Updating With Confidence

Oliver E Richardson, Joseph Y Halpern

July 19, 2022

Contents

1 Introduction

The ability articulate a "degree of confidence" is an important aspect of knowledge representation. Of course, there are many well-established ways of representing uncertainty, probability chief among them. Indeed, an informal poll of our colleagues suggests that most computer scientists view "confidence" as a synonym for probability. Although this use of the word is perfectly reasonable, it seems to have shadowed another conception of confidence—one that is fundementally different, if at first sublty so.

Why not use 0-1, and make things easier for the reader …

For us, confidence is a measure of trust. Like probability, it is a scale between two extremes. While probability ranges from untenable (0) to undeniable (1), confidence ranges from completely untrustworthy ($\bot$) to fully trusted ($\top$). This paper explores how confidence works in the context of learning. Why learning? In this setting, one has some belief state, and recieves inputs, which one might have some degree of confidence in, which is used to modify one's belief state. Our use of confidence can be viewed as a measure of how seriously to take an input in updating our beliefs.

High confidence is in many ways like high probability: if we really trust a statement A, we should fully incorporate it into our beliefs, and thereby come to believe it with high probability. Similarly, it only makes sense to be extremely confident in a A if you believe that A is extremely likely to be true. Low confidence, on the other hand, is quite different from low probability. If we have little trust in A, we should *ignore* A, rather than coming to believe that A is unlikely. For example, if an adversary tells you something that you happen to already believe, you might say you have low confidence in their statement, but nevertheless ascribe it high probability.

1.1 Other Conceptions of Confidence.

You need to think about how to make a story out of this section

Probability. Some people do use "confidence" to mean the same thing as probability. When they say they have low confidence in ϕ, they mean that they think ϕ is unlikely.

One of the biggest shortcomings of probability is its inability to represent a truly neutral attitude towards a proposition. A value of $\frac{1}{2}$ may be equally far from zero as it is from one, but is by no means a neutral assessment in all cases: hearing that your favored candidate has a 50% chance of winning is big news if a win was previously thought to be inevitable. For this reason, telling someone the odds are 50/50 is quite different from saying you have no idea. By contrast, zero confidence represents something truly neutral: a statement made with zero confidence does not stake out a claim, and a statement recieved with zero confidence does not affect the recipient's beliefs. Nevertheless, in some contexts, we will see that confidences correspond to to probabilities.

Opacity. To use a graphical metaphor, think of certainty as black or white. Probability describes shades of gray, while confidence describes opacity. If we are painting with black and start with a white canvas, there is a precise correspondence between the opacity and the resulting shade of gray.

misplaced

Upper and Lower Probabilities. Upper and lower probabilities can describe a neutral attitude towards a proposition, but they are not really a specification of trust, but rather a direct specification of a belief state. It isn't immediately clear how to use these representations of uncertainty to update, and they're a little too complex to function effectively as the primitive measure of trust that we're after.

What's the story here? These see like random thoughts about upper/lower probability

Shafer's Weight of Evidence. Shafer's "weight of evidence" is precisely the same concept we have in mind. Our analysis precsely reduces to his, in the setting where belief states are Belief functions (which generalize probabilities, but not, say, neural network weights), and observations are events. Thus, this paper can be viewed as generalizing this concept to a broader class of settings, without requiring that one adopt Shafer's conception of a belief state or an observation.

Variance and Entropy. The inverse of variance, sometimes known as precision, is also commonly used to measure confidence. If a sensor is unreliable and can give a range of answers, the variance of

references?

so why mention precision?

the estimate is a very common way of quantifying this reliablility. If measurements have zero variance, in some sense one has absolute confidence ($\top$) in the sensor. If measurements have infinite variance, then in some sense one has no confidence in the sensor, since individual samples convey no information about the true value of the quantity measured. As with probability, inverse variance will coincide with confidence in some settings; we will see how in **??**.

not clear. What if the mean is finite?

Entropy, like variance, is a standard way of measuring uncertainty, and in some settings, confidence coincides with entropy (see **??**). The assumption underlying both approaches is that there's some "true" value of the variable, and that the randomness is epsistemic (due to sensor errors) rather than aleotoric (inherrent in the quantity being measured).

Confidence Intervals and Error Bars. Another notion of the word "confidence" comes from the term "confidence interval". This concept arises in settings involving a probability distribution $\Pr(X)$ over a metric space X, typically $X = \mathbb{R}$. A 95% confidence interval is the (largest) ball containing 95% of the probability, and its size is a geometric measurement of how . This intuition behind this reading of the word confidence is the same as

You also need to add a story for our conception of confidence

2 A Formalism for Updating

Suppose you you have some belief state $\theta \in \Theta$. It might be a probability distribution, a parameter setting to some parameteric family of probabilities, a set of probabilities, a Dempster-Shafer belief function, or even just a setting of weights for a neural network. Now, suppose you recieve some input information $\phi \in \Phi$, such as an event, or a sample from a dataset, which you may use to update your belief state. Before we get to confidence, let's first see how far we can get without it. The first thing we need to assume is that the updating process can be captured in functional form.

This is a huge assumption (which I agree is one we need). You *must* discuss how reasonable it is and when it might apply.

F. There exists a function $F : \Phi \to (\Theta \to \Theta)$, which, given the old belief state θ and new information ϕ, produces the belief state $F_\phi\theta$ that corresponds to the result of observing ϕ in state θ.

Given such a function F and a statement ϕ, we call $F_\phi : \Theta \to \Theta$ an *update*. If purpose of F is to *fully* incorporate the new information into our beliefs (and without a notion of confidence, we can't very well specify how much to incorporate), then two successive updates with the same information ought to have the same effect as a single one. Intuitively, this is because if we have just *fully* updated our beliefs to be consistent with the information ϕ, then a second observation of ϕ will require no further alterations of our belief state. In this case, we call F an *update rule*, or more precisely, a Θ-*update rule on* Φ, and insist that

Again, you *must* discuss how reasonable this is. Maybe your belief state isn't rich enough to fully incorporate the information.

This is misplaced. We can do much of what we're doing without this assumption, and it's a major assumption.

UR. For all $\phi \in \Phi$, the update F_ϕ is idempotent (i.e., $F_\phi \circ F_\phi = F_\phi$).

Once Θ, Φ, and any implicit structure in them is specified, there is typically a natural choice of update rule. To illustrate, we now consider three different probability-update rules for different choices of Φ. In each case, the possible belief states $\Theta := \Delta W$ be the set of all probability distributions over a finie set $W = \{w_1, \ldots, w_n\}$ of "possible worlds".

1. **Conditioning.** First, consider the case where observations are events, i.e., $\Phi := 2^W$. Here, the appropriate rule seems to be conditioning:

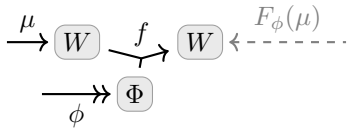
$$(-)\Big|(\,\cdot\,) : \qquad \begin{aligned} 2^W &\to (\Delta W \to \Delta W) \\ A &\mapsto (\ \mu \ \mapsto \mu \mid A\ ), \end{aligned}$$

where the conditional measure $\mu \mid A$ is given by $(\mu \mid A)(B) = \mu(B \cap A)/\mu(A)$, provided $\mu(A) > 0$, and otherwise is just equal to μ. Observe:

- Provided $\mu(A) > 0$, then $(\mu \mid A) \mid A = \mu \mid A$, so conditioning is an update rule.
- If $\mu(A \cap B) > 0$, then $(\mu \mid A) \mid B = \mu \mid (A \cap B) = (\mu \mid B) \mid A$, so the order that information is recieved does not matter, so long as that information is consistent with one's beliefs.

2. **Imaging.** A second example of an update rule is the "imaging" approach of David Lewis (Lewis 1976). Suppose that, for some set Φ, that we already have a W-update rule $f : \Phi \to (W \to W)$, which we interpret as assigning, to each statement $\phi \in \Phi$ and $w \in W$, a unique world $f_\phi w \in W$ which is the world "most similar to w, in which ϕ is true" (Gardenfors 1982). In this case, idempotence of f_ϕ amounts to the (very reasonable) requirement that the world most similar to $f_\phi w$ in which ϕ is true, is $f_\phi w$ itself. From f, we can construct a ΔW-update rule by

$$\begin{aligned} F_\phi(\mu) &:= f^\sharp(\mu) \\ &= A \mapsto \mu(\{w : f(w) \in A\}) \end{aligned}$$



which intuitively moves the probability mass on each world w to the $f_\phi w$, the closest world to w in which ϕ is true. And, since f is idempotent, F will be as well.

