The Logic of Hebbian Learning

Caleb Kisby and **Saúl A. Blanco**
Department of Computer Science, Indiana University, Bloomington, IN 47408, USA
{cckisby, sblancor}@indiana.edu

Lawrence S. Moss
Department of Mathematics, Indiana University, Bloomington, IN 47405, USA
lmoss@indiana.edu

Abstract

We present the logic of Hebbian learning, a dynamic logic whose semantics are expressed in terms of a layered neural network learning via Hebb's associative learning rule. Its language consists of modality $\Box\varphi$ (read "the propagation of signal φ") as well as dynamic modalities $[\varphi^+]\psi$ (read "evaluate ψ after performing Hebbian update on φ"). We give axioms and inference rules that are sound with respect to the neural semantics; these axioms characterize Hebbian learning and its interaction with propagation. The upshot is that this logic describes a neuro-symbolic agent that both learns from experience and also reasons about what it has learned.

Introduction

Artificial intelligence has long been marked by a schism between two of its major paradigms: symbolic reasoning and connectionist learning. Neural systems have had wild success with learning from unstructured data, whereas symbolic reasoners are notorious for their rigidity. On the other hand, symbolic systems excel at sophisticated (static) reasoning tasks that neural systems cannot readily learn. Symbolic modules also tend to have more explainable reasoning, thanks to their use of explicit inferences in an intuitive language. Moreover, due to their connection with logic, it is straightforward to compare the relative power and complexity of different symbolic reasoners.

But as Valiant famously put it, intelligent cognitive agents must have *both* "the ability to learn from experience, and the ability to reason from what has been learned" (Valiant 2003). *Neuro-symbolic artificial intelligence* has emerged in the last few decades to meet this challenge — a monumental effort to integrate neural and symbolic systems, while retaining the advantages of both (see (Bader and Hitzler 2005) and (Sarker et al. 2021), two surveys that span the decades). Despite the cornucopia of neuro-symbolic proposals, the field has not yet agreed on an interface between the two that satisfyingly preserves both flexible learning and expressive reasoning.

Following the path set out by (Balkenius and Gärdenfors 1991) and (Leitgeb 2001; 2003), we advance the following proposal for the neuro-symbolic interface. Rather than viewing the neural and symbolic as two different systems to be combined, we view them as two ways of interpreting the same agent. More precisely, we view the dynamics of neural networks as the semantics to a formal logic. This will immediately give us a formal correspondence between the chosen neural network model and this logic.

Previous work, particularly (Leitgeb 2001), has considered how forward propagation in feed-forward nets forms a sound and complete semantics for the (static) conditional logic **CL** (*loop-cumulative*). The novelty of our paper is that we extend this logic by viewing a simple learning policy — Hebbian update ("neurons that fire together wire together") — as a dynamic modality. By doing so, we demonstrate that the dynamics of Hebbian learning (in feed-forward nets) directly corresponds to a particular multimodal logic that we call *the logic of Hebbian learning*. This logic meets Valiant's challenge: It characterizes a cognitive agent that can learn from experience and also reason about what it has learned.

Our main result is the soundness of axioms and inference rules that characterize Hebbian learning. The most interesting axioms involve the interaction between Hebbian update and forward propagation. We also offer intuitive readings for the two modalities in terms of *preference change*. And although we leave the question of completeness open, we close by considering the importance of completeness for logics of this kind.

Related Work

Logics with Neural Semantics. As mentioned above, this work builds on (Balkenius and Gärdenfors 1991), (Leitgeb 2001; 2003), which formally characterize the dynamics of a variety of inhibitory neural networks as

corresponding conditional logics. Similarly, (Blutner 2004) demonstrates that the Hopfield networks correspond to the logic of what he calls "weight-annotated Poole systems." But the idea that we can view neural networks as the semantics for symbolic reasoning dates back to (McCulloch and Pitts 1943). To date, no such neural semantics has tackled the issue of learning — doing this for Hebbian update is precisely the contribution of our paper.

Neuro-Symbolic AI. Across the neuro-symbolic literature, an ubiquitous premise is that integration involves combining or composing otherwise distinct neural and symbolic modules. In contrast, this paper presents the neural and symbolic as two perspectives we can have about the same agent.

To our knowledge, the combined work work of (Garcez, Broda, and Gabbay 2001) and (Garcez, Lamb, and Gabbay 2008) is the only neuro-symbolic proposal (besides neural semantics, see above) that exhibits this intimate interface between the two. The former gives a formally sound method for extracting nonmonotonic rules from a network and the latter gives a method for build neural network models from rules (in a variety of different logics). When combined, we can freely translate between a neural network and its beliefs. But once again, unlike our work this framework does not offer a logical account of the neural network's learning.

