Where did this come from?

Learner's Confidence

Abstract

We introduce a new notion of confidence, which arises when learning or updating beliefs.

You need a story here: There's another well-known way of representing uncertainty. Here's the intuition for it a and here's why it's relevant to us …

1 INTRODUCTION

What should it mean to say that one has a high degree of confidence in a statement ϕ? It is often taken to mean that we think ϕ is likely. Here we argue that there is a related but more useful conception of confidence that arises when learning—one that complements likelihood and, moreover, unifies several different concepts in the literature.

For us, confidence is a measure of *trust*, rather than likelihood. In particular, the degree of confidence that one has in a piece of information ϕ quantifies how seriously to take ϕ in updating one's beliefs. So at one extreme, if we observe ϕ but have no confidence in it, we should not change our beliefs at all; at the other, if we have full confidence in ϕ, we should fully incorporate it into our beliefs.

Example 1. Suppose our belief state is a probability measure, and ϕ is an event. A full-confidence update then amounts to conditioning on ϕ, after which ϕ has probability 1, and so cannot be further incorporated. Here is an obvious way to describe intermediate degrees of confidence: if we learn ϕ with confidence $\alpha \in [0,1]$ and start with prior probability Pr, then we end up with the posterior $(1-\alpha)\Pr + \alpha(\Pr \mid \phi)$. Thus, having high confidence in ϕ leads to posterior beliefs that give ϕ high probability. The converse is false, however, so confidence and probability can be quite different. 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. Prior probabilty is further decoupled from confidence: if we learn a surprising fact ϕ from a trusted source, we have high confidence in ϕ despite it having low prior probability. □

What two styles are you talking about?

In some contexts, both styles of measuring confidence are used. To illustrate, we turn to another class of examples of confidence based on the work of Shafer, whose 1976 book was written largely to develop a theory of what we have been calling confidence, based on (and tailored to) to his preferred representation of uncertainty.

Example 2. Let W be a finite set of possible worlds, and suppose our belief state is a *mass function* $m : 2^W \to [0,1]$ satisfying $\sum_{U \subseteq W} m(U) = 1$ and $m(\emptyset) = 0$. From m, one can obtain a Dempster-Shafer belief function, which is a generalization of a finite probability measure over W, by $\mathrm{Bel}_m(U) = \sum_{V \subseteq U} m(V)$ [Shafer, 1976].

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 another mass function s, called a *simple support function*, that places α of its mass on ϕ, and the rest on the trivial event W. To combine our prior belief m with the new evidence s, Shafer argues we should use Dempster's rule of combination, to obtain a posterior $m' := m \oplus s$, given in this case by:

$$m'(U) = \frac{1}{1 - \alpha \sum_{V \subseteq (W \setminus \phi)} m(V)} \Big((1-\alpha) m(U) + \alpha \sum_{\substack{V \subseteq W \\ V \cap \phi = U}} m(V) \Big).$$

It is easy to verify that when $\alpha = 0$, the posterior beliefs are the same as the prior ones, and that when $\alpha = 1$, all mass is assigned to subsets of ϕ. It follows that, after the update, $\mathrm{Bel}_{m'}(\phi)$. So again, we have two extremes in confidence, continuously parameterized by a value $\alpha \in [0,1]$.

We now look at some special cases. Suppose that Bel_m is a probability measure Pr, or equivalently, that m only assigns mass to singletons. Then m' also only assigns mass to singletons, and is given by:

$$m'(\{x\}) = \frac{\alpha \ \Pr(\{x\} \cap \phi) + (1-\alpha) \ \Pr(\{x\})}{1 - \alpha + \alpha \ \Pr(\phi)}. \quad (1)$$

Thus, as a function of $\alpha \in [0,1]$, m' is a path that begins at Pr, ends at $(\Pr \mid \phi)$, and can even be viewed as a "proportion

Cut this. In the previous you argued that confidence is not the same as posterior probability. Now you suddenly bring in prior probability. The story here is completely unclear.

of the way to incorporation", just like in Example 1—yet it is parameterized differently. Therefore, to appropriately determine a numerical value of confidence, you need to know something more about the updating procedure.

Alternatively, suppose that m is not a probability but rather another simple support function on ϕ. Then so is $m' = m \oplus s$. How much total evidence for ϕ does m' represent? It is overwhelmingly standard to have a measurement that combines additively: if you had three (distinct) gallons of water and get another, you now have four; if you had six (independent) random bits and get three more, you now have nine. What would be necessary to get an additive measurement confidence, for simple support functions? Shafer calls such a quantity *weight of evidence*, and proves that that of s must be of the form $w = k \log(1 - \alpha)$ for some $k < 0$ [Shafer, pg 78]. Note that this is precisely the expression for t in (2), because a choice of $\iota < 1$ is equivalent to a choice of $k = \log(1 - \iota) < 0$. Weight of evidence is not just additive; it also plays a fundemental role in the theory of belief functions. For example, it provides a canonical (and minimal) way of decomposing combined evidence into simple support functions [Shafer, Theorem 5.5]. □

Cut this; it's irrelevant to your story and takes up valuable space.

Shafer's theory addresses two problematic aspects of the Bayesian epistemic picture. It (1) allows for belief states that represent ignorance, and (2) for observations other than those that "establish a single proposition with certainty" [Shafer, 1976, Chapter 1: §7,§8]. The theory is effective on both counts, but we are only interested in the second: we have no prescription for how beliefs ought to be represented, but rather a theory of uncertain evidence that applies no matter how they are represented. Shafer's theory, because it addresses (1), applies only in an esoteric paradigm in which one's belief state is a Dempster-Shafer belief function. The notion of confidence we present in this paper can be thought of a vast generalization of how Shafer handles issue (2).