3. **Jeffrey's Rule.** Next, consider a more general form of observation, in which observations themselves are probabilities. Formally, suppose Φ consists of distributions $\pi(X)$, where $X : W \to S$ is a random variable, and $\pi \in \Delta S$ is a distribution over the possible values that X can take. Jeffrey's rule, given by

$$\begin{aligned} \mathrm{J}_{\pi(X)}(\mu) &:= \sum_{x \in S} \pi(X = x)\ \mu \,\Big|\, \{w : X(w) = x\} \\ &= A \mapsto \sum_{x \in S} \pi(X = x)\, \mu(A \mid X = x) \end{aligned}$$

When $\pi(X) = \delta_x$ is a point mass $X = x$, then Jeffrey's Rule simply conditions on the event $X = x$. For this reason, Jeffrey's Rule is sometimes seen as a way of making an update with uncertainty (i.e., a "low-confidence" update), but as we will see, it instead is perhaps better thought of as a high-confidence update on a more expresive class of observations.

Note that if $\mu' := J_{\pi(X)}\mu$ is the result of applying Jeffrey's rule to $\pi(X)$ and μ, then $\mu'(X) = \pi(X)$, so $\pi(X)$ has been fully incorporated into μ', and the old beliefs $\mu(X)$ about X have been completely destroyed by the update.

If we take a step back, fully incorporating information is really quite extreme. An agent that updates with conditioning for instance, is forever committed to fully believing A, and consequently, learns nothing from observing A agin in the future. Humans don't work this way. The effectiveness of flash cards as a learning tool demonstrates this clearly: if we were using an update rule, two cycles through a deck of flash cards would be no different from one. Similarly, artificial neural networks are trained with many incremental updates, and cycle through training data more than once. Indeed, this is one biggest differences between modern machine learning algorithms and the older rule-based ones: the new ones update parameters little-by-little, rather than fully incorporating input information. How shall we alter our picture to account for less extreme belief alterations, in which information is only partially incorporated? This is where confidence comes in.

Let $\mathcal{C}$ be the set of possible confidences, which, for now, we will take to be the interval $[\bot := 0, \top := 1]$. We are now in a position to take confidence into account in our updates. As before, our first axiom is that we can capture the updating process in functional form.

CF0. There exists some function

$$F : \mathcal{C} \to (\Phi \to (\Theta \to \Theta))$$

which, given a confidence and new information ϕ, in addition to a prior belief state θ, produces the belief state $F^c_\phi\theta$ that corresponds to the result of observing ϕ in state θ.

Historically speaking, CF0 has not proved as anodyne as it looks. Some might object that it's not possible to write such a function that is appropriate in all circumstances. For example, Shafer argues for Dempster's rule of combination as a way of incorporating information, but is very careful to emphasize that it ought to be used only on *independent* information, for reasons illustrated below.

Example 1. You have initial belief state θ_0. Now, someone comes up to you and tells you that ϕ is true, a statement that you trust to some intermediate degree of confidence $c \notin \{\bot, \top\}$. So, in accordance with CF0, you use F to transform your beliefs, partially incorporating the information to arrive at some belief state $\theta_1 := F^c_\phi(\theta_0)$. Immediately afterwards, your friend repeats what they just said: ϕ is true. Your confidence in the statement remains the same, and so according to CF0, you again update your beliefs, arriving at $\theta_2 := F^c_\phi(\theta_1)$. Except in very special circumstances (e.g., you already know that ϕ is true, or $c \in \{\bot, \top\}$), typically θ_2. And yet, it seems your your attitude towards ϕ ought to be the same whether you've heard it twice or only once. △

Now, it's important to mention that we're not quite in the same position as Shafer. Shafer was prescribing a concrete representation of Θ (a belief function) and a concrete update rule F (Dempster's rule of combination), and so he needed to defend these choices. We only need to defend something much more modest: we only need to defend the assumption that, if Θ and Φ properly model the relevant aspects of the scenario at hand, then there exists *some* function F which performs updates appropriately. Descriptively speaking, we're also in good shape: for synthetic agents, it suffices to point out that learning algorithms represent functions, which given a state, an input, and a number of iterations (confidence), produce an output. And, supposing that Θ, Φ, and C all capture the relevant respective aspects of a human's belief state, input information, and attitude towards it, how could it be that a human does otherwise? In any case, keeping Example 1 in mind, here are three ways to proceed.

The \Theta's that we'll thypically use apparently *don't* model all the relevant aspects.

I1. Accept Severe Limiations. Like Shafer, we could be careful to claim nothing about the belief updating process except in the (unusual) case where information recieved is independent. This would be a severe limiation to the theory, and much less necessary than it was for Shafer. Imagine that we are writing code that describes how a synthetic agent updates its beliefs. Shafer's approach is to package any such code with a warning against running it unless assured that observations will always be independent. But independece is notoriously difficult to establish; are we to simply accept that the code will not behave correctly in any realistic scenario?

In practice, many theoretical properties of standard statistical learning algorithms are heavily dependent on indepencence assumptions (most commonly, that one recieves independent, identically distributed samples). This warning label not seem to keep them from being applied in settings where practitioners readily admit samples are not really independent at all—nor indeed performing well emperically in those settings (**???**).

That fact that they are applied without worrying about the warning is not necessarily a good thing!

I2. Appropriately Enrich Domains. In Example 1, it seems obvious that we ought to ignore the second copy of the information, because it has already been accounted for. However, this intuition is highly contingent on the implicit supposition that we *know* the second input to be a replica of the first. Were we ignorant to the nature of the second piece of information, perhaps it would not be so unreasonable to incroproate it again, even without a proof of independence. So, if we would like our agent to make the same decisions that we did, it seems only fair to give it access to the knowledge that we needed to get there. One way of doing this is to extend the belief state so that it also tracks what information has been incorporated.

For Example 1 to work, it is critical that we are able to discern that the two inputs were identical. As a result, it seems that the relevant description of the input information was not just ϕ, but a pair (ϕ, id) that also a description of its identity. It is also critical that we remember the identity

of previously incorporated information, so we would also be better off with a belief space Θ reflects this. With these two modifications, any commitment function can be straightforwardly modified to avoid the issue in Example 1.

And if you do so, you lose functionality.

We submit that it is always possible to enrich the space of beliefs and observations in this way to track the relevant information, to resolve the issue. With a few more assumptions later on, we will be able to formalize the construction we just alluded to (Example 7).

I3. An Incremental Interpretation of Confidence. Finally, we can get around the issue by interpreting a confidence $c \in \mathcal{C}$ not as an absolute measurement of confidence, but rather an incremental one. This means viewing $c \in \mathcal{C}$ as the degree of *additional* confidence we have in ϕ, beyond whatever we have already incorporated into our beliefs.

This proposal might be concerning. One might worry that it's harder to make sense of "incremental confidence" than an absolute notion. How ought we to numerically describe the confidence of an update? Suddenly this becomes much more subjective, for to assign a number, not only must we describe how much trust we have in the new information, but we must also take history or current belief state into account. Furthermore, the words "incremental" and "additional" suggest that we will need a formal description of how to aggregate confidences—the very concept of which we will need to defend.

Even modulo these concerns, the incremental interpretation still leaves us in a strictly better place than we were before. To begin, in situations where inputs are independent (i.e., the only cases where we would have been allowed to apply the commitment function according to Item I1), the two notions coincide. More explicitly: if the new information ϕ is indepenendent of everything we've previously seen, then an absolute measurement of our confidence in it is no different from a measrurement of how much we ought to increment it from having no confidence. Already, though, we can do more. In the situation described by Example 1, for instance, the second utterance induce no *additional* confidence ($\bot$), and so applying F with no confidence clearly gives the desired result of ignoring the new information (per CF1). And even in general, the prospect of having to numerically estimate a fuzzy quantity seems more promising than red tape requiring that F only be used (in good conscience) on independent information.

Given a confidence $c \in \mathcal{C}$ and a statement $\phi \in \Phi$, we write $F^c_\phi : \Theta \to \Theta$ for the update prescribed by the commitment function F. Furthermore, we will insist that commitment functions respect our interpretation of confidence at the two extremes.

CF1. For all $\theta \in \Theta$ and $\phi \in \Phi$, $F^\bot_\phi(\theta) = \theta$. **(neutrality)**

CF2. For all ϕ, $F^\top_\phi : \Theta \to \Theta$ is an idempotent update.
Equivalently, $F^\top : \Phi \to (\Theta \to \Theta)$ is an update rule. **(certainty)**
We call F a *refinement* of the update rule $F^\top$.

