Logical Dynamics of Neural Network Learning

Caleb Schultz Kisby

Collab. with Saúl Blanco & Larry Moss
Indiana University

1st GALAI Workshop
January 26, 2024

Foundations for Neuro-Symbolic AI

From van Harmelen (2022):[1]

> "What are the possible interactions between knowledge and learning? Can reasoning be used as a symbolic prior for learning ... Can symbolic constraints be enforced on data-driven systems to make them safer? Or less biased? Or can, vice versa, learning be used to yield symbolic knowledge? And if so, how to manage the inherent uncertainty that comes with such learned knowledge..."
>
> "... neuro-symbolic systems currently lack a theory that even begins to ask these questions, let alone answer them. All too often, new conference papers and ArXiv manuscripts simply propose a new neuro-symbolic architecture, or a new algorithm, without even discussing which of the above questions (or any others, for that matter) they aim to address."

[1] F. Harmelen. "Preface: The 3rd AI Wave Is Coming, and It Needs a Theory". In: *Neuro-Symbolic Artificial Intelligence*. Ed. by P. Hitzler and M. Sarker. IOS Press BV, 2022.

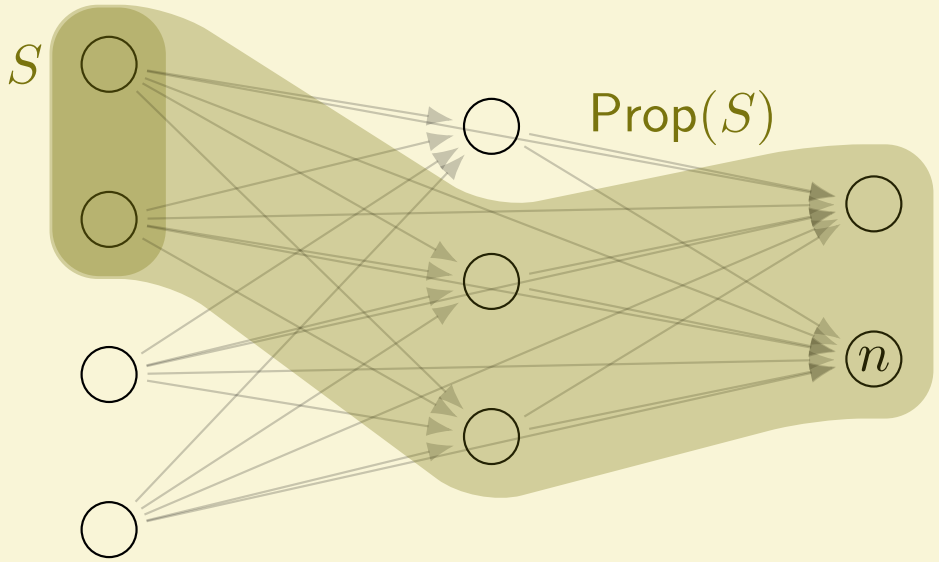
Neural Network Semantics

Neural Network Semantics

$$\mathcal{N} = \langle N, E, W, A, \eta, [\![\cdot]\!] \rangle$$

$$\mathsf{Prop}_{\mathcal{N}}(S) : \mathcal{P}(N) \to \mathcal{P}(N)$$

$\mathsf{Prop}_{\mathcal{N}}(S) =$ the set of all nodes that are eventually activated on input S



$$\mathcal{N} \models \varphi \Rightarrow \psi \text{ iff } \mathsf{Prop}_{\mathcal{N}}([\![\varphi]\!]) \supseteq [\![\psi]\!]$$

A Brief Timeline

1991. Balkenius & Gärdenfors realize that Prop behaves like a nonmonotonic conditionals
2001. Leitgeb proves this, via a completeness theorem
2003. Leitgeb extends completeness to various neural network architectures
2022. Giordano, Gliozzi, & Theseider Dupré prove soundness for fuzzy activation functions
2022. Schultz Kisby, Blanco, & Moss prove soundness for a basic learning policy (Hebbian learning)
2022. Odense & d'Avila Garcez start writing a survey of approaches like this (they call this "semantic encoding")
2024. Our new result: Completeness for Hebbian Learning! (Accepted paper at AAAI 2024)