Why take up valuable space to mention the first if you're only interested in the second?

Cut this; (1) it's irrelevant to our story; (2) I see no reason that it's true.

It's not so "vast".

This notion of confidence applies far more broadly. We now give an example with the same critical elements, but a very different flavor.

Example 3. Consider a neural network, whose "belief state" is a setting of weights. For definiteness, suppose we are talking about a classifier, so that for every setting of weights θ, there is a function $f_\theta : X \to \Delta Y$ that maps inputs $x \in X$ to distributions $f_\theta(x)$ over possible class lables Y. Modern learning algorithms (like gradient descent) are iterative procedures that repatedly make incremental changes to the weights. Therefore, if we perform one iteration of such a procedure to update θ using a labeled training example $\phi = (x, y)$ to obtain new weights θ', there is no guarantee that $f_{\theta'}(x)$ gives high probability to y—only that it is higher than it was before. In other words, such algorithms (in contrast to their historical counterparts like conjunction learning algorithms [**?**]) do not take any one encounter with a training example too seriously—that is, they make low-confidence updates to the weights. This relative distrust of individual data points is arguably what makes the training process robust to noisy or contradictory observations.

Nevertheless, if we train by repeatedly updating on ϕ, the weights do eventually converge. Adopting (and spending the resources to compute) these convergent weights is unorthodox, and appropriate only if we have complete trust in ϕ, meaning we find it critical that x always be classified as y. At the opposite extreme, if we have no trust in ϕ, we should simply discard it without changing our weights. Our definition of a full-confidence update also suggests what to do for intermediate levels of confidence: simply stop the training process before convergence. Thus, the number of training iterations n functions as a description of intermediate levels of confidence: it interpolates between no confidence (zero iterations) and full confidence (infinitely many iterations).

In the simplest settings, training examples do not come with confidence annotations, so typically one effectively assigns them all the same (small) confidence value, a number which is closely related to the learning rate. In richer settings, per-example confidences often arise explicitly: perhaps from agreement between annotators, or confidence scores in self-training [Zou et al., 2019]). We would also like to point out that the trust conveyed by a confidence level encompasses more than just accurcacy. Suppose, for example, that the classifier is intended to aid in hiring, and that we would like to change our current discriminatory hiring practices. In this case, we ought to have low trust in historical training data, not because it is inaccurate, but because we don't want it to take it too seriously in forming our new hiring practice. □

As Ii've said many times before, this is just the wrong story at this point in the paper.

We have now seen two very different ways of describing confidence. Are they related? Should we prefer one style over the other? The number $\alpha \in [0, 1]$ used in Example 1 appears to have an important advantage: an intermediate value can be clearly interpreted as a "proportion of incorporation", while the number of training iterations n used in Example 3 only really has meaning relative to other values of n for the same training algorithm. On the other hand, the apprach taken in the second example appears to be more general. It is not clear that a proportion $\alpha \in [0, 1]$ is meaningful in Example 3, but Example 1 can be readily captured by a number of iterations n, as follows. If we fix some "unit confidence", say $\iota = 0.01$, then for any $n \in \mathbb{N}$, sequentially performing n updates of confidence ι is equivalent to a single one of confidence $\alpha = 1 - (0.99)^n$. Or, inverting: to specify confidence α, it suffices to instead give a number

$$n = \log(1 - \alpha) / \log(1 - \iota) \quad \in [0, \infty] \tag{2}$$

of confidence-ι updates to be performed in sequence.

I know that you're enamored with additivity, but it's not the right story here.

In three places now we have seen evidence suggesting that additive descriptions of confidence do not have objective meaning, but rather are only meaningful relative to some inscrutible choice. The meaning of n in Example 3 depends on

I said this before, but let me be a bit more blunt. This story is horrible. You don't ever formally describe how updating with confidence works with neural networks. I thought you had agreed to rewrote this. I stopped reading here, but I will pint out that I still think it's important to add Kalman filters here.

the learning algorithm, the quantity in (2) requires an irrelevant choice of ι, and the weight of evidence is only unique up to the constant k. However, we will see in Section 3.1 that insofar as there is an α with an objective meaning, there is also a natural additive scale for confidence: the unique one for which small values of n and small values of α are indistinguishable, or equivalently, the unique base for which derivatives of α and n with respect to one another (or any third quantity) are identical. Thus, there is a special scale of additive confidence $n \in [0, \infty]$ for the same reason that there is natural base for exponentiation: it is the only base whose differential at 0 is the identity. (Indeed in these simple cases, this amounts to taking $k = -\log e$ and $\iota = 1 - \frac{1}{e}$ so that all logarithms are base e—but at a deeper level, this is a consequence of the uniqueness of the exponential map of the Lie algebra implicit in our framework.) To summarize: there is a unique way to represent a confidence $\alpha \in [0, 1]$ as a confidence $\beta \in [0, \infty]$ such that the two scales behave identically when both are small. But the intuition behind additive confidence $\beta \in [0, \infty]$ generalizes past where the intuition for $\alpha \in [0, 1]$ will take us.

We conclude with a toy example that showcases an assortment of other features and themes that can be captured with our definition of confidence.

Example 4. Jugo is an impartial juror. Like the other jurors, she has two buttons in front of her, labeled G and N. Her instructions are to listen to evidence, and press G to increase the probability of a guilty verdict, and N to increase the probability of a not-guilty verdict.