Logics for Learning. Two recent papers, (Baltag et al. 2019) and (Baltag, Li, and Pedersen 2019), also present dynamic modal logics that characterize learning. The former models an individual's learning in the limit, whereas the latter models supervised learning as a game played between student and teacher. But it is unclear how learning policies expressed in these logics might relate to specific neural implementations of learning, such as Hebbian update and backpropagation.

Background

Neural Network Models

A model of the logic of Hebbian learning is just a special type of artificial neural network that we call a *binary feedforward neural network* (BFNN).

Definition 1. A BFNN is a pointed directed graph $\mathcal{N} = \langle N, E, W, T, A, \eta \rangle$, where

- N is a finite nonempty set (the set of neurons)
- $E \subseteq N \times N$ (the set of excitatory connections)
- $W : N \times N \to \mathbb{R}$ (the weight of a given connection)
- T is a function which maps each $n \in N$ to $T^{(n)} \in \mathbb{R}$ (the threshold for n)
- A is a function which maps each $n \in N$ to $A^{(n)} : \mathbb{R}^k \to \mathbb{R}$ (the activation function for n, where k is the indegree of n)
- $\eta \in \mathbb{R}, \eta \geq 0$ (the learning rate)

Moreover, BFNNs are *feed-forward* in the sense that they do not contain cycles of edges with all nonzero weights. BFNNs are *binary* in the sense that the *output* $O^{(n)} : \mathbb{R} \to \{0, 1\}$ of a neuron n is given by $O^{(n)}(x) = 1$ iff $x \geq T^{(n)}$.

We further require that activation functions $A^{(n)}$ are

- **(Nondecreasing)** For all $\vec{x}, \vec{y} \in \mathbb{R}^k$,
$$\text{if } \vec{x} \leq \vec{y} \text{ then } A^{(n)}(\vec{x}) \leq A^{(n)}(\vec{y})$$
- **(Subadditive)** For all $\vec{x}, \vec{y} \in \mathbb{R}^k$,
$$A^{(n)}(\vec{x} + \vec{y}) \leq A^{(n)}(\vec{x}) + A^{(n)}(\vec{y})$$

This class of activation functions includes in particular the rectified linear activation function (ReLU).

We write W_{ij} to mean $W(i, j)$, for $(i, j) \in E$. When m_i is drawn from a sequence $m_1, \ldots, m_k$, we can write $A^{(n)}(\overrightarrow{W}(m_i, n))$ as shorthand instead of the full expression $A^{(n)}(W(m_1, n), \ldots, W(m_k, n))$.

The Dynamics of Propagation

Of course, BFNNs are not merely static directed graphs, but are dynamic in nature. When a BFNN receives a signal (which we model as the initial state), it propagates that signal forward until the state of the net stabilizes. This stable state of the net is considered to contain the net's response (answer) to the given signal (question). We model forward propagation as follows, drawing heavily from the approach proposed by (Leitgeb 2001; 2018).[1]

We consider a neuron n active if its activation $A^{(n)}$ exceeds its threshold $T^{(n)}$ (intuitively, if the neuron fires). Since our BFNNs are binary, either a given neuron is active (1) or it is not (0). So we can identify the state of $\mathcal{N}$ with a set of neurons that are active. For a given BFNN $\mathcal{N}$, let

$$\mathsf{Set} = \{S \mid S \subseteq N\}$$

be its set of states.

A state $S \in \mathsf{Set}$ can further activate more neurons when the outputs of neurons in S cause these new neurons' activations to exceed their thresholds. Eventually, the state of $\mathcal{N}$ stablizes. We call this final state of affairs $\mathsf{Prop}(S)$, the *propagation* of S.

Definition 2. Let $\mathsf{Prop} : \mathsf{Set} \to \mathsf{Set}$ be defined recursively as follows: $n \in \mathsf{Prop}(S)$ iff either

(Base Case) $n \in S$, or

(Constructor) There exist $m_1, \ldots, m_k \in \mathsf{Prop}(S)$ where each $(m_i, n) \in E$ and

$$A^{(n)}(\overrightarrow{W}(m_i, n)) \geq T^{(n)}$$

[1] Note that (Leitgeb 2001; 2018) defines propagation for *inhibition nets*, i.e. weightless binary feed-forward networks whose dynamics are determined by both excitatory and inhibitory connections. But later in this same paper, Leitgeb proves that inhibition nets and binary feed-forward networks (BFNNs) are equivalent with respect to their propagation structure. And so we import his ideas and results here.

