{\phi \in A} \exp(-u(\phi, h)) \prod_{\phi \in \bar{A}} (1 - \exp(-u(\phi, h))).$$

It is not hard to see that this implies

$$P(X{\supseteq}A \mid h) = \prod_{\phi \in A} \exp(-u(\phi, h)) = \exp(-\sum_{\phi \in A} u(\phi, h)).$$

By viewing an observation ϕ as the event $X \supseteq \{\phi\}$, we now have an event whose (strictly positive) likelihood corresponds to the potential $u(\phi, -)$. This establishes the reverse direction of the theorem. □

A.1 THE ADDITIVE REPRESENTATION THEOREM

The proof of Theorem 4 is a bit more techncial than the others. We will first need a technical result about differential geometry. In this section we assume that is a continuum (a one-dimensional, totally ordered confidence domain), and that $F : [\bot,\top] \times \Theta \to \Theta$ is a commitment function (satisfying L1–5).

Now a few definitions. A point $p = (\chi, \theta) \in [\bot,\top] \times \Theta$ is called *active* if $\frac{\partial F}{\partial \chi}|_p \neq 0$. p is a *submersion point*, or *submersive*, if $dF|_p : T_p([\bot,\top] \times \Theta) \to T_{F(p)}\Theta$ is surjective. (That is, if F is a submersion at p.)

Lemma 11. *For all $\theta \in \Theta$, if there exists an active point p in the fiber $F^{-1}(\theta)$, then there also exists an active point $\hat{p}$ in the fiber that is a submersion point.*

Proof. For the sake of contradition, suppose otherwise—that $p^* = (\chi^*, \theta_0) \in F^{-1}(\theta)$ is an active point in the fiber $F^{-1}(\theta)$, but no submersion point in the fiber is active (i.e., $\frac{\partial F}{\partial \chi}|_p = 0$).

Select a sequence of strictly increasing confidences $(\chi_n) \in [\bot,\top]^{\mathbb{N}}$ that approach χ^* from below. (So $(\chi_n) \to \chi^*$.) For each n, define $\theta_n := F(\chi_n, \theta_0)$. Since F is continuous, $(\theta_n) \to F(\chi^*, \theta_0) = \theta$. By L3, since $\chi_n < \chi^*$, we are

guaranteed that there exists some $\delta_n \leq \chi^*$ such that $F(\delta_n, F(\chi_n, \theta_0)) = \theta$, which we use to define the sequence $(\delta_n)_{n\in\mathbb{N}}$. Defining $p_n := (\delta_n, \theta_n)$ gives a sequence of points, each lying in the fiber $F^{-1}(\theta)$ owing from the definitions of θ_n and δ_n. Note that $(p_n) \to (\delta_{\lim}, \theta)$. Since $[\bot, \top]$ is homeomorphic to an interval, it is bounded, so by the Bolzano-Weierstrass theorem, (δ_n) has a convergent subsequence; let (δ_m) be such a subsequence limiting to the smallest possible value (i.e., $\lim_{m\to\infty} \delta_m = \liminf_{n\to\infty} \delta_n =: \delta_{\lim}$).

Define also the sequence $(q_n = (\chi_n, \delta_n))_{n\in\mathbb{N}}$. Intuitively, each $q_n = (\chi_n, \delta_n)$ is a different way of splitting up the effective total confidence $\chi_n * \delta_n \cong \chi^*$.

Intuitively, as χ_n approaches χ^*, the remaining residual confidence δ_n required to effectively get there should decrease to $\bot$. Indeed,

$$\lim_{n\to\infty} \delta_n = \lim_{n\to\infty} s(\chi^*, \chi_n) = s(\chi^*, \lim_{n\to\infty} \chi_n) = s(\chi^*, \chi^*) = \bot.$$

The point $p_\bot = (\bot, \theta)$, which is obviously in the fiber $F^{-1}(\theta)$, is a submersion point—since $F(\bot, \cdot) = \mathrm{id}_\Theta$ is the identity map on Θ, it follows that $\frac{\partial F}{\partial \theta}|_{p_\bot}$ is the identity map on $T_\theta\Theta$ (i.e., the identity matrix in any coordinate representation). This is a sufficient condition for the differential of F to be surjective at this point, even if the derivative with respect to χ is zero. Furthermore, since the set of invertable matrices is open and F is C^1 along the line $\{\bot\} \times \Theta$, it follows that any point sufficiently close to that line (i.e., with small enough value of χ) will be a submersion point as well.