More concretely, the system works as follows. There are J jurors, labeled $\{1, \dots, J\}$; let $pressed(j, \mathsf{B}, t)$ be a variable that is equal to one if juror $j \in$ is pressing button B button at time t, and zero otherwise. The "belief state" of this automated system is a single number $g \in [0, 1]$, representing the probability of a guilty verdict. When a single juror presses G, g approaches 1 exponentially, and if they instead press N, g decays to zero. In the first case (G is pressed) the system evoves according to $\frac{\mathrm{d}g}{\mathrm{d}t} = (1 - g)$ while in the second, $\frac{\mathrm{d}g}{\mathrm{d}t} = -g$. The first is the vector field associated with the G button, and the second is the vector field associated with N. The total effect of all buttons is then the sum of that of all buttons across all vector fields, when they are active:

$$\frac{\mathrm{d}g}{\mathrm{d}t} = \sum_{j=1}^{J} pressed(j, \mathsf{G}, t)(1 - g) - pressed(j, \mathsf{N}, t)(g) \ ,$$

so that g exponentially approahces 1 when more G buttons are pressed than N buttons, and symmetricaly, exponentially approaches 0 when more N buttons than G buttons are pressed. At the end of the trial, the defendant is convicted with probability equal to the final value of g.

Let ϕ represent a piece of evidence suggesting guilt, presented by the prosecution from time t_1 to time t_2, and suppose for now that only buttons labeled G are pressed in this interval. The system measures j's confidence in ϕ by

$$w_j := \int_{t_1}^{t_2} G_j(t)\mathrm{d}t = \text{ total time } j \text{ presses G during } \phi,$$

Note that $w_j = 0$ if and only if j does not press any buttons, which (a) indicates that j does not trust the evidence ϕ, and (b) communicates this fact to the system, by telling it to ignore the evidence. Note that this is an additive representation of confidence, since pressing the button for four seconds, and then three more later, is by definition the same as pressing it for seven. While the maximum possible confidence of w_j is $(t_2 - t_1)$, this system does not allow a juror to express *full* confidence in ϕ because no finite amount of G-pressing will result in a guilty verdict with probability one; it is always possible to increase the value of g through additional evidence.

Altogether, the system's confidence in ϕ can be measured by as the unique value W for which

$$\int_{t_1}^{t_2} W(1 - g(t))\mathrm{d}t = g(t_2) - g(t_1),$$

which, so long as only G buttons are pressed, equals $W := \sum_j w_j$, so this measure of confidence is additive across jurors as well as across time. This is appropriate, since the jurors are independent and not communicating with each other. As before, $W = 0$ if and only if no juror presses any buttons between times t_1 and t_2, indicating zero trust leant to ϕ. In such a case, the system ignores ϕ in updating its beliefs. And just as no individual juror can send a full-confidence update to the system, the system cannot recieve a full-confidence from the jurors as a whole.

The picture gets significantly more complicated if we consider the possibility that jurors might press the N button. For example, if ϕ, which was intended as evidence of guilt, has the effect of getting jurors to press N, there is a sense in which they have *negative* confidence in ϕ, since the belief update happened in the opposite direction of what ϕ represents; rather than *no* trust, this is represents *distrust*. Small negative updates are always possible except at the boundary of belief space, but in this paper, we focus almost entirely on positive confidence updates.

The introduction of the second button also uncovers a significant source of complexity: unlike Examples 1 to 3, the order that evidence is presented matters, when there is more than one possible response to it. Evidence presented later has a larger effect, meaning that this system exhibits a recency bias.

Now consider a variant of this system that does not trust all jurors equally; rather, it trusts each juror j to a degree $\beta_j \in [0, \infty]$, and now g evolves according to

$$\frac{\mathrm{d}g}{\mathrm{d}t} = \sum_{j=1}^{J} \beta_j \Big(G_j(t)(1 - g) - gN_j(t) \Big).$$

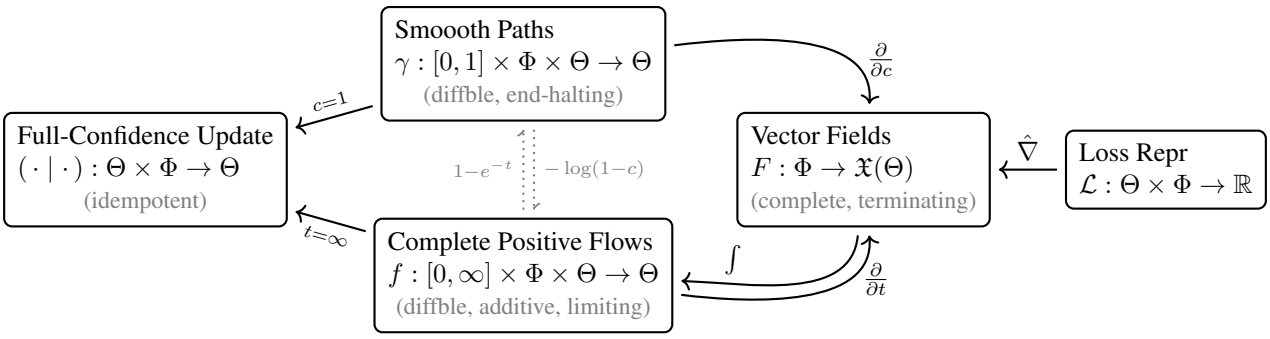


Figure 1: Different representations of update functions, and the relationships they have with one another. Sections 3 and

In this case, the system can be said to have trust β_j in juror j, since j's buttons are ignored when $\beta_j = 0$. When $\beta_j = \infty$ (an expression of full confidence in j), g immediately jumps to 0 when j presses N, or to 1 if j presses G (unless canceled by another full-confidence juror pressing the opposite button). If all jurors have full confidence, then the verdict of this system is a majority vote at the last moment a button was pressed. Thus, the weights attached to weighted combinations are (additive) expressions of confidence as well. □

Example 4 illustrates how a (sufficiently nice) vector field, which is simpler than a smooth path for every starting point, is enough to define an additive notion of confidence, via its integral curves. It may seem strange to define confidence via a vector field, which does not mention confidence at all—but in a sense, it works because a vector field captures precisely everything about the update *except* for the confidence. We do this formally in Section 3.1.