CF1 captures the intuition that we should ignore information in which we have no confidence, while CF2 formalizes the intuition that a full-confidence updates act as we imagined. At this point, we would like to point out that those who find CF1 reasonable have implicitly either accepted either Item I1 or Item I3.

Example 2. Suppose you first hear ϕ from a partially trusted source, and incroproate it into your beliefs appropriately. Then, the same source sends you a second message, which is obviously spam. In an absolute sense, you now have no confidence ($\bot$) in anything this source tells you, including (in retrospect) both messages. It seems appropriate to excise ϕ from your belief state in response, rather than leaving your belief state unchanged, as CF1 would prescribe.

Note that in this scenario, while it seems that we ultimately have no confidence in ϕ, it does not seem

to be the case that we have no incremental confidence in ϕ. Rather, the incremental confidence seems to be the inverse of the original confidence. △

From this point forward, we use the incremental interpretation of confidence, with the understanding that it also admits a more conservative reading (in that it is less widely applicable), in which confidence is measured absolutely, and also all applications of the function F are independent.

(depricated)

and so for most of this paper, we take $\mathcal{C} := \mathbb{R}_+$ to be the group of extended nonnegative real numbers under addition. With this choice of confidence domain, CF8 begins to have more bite, although, as we will see, the effect is more to pin down a coherent system of measurement, and does not appear to restrict modeling expressivness.

Here are some more abstract examples of commitment functions, with confidence domain $\mathcal{C} := \mathbb{R}_+$.

1. Once again, suppose W is a finite set, $\Theta := \Delta W$, and $\Phi := 2^W$. Here are two natural commitment functions for this scenario, both of which are refinements of conditioning.
   - $(F1^c_A\mu)(B) = (1 - e^{-c})\mu(B|A) + e^{-c}\mu(B)$
   - $(F2^c_A\mu)(B) \propto \mu(B|A)^{(1-e^{-c})}\mu(B)^{e^{-c}}$

   The first commitment function, $F1$, linearly interpolates between the result of ignoring the information contained in the event A (i.e., leaving the belief state μ unchanged) and conditioning on A. By contrast, $F2$ does a similar interpolation, but multiplicatively.
2. Suppose that Θ is the set of possible parameter settings for a neural network, which aims to predict an element of $Y \subset \mathbb{R}^m$ given an in put from $X \in \mathbb{R}^n$. So, for each $\theta \in \Theta$, we have a function $f_\theta : X \to Y$, and for fixed $x \in X$, the function $\theta \mapsto f_\theta(x) : \Theta$ is differentiable.
3. Again consider a finite set W and suppose Θ consists of all Dempster-Shafer belief functions

We are particularly interested in the setting where Θ parameterizes a family of probaility distributions. To that end, suppose that $\mathcal{X} = (X, \mathcal{A})$ be a measurable space, so that X is a set and $\mathcal{A}$ be a σ-algebra over it, let $\Delta\mathcal{X}$ denote the set of probability measures over $\mathcal{X}$, and keep in the back of our heads an indexed family $\mathcal{P} = \{p_\theta \in \Delta\mathcal{X} \mid \theta \in \Theta\}$ of probability distributions.

2.1 Differentiability and Agregating Incremental Updates

Confidence is meant to interpolate between fully incorporating information and ignoring it. Such an interpolation becomes more useful if it is continuous, and more useful still if it is differentiable. Next, we present two variants of a differentiability axiom, depending on the structure one has in hand.

CF3. **(differentiability)**

1. Θ has a manifold structures, and for all θ and ϕ, the function $\beta \mapsto F^\beta_\phi(\theta) : \mathcal{C} \to \Theta$ is continuously differentiable at $\beta = \bot$.
2. Θ parameterizes a family of probabilities over $(\mathcal{X}, \mathcal{A})$, via $\{\Pr_\theta\}_{\theta\in\Theta}$. for all $\theta \in \Theta$, $\phi \in \Phi$, and $A \in \mathcal{A}$, the function $\beta \mapsto \Pr_{F^\beta_\phi(\theta)}(A) : \mathcal{C} \to [0, 1]$ is continuously differentiable at $\beta = \bot$.

If Θ is a differentiable manifold and $\Pr : \Theta \to \Delta\mathcal{X}$ is a differentiable map, then the second follows from the first. It's simpler to assume that Θ carries a differentiable structure, so we will assume this when possible.

When we have $\mathcal{C} := \mathbb{R}_+$,

⟨ INCOMPLETE ⟩

CF4. For all $\beta_1, \beta_2 \in \mathbb{R}_+$, $F_\phi^{\beta_1} \circ F_\phi^{\beta_2} = F_\phi^{\beta_1+\beta_2}$ **(additivity)**

Proposition 1. *If F is a differentiable commitment function with confidence domain $\mathbb{R}_+$,then there is a unique update rule G with the same confidence domain, that behaves approximately like F for small increments of confidence, and is also additive (CF8).*

2.2 Optional Axioms for Update Rules

We now detail some other properties we might want an update rule to have.

Symmetry and Commutativity Some update rules, such as the one in **??** have a particularly conveneint property: the result of applying many updates does not depend on the order in which the information arives.

CF5. For all $\phi_1, \phi_2 \in \Phi$, $c_1, c_2 \in \mathcal{C}$, and $\theta \in \Theta$, we have that $F_{\phi_1}^{c_1}(F_{\phi_2}^{c_2}(\theta)) = F_{\phi_2}^{c_2}(F_{\phi_1}^{c_1}(\theta))$. **(Commutativity)**

We will not always want to insist on commutativity. Human belief updates, for example, are notoriously non-commutative, in part due to confirmation bias: if you are already fairly certain that P is false, you are likely to disregard information that P is true. Thus earlier updates tend to be more impactful.[1]

We would also like update rules to preserve any symmetries shared by the state space $\mathcal{X}$ and the assertion language Φ, so that updates are not sensitive to irrelevant labelings of points. Concretely, let $\mathrm{Aut}(\Theta, \Phi)$ be the set of automorphisms $\sigma : \Theta \to \Theta$ (say, rotations of the simplex of distributions), that also have an associated action on assertions, so that $\sigma\phi \in \Phi$ is the corresponding relabeling of ϕ under σ. The symmetry condition can now be captured by:

CF6. For all $\sigma \in \mathrm{Aut}(\Theta, \Phi)$, we have $F_{\sigma\phi}^{\beta}(\sigma(\theta)) = \sigma\Big(F_\phi^\beta(\Pr)\Big)$. **(symmetry)**

Intuitively, this axiom states that doing an update is equivalent to changing to an equivalent representaton, doing the appropriately transformed update, and then transforming back.

⟨under construction⟩

Compatibility with Divergences and Cost Functions

Suppose that, in addition, we have a "divergence" function $d : X \times X \to \mathbb{R}^+$ on X, with the property that $d(x, y) = 0$ iff $x = y$. For sets $A \subset X$, let $d(x, A) := \inf_{a \in A} d(x, a)$ be the smallest possible divergence between x and any member of A; symmetrically, define $d(A, x) := \inf_{a \in A} d(a, x)$.
If we drop the requirement that the smallest possible value of $d(A, x)$ must equal zero, we obtain the more general notion of a cost (or loss) function, $c : \mathcal{A} \times X \to \mathbb{R}_{\geq 0}$ satisfying

$$\forall A \in \mathcal{A}.\ \arg\min_x c_A(x) = A.$$

The idea is that, if $c(x)$ is some cost that "incentivises" membership in A, then increasing confidence in A ought decrease expected cost.

CF7. For all $\beta > 0$ and $A \in \mathcal{A}$, we have $\mathbb{E}_{F_A^\beta(\Pr)}[c_A] \leq \mathbb{E}_{\Pr}[c_A]$ **(c-monotonicity)**

We can also use c to define a notion of independence.

[1] It is also possible to model this effect with a commutative update rule, by expressing reduced confidence in later inputs.

Definition 1 (c-independence)**.** For $A, B, Z \subset X$, we say that A and B are c-independent given Z iff for all $z \in Z$, we have that $c_{A\cap B}(z) = c_A(z) + c_B(z)$. □

To give a few examples:

1. Every set $A \in \mathcal{A}$ is independent of the trivial event X, since $c(z, A \cap X) = c(z, A)$ and $c(z, X) = 0$.
2. Suppose $\mathcal{X}$ is a subset of $\mathbb{R}^n$, so that $x = (x_1, \ldots, x_n)$, and cost is L1 distance, given by $c(x, A) = \inf_{a\in A} \sum_{i=1}^n |a_i - x_i|$. Now for $i \neq j$, the sets $A_i(b) := \{x : x_i = b\}$ and $A_j(b') := \{x : x_j = b'\}$ are unconditionally c-independent. This makes sense, since they are orthogonal hyperplanes.