Define the function $H(\chi, \delta) := F(\delta, F(\chi, \theta_0)) : [\bot, \top]^2 \to \Theta$, whose utility we will see shortly. The level set $H^{-1}(\theta)$ consists of confidence pairs (χ, δ) for which $F(\delta * \chi, \theta_0) = \theta$ for which sequential application leads to our target. At the point p_n, what direction keeps us within this set? Taking the differential of H at the point p_n, by the chain rule, we find:

$$dH|_{p_n}(v) = v_\delta \Big(\frac{\partial F}{\partial \chi}(\delta_n, \theta_n)\Big) + v_\chi \Big(\frac{\partial F}{\partial \theta}(\delta_n, \theta_n)\frac{\partial F}{\partial \chi}(\chi_n, \theta_0)\Big), \tag{3}$$

for a vector $v = v_\delta \frac{\partial}{\partial \delta} + v_\chi \frac{\partial}{\partial \chi} \in T_{p_n}[\bot, \top]^2$ tangent to p_n. We are looking for vectors in the kernel of $dH|_{p_n}$ (i.e., for which $dH_{p_n}(v) = 0$); these are the ones that lie tangent to the level set of interest.[1] Remarkably, this relates the conditions of activeness and submersiveness at p_n to activeness at the point p^*, which was guaranteed by assumption!

- By our assumption that $p^* = (\chi^*, \theta_0)$ is active, the derivative $\frac{\partial F}{\partial \chi}|_{(\chi^*, \theta_0)} =: v^*$ exists and is a nonzero tangent vector; moreover, that nonzero value is the limit of the sequence $(\frac{\partial F}{\partial \chi}(\chi_n, \theta_0))_{n\in\mathbb{N}}$. Therefore, for $\epsilon > 0$ there exists an integer N_1 for which $\frac{\partial F}{\partial \chi}(\chi_n, \theta_0)$ is within ϵ of v^* (for any choice of coordinates) for all $n > N_1$.
- Since $\delta_{\lim} = \bot$, we know that $(p_n) = (\delta_n, \theta_n) \to (\bot, \theta)$. Therefore, there exists an integer N_2 for which $n > N_2$ implies p_n is in the a neighborhood of $p_\bot$ where $\frac{\partial F}{\partial \theta}$ is within ϵ of the identity matrix (say for the same choice of coordinates and ϵ) and in particular invertible. Therefore, p_n is submersive; since we assumed for contradiction that there are no active submersive points in the fiber, we must conclude that $\frac{\partial F}{\partial \chi}(p_n) = \frac{\partial F}{\partial \chi}(\delta_n, \theta_n) = 0$. So the first term of (3) is zero.

From these two observations, we deduce that, for all $n > \max(N_1, N_2)$, the quantity $w_n := \frac{\partial F}{\partial \theta}(p_n)\frac{\partial F}{\partial \chi}(\chi_n, \theta_0)$ on the right side of (3), is the product of an invertable matrix (whose trace is bounded away from zero) and a vector bounded away from zero, and hence itself a vector w_n bounded away from zero. This forces $v_\chi = 0$. Furthermore, this same line of reasoning applies not only for the points p_n and p_{n+2}, but for the entire curve they lie on. Parameterizing this curve as a path $\gamma(t)$ along this curve starting at p_n and ending at p_{n+2}, we find that the kernel of $dH|_{\gamma(t)}$ has a zero χ-component for all t along this segment. Thus the curve $\gamma(t)$ must have zero derivative in its first component (χ), and χ must be constant along it. And yet $\chi_n < \chi_{n+1} < \chi_{n+2}$ are strictly increasing coordinates! This is a contradiction. □

Theorem 4. *If $[\bot, \top]$ is a continuum and $F : [\bot, \top] \times \Theta \to \Theta$ is a commitment function (i.e., satisfies L1–5), then there exists a continuous "translation" function $g : [\bot, \top] \times \Theta \to [0, \infty]$, and a commitment flow ${}^+F$ such that $\forall \theta, \chi.\ {}^+F(g(\chi, \theta), \theta) = F(\chi, \theta)$.*

[1] In more detail: since this differential has constant rank at a neighborhood of the limiting point (as we are about to show), the points lie on a smooth sub-manifold, by the constant rank level subset theorem. That submanifold is a one-dimensional curve the primary argument in the proof of Theorem 4—from L2, L3 and L5, it follows that all $\frac{\partial F}{\partial \chi}$.

Proof. For each $\theta \in \Theta$, let

$$Dir(\theta) := \left\{ \frac{\partial}{\partial \chi} F(\chi, \theta_0) : \theta_0 \in \Theta, \chi \in [\bot, \top], F(\chi, \theta_0) = \theta \right\} \subseteq T_\theta \Theta$$

be the tangent subspace at θ spanned by derivatives of F at various starting points. The key to proving the theorem is to show that the elements of $Dir(\theta)$ are all parallel and oriented the same direction; this will allow us to use it to define a vector field which locally captures updating with F (up to re-scaling) regardless of the "original" starting belief state θ_0. At this point, we can recover an additive representation from the integral curves of this vector field.