In **??**, we demonstrate that it is typically possible to get an even more compact representation of the updating process, by representing the vector field implicitly as gradients of some "loss function". To do this at full generality, we need to make sense of a gradient, which requires more structure, in the form of a Riemannian metric. It turns out that, up to a multiplicative constant, there is a unique natural Riemannian metric on any parameterization of a probability distributions Chentsov [1982]; taking gradients with respect to this geometry, show how familiar loss functions on probability measures correspond to different standard notions of confidence in the other representations.

Once we have the formalism, fully in place, we give further examples of how confidence works in exponential families, in particular showing how Kalman gain and inverse variance can be viewed as confidence as well.

2 UPDATE FUNCTIONS AND THEIR PATH REPRESENTATIONS

Throughout, we use Θ to refer to some space of possible belief states, and Φ for a set of possible inputs. Now, suppose we have some initial belief state θ, and observe ϕ with confidence $\alpha \in [0,1]$. How do our beliefs change as a result? To study this mathematically, we assume that the process can be captured in functional form:[1]

[CF0] There exists some function $F : \Phi \times [0,1] \times \Theta \to \Theta$ which, given input $\phi \in \Phi$, confidence $\chi \in [0,1]$, and a prior belief state $\theta \in \Theta$, produces the appropriate posterior belief state $F(\phi, \chi, \theta) \in \Theta$.

Given such a function F, as well as a statement $\phi \in \Phi$ and confidnece $\chi \in [0,1]$, we call $F^{\chi}_{\phi} = F(\phi, \chi, -) : \Theta \to \Theta$ an *update*. To ensure F captures the belief updating process, we also require that it satisfy some axioms, characterizing the important shared features between Examples 1 to 3. First, we would like no-confidence updates to leave our belief state unchanged.

[CF1] For all $\theta \in \Theta$ and $\phi \in \Phi$, $F^0_{\phi}(\theta) = \theta$.

Text The appropriate way to deal with full confidence, on the other hand, depends on the relationship between Θ and Φ, but it still characterized by an important property.

As I thought we agreed, this long discussion of full confidence should be greatly shortened.

2.1 UPDATING WITH FULL CONFIDENCE

Since the purpose of F^1_{ϕ} is to *fully* incorporate ϕ into our beliefs, 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 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.

[1] It should be straightforward to extend our theory so as to handle randomized updates as well; the point is that the belief state, observation, and confidence must together contain enough information to describe the updating process.

This is problematic, because the issue of succiessive updates is, in general, problematic. You haven't discussed it at all. Indeed, you haven't even hinted thjat there's an issue.

Definition 1. A *full-confidence (Θ-)update rule* (for Φ) is a mapping $P : \Phi \times \Theta \to \Theta$ such that for all $\phi \in \Phi$, $P_\phi = (\theta \mapsto P(\phi, \theta)) : \Theta \to \Theta$ is idempotent. That is, $P_\phi(P_\phi(\theta)) = P_\phi(\theta)$ for all $\phi \in \Phi$ and $\theta \in \Theta$. □

[CF2] Full-confidence updates are idempotent. Or, equivalently, $F^1 = (\phi, \theta) \mapsto F(\phi, 1, \theta) : \Phi \times \Theta \to \Theta$ is a full-confidence update rule.

Once Θ, Φ, and any implicit structure in them is specified, there is often a natural choice of full-confidence update rule. To illustrate, we now consider three different 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".

2.1.1 Conditioning. First, consider the case where observations are events, i.e., $\Phi := 2^W$. Here, the appropriate rule seems to be conditioning: starting with $\mu \in \Delta W$, 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 a full-confidence update.
- 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 it is consistent with one's beliefs).

There are well-known issues with conditioning μ on A when $\mu(A) = 0$, typically this operation is undefined. Note that to satisfy CF2, the result must either be μ itself or a distribution that gives probability 1 to A.

2.1.2 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 full-confidence update rule $f : \Phi \times W \to W$ on individual worlds, which we interperet as assigning each statement $\phi \in \Phi$ and $w \in W$ an element $f(\phi, w) \in W$ which is the unique "world most similar to w, in which ϕ is true" [Gardenfors, 1982]. In this case, idempotence of $f_\phi : W \to W$ 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 full confidence update rule F for ΔW with the pushforward

$$F(\phi, \mu) = A \mapsto \mu(\{w : f(w, \phi) \in A\})$$

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.

Why are we considering this here? What's the story?

2.1.3 Jeffrey's Rule. Next, consider a more general form of observation, in which observations themselves are probabilities. Formally, let Φ be the set of pairs (X, π), where $X : W \to S$ is a random variable taking values in some set S, and $\pi \in \Delta S$ is a probability on S. Jeffrey's rule is then:

$$\mathrm{J}_{(X,\pi)}(\mu) := \sum_{x \in S} \pi(X{=}x)\ \mu\big|(X{=}x)$$

When π places all mass on some $x \in S$, Jeffrey's Rule amounts to conditioning on $X{=}x$. For this reason, Jeffrey's Rule is sometimes often thought of as a generalization of conditioning that admits for less that complete certainty. However, it is still a full-confidence update rule—just one that handles observations that can be uncertain.