Armed with this notion of independence, we can articulate one further axiom for udpate rules:

CF8. If A and B are independent given Supp(Pr), then $F^\beta_A \circ F^\beta_B(\Pr) = F^\beta_{A\cap B}(\Pr)$. **(decomposition)**

The idea here is that if A and B are independent, then it is equivalent to learn them in either order, or both at once.

2.3 Combining Updates

Proposition 2 (closure under rescaling)**.** *If F is a commitment function for $(\Theta, \Phi, \mathbb{R}_+)$, then so is $F^k := (\phi, \theta, \beta) \mapsto F^{k\beta}_\phi(\theta)$, for every positive real number $k > 0$.*

Sequences. A Θ-update rule F on Φ also suggests an extension to updates on the more expressive set

$$(\Phi \times \mathbb{R})^* := \left\{ \text{finite sequences } \; (\phi_i, \beta_i)_{i\in[n]} \;\middle|\; \phi_i \in \Phi, \beta_i \in \mathbb{R} \text{ for all } i \in [n] \right\}$$

via sequential composition of the underlying updates:

$$\bar{F}^k_{\boldsymbol{\varphi}\cdot\boldsymbol{\beta}}(\mu) := F^{k\beta_1}_{\phi_1} \circ F^{k\beta_2}_{\phi_2} \circ \cdots \circ F^{k\beta_n}_{\phi_n}(\mu).$$

$\bar{F}$ satisfies CF1, CF3 and CF6 if F does. In general, it will not in general be an update rule, as it may not be additive (CF8). But if F is commutative (CF5), then $\bar{F}$ *does* satisfy CF8, and is also commutative (CF5) itself.

We will see that even non-commutative update rules, for which the order is important, can also be naturally combined in an unordered way.

2.4 Vector Field Representations of Differentiable commitment functions

For a smooth manifold M (such as the space $\Delta\mathcal{X}$ of distributions over $\mathcal{X}$), and a point $p \in M$, we follow convention by writing T_pM for the tangent space to M at point p (Lee 2013), and $TM := \sum_{p\in M} T_pM$ for the full tangent bundle over M. A vector field over M is a smooth map $\mathbf{v} : M \to TM$ assigning a tangent vector $\mathbf{v}(p) \in T_pM$, to every point $p \in M$.

Theorem 3. *Every update rule $F : \Phi \times \mathbb{R} \to (\Theta \to \Theta)$ satisfying CF1, CF3 and CF8 corresponds to a unique Φ-indexed collection of vector fields $F' : \Phi \times \Theta \to T\Theta$*

In the language of

Corollary 3.1. *There is a bijective correspondence between udpate rules satisfying CF1, CF3 and CF8 and Φ-indexed collections of **complete** vector fields.*

We call F' the *vector field representation* of a differentiable update rule F.

One defining feature of vector fields is closure under linear combination. Because they are in natural correspondance with differentiable additive update rules, update rules also inherit this structure.

In particular, given two update flows $F, G : \mathbb{R} \to \Theta$, we can define $F \oplus G$ via the vector field $(F \oplus G)' = F' + G'$.

⟨under construction⟩

Interaction With Certainty Axioms.

⟨ TODO: prove impossible to individually they satisfy certainty, but not together. ⟩

Definition 2. For an assertion language Φ, let $\overline{\Phi}$ denote the space of weighted formal sums of elements of Φ. □

Proposition 4. *Every update rule F on Φ, can be naturally extended to an update rule $\bar{F}$ on $\overline{\Phi}$ via the total vector field*

$$\bar{F}'_{\sum_i a_i \phi_i}(\theta) := \sum_i a_i F'_{\phi_i}(\theta).$$

If Φ is itself a measurable space, we can extend this further: Every update rule F on Φ, can be naturally extended to an update rule $\bar{F}$ on the space $\mathcal{M}(\Phi)$ of measures over Φ, via

$$\bar{F}'_{\beta(\Phi)}(\theta) := \int_\Phi F'_\phi(\theta) \mathrm{d}\beta.$$

2.5 Linear Update Rules

There are many definitions of linear update rules:

Definition 3. Let F be a differentiable update rule on Θ. We say that F is ...

- *linear* if Θ is a vector space over $\mathbb{R}$, and the vector field F'_ϕ is a linear operator, i.e., for all $a, b \in \mathbb{R}$, we have that
$$F'_\phi(a\theta_1 + b\theta_2) = aF'_\phi(\theta_1) + bF'_\phi(\theta_2).$$
- *cvx-linear* if $\Theta \subset \mathbb{R}^n$ is a convex set, and, for all $a \in [0, 1]$, we have that
$$F'_\phi(a\theta_1 + (1-a)\theta_2) = aF'_\phi(\theta_1) + (1-a)F'_\phi(\theta_2).$$
- *$\mathcal{L}$-cvx-linear* if $\Theta \subset \mathbb{R}^n$ and F is an optimizing update rule with a loss representation $\mathcal{L}$ linear in its first argument, i.e.,
$$\mathcal{L}(a\theta_1 + (1-a)\theta_2, \varphi) = a\mathcal{L}(\theta_1, \varphi) + (1-a)\mathcal{L}(\theta_2, \theta).$$

□

Proposition 5. *If F is a $\mathcal{L}$-cvx-linear, then it is also cvx-linear.*

In fact, the first condition is much stronger;

Proposition 6. *if F is a nontrivial $\mathcal{L}$-cvx-linear optimizing UR, then Θ equals cone generated by the rays $\{F'_\varphi \theta : \theta \in \Theta, \varphi \in \Phi\}$. In particular, if there is some θ such that 0 is in the interior of the convex hull $\mathrm{conv}(\{F'_\phi \theta\}_{\phi \in \Phi})$, then $\Theta = \mathbb{R}^n$.*

Proposition 7. *Every linear update rule is of the form $F^\beta_\phi(\theta) = \theta^T \exp(\beta V)$, where $\exp(\beta V)$ is the matrix exponential.*[2]

[2] Concretely, if $V = U^T \mathrm{Diag}(\lambda_1, \ldots \lambda_n) U$ is an eigendecomposition of V, then $\exp(V) = U^T \mathrm{Diag}(e^{\beta\lambda_1}, \ldots e^{\beta\lambda_n}) U$.

Proposition 8. *A linear update rule F is commutative iff, for every pair of statements $\phi, \phi' \in \Phi$, the matrices V_ϕ and $V_{\phi'}$ commute.*

3 Optimizing Update Rules

Suppose we have:

1. A differentiable loss function $\mathcal{L} : \Theta \times \Phi \to \mathbb{R}$, which intuitively measures the "incompatibility" between a belief state θ and an assertion φ, and
2. A way of taking the gradient of $\mathcal{L}$ with respect to θ.[3]

Then we can define an update rule $\texttt{GF}[\mathcal{L}]$ that reduces inconsistency by gradient descent. Concretely, such an update rule has a vector field:

$$\texttt{GF}[\mathcal{L}]'_\phi(\theta) = -\nabla_\theta \mathcal{L}(\theta, \phi).$$

Proposition 9. *An update rule F on a Riemannian manifold Θ is optimizing update rule if and only if $(F')^\flat$ is a conservative co-vector field. Lee 2013, Prop 11.40*

Example 3: Weighted Average

$$\Theta = \mathbb{R}^n \times (\mathbb{R}_{>0} \cup \{\infty\}); \qquad \Phi = \mathbb{R}^n$$

A belief state $(\mathbf{x}, w) \in \Theta$ consists a current estimate $\mathbf{x}$ of the quantity of interest, and a weight w of the total internal confidence in the estimate.

