Problem Statement

Consider the dynamic epistemic language

$$p \mid \neg\varphi \mid \varphi \wedge \psi \mid \mathbf{K}\varphi \mid \mathbf{T}\varphi \mid [P]\varphi$$

K is knowledge. **T** is more interesting — $\mathbf{T}\varphi$ says that the current world is 'minimal' or 'most typical' over worlds satisfying φ. (As far as I can tell, this is not quite the same as the [best] operator, see Remark 10 in [2]). $[P]$ is some dynamic update given by $\mathcal{M} \to \mathcal{M}^\star_P$ (this is a free variable; the problem will be to find the right update).

For the static part of the logic, choose your favorite semantics — plausibility models, evidence models, etc. For now, I'll take Johan's approach from [3], which I've been using as a desk reference for all this. Let's assume we have a single-agent plausibility model, with an extra accessibility relation R for knowledge: $\mathcal{M} = \langle W, R, \leq, V\rangle$. $\leq$ is uniform over all states; we do not have a different plausibility relation $\leq_s$ for each state. As usual, $x \leq y$ reads "the agent finds x at least as plausible as y."

Definition 1. The semantics are given by

$$\begin{array}{lll}
\mathcal{M}, w \Vdash p & \text{iff} & w \in V(p) \\
\mathcal{M}, w \Vdash \neg\varphi & \text{iff} & \mathcal{M}, w \nVdash \varphi \\
\mathcal{M}, w \Vdash \varphi \wedge \psi & \text{iff} & \mathcal{M}, w \Vdash \varphi \text{ and } \mathcal{M}, w \Vdash \psi \\
\mathcal{M}, w \Vdash \mathbf{K}\varphi & \text{iff} & \text{for all } u \text{ with } wRu, \mathcal{M}, u \Vdash \varphi \\
\mathcal{M}, w \Vdash \mathbf{T}\varphi & \text{iff} & w \text{ is } \leq\text{-minimal over } \{u \mid \mathcal{M}, u \Vdash \varphi\} \\
\mathcal{M}, w \Vdash [P]\varphi & \text{iff} & \mathcal{M}^\star_P, w \models \varphi
\end{array}$$

I will use the shorthand $[\![\varphi]\!]_{\mathcal{M}} = \{u \mid \mathcal{M}, u \Vdash \varphi\}$, and drop $\mathcal{M}$ when it's understood from context. I should also point out that, by definition of $\leq$-minimal, we have the validities $\mathbf{T}\varphi \to \varphi$ (Refl) and $\mathbf{T}\varphi \to \mathbf{TT}\varphi$ (Trans).

Iterated Hebbian learning, formalized as a dynamic update on neural network models, can be reduced to this language [1]. The reduction axioms are:

$$\begin{array}{lll}
[P]p & \leftrightarrow & p \quad \text{for propositions } p \\
[P]\neg\varphi & \leftrightarrow & \neg[P]\varphi \\
[P](\varphi \wedge \psi) & \leftrightarrow & [P]\varphi \wedge [P]\psi \\
[P]\mathbf{K}\varphi & \leftrightarrow & \mathbf{K}[P]\varphi \\
[P]\mathbf{T}\varphi & \leftrightarrow & \mathbf{T}([P]\varphi \wedge (\mathbf{T}P \vee \mathbf{K}(\mathbf{T}P \vee \mathbf{T}[P]\varphi)))
\end{array}$$

I would like to understand what neural network updates are doing "classically," i.e. for each neural network update, what is an "equivalent" update over possible worlds / plausibility / evidence models? For iterated Hebbian learning, my question for you is:

Question. Is there a dynamic model update (over your classical model of choice) that satisfies these reduction axioms?

I've been stuck on this since November, and it's much trickier than I initially thought.