Alternatively, consider a finite automaton with state space Set and transition function $F_{S^*} : \mathsf{Set} \to \mathsf{Set}$ tracking the propagation of an initial state S^* through $\mathcal{N}$. We can view $\mathsf{Prop}(S^*)$ as a fixed point of F_{S^*} (Leitgeb 2001).

The key insight of (Leitgeb 2001) is that we can neatly characterize the algebraic structure of Prop as a closure operator.

Theorem 1. Let $\mathcal{N} \in \mathsf{Net}$. For all $S, S_1, S_2 \in \mathsf{Set}$, Prop satisfies

- **(Inclusion)** $S \subseteq \mathsf{Prop}(S)$
- **(Idempotence)** $\mathsf{Prop}(S) = \mathsf{Prop}(\mathsf{Prop}(S))$
- **(Distributivity)** $\mathsf{Prop}(S_1) \cup \mathsf{Prop}(S_2) \subseteq \mathsf{Prop}(S_1 \cup S_2)$
- **(Cumulative)** If $S_1 \subseteq S_2 \subseteq \mathsf{Prop}(S_1)$ then $\mathsf{Prop}(S_1) = \mathsf{Prop}(S_2)$
- **(Loop)** If $S_1 \subseteq \mathsf{Prop}(S_0), \ldots, S_n \subseteq \mathsf{Prop}(S_{n-1})$ and $S_0 \subseteq \mathsf{Prop}(S_n)$, then $\mathsf{Prop}(S_i) = \mathsf{Prop}(S_j)$ for all $i, j \in \{0, \ldots, n\}$

Crucially, Prop is not monotonic — it is not the case that for all $S_1, S_2 \in \mathsf{Set}$, if $S_1 \subseteq S_2$, then $\mathsf{Prop}(S_1) \subseteq \mathsf{Prop}(S_2)$. The reason for this is that a BFNN's weights W_{ij} can be negative, which allows S_2 to to inhibit the activation of new neurons that were otherwise activated by S_1. But in place of monotonicity, Prop is *loop-cumulative* (in the terminology of (Kraus, Lehmann, and Magidor 1990)).

From Hebbian Learning to Logic

The Dynamics of Hebbian Learning

The plan from here is to extend this logic of propagation by providing an account of Hebbian learning. Our goal is to cast Hebbian update as a dynamic modality, so that we can explore its interactions with Prop in symbolic language. As with Prop, we start by outlining the algebraic structure of Hebbian update.

Hebb's classic learning rule (Hebb 1949) states that when two adjacent neurons are simultaneously and persistently active, the connection between them strengthens. In contrast with, e.g. backpropagation, Hebbian learning is errorless and unsupervised. Another key difference is that Hebbian update is local — the change in a weight ΔW_{ij} depends only on the activation of the immediately adjacent neurons. For this reason, the Hebbian family of learning policies has traditionally been considered more biologically plausible than backpropagation. There are many variations of Hebbian learning, but we only consider the most basic (unstable) form of Hebb's rule: $\Delta W_{ij} = \eta x_i x_j$, where η is the learning rate and x_i, x_j are the outputs of adjacent neurons i and j, respectively.

In order to incorporate Hebb's rule into our framework, we introduce a function Inc ("increase the weights") to strengthen those neurons in a BFNN $\mathcal{N}$ that are active when we feed $\mathcal{N}$ a signal (initial state) $S \in \mathsf{Set}$.

Definition 3. For $S \in \mathsf{Set}$, let $\chi_S : N \to \{0, 1\}$ be given by $\chi_S(n) = 1$ iff $n \in S$

Definition 4. Let $\mathsf{Inc} : \mathsf{Net} \times \mathsf{Set} \to \mathsf{Net}$ be such that $\mathsf{Inc}(\langle N, E, W, T, A, \eta\rangle, S) = \langle N, E, W^*, T, A, \eta\rangle$, where

$$W^*_{ij} = W_{ij} + \eta \cdot \chi_{\mathsf{Prop}(S)}(i) \cdot \chi_{\mathsf{Prop}(S)}(j)$$

Notice that we propagate S before getting the active status of neurons. This is because otherwise we would never strengthen connections beyond the input layer.

We were able to formulate the algebraic properties of Prop in terms of Set containment. Similarly, we express the properties of Inc in terms of Net containment.