Soundness & Completeness

- Leitgeb (2001)[2] showed that the neural semantics for $\varphi \Rightarrow \psi$ are completely axiomatized by:

$$\text{Refl.}\quad \varphi \Rightarrow \varphi$$

$$\text{LLE.}\quad \frac{\varphi \leftrightarrow \psi \qquad \varphi \Rightarrow \rho}{\psi \Rightarrow \rho}$$

$$\text{Weak.}\quad \frac{\varphi \rightarrow \psi \qquad \rho \Rightarrow \varphi}{\rho \Rightarrow \psi}$$

$$\text{CC.}\quad \frac{\varphi \wedge \psi \Rightarrow \rho \qquad \varphi \Rightarrow \psi}{\varphi \Rightarrow \rho}$$

$$\text{CM.}\quad \frac{\varphi \Rightarrow \psi \qquad \varphi \Rightarrow \rho}{\varphi \wedge \psi \Rightarrow \rho}$$

$$\text{Loop.}\quad \frac{\varphi_0 \Rightarrow \varphi_1 \qquad \cdots \varphi_{k-1} \Rightarrow \varphi_k \qquad \varphi_k \Rightarrow \varphi_0}{\varphi_i \Rightarrow \varphi_j}$$

- This is classified as a "Loop-Cumulative" conditional by Kraus, Lehmann, and Magidor (1990)[3]

[2] H. Leitgeb. "Nonmonotonic reasoning by inhibition nets". In: *Artificial Intelligence* 128.1-2 (2001).

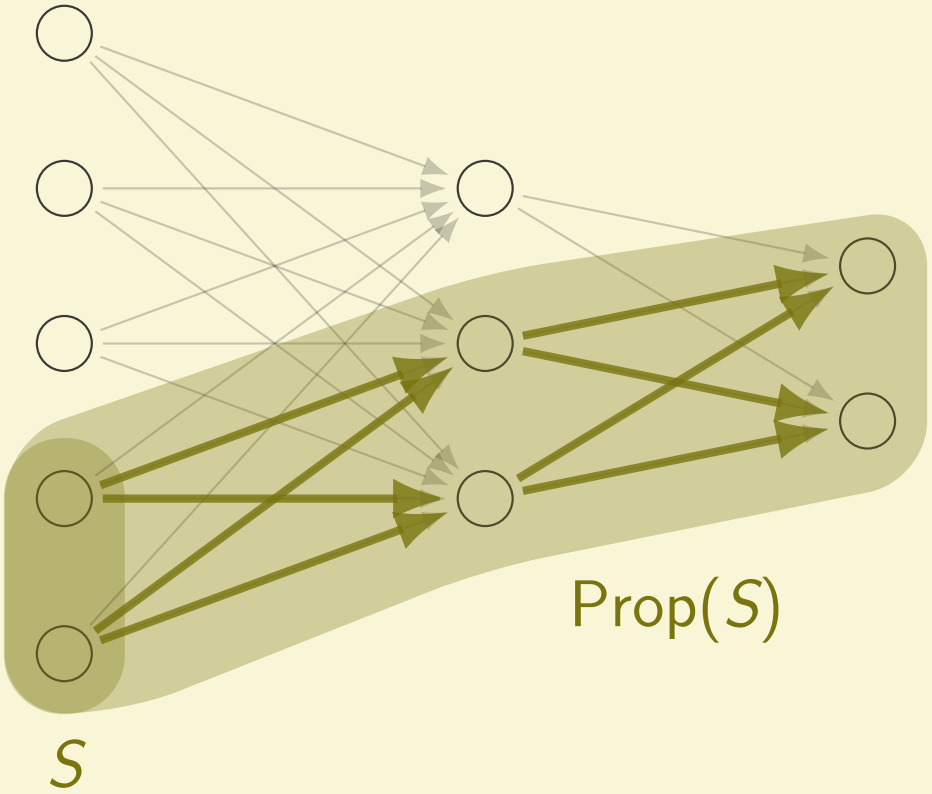
[3] S. Kraus, D. Lehmann, and M. Magidor. "Nonmonotonic reasoning, preferential models and cumulative logics". In: *Artificial intelligence* 44.1-2 (1990).