Updating proceeds by taking a weighted average of the previous estimate and the new input, weighted by their respective confidences, which is captured by:

$$F^\beta_{\mathbf{y}}(\mathbf{x}, w) = \left(\frac{w\mathbf{x} + \beta\mathbf{y}}{w + \beta}, w + \beta\right) \quad \text{and} \quad F^\beta_{\mathbf{y}}(\mathbf{x}, \infty) = (\mathbf{x}, \infty)$$

It is additive, since

$$\begin{aligned} F^{\beta_2}_{\mathbf{y}} \circ F^{\beta_1}_{\mathbf{y}}(\mathbf{x}, w) &= \left(\frac{(w + \beta_1)\frac{w\mathbf{x} + \beta_1\mathbf{y}}{w + \beta_1} + \beta_2\mathbf{y}}{(w + \beta_1) + \beta_2}, (w + \beta_1) + \beta_2\right) \\ &= \left(\frac{w\mathbf{x} + (\beta_1 + \beta_2)\mathbf{y}}{w + (\beta_1 + \beta_2)}, w + (\beta_1 + \beta_2)\right) = F^{\beta_1 + \beta_2}_{\mathbf{y}}(\mathbf{x}, w) \end{aligned}$$

And it is clearly differentiable, with a simple calculation revealing that $F'_{\mathbf{y}}(\mathbf{x}, w) = \left(\frac{\mathbf{y} - \mathbf{x}}{w}, 1\right)$.

Observations:

- The update rule cannot be extended differentiably to states $\theta = (\mathbf{x}, w)$ with $w = 0$. Intuitively, we need to have some estimate with positive confidence to update beliefs in a differentiable way. This is related to the fact that plain emperical risk minimization (ERM) is unstable, but stable with even a small amount of regularization.
- The certainties are given by

$$\lim_{\beta \to \infty} F^\beta_{\mathbf{y}}(\mathbf{x}, w) = (\mathbf{y}, \infty)$$

[3] such as a tangent-cotangent isomorphism $(-)^\sharp : T^*_p\Theta \to T_p\Theta$, perhaps coming from an affine connection, in turn perhaps coming from a Riemmannian metric.

- F is commutative, invertible, and symmetric with respect to permutation of the dimensions, but it is not conservative: if we had $U(\mathbf{x}, w, \mathbf{y})$ twice differentiable such that $\nabla_{\mathbf{x},w} U = F'$, then we would have

$$\frac{\partial^2}{\partial w \partial x_i} U = \frac{\partial}{\partial w} \frac{y_i - x_i}{w} = \frac{x_i - y_i}{w^2} \qquad \text{but} \quad \frac{\partial^2}{\partial x_1 \partial w} U = \frac{\partial}{\partial x_i} 1 = 0$$

  violating Clairaut's theorem on equality of mixed partials. Therefore, F is not an optimizing update rule.

Natural Gradients for Probability Distributions. When Θ parameterizes a family of probability distributions, via some $\Pr : \Theta \to \Delta\mathcal{X}$, there is a particularly natural metric on Θ, called the Fisher information metric. This metric is the unique one on Θ that is independent of the representation of $\mathcal{X}$ (Chentsov 1982), in the following sense. If there are cpds $p(Y|X)$ and $q(X|Y)$ such that, for all $\theta \in \Theta$, the distribution $\Pr_\theta(X)$ is unchanged after converting to Y and back again X (via p and q respectively), as depicted by the following commutative diagram,

$$\begin{array}{ccc} \Theta & \xrightarrow{\Pr} & X \\ {\scriptstyle \Pr}\downarrow & & \uparrow{\scriptstyle q} \\ X & \xrightarrow[p]{} & Y \end{array}$$

then clearly the family $\Pr(Y|\Theta) := p \circ \Pr_\theta$ carries the same information about the parameters (and in particular how best to update them) as $\Pr_\theta$. Chentsov's theorem says, that, up to a multaplicative constant, the Fisher information metric is the only metric on Θ, as a function of the parameterization $\Pr$, which gives identical geometry in both cases.

At each point Θ, the components of the Riemannian metric form a matrix—in this case, the Fisher information matrix $\mathcal{I}(\theta)$—which allow us to now compute the gradient in the natural geometry from the coordinate derivatives as

$$\texttt{NGF}[\mathcal{L}]'_\phi(\theta) = -\hat{\nabla}_\theta \mathcal{L}(\theta, \phi) = \mathcal{I}(\theta)^\dagger \nabla \mathcal{L}(\theta, \phi)$$

where $\mathcal{I}(\theta)^\dagger$ denotes the Moore-Penrose psuedoinverse of the matrix $\mathcal{I}(\theta)$, and $\nabla\mathcal{L}$ is the gradient for the euclidean metric i.e., the vector of partials $[\frac{\partial \mathcal{L}}{\partial \theta_1}, \ldots, \frac{\partial \mathcal{L}}{\partial \theta_n}]^\mathsf{T}$.

3.1 Expected Utility Maximization Update Rules

Suppose, for each $\phi \in \Phi$, we have a utility function $U_\phi : X \to \mathbb{R}$ on the underlying set X. We can use this to define an update rule

$$\begin{aligned} \text{Bolz}[U] &: (\mathbb{R} \times \Phi) \to \Delta\mathcal{X} \to \Delta\mathcal{X} \\ \text{Bolz}[U]^\beta_\varphi(\mu) &:\propto \mu \exp(-\beta U_\phi) \\ &= A \mapsto \frac{1}{\mathbb{E}_\mu[\exp(-\beta U_\phi)]} \int \exp(-\beta U_\phi) \mathbb{1}_A \mathrm{d}\mu \end{aligned}$$

Proposition 10. *Bolzmann Update Rules are additive, zero, differentiable, invertable, and commutative.*

Proof. **Commutativity.** For some normalization factors Z, Z', Z'', we have:

$$\begin{aligned} F_\phi^\beta(F_{\phi'}^{\beta'}(\mu)) &= F_\phi^\beta\Big(\frac{1}{Z}\,\mu\,\exp(-\beta' c_{\phi'})\Big) \\ &= \frac{1}{Z'}\frac{1}{Z}\,\mu\,\exp(-\beta' c_{\phi'})\exp(-\beta c_\phi) \\ &= \frac{1}{Z''}\,\mu\,\exp(-\beta' c_{\phi'} - \beta c_\phi) \end{aligned}$$

which is the same expression when we exchange (ϕ, β) and (ϕ', β'). □

Note that this is true even for costs generated by asymmetric distances $c_{\{y\}}(x) = d(y, x) \neq d(x, y) = c_{\{x\}}(a)$.

Remark 1. *Regarding $U_\varphi : \mathcal{X} \to \mathbb{R}$ as a potential energy over X,* $\mathrm{Bolz}[U]_\varphi^\beta(\mathrm{Unif})$ *is the Boltzmann distribution at inverse temperature (thermodynamic coldness) β In the thermodynamic analogy, as temperature decreases, one becomes more certain that particles are in their most favorable states.*

The certainties of $\mathrm{Bolz}[U]$ are the minimizers of U.

⟨under construction⟩

As a reminder, we have $\Theta = \Delta\mathcal{X}$, and suppose we have $c : X \times \Phi \to \mathbb{R}$. Suppose $U(\theta, \varphi) = \mathbb{E}_\theta[c_\varphi]$ (linearity, Definition 3). Then $(\mathrm{Boltz}\ U)'_\varphi\theta = \theta(\mathbb{E}_\theta[c_\varphi] - c_\varphi)$,while

$$\begin{aligned} (\texttt{GD}\ U)'_\varphi\theta &= -\nabla_\theta\,\mathbb{E}_\theta[c_\varphi] \\ &= -c_\varphi. \end{aligned}$$

This second expression, though, doesn't seem quite right — it isn't even a tangent vector to the probability simplex, since its components don't sum to zero. This issue is in our naive computation of the gradient. We have computed the collection of partial derivatives $\frac{\partial}{\partial\theta_i}(\theta \cdot c_\varphi) = (c_\varphi)_i$, which is technically co-vector field, not a vector field.[a]
For simplicity, suppose that $X = \{1, \ldots, n\}$, in the discussion that follows. If our parameter space were all of $\mathbb{R}^n$, we could simply collect these terms and take a transpose, to get

$$\nabla_\theta\,\underset{\theta}{\mathbb{E}}[c_\varphi]$$

There are two ways to proceed from here. The first makes use of manifold theory: for each point $p \in \Delta X$, begin by identifing a neighborhood $U \ni p$ with an open subset of $\mathbb{R}^{(n-1)}$, and define an inner product (a metric tensor) $g_p(\cdot, \cdot)$ on tangent vectors $v \in T_p\Delta X$, making $\Delta\mathcal{X}$ into a Riemannian Manifold, and then compute the gradient in the standard way, using the inverse of the metric tensor g in order to convert covectors to vectors in a natural way.
The second, which is computationally simpler, is to take the metric induced by an embedding in Euclidean space. This approach is equally general, because Nash's Theorem (Nash 1956) tells us that any n-dimensional Riemannian manifold may be isometrically embeded in $\mathbb{R}^{2n+1}$.