Definition 5. Let $\mathcal{N}_1, \mathcal{N}_2 \in \mathsf{Net}$ differ only in their weight and threshold functions. We write

$$\mathcal{N}_1 \preceq \mathcal{N}_2$$

to mean that for all $S \in \mathsf{Set}$, $\mathsf{Prop}_{\mathcal{N}_1}(S) \subseteq \mathsf{Prop}_{\mathcal{N}_2}(S)$. We use $\cong$ in the usual way to express that $\mathcal{N}_1 \preceq \mathcal{N}_2$ and $\mathcal{N}_2 \preceq \mathcal{N}_1$.

For example, $\mathsf{Inc}(\mathcal{N}, S)$ is a supernet of $\mathcal{N}$ because strengthening weights via the Inc operation only has the potential to *expand* future propagations. To further cement this intuition, consider the least upper bound $\mathcal{N}^{lub}$ of $\preceq$. $\mathcal{N}^{lub}$ is that net whose weights have been "maximally" strengthened, and so every propagation $\mathsf{Prop}(S)$ results in the entire set N.

We have the following test to determine if $\mathcal{N}_1$ is a subnet of $\mathcal{N}_2$.

Lemma 2. Suppose $\mathcal{N}_1$ and $\mathcal{N}_2$ are the same except for their weight and threshold functions, and let $S \in \mathsf{Set}$. Then $\mathsf{Prop}_{\mathcal{N}_1}(S) \subseteq \mathsf{Prop}_{\mathcal{N}_2}(S)$ iff for all $m_1, \ldots, m_k, n \in \mathsf{Prop}_{\mathcal{N}_1}(S)$ with $(m_i, n) \in E$,

$$A^{(n)}(\overrightarrow{W}_{\mathcal{N}_1}(m_i, n)) \geq T^{(n)}_{\mathcal{N}_1} \quad \text{implies} \quad A^{(n)}(\overrightarrow{W}_{\mathcal{N}_2}(m_i, n)) \geq T^{(n)}_{\mathcal{N}_2} \qquad (*)$$

Corollary 1. Let $\mathcal{N}_1, \mathcal{N}_2$ be the same except for their weight and threshold functions. Then $\mathcal{N}_1 \preceq \mathcal{N}_2$ iff for all $m_1, \ldots, m_k, n \in N$, $(*)$ holds.

Notice that the right-hand side of this biconditional does not mention Prop; this test reduces the *dynamic* condition ($\mathcal{N}_1 \preceq \mathcal{N}_2$) to a *static* condition (a statement about the relative weights and thresholds between $\mathcal{N}_1$ and $\mathcal{N}_2$).

These two facts are exactly what we need for the following following algebraic characterization of Inc.

Theorem 3. For all $\mathcal{N}, \mathcal{N}_1, \mathcal{N}_2 \in \mathsf{Net}$ and $S, S_1, S_2 \in \mathsf{Set}$, Inc satisfies

- **(Inclusion)** $\mathcal{N} \preceq \mathsf{Inc}(\mathcal{N}, S)$
- **(Absorption)** $\mathsf{Inc}(\mathcal{N}, \mathsf{Prop}(S)) \cong \mathsf{Inc}(\mathcal{N}, S)$
- **(Monotonicity in $\mathcal{N}$)** if $\mathcal{N}_1 \preceq \mathcal{N}_2$ then $\mathsf{Inc}(\mathcal{N}_1, S) \preceq \mathsf{Inc}(\mathcal{N}_2, S)$
- **(Iteration)** $\mathsf{Inc}(\mathcal{N}, S_1 \cup \mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S_1)}(S_2)) \preceq \mathsf{Inc}(\mathsf{Inc}(\mathcal{N}, S_1), S_2)$
- **(Distributivity)** $\mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S_1)}(S) \cup \mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S_2)}(S) \subseteq \mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S_1 \cup S_2)}(S)$
- **(Local)** $\mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S_2)}(S_1) \subseteq \mathsf{Prop}_{\mathcal{N}}(S_1) \cup \mathsf{Prop}_{\mathcal{N}}(S_2)$
- **(Cumulative)** If $\mathsf{Prop}_{\mathcal{N}}(S_1) \subseteq \mathsf{Prop}_{\mathcal{N}}(S_2)$ and $\mathsf{Prop}_{\mathcal{N}}(S_2) \subseteq \mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S)}(S_1)$, then $\mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S)}(S_1) = \mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S)}(S_2)$
- **(Loop)** If $\mathsf{Prop}_{\mathcal{N}}(S_1) \subseteq \mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S)}(S_0)$, $\ldots$, $\mathsf{Prop}_{\mathcal{N}}(S_n) \subseteq \mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S)}(S_{n-1})$, and $\mathsf{Prop}_{\mathcal{N}}(S_0) \subseteq \mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S)}(S_n)$, then $\mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S)}(S_i) = \mathsf{Prop}_{\mathsf{Inc}(\mathcal{N},S)}(S_j)$ for all $i, j \in \{0, \ldots, n\}$