Learning Wrecks the Model!

[See board]

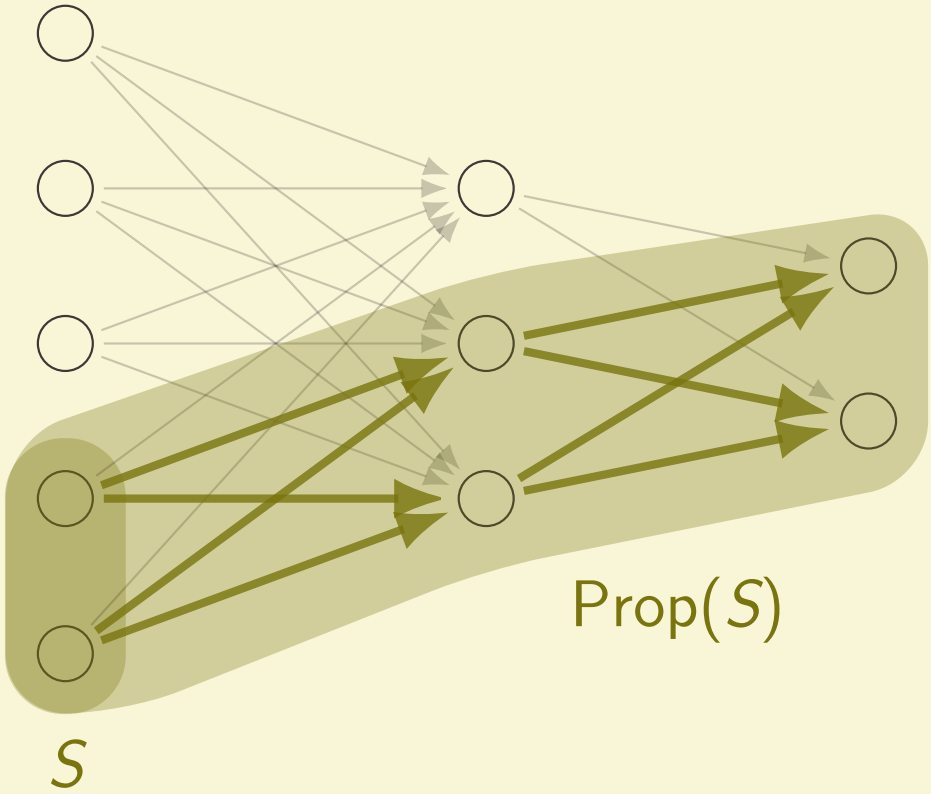
Hebbian Learning

Neurons that fire together wire together



Iterated Hebbian Learning

Neurons that fire together wire together



Repeat this update until a fixed point!
i.e. until the weights are “maximally high”

Language and Semantics

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

We define duals $\langle\mathbf{K}\rangle$, $\langle\mathbf{T}\rangle$ as usual.

$$\begin{array}{lcl}
[\![p]\!] & \in & \mathcal{P}(N) \\
[\![\neg\varphi]\!] & = & [\![\varphi]\!]^{\mathsf{C}} \\
[\![\varphi \wedge \psi]\!] & = & [\![\varphi]\!] \cap [\![\psi]\!] \\
[\![\langle\mathbf{K}\rangle\varphi]\!] & = & \mathsf{Reach}([\![\varphi]\!]) \\
[\![\langle\mathbf{T}\rangle\varphi]\!] & = & \mathsf{Prop}([\![\varphi]\!]) \\
[\![[\varphi]\psi]\!]_{\mathcal{N}} & = & [\![\psi]\!]_{\mathsf{Hebb}(\mathcal{N},[\![\varphi]\!])}
\end{array}$$

We also define $\mathcal{N} \models \varphi$ iff $[\![\varphi]\!]_{\mathcal{N}} = \mathsf{N}$