[a] it acts on a vector field V by $-c_\varphi(V) = -V(c_\varphi)$.

Proposition 11. *The associated vector field is given by* $(\mathrm{Boltz}\,U)'_\varphi p = p(\mathbb{E}_p[U_\varphi] - U_\varphi)$.

Proof. Let $f(X) := \exp(-\beta U(X, \varphi))$, and $g(X) := U(X, \varphi)$.

$$\begin{aligned}
\mathrm{Boltz}'_\varphi \theta &= \frac{\partial}{\partial \beta} \mathrm{Boltz}^\beta_\varphi(p)\Big|_{\beta=0} \\
&\quad \langle \texttt{TODO: finish typesetting algebra} \rangle \\
&= x \mapsto p(x) \frac{f(x)}{\mathbb{E}_p[f]} \left(\mathop{\mathbb{E}}_p \left[\frac{f}{\mathbb{E}_p[f]} g \right] - g(x) \right) \Big|_{\beta=0} \\
&= \frac{pf}{\mathbb{E}_p[f]^2} \left(\mathbb{E}_p[fg] - g\, \mathbb{E}_p[f] \right) \Big|_{\beta=0} \\
&= x \mapsto p(x)(\mathbb{E}_p[g] - g(x)) \qquad \text{since } f(X) = 1 \text{ when } \beta = 0
\end{aligned}$$

As a sanity check, note that the sum over all components is

$$\sum_{x \in X} ((\mathrm{Boltz}\, U)'_\varphi\, \theta)_x = \sum_{x \in X} p(x)(\mathbb{E}_p[g] - g(x)) = \mathbb{E}_p[\mathbb{E}_p[g]] - \mathbb{E}_p[g] = 0,$$

so indeed it lies within the tangent space. □

Proposition 12. *The Bolzmann update rule with potential $U(X)$ is the natural gradient flow update rule for expected value of U, i.e.,* $\mathrm{Bolz}[U] = \texttt{NGF}[\mu \mapsto \mathbb{E}_\mu\, U]$.

Example 4: Gaussian NGD

Consider the case where $\Theta = \{(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\}$ is the half-space of parameters to a Gaussian over some real variable X, and $\Phi \cong \mathbb{R}$ consists of possible observations of X.

One natural loss function is negative log likelihood (differential surprisal) of the observation x according to your belief state $\theta = (\mu, \sigma^2)$:

$$\mathcal{L}(x, \mu, \sigma^2) = -\log \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 = \left\langle\!\!\left\langle \begin{array}{c} \mu \twoheadrightarrow \boxed{\mu} \\ \sigma^2 \twoheadrightarrow \boxed{\sigma^2} \end{array} \overset{\mathcal{N}}{\rightarrowtail} \boxed{X} \overset{x}{\twoheadleftarrow} \right\rangle\!\!\right\rangle.$$

The fisher information for a normal distribution is given by

$$\mathcal{I}(\mu, \sigma) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}$$

The natural gradient update rule is given by

$$F'_x(\mu, \sigma^2) = -\hat{\nabla}_{\mu, \sigma^2} \mathcal{L}(x, \mu, \sigma^2) = \mathcal{I}(\mu, \sigma^2)^{-1} \begin{bmatrix} \frac{x - \mu}{\sigma} \\ \frac{-\sigma^2 + (x-\mu)^2}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} x - \mu \\ (x - \mu)^2 - \sigma^2 \end{bmatrix}$$

Note that:

- $\mathbb{E}_{x \sim \nu}[F'_x(\mu, \sigma^2)] = \mathbf{0}$ if and only if ν has mean μ and variance σ^2. Moreover, this point is the unique global attractor. This means that,

1. If observations are drawn from a fixed distribution $\nu(X)$, and we repeatedly use F to update $\theta = (\nu, \sigma)$ with small confidence ϵ, then μ will approach the mean $\mathbb{E}_\nu[X]$ of ν and σ^2 will approach the variance $\mathbb{E}_\nu[X^2] - \mathbb{E}_\nu[X]^2$.
2. If we perform a single high-confidence update on the extended observation $\varphi \propto \nu$, in which each x has relative confidence $\nu(x)$, the result will be a Gaussian with the mean and variance of ν, i.e.,

$$\forall \theta. \quad \lim_{c \to \infty} \Pr_{F^c_\nu(\theta)} = \mathcal{N}(\mathbb{E}_\nu[X], \mathrm{Var}_\nu[X])$$

In this sense, relative confidence acts like probability.

- If we update with the observation $x = \mu$ of our estimate with confidence c, the mean is unchanged, and our estimate of the variance becomes the harmonic mean of our previous variance σ_0^2 and the inverse confidence $\frac{1}{c}$. That is,

$$F^c_\mu(\mu, \sigma_0^2) = \left(\mu, \frac{1}{c + \frac{1}{\sigma_0^2}}\right).$$

In particular, if σ_0^2 is very large, so that our initial beliefs say very little, updating with confidence c results in variance $\frac{1}{t}$. In this sense, the magnitude of confidence acts as the inverse of variance.

4 Internal Confidence

a Fix a set Φ of assertions. We now study a natural way of constructing a space of beliefs Θ and a commitment function$F$ for it, which might be summarized as an internal log of changes in confidence.

First, suppose Φ is a finite set $\{\phi_1, \ldots \phi_n\}$, and $\mathcal{C} = \mathbb{R}_+$. Then, define

$$\Theta := \Theta^{\mathrm{FIN}}_{\Phi,B} := \left\{\text{vectors } v \in \mathbb{R}^n\right\}$$

In fancier mathematical language, the material of this section may be seen as studying the final coalgebras of the signature $G := (-)^{\Phi \times C} \times B$ modulo the co-equations **??** and CF1 in the category of manifolds and differentiable maps.

Proposition 13. *The final coalgebra*

5 Settings

5.1 Probabilities and Events

5.2 Probabilities and Samples

First, let's take the where belief state is an explicit representation of a finite probability distribution, and observations are samples of it. Concretely, this means $\Theta := \Delta W$, and $\Phi := W$.

Intuitively, we would like it to be the case, that if we observe samples drawn from our

CF7. $\bar{F}'_\mu \mu = \vec{0}$. That is, $\mathbb{E}_{x \sim \mu}[F'_x \mu] = \vec{0}$. **(calibration)**

Proposition 14. *Expected utility maximizing update rules cannot be calibrated for more than one distribution.*

Proof. Suppose we have a commitment function given by $\mathcal{L}^F(\mu, x) = \mathbb{E}_{y\sim\mu} U(x, y)$. Then

$$F'_x\mu = \mu \odot (\vec{U}_x - \underset{\mu}{\mathbb{E}} U_x)$$

It is calibrated iff

$$\begin{aligned}
\vec{0} = \bar{F}'_\mu(\mu) = \sum_x \mu(x) F'_x(\mu) = \sum_x \mu(x)\Big(\vec{\mu} \odot \vec{U}_x - (\mathbb{E}_\mu U_x)\vec{\mu}\Big) \\
\iff \qquad \forall y \in X.\ \ 0 = \sum_x \mu(x)\mu(y)(U(x,y) - \mathbb{E}_\mu U_x) \\
\iff \qquad \forall y \in X.\ \ \underset{x\sim\mu}{\mathbb{E}}\underset{z\sim\mu}{\mathbb{E}}[U(x,z)] = \underset{x\sim\mu}{\mathbb{E}}[U(x,y)]
\end{aligned}$$

Choosing $\mu = \delta_w$, we this means

$$\forall y.\ \ U(w,w) = U(w,y)$$

but this means updating with respect to x does nothing at all, no matter what state you're at. □

6 Examples

6.1 Assorted

Example 5: Internal Confidence

C

6.2 Relaxations of Bayes Rule (Updating By Conditioning)

Fix some measurable space $\mathcal{X} = (X, \mathcal{A})$ of possible outcomes, and consider an update rule for probability distributions $\Theta := \Delta\mathcal{X}$ on measurable events $\Phi = \mathcal{A}$.

