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 $[\mathsf{best}]$ operator, see Remark 10 in [2]). $[P]$ is some dynamic update given by $\mathcal{M} \to \mathcal{M}_P^{\star}$ (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

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

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.

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? In this case, 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 (I probably should have reached out sooner).

Progress So Far

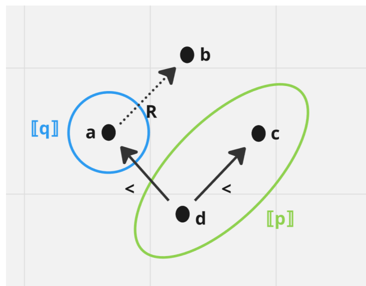
I've somewhat misled you by talking in terms of plausibility models. In fact, the reduction above is *invalid* for relational plausibility upgrades.

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

Proof. Let $\mathcal{M} \to \mathcal{M}^\star$ be any plausibility upgrade. 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



Note that $\mathcal{M} \to \mathcal{M}^\star$ only modifies $\leq$. This means that $[\![q]\!]_{\mathcal{M}} = [\![q]\!]_{\mathcal{M}_p^\star}$. So in particular, $[\![q]\!]_{\mathcal{M}_p^\star}$ is finite and nonempty. So there is some w that is $\leq$-minimal over $[\![q]\!]_{\mathcal{M}_p^\star}$. So $\mathcal{M}, w \Vdash [p]\mathbf{T}q$.

WHOOPS, THINK ABOUT IT MORE

I will now show that no choice of w can satisfy $\mathbf{T}(q \wedge (\mathbf{T}p \vee \mathbf{K}(\mathbf{T}p \vee \mathbf{T}q)))$. Well, the only u such that $\mathcal{M}, u \Vdash q$ is a. But $\mathcal{M}, a \nVdash \mathbf{T}p$ (since a is not a $\leq$-minimal element of $[\![p]\!]$). Additionally, 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 $\mathcal{M}, a \nVdash \mathbf{KT}p \vee \mathbf{T}q$.

I will now show that $\mathcal{M}, w$ does not satisfy either of the disjuncts. In particular:

1. $\mathcal{M}, w \nVdash \mathbf{T}(p \wedge q)$, since $[\![p]\!] \cap [\![q]\!] = \emptyset$ (and so there is no $\leq$-minimal element of $[\![p]\!] \cap [\![q]\!]$). So $\mathcal{M}, w$ does not satisfy the left disjunct.
2. $\mathcal{M}, w \nVdash \neg\Diamond(p \wedge \langle\mathbf{T}\rangle q)$, since there *does* exist b such that $\mathcal{M}, b \Vdash p \wedge \langle\mathbf{T}\rangle q$ (observe: $b \in [\![p]\!]$ and b is not minimal in $[\![q]\!]^{\complement}$). So $\mathcal{M}, w$ does not satisfy the right disjunct. $\square$

The crucial step of this proof is finding this $\leq$-minimal w in $[\![q]\!]_{\mathcal{M}_p^\star}$. Note that this step does not rely on the well-foundedness of $\leq$ — we can construct a similar model that is not well-founded if we like. But it *does* rely on the fact that $[p]q \leftrightarrow q$ is valid: re-ordering $\leq$ *cannot add or remove* elements from $[\![q]\!]$. In particular, the proof would break if our update could make $[\![q]\!]$ empty or make $[\![q]\!]$ include an infinite descending chain. (But I can't figure out an update that would do these in the right way…)

Corollary 1. No plausibility upgrade can make Axiom B valid.

The proof is a simple extension of the above proof, replacing p with $\langle\mathbf{T}\rangle p$. We can show the same for axiom C, by modifying the construction slightly.

Proposition 2. No plausibility upgrade can make Axiom C valid.

Proof. Consider the propositional instance of Axiom C:

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

Let $\mathcal{M} \to \mathcal{M}^\star$ be any plausibility upgrade. This time, let $\mathcal{M}$ be

[PICTURE]

Again, $\mathcal{M} \to \mathcal{M}^\star$ only modifies $\leq$, and in particular $[\![q]\!]_{\mathcal{M}_p^\star}$ is finite and nonempty. So there is some w that is $\leq$-minimal over $[\![q]\!]_{\mathcal{M}_p^\star}$. So $\mathcal{M}, w \Vdash [p]\mathbf{T}q$.

TODO $\square$

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.