Progress So Far

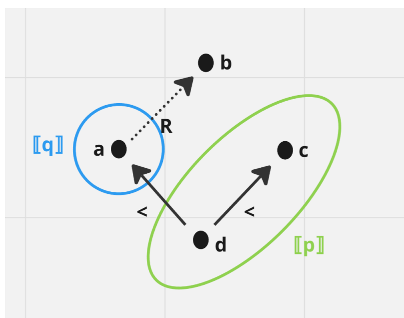
I've somewhat misled you by talking in terms of plausibility models. In fact, the reduction above is *invalid* for relational plausibility upgrades (where the only thing we're changing is $\leq$).

Proposition 1. No plausibility upgrade $\mathcal{M} \to \mathcal{M}^\star$, where $\mathcal{M} = \langle W, R, \leq, V \rangle$ and $\mathcal{M}^\star = \langle W, R, \leq^\star, V \rangle$ can make the axioms for iterated Hebbian learning valid.

Proof. Let $\mathcal{M} \to \mathcal{M}^\star$ be any plausibility upgrade, and suppose that the first four axioms are valid for this upgrade. I will show that the very last axiom cannot hold for all $\mathcal{M}, w$; specifically, this propositional instance will fail:

$$[p]\mathbf{T}q \leftrightarrow \mathbf{T}(q \wedge (\mathbf{T}p \vee \mathbf{K}(\mathbf{T}p \vee \mathbf{T}q)))$$

Let's construct a $\mathcal{M}$ and w that make it fail. Let $\mathcal{M}$ be



Since $[p]q \leftrightarrow q$ is valid, $[\![q]\!]_{\mathcal{M}_p^\star} = [\![q]\!]_{\mathcal{M}} = \{a\}$. We pick w to be a: a is $\leq$-minimal over $[\![q]\!]_{\mathcal{M}_p^\star}$, and so $\mathcal{M}, a \Vdash [p]\mathbf{T}q$.

However, $\mathcal{M}, a \nVdash \mathbf{T}(q \wedge (\mathbf{T}p \vee \mathbf{K}(\mathbf{T}p \vee \mathbf{T}q)))$. If it did, then by reflexivity of $\mathbf{T}$ we would have $\mathcal{M}, a \Vdash q \wedge (\mathbf{T}p \vee \mathbf{K}(\mathbf{T}p \vee \mathbf{T}q))$. But a does not satisfy the right conjunct. First, $\mathcal{M}, a \nVdash \mathbf{T}p$ (since a is not a $\leq$-minimal element of $[\![p]\!]$). And second, there *is* b with aRb such that b is not a $\leq$-minimal element of either $[\![p]\!]$ *or* $[\![q]\!]$. So $\mathcal{M}, b \nVdash \mathbf{T}p \vee \mathbf{T}q$, and thus $\mathcal{M}, a \nVdash \mathbf{K}(\mathbf{T}p \vee \mathbf{T}q)$. $\square$

Discussion. This immediately rules out lexicographic upgrade, conservative upgrade, and other variants, since these update policies just re-order the plausibility relation $\leq$. The proof also shows that $[P]\varphi \leftrightarrow \varphi$ rules out any upgrade that just re-assigns propositions (i.e. modifies V). We might instead consider updates that add or remove states or change the knowledge relation R, but we have to be very careful — for example, conditionalization (i.e. public announcement $[!P]$ for a single agent) is also ruled out by $[P]\varphi \leftrightarrow \varphi$ as well as $[P]\mathbf{K}\varphi \leftrightarrow \mathbf{K}[P]\varphi$.

I've considered looking at neighborhood models (e.g. evidence models) to model this update. It's not clear to me what the neighborhood semantics for $\mathbf{T}$ should be, and I'm currently trying to find out. What are your thoughts?

References

[1] Caleb Schultz Kisby, Saúl A Blanco, and Lawrence S Moss. "What Do Hebbian Learners Learn? Reduction Axioms for Iterated Hebbian Learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 13. 2024, pp. 14894–14901.

[2] Johan Van Benthem. "Dynamic logic for belief revision". In: *Journal of applied non-classical logics* 17.2 (2007), pp. 129–155.

[3] Johan Van Benthem. *Logical dynamics of information and interaction*. Cambridge University Press, 2011.