The function

$$\text{Bayes}^\beta_A(\mu) := \begin{cases} \mu \mid A & \text{if } \beta > 0 \\ \mu & \text{if } \beta = 0 \\ \mu \mid \bar{A} & \text{if } \beta < 0 \end{cases}$$

satisfies CF1 and CF8, so it is an update rule. But it is not differentiable in β (it doesn't satisfy CF3). However, this behavior does arise as the limit of large-magnitude β for a differentiable update rule. Consider the update rule

$$\begin{aligned}
F^\beta_A(\mu) &:\propto \mu \exp(\beta \mathbb{1}[A]) \\
F^\beta_A(\mu)(B) &:= \frac{1}{\mathbb{E}_\mu[e^{\beta\mathbb{1}[A]}]} \int_X \mathbb{1}[B] \exp(\beta\mathbb{1}[A]) \mathrm{d}\mu \\
&= \frac{\mu(B\cap A)e^\beta + \mu(B\cap\bar{A})}{\mu(A)e^\beta + \mu(\bar{A})}.
\end{aligned}$$

Note that

$$\lim_{\beta\to\infty} F_A^\beta(\mu) = B \mapsto \frac{\mu(B\cap A)}{\mu(A)} = \mu \mid A;$$
$$F_A^0(\mu) = B \mapsto \mu(B); \qquad \text{and}$$
$$\lim_{\beta\to-\infty} F_A^\beta(\mu) = B \mapsto \frac{\mu(B\cap \bar{A})}{\mu(\bar{A})} = \mu \mid \bar{A}.$$

Thus, Bayes $= \lim_{k\to\infty} F^k$, where F^k is the update rule given by rescaling, as in Proposition 2.

⟨under construction⟩

Is this the only such relaxation of Bayes Rule?
Suppose F is another such relaxation, satisfying CF1, CF3, CF5, CF6, CF8 and CF9 and **??**.
Proposition 15. *F is unique up to curve reparameterization.*
By differentiability (CF3) and Corollary 3.1, we know that F is generated by vector field $F' : \mathcal{A} \times \Delta\mathcal{X} \to \Delta\mathcal{X}$. We start by investigating the action on singletons.

$$F'_{\{x\}}(\mu)$$

Assuming that the vector field is conservative,
Conjecture 16. *F is the unique update rule, up to multiplicative scaling, satisfying CF1, CF3, CF5, CF6, CF8 and CF9 and* **??**

6.3 Confidence as a Belief State: Mixture of Experts

6.4 Weight of Evidence

6.5 Loss Functions

6.6 Bolzman Rationality

Example 6

Consider a neural network whose output is a c

7 Confidence in PDGs

7.1 Quantitative PDGs

7.1.1 Updating Second-Order Distributions with PDGs

Let $\mathcal{N}$ be a set of variables, where each variable $X \in \mathcal{N}$ can take possible values in the set $\mathcal{V}(X)$, and let $\mathcal{V}(\mathcal{N})$ be the set of possible joint settings of all variables in $\mathcal{N}$.

$$\Theta := \left\{\text{Second-order probabilities } \Pr \in \Delta^2\mathcal{V}(\mathcal{N})\right\}$$
$$\Phi := \left\{\text{cpds } p(Y|X) \text{ where } X, Y \subset \mathcal{N}\right\}$$

Now, consider the Bolzmann update rule

$$F_L^\beta(\Pr)(\mu) \propto \Pr(\mu)\exp(-\beta \boldsymbol{D}(\mu \parallel \mathbf{p}_L))$$

Note, that the free extension of Φ is isomorphic to the set of "purely quantitative" PDGs, i.e., those with $\boldsymbol{\alpha} = \mathbf{0}$, over the variables $\mathbf{X}$:

$$\bar{\Phi} = \left\{\text{PDGs } \mathcal{M} = (\mathbf{X}, \mathcal{E}, \mathbf{P}, \mathbf{0}, \beta)\right\}$$

and it is easily verified that update rule from before extends to

$$F^k_{\mathcal{M}}(\Pr)(\mu) \propto \Pr(\mu) \exp(-k[\![\mathcal{M}]\!]_0(\mu))$$

Note that

$$\begin{aligned}\lim_{k\to\infty} F^k_{\mathcal{M}}(\Pr)(\mu) &\propto \Pr(\mu)\mathbb{1}[\mu \in [\![\mathcal{M}]\!]^*_0] \\ &= \Pr \mid [\![\mathcal{M}]\!]^*_0 \\ &= \Pr \mid \{\!\{\mathcal{M}\}\!\} \text{ if } \mathcal{M} \text{ is consistent.}\end{aligned}$$

That is to say, applying the update with observation $\mathcal{M}$ and infinite certainty is like conditioning the higher-order distribution Pr on those distributions μ consistent with $\mathcal{M}$. Now we turn to the opposite extreme, when confidence is low. When $k = 0$, clearly $F^k_{\mathcal{M}}(\Pr) = \Pr$, so F satisfies CF1. And in the limit, as $\gamma \to 0$,

$$\begin{aligned}\lim_{k\to 0} F^k_{\mathcal{M}}(\Pr)(\mu) &= \Pr \mid \{Inc_{\mathcal{M}}(\mu) < \infty\} \\ &= \Pr \mid \{\mu : \forall L.\ \mu(Y|X) \ll \mathbf{p}_L\}.\end{aligned}$$

So F can only be continuous at $k = 0$ for Pr (and hence can only satisfy CF3) if, for all edges $L \in \mathcal{E}$, we have that $\Pr(\mu(Y|X) \ll \mathbf{p}_L) = 1$. So one easy way to ensure differentiability for all update rules is to require strictly positive probabilities in theprobabilities in the cpds.

7.1.2 Updating Distributions with PDGs and Incompatibility

$$\begin{aligned}\Theta &:= \left\{\text{Joint probabilities } \mu \in \Delta\mathcal{V}(\mathcal{N})\right\} \\ \Phi &:= \left\{\text{cpds } p(Y|X) \text{ where } X, Y \subset \mathcal{N}\right\}\end{aligned}$$

Given a cpd $p(Y|X)$ with $X, Y \subset \mathbf{X}$, define an update rule on joint distributions μ by

$$\left(\texttt{CPD\_UR}^{\beta}_{p(Y|X)}\mu\right)(\mathbf{X}) \propto \mu(\mathbf{X}) \left(\frac{p(Y|X)}{\mu(Y|X)}\right)^{1-e^{-\beta}}$$

Writing this in a where $\epsilon := 1 - e^{-\beta}$ and $Z := \mathbf{X} \setminus \{X, Y\}$.

Proposition 17. $\texttt{CPD\_UR}_{p(Y|X)}$

Proposition 18. *The vector field of* $\overline{\texttt{CPD\_UR}}$ *is the natural gradient of* $Inc_{\mathcal{M}}$*, i.e.,* $\overline{\texttt{CPD\_UR}}'_{\mathcal{M}} = -\hat{\nabla} Inc_{\mathcal{M}}$*, so* CPD_UR *is the update rule on joint distributions which reduces incompatibility by locally moving in the direction of steepest descent.*

Proposition 19. *Updating* $\mathcal{M}$ *with absolute certainty computes the information projection onto the convex set of distributions consistent* $\mathcal{M}$ *i.e.,*

$$\texttt{CPD\_UR}^{\infty}_{\mathcal{M}}(\mu) = \operatorname*{argmin}_{\mu' \in \{\!\{\mathcal{M}\}\!\}} \boldsymbol{D}(\mu' \parallel \mu)$$

7.1.3 Updating PDGs with PDGs, via inconsistency reduction

$$\Theta := \Big\{\text{PDGs}\Big\}; \qquad \Phi := \Big\{\text{PDGs}\Big\}$$

We now consider the more general setting, in which both the observations Φ and the belief states Θ are Here's an update rule that simply adds the new data to a PDG.

$$F^{\beta}_{p(Y|X)}(m) = m \; + \; \boxed{\text{X}} \xrightarrow[(\beta)]{p} \boxed{\text{Y}}$$

It's extension is the update rule that incorporates data of the new PDG to the old one, where all confidences in the new PDG are scaled by k.

$$\bar{F}^{k}_{m'}(m) = m \; + \; k\, m'$$

The colletion of all PDGs, even on a finite set of finite variables, does not naturally form a manifold of fixed finite dimension. But if we restrict our attention to PDGs with a certain fixed structure, then it does. But, per (Richardson 2022), we already have a natural choice of a loss $\mathcal{L} = \langle\!\langle - \rangle\!\rangle : \text{PDG} \times \text{PDG} \to \mathbb{R}$: the inconsistency of the joint PDG.

Still, there is more than one reasonable way to perform updates: we can adjust the internal confidences $\boldsymbol{\beta}$ of our PDG, or its cpds $\mathbf{p}$.

$$\bar{G}'_{m'}(m) = -\hat{\nabla}_{\mathbf{p}} \langle\!\langle m + m' \rangle\!\rangle$$

Open Question 1. How do $\bar{F}$ and $\bar{G}$ relate to one another?