Let $\mu' := J_{\pi(X)}(\mu)$ be the result of applying Jeffrey's rule for (X, π) to μ. Note that $\mu'(X) = \pi(X)$, so $\pi(X)$ has been fully incorporated into μ', while all information about the old prior belief about X has been destroyed by the update.

Please refrain from using idiosyncratic expressions like "middling confidence

2.2 THE PATH OF MIDDLING CONFIDENCE

Full updates are quite extreme. An agent that updates by conditioning, for instance, permanently commits to believing everything it ever learns, and gains nothing from making the same observation twice. Clearly humans are not like this; revisiting information improves our learning [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 for less extreme belief alterations, in which information is only partially incorporated. This is the role of intermediate values of confidence.

I don't understand this notation. Please use something more standard.

Since confidence is supposed to interpolate between prior beliefs and full update, we would like each $\chi \mapsto F(\theta, \chi, \phi) : [0, 1] \to \Theta$ to be a continuous path in belief states, from the initial belief to full incorporation.

[CF3] Θ comes with a topology, with respect to which the restriction $F_\phi|_\theta : [0, 1] \to \Theta$ is continuous for all $\phi \in \Phi$ and $\theta \in \Theta$.

I realize that you mean the | to represent restriction, but you haven't defined F \phi, the function you're restricting.

It would be particularly nice if updates were continuous in our initial beliefs as well: after all, similar priors typically result in similar posteriors beliefs. This would allow us to strengthen CF3 to something simpler:

[CF3′] Θ comes with a topology, with respect to which $F_\phi : [0, 1] \times \Theta \to \Theta$ is continuous for all $\phi \in \Phi$.

Unfortunately, this is too strong to handle extreme beliefs. In the probabilistic case, for instance:

Proposition 1. *There is no continuous extension of conditioning to a function F satisfying CF3′.*

I don't understand what you mean by this.

This is a consequence of the fact that there's no way to extend conditioning continuously to zero-probability events. What we can do is to define a set Θ_ϕ of belief states that do not outright contradict ϕ, and ask that F_ϕ is continuous when restricted to this set. Better yet, we can do the reverse, and define Θ_ϕ as the set of beliefs on which F_ϕ is continuous.

Proposition 2. *There is a maximal open set $\Theta_\phi \subseteq \Theta$ such that the restriction $F_\phi|_{\Theta_\phi} : [0, 1) \times \Theta_\phi \to \Theta$ is continuous.*

Sorry; you're way down in the weeds here. I strongly suspect you've lost all readers at this point. They've gotten bored and don't see how this fits into the story.

 Why do we care about \Theta_\phi

In Example 1, $\Theta_\phi = \{\mu \in \Delta W : \mu(\phi) > 0\}$. In Example 3, Θ_ϕ consists of all weights $\mathbf{w}$ such that the loss $\mathcal{L}(\mathbf{w}, \phi) < \infty$ that the training algorithm minimizes is finite.

Confidence interpolates between ignoring and fully incorporating information. This interpolation becomes more useful if it is not just continuous, but differentiable as well.

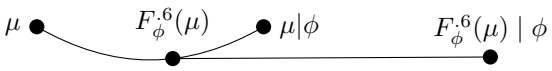
[CF4] Θ carries a differentiable structure,[2] and for all $\phi \in \Phi$ and $\theta \in \Theta$, the function $[0, 1) \ni \chi \mapsto F(\theta, \chi, \phi) \in \Theta$ is differentiable.

This is quite nonstandard; please do not use this notation.

To motivate our next axiom, suppose that we learn ϕ with some small degree of confidence χ, but then immediately realize that we should have updated with higher confidence. To fix this, it seems clear that we need to gain additional confidence in ϕ, which can be done with a second update (with a smaller residual degree of confidence).

[CF5] If $0 < \chi_1 < \chi_2 < 1$, then for all $\theta \in \Theta$, there exists $\chi_3 \in (0, \chi_2)$ such that $F_\phi^{\chi_3}(F_\phi^{\chi_1}(\theta)) = F_\phi^{\chi_2}(\theta)$.

The point of the monotonicity condition CF5 is to rule out cases such as the one depicted below, in which "resuming" an update results in a totally different set of belief states.



CF5 also ensures that a high confidence update can be decomposed into several updates of lower confidence.

Finally, we impose a mild regularity condition that can be thought of as a limit CF5 when confidences χ_1 and χ_2 become infinitessimally close; we make this precise via differentiability (CF4).

CF6 also says that there are no saddle points. Why is that reasonable?

[CF6] If $\frac{\partial}{\partial \chi} F_\phi^\chi(\theta) = 0$ and $\chi' > \chi$, then $F_\phi^{\chi'}(\theta) = \theta$.

Intuitively, CF6 amounts to requiring that updates do not "temporarily pause" for intermediate values of confidence, and then resume for higher values of confidence. The culmination of these axioms results in our first definition of an update function that takes confidence into account.

Definition 2. A function $F : \Phi \times [0, 1] \times \Theta \to \Theta$ satisfying CF1–6 is a *path update function*. □

I don't like this story (a number \chi \in [0,1] is not the only useful way to quantify confidence). Which concepts are you referring to when you say "these concepts"?

Path update functions capture some important aspects of uncertain updating—but a number $\chi \in [0, 1]$ is not the only useful way to quantify confidence. For the full picture of of how these concepts are related, we will need to look at confidence from several more angles.

[2] That is, either Θ is either a subset of $\mathbb{R}^n$, or a differentiable manifold (with corners) [Lee, 2013, Joyce, 2009].