Suppose (χ_1, θ_1) and (χ_2, θ_2) are such that $F(\chi_1, \theta_1) = F(\chi_2, \theta_2) = \theta$. To show that the corresponding directions in $Dir(\theta)$ are parallel, it suffices to show that the sub-tangent spaces of $T_\theta \Theta$ generated by infinitesimal perturbations of χ_1 and χ_2, respectively, are the same. For all $\chi_1' > \chi_1$, we know (by L3) that

$$\exists \tilde{\chi}_1.\ F(\chi_1', \theta_1) = F(\tilde{\chi}_1, F(\chi_1, \theta_1)) = F(\tilde{\chi}_1, F(\chi_2, \theta_2)).$$

Thus, for all $\chi_1' > \chi_1$, there exists some $\chi_2' := \tilde{\chi}_1 * \chi_2 \geq \chi_2$ such that $F(\chi_2', \theta_2) = F(\chi_1', \theta_1)$. Symmetrically, for all $\chi_2' > \chi_2$, there exists a corresponding $\chi_1' \geq \chi_1$ with the same property. In particular, this is true for χ_1' and χ_2' that are infinitesimally close to χ_1 and χ_2, and thus the ray in the tangent space $T_\theta \Theta$ generated by positive perturbations of χ_1 and χ_2 are the same (if nonzero). Formally speaking, this argument establishes that either

$$\{dF(v, \theta_1) : v \in T_{\chi_1}[\bot, \top]\} = \{dF(v, \theta_2) : v \in T_{\chi_2}[\bot, \top]\},$$
$$\text{or one of the two equals the singleton } \{\mathbf{0}\}.$$

(Recall that $T_\chi[\bot, \top]$ is the tangent space at $\chi \in [\bot, \top]$, and has the same dimension as $[\bot, \top]$.) It follows that the dimension of $\operatorname{span}(Dir(\theta))$ is at most the dimension of the confidence domain $[\bot, \top]$ itself—and since that domain was assumed to be one-dimensional, we have shown that $\dim \operatorname{span}(Dir(\theta))$ is equal either to one or to zero. Moreover, we have shown that all (nonzero) tangent vectors in $Dir(\theta)$ point in the same direction.

Define a vector field $X(\theta)$ by a continuous selection from $Dir(\theta)$ that is nonzero whenever $Dir(\theta)$. Such a continuous selection exists because F itself is twice continuously differentiable (C^2) when restricted to Θ_ϕ.

For each point θ: if $Dir(\theta) \neq \{\mathbf{0}\}$, then select any $(\theta_0, \chi) \in F^{-1}(\theta)$ for which $\frac{\partial}{\partial \chi} F(\theta_0, \chi) \neq 0$. Applying Lemma 11, this guarantees the existence of an active submersion point $\hat{p}$; in turn, by the submersion theorem, this guarantees the existence of a C^1 local section $\sigma_\theta : U_\theta \to [\bot, \top] \times \Theta$ on some neighborhood $U_\theta \ni \theta$. We then define a local vector field on U_θ according to $Y_\theta(\theta') := \frac{\partial F}{\partial \chi}(\sigma(\theta'))$. Since $\{U_\theta\}_{\theta \in \Theta}$ is an open cover of Θ, we know there exists a partition of unity $R = \{\rho_\theta : U_\theta \to [0, 1]\}$ subordinate to it—meaning that this indexed family has the following properties [Lee, 2013, Thm 2.23]:

1. for all $\theta \in \Theta$, $\rho_\theta(\theta') = 0$ when $\theta' \notin U_\theta$.
2. every point $\theta' \in \Theta$ has a neighborhood that intersects the support of ρ_θ for only finitely many values of θ.
3. $\forall \theta' \in \Theta. \sum_\theta \rho_\theta(\theta') = 1$.

Finally, this allows us to define our vector field as

$$X(\theta') := \sum_{\theta \in \Theta} \rho_\theta(\theta') Y_\theta(\theta). \tag{4}$$

This is continuous because each Y_theta is smooth, and only finitely many terms ρ_θ are nonzero.

For $\theta \in \Theta$ and any vector field $V \in \mathfrak{X}(\Theta)$, we use the standard notation $\exp_\theta(V) := y(1)$ for the unique solution to the differential equation $\frac{\mathrm{d}y}{\mathrm{d}t} = V(y)$ with initial condition $y(0) = y_0$, evaluated at $t = 1$. By the rescaling lemma [e.g., Lee, 2013, Lemma 9.3], $\exp_\theta(tV) = y(t)$ is the result of starting at θ and following the vector field V for time $t \geq 0$. Since scaling a vector field by a positive scalar field results in the same (or truncated) integral curves after reparameterization, for all $\theta \in \Theta$ and $\chi \in [\bot, \top]$, there exists some $t_{(\theta,\chi)} \in [0, \infty]$ such that $\exp_\theta(t_{(\theta,\chi)} X) = F(\chi, \theta)$.