Note that we can express
Leitgeb's $\varphi \Rightarrow \psi$ as $\mathbf{T}\varphi \rightarrow \psi$

Soundness for $\mathbf{K}$ and $\mathbf{T}$

- First, Reach is a standard monotonic closure operator:

Nec.	From $\vdash \varphi$ we can infer $\vdash \mathbf{K}\varphi$
Dual.	$\langle \mathbf{K} \rangle \varphi \leftrightarrow \neg \mathbf{K} \neg \varphi$
Refl.	$\mathbf{K}\varphi \rightarrow \varphi$
Trans.	$\mathbf{K}\varphi \rightarrow \mathbf{K}\mathbf{K}\varphi$
Distr.	$\mathbf{K}(\varphi \rightarrow \psi) \leftrightarrow (\mathbf{K}\varphi \rightarrow \mathbf{K}\psi)$

- Prop is non-monotonic, but is "Loop-Cumulative":

Nec.	From $\vdash \varphi$ we can infer $\vdash \mathbf{T}\varphi$
Dual.	$\langle \mathbf{T} \rangle \varphi \leftrightarrow \neg \mathbf{T} \neg \varphi$
Refl.	$\mathbf{T}\varphi \rightarrow \varphi$
Trans.	$\mathbf{T}\varphi \rightarrow \mathbf{T}\mathbf{T}\varphi$
Cumulative.	$(\varphi \rightarrow \psi) \wedge (\mathbf{T}\psi \rightarrow \varphi) \rightarrow (\mathbf{T}\varphi \rightarrow \psi)$
Loop.	$(\mathbf{T}\varphi_0 \rightarrow \varphi_1) \wedge \cdots \wedge (\mathbf{T}\varphi_k \rightarrow \varphi_0) \rightarrow (\mathbf{T}\varphi_0 \rightarrow \varphi_k)$

A Complete Description of Hebb*

[See board]

Reduction Axioms for $[\varphi]$

Theorem. The following axioms are sound:

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

Completeness & Model Building

Theorem. **Assuming** model building for the base language: For all consistent $\Gamma \subseteq \mathcal{L}$ there is a net $\mathcal{N}$ such that $\mathcal{N} \models \Gamma$.

Theorem. **Assuming** completeness for the base language: $[\varphi]$ is completely axiomatized by the reduction axioms from before.

Consequences for AI Alignment

[See board]

Conjecture & Speculation

[See board]

Future Work

- Completeness for fuzzy nets
- Stabilized Hebbian Learning
- Single-step update
- What kind of preference upgrade is backpropagation?

Contact:
Caleb Schultz Kisby
cckisby@iu.edu
`https://ais-climber.github.io/`

References

Balkenius, C. and P. Gärdenfors. "Nonmonotonic inferences in neural networks". In: *KR*. Morgan Kaufmann, 1991, pp. 32–39.

Giordano, L., V. Gliozzi, and D. Theseider Dupré. "A conditional, a fuzzy and a probabilistic interpretation of self-organizing maps". In: *Journal of Logic and Computation* 32.2 (2022), pp. 178–205.

Harmelen, F. "Preface: The 3rd AI Wave Is Coming, and It Needs a Theory". In: *Neuro-Symbolic Artificial Intelligence*. Ed. by P. Hitzler and M. Sarker. IOS Press BV, 2022.

Kisby, C., S. Blanco, and L. Moss. "The Logic of Hebbian Learning". In: *The International FLAIRS Conference Proceedings*. Vol. 35. 2022.

Kraus, S., D. Lehmann, and M. Magidor. "Nonmonotonic reasoning, preferential models and cumulative logics". In: *Artificial intelligence* 44.1-2 (1990), pp. 167–207.

Leitgeb, H. "Nonmonotonic reasoning by inhibition nets". In: *Artificial Intelligence* 128.1-2 (2001), pp. 161–201.

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

Odense, S. and A. S. d'Avila Garcez. "A Semantic Framework for Neural-Symbolic Computing". In: *ArXiv* abs/2212.12050 (2022).