Like Prop, Inc is loop-cumulative (in S). But also like Prop, Inc is not monotonic in S. Consider a BFNN $\mathcal{N}$ with $N = \{a, b, c, d, e\}$, $W_{ab} = W_{ce} = 2$, $W_{bc} = -2$, $W_{de} = -1$, and $T = 2$. We have $\mathsf{Prop}_{\mathcal{N}_1}(S_1) = \{a, c, e\}$, whereas $\mathsf{Prop}_{\mathcal{N}_2}(S_2) = \{a, b\} = S_2$. This means that $\mathsf{Inc}(\mathcal{N}, S_1)$ increases the weights of $a \to c, c \to e$, while $\mathsf{Inc}(\mathcal{N}, S_2)$ has no effect on the weights of $\mathcal{N}$. And so $S_1 \subseteq S_2$, yet $\mathsf{Prop}_{\mathsf{Inc}\mathcal{N},S_1}(S) = \{c, d, e\} \not\subseteq \{c, d\} = \mathsf{Prop}_{\mathsf{Inc}\mathcal{N},S_2}(S)$.

Syntax and Semantics

We can now introduce the logic of Hebbian learning. Let $p, q, \ldots$ be finitely many propositional variables. These represent fixed, 'ontic' states, i.e. established choices of neurons that correspond to features in the external world. For example, p might be the set of neurons that encapsulates the color *pink*. We presume that we already agree on these states, although we acknowledge that this is a major unresolved empirical issue. As for more complex formulas:

Definition 6. The formulas of our language $\mathcal{L}$ are given by

$$\varphi ::= p \mid \top \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \to \varphi \mid \Box\varphi \mid [\varphi^+]\varphi$$

where p is any propositional variable. We define $\bot, \vee, \leftrightarrow$ and the dual modalities $\Diamond, \langle\varphi^+\rangle$ in the usual way.

The modalities $\Box$ and $[\varphi^+]$ reflect our two operations Prop and Inc, respectively. We intend for $\Box\varphi$ to denote "the propagation of signal φ," and for $[\varphi^+]\psi$ to denote "after performing Hebbian update on φ, evaluate ψ." We will soon offer more classical, truth-conditional readings of $\Box$ and $[\varphi^+]$, but for now we content ourselves with these neural denotations.

A model of our logic is just a BFNN $\mathcal{N}$ equipped with an interpretation function $[\![\cdot]\!] : \mathcal{L} \to \mathsf{Set}_{\mathcal{N}}$.

Definition 7. Let $\mathcal{N} \in \mathsf{Net}$. Our semantics are defined recursively as follows:

$[\![p]\!]$		$\in \mathsf{Set}$ is fixed, nonempty
$[\![\top]\!]$	$=$	$\emptyset$
$[\![\neg\varphi]\!]$	$=$	$\overline{[\![\varphi]\!]}$
$[\![\varphi \wedge \psi]\!]$	$=$	$[\![\varphi]\!] \cup [\![\psi]\!]$
$[\![\varphi \to \psi]\!]$	$=$	$[\![\top]\!]$ iff $[\![\varphi]\!] \supseteq [\![\psi]\!]$
$[\![\Box\varphi]\!]$	$=$	$\mathsf{Prop}([\![\varphi]\!])$
$[\![[\varphi^+]\psi]\!]$	$=$	$[\![\psi]\!]_{\mathsf{Inc}(\mathcal{N},[\![\varphi]\!])}$

It may seem odd that we interpret $\wedge$ as union (instead of intersection), $\to$ as superset (instead of subset), and $\top$ as $\emptyset$ (instead of N). But this choice reflects the intuition that neurons act as "elementary-feature-detectors" (Leitgeb 2001). For example, say $[\![\varphi]\!]$ represents those neurons that detect whether an object is a car, and $[\![\psi]\!]$ represents those neurons that detect the color yellow. If the net observes a yellow car ($\varphi \wedge \psi$), *both* sets of neurons $[\![\varphi]\!]$ and $[\![\psi]\!]$ activate. As for $\to$, "every car is yellow" ($\varphi \to \psi$) means that whenever a car is detected, the color yellow must be detected as well. That is, the set of neurons detecting car-features $[\![\varphi]\!]$ must *contain* those neurons detecting yellow-features $[\![\psi]\!]$. This justifies us reading propositional connectives classically, despite the backwards flavor of the semantics.