No! You can't presume that the reader knows what this is, and if it's important, you can't get away with pointing to a reference. If it's not important, just assume that \Theta is a a subset of R^n in CF4, and then say (in a footnote) that this applies more generally, and in particular, if \Theta is a differentiable manifold with corners (whatever that is).

3 ADDITIVE REPRESENTATIONS OF CONFIDENCE, AND THEIR VECTOR FIELDS

Bad story. Just say that you want additivity.

Recall that the number of training iterations n in Example 3 and Shafer's weight of evidence w in Example 2 are measurements of confidence that do not lie in $[0, 1]$, but rather in $[0, \infty]$. In such cases, the appropriate analogue is a function

$$F : \Phi \times [0, \infty] \times \Theta \to \Theta \tag{3}$$

satisfying modified versions of CF1–6 that use ∞ in place of 1 for the upper limit of confidence. One reason to prefer this scale is that it allows us to represent degree of confidence in a way that combines additively. Nearly all measurable quantities used in science and everyday life can be described additively: if you have six (feet/meters/galons/joules/people/votes/dollars), and then you find seven additional (distinct) ones, then you have thirteen altogether. We would like a measure of confidence that also works this way.

Text

Again, this needs a serious discussion of what it means to have consecutive updates.

[CF7] For all $\chi_1, \chi_2 \in [0, \infty]$, $F_\phi^{\chi_1} \circ F_\phi^{\chi_2} = F_\phi^{\chi_1 + \chi_2}$.

Recall how CF5 allows us to decompose high-confidence updates into sequences of low-confidence ones; CF7, which implies CF5, describes a particularly convenient way that the decomposition might work.

Definition 3. A *flow update function* is a function $F : \Phi \times [0, \infty] \times \Theta \to \Theta$ satisfying the appropriate analogues of CF1–4 and CF7. □

This is not a definition, since you haven't defined "appropriate analogues".

To some, CF7 might already be pallatable, but it is clearly a nontrivial assumption, and looks like it might severely restrict the expressiveness of our update formalism. Fortunately, this is not the case. While CF7 does significantly pin down how confidence can be measured, it has no effect on what confidence can express.

You have to explain what this means and why anyone would care.

Theorem 3. *If F satisfies CF0–6, then there is a unique flow update rule ${}^+F$ that behaves like F for low confience updates (and is also additive: CF7). Furthermore, there exists a function g such that, for all θ, ϕ, and χ,*

$$F(\phi, \chi, \theta) = {}^+F(\phi, g(\phi, \chi, \theta), \theta).$$

Thus, updates performed with F are equivalent to updates performed with ${}^+F$, except that the degree of confidence needs to be translated appropriately (via g).

3.1 VECTOR FIELD REPRESENTATIONS

We now turn to another representation of additive update functions, as vector fields. This representation allows us to extend the set Φ of possible observations to to a larger set $\overline{\Phi} \supseteq \Phi$ with some algebraic operations.

Please add parens!

In CF4, we assumed that Θ has a differentiable structure; thus, 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 over Θ is then a smooth map $X : \Theta \to T\Theta$ assigning a tangent vector $X(\theta) = (\theta, \mathbf{v}) \in TM$ to every $\theta \in \Theta$; the set of all vector fields over Θ is denoted $\mathfrak{X}(\Theta)$ and is closed under linear combination [Lee, 2013, §8]. There is a close relationship between additive confidence and such vector fields. Given a flow update function F, and observation ϕ, the differential of F is a vector field

Define "smooth"

What's TM?

I'm going over a cliff. Why are you doing this? Why should I care? What does this have to do with confidence?

$$F'_\phi := \frac{\partial}{\partial \chi} F^\chi_\theta \Big|_{\chi=0} \in \mathfrak{X}(\Theta). \tag{4}$$

Moreover, we can recover F_ϕ as the integral curves of F'_ϕ.

Proposition 4. *Let F be a flow update function, and fix the vector field F'_ϕ. F_ϕ is the only flow update satisfying* (4).

It's not counterintuitive because I have no intuition for it (because you haven't given me any).

Thus, every flow update function can be equivalently represented by its differential. It may seem counter-intuitive that F'_ϕ, which no longer mentions confidence at all, can capture confidence. In a sense, it does so by specifying everything about the update *except* for the degree of confidence. Having separated the confidence from the mechanics of the update, this vector field representation allows us to

??? This is exactly what I'm missing. What does the vector field reprsentation allow us to do? This should have come first!

3.2 ORDERLESS COMBINATION OF OBSERVATIONS

One defining feature of vector fields is that they can be linearly combined to form new vector fields. Therefore flow update rules, which are in natural correspondance with differentiable additive update rules, also inherit this structure.

The first way of combining propositions is to rescale them. For $\phi \in \Phi$ and a scalar $k \in [0, \infty)$, we can extend F a new input $k \cdot \phi \in \overline{\Phi}$ by

What's the intuition for this new input?

$$F^\chi_{k\cdot\phi}(\theta) := F^{k\chi}_\phi(\theta).$$

In this way, the set Φ inherits the additivity of the update rule in the form of scalar multiplication. It turns out more is possible: updates inherit the entire vector space structure.

I have no idea of how to think of \phi_2 \oplus \phi_2 if \phi_1 and \phi_2 aren't independent (and, of course, you haven't told me).