Conjecture 20. `CPD_UR` *should be a special case.*

7.1.4

$$\Theta := \Big\{\text{PDGs over variables } \mathbf{X}\Big\}; \qquad \Phi := \Big\{\text{Probabilities } \mu \in \Delta\mathcal{V}(\mathbf{X})\Big\}$$

8 Other Confidence Domains

To describe a degree of partial incorporation, we will need a domain of possible confidence values. Mostly, we will stick to using real numbers, but it will clarify things to stay more general for now, so that we can see the properties we actually need. Formally, a *confidence domain* is a tuple $(\mathcal{C}, \oplus, \bot, \top)$, where $(\mathcal{C}, \oplus, \bot)$ is a monoid with operation $\oplus$ and neutral element $\bot$, and $\top \in \mathcal{C}$ is an absorbing element—i.e., $\top \oplus c = \top$ for all $c \in \mathcal{C}$. In terms of confidence, we interpret the components as follows:

- The elements of $\mathcal{C}$ are the possible degrees of confidence.
- The monoid operation $\oplus : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$ describes how to combine two (independent) confidences in some statement, to obtain a new confidence in that statement.
- The neutral element $\bot \in \mathcal{C}$ indicates "no confidence" in an observation. The monoid identity laws, which assert that $\bot \oplus c = c = c \oplus \bot$ for all $c \in \mathcal{C}$, reflect the intuition that we should ignore untrusted information in combining confidences.
- The absorbing element $\top$ indicates "full confidence". The absorbtion property corresponds to the intuition that, definitive information that ϕ is true, when combined with other (perhaps less reliable) information that ϕ is true, is still definitive.

In this more general setting, the analogue of additivity (CF8) becomes:

CF8. For all $c_1, c_2 \in \mathcal{C}$, $F_\phi^{c_1} \circ F_\phi^{c_2} = F_\phi^{c_1 \oplus c_2}$ **(combination)**

CF8 looks like it could be problematic, but it simply states that commitment functions respect the combination operation. If we fix an assertion ϕ, then an update with confidence c_1 followed by an update with confidence c_2 is equivalent to an update with confidence $c_1 \oplus c_2$, which is, by definition, the result of comining confidences c_1 and c_2. On its own, so long as we have the freedom to choose $\mathcal{C}$, CF8 has no teeth.

Proposition 21. *If $F : \mathcal{C} \to (\Phi \to (\Theta \to \Theta))$ satisfies CF1 and CF2, then we can construct a new update function for Θ on Φ, that behaves in exactly the same way, except that it is exteneded to a larger confidence domain, for which which it does satisfy CF8.*

Proof. Consider the new confidence domain

$$\mathcal{C}' := \Big\{ \text{ finite lists } [c_1, \dots c_n] \text{ with each } c_i \in \mathcal{C}, \quad ::, \quad [\,], \quad [\top] \Big\},$$

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

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

Concatenating the empty list $[\,]$ on either side has no effect, by construction, for all $L \in \mathcal{C}'$, we have $[\top] :: L = [\top] = L :: [\top]$, and :: is clearly associative, so $\mathcal{C}'$ is also a confidence domain.

The new update rule for this confidence is given by:

$$AF_\phi^{[c_1,\dots,c_n]}(\theta) := (F_\phi^{c_n} \circ \dots \circ F_\phi^{c_1})(\theta).$$

AF has the same behavior as F on the elements that correspond to the original confidence domain, since $AF_\phi^{[c]}(\theta) = F_\phi^c(\theta)$, and it is additive by construction, since

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

□

For convenience of measurement, and so that we may better study confidence as a *smooth* interpolation between ignoring and fully incorporation, we shall focus primarily on cases where confidence can be measured as a real number. We now consider two such confidence domains.

- First, we consider the zero-one confidence domain

  $$\text{ZO} := \Big([0,1], \quad a \star b := a + b - ab, \quad 1, \quad 0 \Big),$$

  which uses the same numerical endpoints as probability; a value of zero represents no confidence, a value of one represents full confidence. For the purposes of updating, we may interpret a confidence of $a \in \text{ZO}$ as the fration of the way between ignoring and fully incorporating information. This motivates the definition of the operator $\star$. If you go 90% of the way to fully

incorporaing some information ϕ, and then 50% of the remaining way, then in total you have gone $90\% + 50\%(100\% - 90\%) = 0.9 + 0.5 - (0.9)(0.5)$ of the way to fully incorporating ϕ.

- We now introduce a second confidence domain based on the real numbers, which is mathematically cleaner, if more difficult to interpret numerically in absolute terms.

$$\mathbb{R}_+ := \Big([0, \infty) \cup \{\infty\}, \quad +, \quad 0, \quad \infty \Big)$$

The use of addition as the combination operator makes it particularly natural to speak of linear combinations of inputs. This point is best illustrated by example.

  - **Voting.** Suppose the elements of Φ correspond to candidates in an election. In a sense, the number of votes a candidate recieves is a measure of how much confidence the electorate has in them—a candidate who recieves no votes is ignored, while a candidate who recieves all of the votes should be listened to exclusively.

    It's hard to say much the raw number of votes a candidate recives in absolute terms, in part becasue it depends on the number of votes recieved by other candidates, and also how many votes you will recieve in the future. Nevertheless, if we are collecting votes, is especially natural to weight candidates by the total number of votes behind them. This way of measuring confidence also applies without change to measure fractional votes.

  - **Chemical Reactions.** Suppose that we have a mixture of nano-bots. Each nano-bot has some type $\phi \in \Phi$, and has the effect of turning matter into bots of type ϕ. For every $\phi \in \Phi$, let β_ϕ be the concentration of bots of type ϕ, say measured in number of bots per liter of solution. In some sense, β_ϕ measure of how much "confidence" the mixture has in ϕ—if the concentration is zero, then that bot type may be ignored, and if all particles are of type ϕ, then

$$\langle \text{ INCOMPLETE } \rangle$$

We will use greek letters $\alpha, \beta, \ldots$ to denote elements of $\mathbb{R}_+$.

Proposition 22. ZO *is isomorphic to* $\mathbb{R}_+$*, but therere is no canonical choice of isomorphism.*

Proof. For every $k > 0$ can construct an isomorphism $\varphi_k : \text{ZO} \to \mathbb{R}_+$ explicitly by $\varphi(a) := -k \log a$. It is a homomorphism, since

$$\varphi(a \star b) = -k \log(ab) = -k \log a - k \log b = \varphi(a) + \varphi(b),$$

while $\varphi(1) = 0$ (so it preserves the identity) and $\varphi(0) = \infty$ (so it preserves the absorbing element). The inverse mapping can also be explicitly by $\varphi^{-1}(r) := \exp(-r/k)$, which is also a homorphism for the same reasons as above. □

A Extra

A.1 Invertable Update Rules

CF9. For all $\phi \in \Phi$, and $\beta \in \mathbb{R}$, the update $F_\phi^\beta : \Theta \to \Theta$ is invertable. **(Invertability)**

This effectively partitions Θ into two

Proposition 23. *If F is a differentiable and invertable update rule (i.e., satisfies CF1, CF3, CF8 and CF9), then for all $\beta \in \mathbb{R}$, $\phi \in \Phi$, the function $F_\phi^\beta : \Theta \to \Theta$ is a diffeomorphism, and its inverse is given by $F_\phi^{-\beta}$, in the sense that*

$$F_\phi^{-\beta}(F_\phi^\beta(\mu)) = \mu = F_\phi^\beta(F_\phi^{-\beta}(\mu)).$$

As a consequence,

Corollary 23.1. *If for any $\beta < \infty$ there exist μ, ϕ, A such that $\mu(A) > 0$ but $F_\phi^\beta(\mu)(A) = 0$, then F is not invertable.*

B

Example 7. Suppose F is an additve update rule. Then, we can explicitly construct a resolution to the problem posed in Example 1 by defining enriched spaces

$$\Phi' := \Phi \times \left\{ \text{ identities } \ id \right\}$$
$$\Theta' := \Theta \times \left\{ \text{histories } L = [(\phi_1, id_1, c_1), \dots (\phi_n, id_n, c_n)] \right\}$$

and new commitment function G by

$$G^\beta_{(\phi,id)}(\theta, L) := \begin{cases} \left(F_\phi^{\beta - \sum_i \beta_i \mathbb{1}[(\phi_i, \mathrm{id}_i) = (\phi, \mathrm{id})]}(\theta), \ L :: (\phi, id, \beta) \right) & \text{if } \beta \neq \bot \\ (\theta, L) & \text{if } \beta = \bot \end{cases}$$

△