Note that Leitgeb's conditional $\varphi \Rightarrow \psi$ is equivalent to the formula $\Box\varphi \to \psi$ (Leitgeb 2001). With our earlier intuition that a propagation contains the answer to the given question, this crucial formula says that the net *classifies* φ as ψ. The reason for our notational facelift is that $\Box$ is more fine-grained than $\Rightarrow$, which allows our language to capture certain subtle interactions between propagation and Hebbian update.

Our interpretation of formulas is completely algebraic, in the sense that formulas denote sets rather than truth-values. But we can consider formulas to have truth-values as follows.

Definition 8. $\mathcal{N} \models \varphi$ iff $[\![\varphi]\!]_{\mathcal{N}} = \emptyset$.

This choice also appears to be strange at its surface. But it is a natural one in light of the fact that we defined $[\![\top]\!] := \emptyset$. For example, consider implication: $\mathcal{N} \models \varphi \to \psi$ holds iff $[\![\varphi \to \psi]\!] = \emptyset = [\![\top]\!]$, which is exactly the condition $[\![\varphi]\!] \supseteq [\![\psi]\!]$ we gave for implication.

A curious consequence is that if $\mathcal{N} \models \varphi$ and φ cannot be written to contain an implication $\to$, then φ must be a tautology. But we do not consider this troubling, since it only makes sense to consider a neural network's judgment of φ when given a state $[\![\psi]\!]$ the net is in.

	Basic Axioms
(PC)	All proposotional tautologies
(DUAL)	$\Diamond\varphi \leftrightarrow \neg\Box\neg\varphi$
(M)	$\Box(\varphi \wedge \psi) \rightarrow (\Box\varphi \wedge \Box\psi)$
(N)	$\Box\top$
(T)	$\Box\varphi \rightarrow \varphi$
(4)	$\Box\varphi \rightarrow \Box\Box\varphi$
	Inference Rules
(MP)	$\frac{\varphi \quad \varphi\rightarrow\psi}{\psi}$
($\text{NEC}_\Box$)	$\frac{\varphi}{\Box\varphi}$
(NEC_+)	$\frac{\psi}{[\varphi^+]\psi}$
($\text{C}_\Box$)	$\frac{\varphi\rightarrow\psi \quad \Box\psi\rightarrow\varphi}{\Box\varphi\leftrightarrow\Box\psi}$
(C_+)	$\frac{\psi\rightarrow\rho \quad [\varphi^+]\rho\rightarrow\psi}{[\varphi^+]\psi\leftrightarrow[\varphi^+]\rho}$
($\text{LOOP}_\Box$)	$\frac{\Box\varphi_0\rightarrow\varphi_1 \cdots \Box\varphi_{k-1}\rightarrow\varphi_k \quad \Box\varphi_k\rightarrow\varphi_0}{\Box\varphi_0\rightarrow\varphi_k}$
(LOOP_+)	$\frac{[\varphi^+]\psi_0\rightarrow\psi_1 \cdots [\varphi^+]\psi_{k-1}\rightarrow\psi_k \quad [\varphi^+]\psi_k\rightarrow\psi_0}{[\varphi^+]\psi_0\rightarrow\psi_k}$
	Reduction Axioms
(R_p)	$[\varphi^+]p \leftrightarrow p$
($\text{R}_\neg$)	$[\varphi^+]\neg\psi \leftrightarrow \neg[\varphi^+]\psi$
($\text{R}_\wedge$)	$[\varphi^+](\psi \wedge \rho) \leftrightarrow ([\varphi^+]\psi \wedge [\varphi^+]\rho)$
	Key Axioms
(ITER)	$[\varphi^+][\psi^+]\rho \rightarrow [(\varphi \wedge [\varphi^+]\psi)^+]\rho$
($\text{NEST}_\wedge$)	$[(\varphi \wedge \psi)^+]\rho \rightarrow ([\varphi^+]\rho \wedge [\psi^+]\rho)$
($\text{NEST}_\Box$)	$[\Box\varphi^+]\psi \leftrightarrow [\varphi^+]\psi$
(ND)	$[\varphi^+]\Box\psi \rightarrow \Box[\varphi^+]\psi$
(PP)	$\Box[\varphi^+]\psi \wedge \Box\varphi \rightarrow [\varphi^+]\Box\psi$

Figure 1: A list of sound rules and axioms of the logic of Hebbian learning.