The second way of combining propositions is to "run them concurrently". Given $\phi_1, \phi_2 \in \Phi$, we can form a new input $\phi_1 \oplus \phi_2 \in \overline{\Phi}$, and we extend F on it by adding the vector fields $F'_{\phi_1 \oplus \phi_2} := F'_{\phi_1} + F'_{\phi_2}$. Note that by definition, $\phi_1 \oplus \phi_2 = \phi_2 \oplus \phi_1$, so this is a way of combining observations orderlessly, even in cases where ϕ_1 and ϕ_2 do not commute. And when ϕ_1 and ϕ_2 already do not depend on order, $\phi_1 \oplus \phi_2$ has the same effect as ϕ_1 followed by ϕ_2.

Proposition 5. *If ϕ_1 and ϕ_2 commute (i.e., $F^\chi_{\phi_1} \circ F^\chi_{\phi_2} = F^\chi_{\phi_2} \circ F^\chi_{\phi_1}$ for all χ), then for all χ,*

$$F^\chi_{\phi_1}(F^\chi_{\phi_2}(\theta)) = F^\chi_{\phi_2 \oplus \phi_1}(\theta) = F^\chi_{\phi_1 \oplus \phi_2}(\theta) = F^\chi_{\phi_1}(F^\chi_{\phi_2}(\theta))$$

Intuitively, $\phi_1 \oplus \phi_2$ is a "mixture observation" containing one part ϕ_1 and one part ϕ_2. This intuition is made precise by the following proposition, which

Proposition 6.

Example 5. ML example: dataset = orderless combination. Rescaling = changing learning rate.

⟨ INCOMPLETE ⟩

□

Definition 4. For an assertion language Φ, let $\overline{\Phi}$ denote the space of weighted formal sums of elements of Φ. □

Proposition 7. *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

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

3.3 LINEAR UPDATE RULES

What does this have to do with confidence? What does it tell me about confidence?

There are many definitions of linear update rules:

Definition 5. 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 8. *If F is a $\mathcal{L}$-cvx-linear, then it is also cvx-linear.*

In fact, the first condition is much stronger;

Proposition 9. *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 10. *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.*[3]

Proposition 11. *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.*

4 OPTIMIZING UPDATES, AND THEIR LOSS REPRESNTATION

This is a paper about confidence, not update rules and their loss representation. Either connect it clearly to confidence or cut it! I will not read this until you've done that.

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,[4] so as to obtain a vector field on Θ which optimizes $\mathcal{L}$.

Then we can define an update rule $\mathtt{GF}[\mathcal{L}]$ that reduces inconsistency by gradient flow (the continuous limit of gradient descent). Concretely, such an update rule has a vector field:

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

Proposition 12. *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 6 (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

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

- 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

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

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

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

[4] 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.

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 $[\frac{\partial \mathcal{L}}{\partial \theta_1}, \dots, \frac{\partial \mathcal{L}}{\partial \theta_n}]^\mathsf{T}$.

4.0.1 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 via exponential decay:

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

Proposition 13. *Bolzmann Update Rules are additive, zero, differentiable, invertable, and commutative.*

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

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

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, $\mathrm{Bolz}[U]^\beta_\varphi(\mathrm{Unif})$ 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 5). 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 $X = \{1, \dots, n\}$, 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 14. *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 $g(X) := U(X, \varphi)$.

$$\mathrm{Boltz}'_\varphi\theta = \frac{\partial}{\partial\beta}\mathrm{Boltz}^\beta_\varphi(p)\Big|_{\beta=0}$$

⟨ TODO: finish typesetting algebra ⟩

$$= x \mapsto p(x)\frac{f(x)}{\mathbb{E}_p[f]}\left(\mathop{\mathbb{E}}_p\left[\frac{f}{\mathbb{E}_p[f]}g\right] - g(x)\right)\Big|_{\beta=0}$$
$$= \frac{pf}{\mathbb{E}_p[f]^2}\left(\mathbb{E}_p[fg] - g\,\mathbb{E}_p[f]\right)\Big|_{\beta=0}$$
$$= x \mapsto p(x)(\mathbb{E}_p[g] - g(x))$$

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 15. *The optimizing update rules for $\Theta = \Delta X$ whose loss represntation is linear (i.e., an expected utility), are precisely the Bolzmann update rules. In particular, 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 7 (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 $\theta = (\mu, \sigma^2)$:

$$\mathcal{L}(x, \mu, \sigma^2) = -\log \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{2}\log 2\pi\sigma^2 + \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

$$= \left\langle \begin{matrix} \mu \\ \sigma^2 \end{matrix} \twoheadrightarrow \begin{matrix} \mu \\ \sigma^2 \end{matrix} \overset{\mathcal{N}}{\rightarrowtail} X \twoheadleftarrow x \right\rangle.$$

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.,

since $f(X) = 1$ when $\beta = 0$

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

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,

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

Equivalently, the precision of the resulting distribution is the average of the confidence c and the previous precision $1/\sigma_0^2$, which suggests that confidence is of the same type as precision. Note that if σ_0^2 is very large, so that our initial beliefs are very uncertain, updating with confidence c results in variance $\frac{1}{t}$. In this sense, the magnitude of confidence acts as the inverse of variance.