With these definitions in place, we define ${}^+F(t, \theta) := \exp_\theta(tX)$ for $t \in [0, \infty]$, and $g(\chi, \theta) := t_{(\theta,\chi)}$. □

B DEFERED CALCULATIONS AND FURTHER RESULTS

Beyond the main numbered results of the paper, we have also deferred a few minor calculations to the appendix.

Kalman Combinativity. We claim that pair (K, r^2) forms a confidence domain. With some simple algebra, one can show that the sequence of updates $(K_2, r_2^2) * (K_1, r_1^2)$ is equivalent to a single update with (K_3, r_3^2), where $K_3 = K_1 + K_2 - K_1 K_2$ just as in example 1 and the other examples using the $[0, 1]$ domian, and

$$r_3^2 = \frac{K_2^2 r_2^2 + K_1^2 (1 - K_2)^2 r_1^2}{(K_1 + K_2 - K_1 K_2)^2}.$$

This is the only non-commutative example we have given. In the case where K is chosen optimally, this reduces to a single domain with inverse variance combining additively.

Proposition 12. *If $F : [\bot, \top] \times \Theta \to \Theta$ satisfies L1 and L2, then there exists another commitment function $^{::}F$ (also for beliefs Θ on observations Φ), that accepts confidences in an extended domain $[\bot, \top]' \supseteq [\bot, \top]$, has the same behavior as F when restricted to the orginal confidence domain, and in addition satisfies all axioms L1–5.*

Proof. Consider the new confidence domain

$$\left\{\text{finite lists } [c_1, \ldots, c_n] \text{ with each } c_i \in [\bot, \top], \quad \leqslant \quad :: , \quad [] , \quad [\top] , \quad \mathfrak{g}' \right\}, \quad \text{where}$$

- The operation "::" is list concatenation, except that it collapses instances of $\top$, i.e.,

$$[c_1, \ldots c_n] :: [d_1, \ldots, d_m] := \begin{cases} [\top] & \text{if } \top \in \{c_1, \ldots, c_n, d_1, \ldots, d_m\} \\ [c_1, \ldots, c_n, d_1, \ldots, d_m] & \text{otherwise.} \end{cases}$$

  Concatenating the empty list $[]$ on either side has no effect, by construction, for all $L \in [\bot, \top]'$, we have $[\top] :: L = [\top] = L :: [\top]$, and :: is clearly associative, so $[\bot, \top]'$ is also a confidence domain.
- The order is given by the prefix ordering: $[c_1, \ldots, c_n] \leqslant [d_1, \ldots, d_m]$ iff $n \leq m$ with $d_i = c_i$ for all $i \in \{0, \ldots, n-1\}$ and $c_i \leq d_i$ if $n \geq 1$.
- The geometry $\mathfrak{g}'$ is given through the appropriate disjoint sum of product topologies and differentiabl structures, so they are non-interacting discrete components.

The new update rule for this confidence is given by:

$$^{::}F([c_1, \ldots, c_n], \theta) := (F^{c_n} \circ \cdots \circ F^{c_1})(\theta).$$

$^{::}F$ has the same behavior as F on the elements that correspond to the original confidence domain, since $^{::}F(c, \theta) = F(c, \theta)$, when $c \in [\bot, \top]$ is a member of the original domain, and it satisfies L5 by construction, since

$$\begin{aligned} {}^{::}F_\phi^{[c_1,\ldots,c_n]}({}^{::}F_\phi^{[d_1,\ldots,d_m]}(\theta)) &:= F_\phi^{d_m} \circ \cdots \circ F_\phi^{d_1}(F_\phi^{c_n} \circ \cdots \circ F_\phi^{c_1}(\theta)) \\ &= (F_\phi^{d_m} \circ \cdots \circ F_\phi^{d_1} \circ F_\phi^{c_n} \circ \cdots \circ F_\phi^{c_1})(\theta) \\ &= {}^{::}F_\phi^{[c_1,\ldots,c_n,d_1,\ldots,d_m]}(\theta) \\ &= {}^{::}F_\phi^{[c_1,\ldots,c_n]::[d_1,\ldots,d_m]}(\theta). \end{aligned}$$

Clearly it satisfies L4. Finally, for L3, define subtraction either at the final element (if the final element is greater than the number subtracted) or by ablating elements of the list from the right. This satisfies L3. □