Inference and Axioms

The proof system for our logic is the usual modal one: We have $\vdash \varphi$ iff either φ is an axiom, or φ follows from previously obtained formulas by one of the inference rules. If $\Gamma \subseteq \mathcal{L}$ is a set of formulas, we consider $\Gamma \vdash \varphi$ to hold whenever there exist finitely many $\psi_1, \ldots, \psi_k$ such that $\vdash \psi_1 \wedge \ldots \wedge \psi_k \rightarrow \varphi$.

We list the axioms and inference rules for our logic in Figure 1. Our main result is the soundness of these axioms and rules — we do not claim that this list forms a complete axiomatization (we revisit the question of completeness in the Conclusion).

The static base of our logic is formed by the non-normal modal logic **EMNT4**, along with inference rules ($\text{C}_\Box$), ($\text{LOOP}_\Box$) expressing the loop-cumulativity of Prop. This fragment of the proof system is just the conditional logic **CL** re-cast in terms of $\Box$.

Like $\Box$, we have the inference rules (NEC_+), (C_+), (LOOP_+) for $[\varphi^+]$ (the latter two express the loop-cumulativity of Inc). Since Hebbian update only affects the propagation of states, we have reduction axioms (R_p), ($\text{R}_\neg$), ($\text{R}_\wedge$). In lieu of a full reduction for $[\varphi^+]\Box\psi$, we instead have the weaker axioms (ND) and (PP). Finally, we also have axioms that deal with nested terms ($\text{NEST}_\wedge$)($\text{NEST}_\Box$) and repeated iteration (ITER).

Soundness of the above axioms is just a matter of matching each axiom with its corresponding property of Inc.

Theorem 4. The rules and axioms above are sound, i.e. hold for all $\mathcal{N} \in \mathsf{Net}$.

Hebbian Update as Preference Change

One of the key advantages of symbolic systems is their explainability. Typically, symbolic languages have an intuitive reading that help explain their inferences. But so far, we have not given a clear English reading of $\Box\varphi$ and $[\varphi^+]\psi$, divorced from the dynamics of neural networks. Here we argue for an account of $\Box\varphi$ and $[\varphi^+]\psi$ in terms of *preference change*.

In the literature on nonmonotonic reasoning, $\varphi \Rightarrow \psi$ is interpreted as "typical φ are ψ," where φ and ψ are read as generics (e.g. 'cars' and 'yellow things'). Since in our logic we can express this conditional as $\Box\varphi \rightarrow \psi$, it follows that $\Box\varphi$ by itself represents of the net's 'typical φ.'[2]

Philosophers use the term *preference* to refer to this notion of whatever is more typical, more normal, expected, or better according to the agent (Van Benthem and Liu 2007). It is often formally defined via a preference ordering $<$ that an agent has over possible worlds (Kraus, Lehmann, and Magidor 1990). But in our logic, preference manifests via relative containment $\subseteq$ of propagations. Roughly: Our cognitive agent prefers those signals whose propagations contain many other propagations.

In the same way $\mathsf{Inc}(\mathcal{N}, S^*)$ expands certain future propagations $\mathsf{Prop}(S)$, we can view $[\varphi^+]$ as updating the network's preference in the direction of φ. In other words, the logic of Hebbian learning is a logic of *preference change*, in the sense of (Van Benthem and Liu 2007). Traditionally, logics of this kind introduce some new dynamic operator $[\circ\varphi]$ into the static language of preferences. The new operator then modifies the agent's preference relation $<$ over worlds towards preferring φ. We leave the question of how $[\varphi^+]$ can be viewed classically as modifying a preference relation to future work.

Reading $\Box$ as preference and $[\varphi^+]$ as preference change, we can now understand the inferences made by a Hebbian learner more clearly. Consider the key axiom (PP), i.e. (Preference Preservation)

$$\text{(PP)} \quad \Box[\varphi^+]\psi \wedge \Box\varphi \rightarrow [\varphi^+]\Box\psi$$

[2] As further support for this typicality reading, note that our $\Box$ strongly resembles the typicality operator **T** introduced into the DL $\mathcal{ALC}$ by (Giordano et al. 2009). Similar to our $\Box$, **T** is inter-translatable with conditionals $\varphi \Rightarrow \psi$ via $\mathbf{T}(\varphi) \sqsubseteq \psi$ (notice the subtle flip in the direction of containment).