□

5 FURTHER EXAMPLES, IN DEPTH

5.1 UPDATE RULES FOR DISCRETE PROBABILITES

5.2 UPDATE RULES FOR PARAMETRIC FAMILIES

5.3 KALMAN FILTERS

6 DISCUSSION

References

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

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

Dynamics F	Flow $F^c_A(\mu)$	Vector Field $F'_A(\mu)$	Loss $\mathcal{L}^F(\mu, A)$	F^c_A, F^d_B commute?
LIN	$(1-c)\,\mu + (c)\,\mu\|A$	$\mu\|A - \mu$	$-\log \mu(A)$	if $\mu(A \cap B) > 0$
LLI	$\propto \mu^{1-c}\,(\mu\|A)^c$ $\propto \mu\,\mathbb{1}_A$	N/A	N/A	always
Bolz$[\mathbf{1}]$	$\propto \mu \cdot \exp(\beta \mathbb{1}_A)$ $\propto \mu \cdot \exp(-\beta \mathbb{1}_{\bar{A}})$	$\mu \odot (\mathbb{1}_A - \mu(A))$ $= \mu(A)\,(\mu\|A - \mu)$ $= \mu \odot (\mathbb{1}_A - \mu(A))$	$\mu(\neg A)$	always

Table 1: A comparison of different update rules when $\Theta = \Delta W$ and $\Phi = 2^W$

Dynamics F	Flow $F^c_q(\mu)$	Vector Field $F'_q(\mu)$	Loss $\mathcal{L}^F(\mu, q)$	Properties
LIN	$(1-c)\,\mu + (c)\,q$	$q - \mu$	$\boldsymbol{D}(q \parallel \mu)$	
LLI	$\propto \mu^{1-c}\,q^c$	$\mu \odot \left(\log \frac{\mu}{q} - \boldsymbol{D}(\mu \parallel q)\right)$	$\boldsymbol{D}(\mu \parallel q)$	

Table 2: Different update rules and their representations when $\Theta = \Phi = \Delta W$

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

Dominic Joyce. On manifolds with corners, 2009. URL https://arxiv.org/abs/0910.3518.

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

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

John Nash. The imbedding problem for riemannian manifolds. *Annals of mathematics*, pages 20–63, 1956.

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

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.

A DISCUSSION ON INCREMENTAL CONFIDENCE AND INDEPENDENCE ASSUMPTIONS

Historically, 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 8. 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 $\mathcal{C}$ 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 8 in mind, here are three ways to proceed.

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 [**?**].

I2. Appropriately Enrich Domains. In Example 8, 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 8 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 8.

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 10).

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 8, 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.

We would like to point out that readers who find who find it reasonable to ignore inputs you have no confidence in (per CF1) have implicitly either accepted either Item I1 or Item I3, as the next example shows.

Example 9. 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. □

We state our results with the incremental interpretation of confidence, with the understanding that all of our results also admit a more conservative reading, in which confidence is measured absolutely, and also all applications of the function F are independent.

B 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 (CF7) 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 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 $\mathcal{C}$, CF7 has no teeth.

Proposition 16. *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 CF7.*

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:

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

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

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

□

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 17. 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. □

C EXTRA

C.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 18. *If F is a differentiable and invertable update rule (i.e., satisfies CF1, CF4, CF7 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 18.1. *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.*

D

Example 10. Suppose F is an additve update rule. Then, we can explicitly construct a resolution to the problem posed in Example 8 by defining enriched spaces

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

and new commitment function G by

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

□

E MORE ON PATH UPDATE RULES

Since each ϕ corresponds to a path

Definition 6 (Homotopic update rules)**.** $\phi \sim_F \psi$ iff they behave the same way for full confidence (that is, $F^1_\phi(\theta) = F^1_\psi(\theta)$ for all $\theta \in \Theta$) and there exists a continuous function $H : \Theta \times [0,1] \times [0,1]$ such that, for all $\theta \in \Theta$ and $\chi \in [0,1]$,

1. $H(\theta, \chi, 0) = F(\theta, \chi, \phi)$,
2. $H(\theta, \chi, 1) = F(\theta, \chi, \psi)$

and for all $s \in [0,1]$,

3. $H(\theta, 0, s) = \theta$;
4. $H(\theta, 1, s) = F^1_\phi(\theta) = F^1_\psi(\theta)$, the last two of which are the same by assumption. □

As usual, homotopy is an equivalence relation.

For example, the Dempster-Shafer update rule (1) is homotopic to the linear update rule from Example 1.

F PROOFS

Theorem 3. *If F satisfies CF0–6, then there is a unique flow update rule ${}^+F$ that behaves like F for low confience updates (and is also additive: CF7). Furthermore, there exists a function g such that, for all θ, ϕ, and χ,*

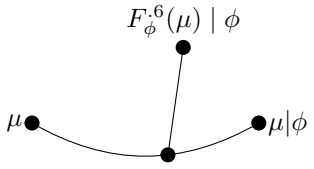
$$F(\phi, \chi, \theta) = {}^+F(\phi, g(\phi, \chi, \theta), \theta).$$

Proof. □

G COUNTER-EXAMPLES

G.1 DROPPING ??

Without CF5, we could have a situation like the below, in which resuming an update could take you to a completely different position if you were interrupted in the middle of the path.



G.2 HALTING UPDATES

Example 11. Suppose $\Theta = \mathbb{R} \cup \{-\infty, \infty\}$, $\Phi = \{\star\}$ is a singleton, and and

$$F^t_\star(\theta) = \theta + t + \sin(t).$$

This is interesting because it "pauses" Clearly F satisfies CF1–4. It also satisfies CF5, since it is increasing in t (but it is not additive CF7). Its vector field is given by

$$\frac{\partial}{\partial t} F^t_\star(\theta) = 1 + \cos(t)\Big|_{t=0} = 2,$$

so its unique additive representation is ${}^+F^t_\star(\theta) = \theta + 2t$. □

Example 12. Now suppose Θ and Φ are as before, but now

$$F^t_\star(\theta) = \theta + t - \sin(t).$$

As before, it satisfies CF1–5, but now its vector field is

$$\frac{\partial}{\partial t} F^t_\star(\theta) = 1 - \cos(t)\Big|_{t=0} = 0,$$

so its unique additive representation does nothing: ${}^+F^t_\star(\theta) = \theta$. □