This says that if our agent prefers for ψ to be true after learning φ, but she also happens to prefer φ, then after learning φ her preference for ψ will be preserved. Informally, this reflects a peculiar kind of cognitive bias whereby a Hebbian agent maintains her prior attitudes when presented with good news.

The key axiom (ND), i.e. (No Disappointment)

$$\text{(ND)} \quad [\varphi^+]\Box\psi \to \Box[\varphi^+]\psi$$

says that if after learning φ, our agent prefers ψ, then she would have preferred ψ to be true after learning φ in the first place. In other words, she will never be disappointed. We loosely take this to be an expression of a Hebbian agent's eagerness to learn.

Conclusion and Future Work

In this paper, we gave sound axioms and rules characterizing the logic of Hebbian learning. This logic interfaces the neuro-symbolic divide by defining its modalities, $\Box$ and $[\varphi^+]$, as the propagation and Hebbian update of a neural network. We provided independent, intuitive readings for $\Box$ and $[\varphi^+]$ in terms of preference and preference change. The upshot of all this is that this logic describes a neuro-symbolic agent that learns associatively and also reasons about what it has learned.

We leave open the question of whether the axioms and rules we list are complete. But we take this opportunity to stress the importance of having strong completeness for logics of this kind. Strong completeness for a *static* neural semantics provides a bridge across which we can extract a set of rules Γ from an interpreted network, and also build an interpreted neural network implementing a set Γ. But once the neural network updates, we lose the interpretations of neurons that allow for these translations. If we had strong completeness for the *dynamic* logic, we could fully track the interpretations while the net learns and preserve this neuro-symbolic correspondence.

Beyond the logic of Hebbian learning, we expect that this framework will be a fruitful way to explore the formal nature of a variety of neuro-symbolic systems. Some exciting future directions include:

1. Mapping more expressive syntax to neural activity
2. Generalizing to a broader class of neural networks
3. Generalizing to a broader class of activation functions
4. Characterizing other learning policies in logical terms

The holy grail of this line of work is to completely axiomatize the (1) first-order logic of (2) nonbinary (fuzzy-valued) neural networks with (3) more varied (e.g. sigmoid) activation functions that (4) learn via backpropagation.

References

Bader, S., and Hitzler, P. 2005. Dimensions of neural-symbolic integration-a structured survey. *arXiv preprint cs/0511042*.

Balkenius, C., and Gärdenfors, P. 1991. Nonmonotonic Inferences in Neural Networks. In *KR*, 32–39.

Baltag, A.; Gierasimczuk, N.; Özgün, A.; Sandoval, A. L. V.; and Smets, S. 2019. A dynamic logic for learning theory. *Journal of Logical and Algebraic Methods in Programming* 109:100485.

Baltag, A.; Li, D.; and Pedersen, M. Y. 2019. On the right path: a modal logic for supervised learning. In *International Workshop on Logic, Rationality and Interaction*, 1–14. Springer.

Blutner, R. 2004. Nonmonotonic inferences and neural networks. In *Information, Interaction and Agency*. Springer. 203–234.

Garcez, A. d.; Broda, K.; and Gabbay, D. M. 2001. Symbolic knowledge extraction from trained neural networks: A sound approach. *Artificial Intelligence* 125(1-2):155–207.

Garcez, A. S.; Lamb, L. C.; and Gabbay, D. M. 2008. *Neural-symbolic cognitive reasoning*. Springer Science & Business Media.

Giordano, L.; Olivetti, N.; Gliozzic, V.; and Pozzato, G. L. 2009. $\mathcal{ALC}+$ T: a preferential extension of description logics. *Fundamenta Informaticae* 96(3):341–372.

Hebb, D. 1949. *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Kraus, S.; Lehmann, D.; and Magidor, M. 1990. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial intelligence* 44(1-2):167–207.

Leitgeb, H. 2001. Nonmonotonic reasoning by inhibition nets. *Artificial Intelligence* 128(1-2):161–201.

Leitgeb, H. 2003. Nonmonotonic reasoning by inhibition nets II. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 11(supp02):105–135.

Leitgeb, H. 2018. Neural Network Models of Conditionals. In *Introduction to Formal Philosophy*. Springer. 147–176.

McCulloch, W. S., and Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5(4):115–133.

Sarker, M. K.; Zhou, L.; Eberhart, A.; and Hitzler, P. 2021. Neuro-symbolic artificial intelligence: Current trends. *arXiv preprint arXiv:2105.05330*.

Valiant, L. G. 2003. Three problems in computer science. *Journal of the ACM (JACM)* 50(1):96–99.

Van Benthem, J., and Liu, F. 2007. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics* 17(2):157–182.