Canonical neural net construction (reference sheet)

1 Neural network models

Definition 1. A neural network model is $\mathcal{N} = \langle N, E, W, A, V \rangle$, where

- N is a finite nonempty set (the set of *neurons*)
- Each $E \subseteq N \times N$ (the *edge relation*)
- $W: E \to \mathbb{Q}$ (the *edge weights*)
- $A: \mathbb{Q} \to \{0, 1\}$ (the binary *activation function*)
- $V: \mathcal{L}_{\text{prop}} \to \mathcal{P}(N)$ (the *valuation function*)

Each choice of E, W, A specifies a transition function from state $S \in \mathsf{State}$ to the next state. Given an initial state S_0, this transition function $F_{S_0}: \mathsf{State}_{\mathcal{N}} \to \mathsf{State}_{\mathcal{N}}$ is given by

$$F_{S_0}(S) = S_0 \cup \left\{ n \mid A\left(\sum_{m \in \text{preds}(n)} W(m,n) \cdot \chi_S(m) \right) = 1 \right\}$$

Postulate 2. I assume that for all states S_0, F_{S_0} applied repeatedly to S_0, i.e.

$$S_0, F_{S_0}(S_0), F_{S_0}(F_{S_0}(S_0)), \ldots, F_{S_0}^k(S_0), \ldots$$

eventually reaches a finite fixed point, and moreover this state is the *only* fixed point under S_0. Formally, this means that for all $S_0 \in \mathsf{State}_{\mathcal{N}}$ there is some $k \in \mathbb{N}$ such that:

1. $F_{S_0}(F_{S_0}^k(S_0)) = F_{S_0}^k(S_0)$. That is, the activation pattern under F_{S_0} will eventually stabilize.
2. $F_{S_0}^k(S_0)$ is the only state $S \in \mathsf{State}_{\mathcal{N}}$ such that $F_{S_0}(S) = S$. In other words, the final state is unique for each initial state S_0.

Let the closure $\mathsf{Clos}: \mathsf{State}_{\mathcal{N}} \to \mathsf{State}_{\mathcal{N}}$ be the function that produces that least fixed point: $\mathsf{Clos}(S) = F_{S_0}^k(S_0)$ for that $k \in \mathbb{N}$ above. Finally, let **Net** be the class of all binary neural network models that satisfy this postulate.

Definition 3. $\mathsf{Reach}: \mathsf{State}_{\mathcal{N}} \to \mathsf{State}_{\mathcal{N}}$, where $n \in \mathsf{Reach}(S)$ iff there exists $m \in S$ with an E-path from m to n.

2 Basic definitions for the logic

Language. The main underlying language is $\mathcal{L}_{\mathbf{C}}$: $\varphi, \psi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \mathbf{A}\varphi \mid \Box\varphi \mid \mathbf{C}\varphi$

Definitions from [1]: The *height* of a formula tracks the maximum nesting level of modal operators. The *order* of a formula is the highest proposition p_i occuring. Let $\mathcal{L}_{h,n}$ be the set of formulas of max height h and max order n: $\mathcal{L}_{h,n} = \{\varphi \in \mathcal{L}_{\mathbf{C}} \mid \text{height}(\varphi) \leq h \text{ and } \text{ord}(\varphi) \leq n\}$.

Semantics. Let $\mathcal{N} \in \mathbf{Net}$, $n \in N$. **Note:** $[\![\varphi]\!] = \{n \in N \mid \mathcal{N}, n \Vdash \varphi\}$

$$\begin{array}{lll}
\mathcal{N}, n \Vdash p & \text{iff} & n \in V(p) \\
\mathcal{N}, n \Vdash \neg\varphi & \text{iff} & \mathcal{N}, n \nVdash \varphi \\
\mathcal{N}, n \Vdash \varphi \wedge \psi & \text{iff} & \mathcal{N}, n \Vdash \varphi \text{ and } \mathcal{N}, n \Vdash \psi \\
\mathcal{N}, n \Vdash \mathbf{E}\varphi & \text{iff} & [\![\varphi]\!] \neq \emptyset \\
\mathcal{N}, n \Vdash \Diamond\varphi & \text{iff} & n \in \mathsf{Reach}([\![\varphi]\!]) \\
\mathcal{N}, n \Vdash \langle\mathbf{C}\rangle\varphi & \text{iff} & n \in \mathsf{Clos}([\![\varphi]\!])
\end{array}$$

Axioms for $\Box$:

(Dual)	$\Diamond\varphi \leftrightarrow \neg\Box\neg\varphi$
(Distr)	$\Box(\varphi \wedge \psi) \leftrightarrow (\Box\varphi \wedge \Box\psi)$
(Refl)	$\Box\varphi \rightarrow \varphi$
(Trans)	$\Box\varphi \rightarrow \Box\Box\varphi$

Axioms for A:

(Dual)	$\mathbf{E}\varphi \leftrightarrow \neg\mathbf{A}\neg\varphi$
(Distr)	$\mathbf{A}(\varphi \rightarrow \psi) \rightarrow (\mathbf{A}\varphi \rightarrow \mathbf{A}\psi)$
(Refl)	$\mathbf{A}\varphi \rightarrow \varphi$
(5)	$\mathbf{E}\varphi \rightarrow \mathbf{A}(\mathbf{E}\varphi)$
(Interact)	$\mathbf{A}\varphi \rightarrow \Box\varphi$

Axioms for C:

(Dual)	$\langle\mathbf{C}\rangle\varphi \leftrightarrow \neg\mathbf{C}\neg\varphi$
(Refl)	$\mathbf{C}\varphi \rightarrow \varphi$
(Trans)	$\mathbf{C}\varphi \rightarrow \mathbf{C}\mathbf{C}\varphi$
(CM)	$\mathbf{A}(\mathbf{C}\varphi \rightarrow \psi) \rightarrow (\mathbf{C}(\varphi \wedge \psi) \rightarrow \mathbf{C}\varphi)$
(Interact)	$\Box\varphi \rightarrow \mathbf{C}\varphi$

Rules of Inference:

(MP)	From $\vdash \varphi \rightarrow \psi$ and $\vdash \varphi$ we can infer $\vdash \psi$
(A-Nec)	From $\vdash \varphi$, we can infer $\vdash \mathbf{A}\varphi$
($\Box$-Rep)	From $\vdash \varphi \leftrightarrow \psi$, infer $\vdash \Box\varphi \leftrightarrow \Box\psi$
(C-Rep)	From $\vdash \varphi \leftrightarrow \psi$, infer $\vdash \mathbf{C}\varphi \leftrightarrow \mathbf{C}\psi$

Figure 1. Sound axioms and rules of inference

Definition 4. The proof system $\vdash$ is as follows: $\vdash \varphi$ iff either φ is valid in propositional logic, or φ is one of the axioms listed above, or φ follows from some previous formulas by one of the inference rules.

Definition 5. $\varphi \in \mathcal{L}_{\mathbf{C}}$ is *consistent* iff $\nvdash \neg\varphi$ (alternatively, iff $\vdash \varphi \rightarrow \bot$)

3 The canonical neural net construction

Canonical formulas. The set of canonical formulas $\mathcal{C}_{h,n}$ is from [1], though we might have to make modifications to this definition.

Definition 6. Let the canonical neural network model (of formulas of height h and order n) be

$$\mathcal{N}^c_{h,n} = \langle N^c_{h,n}, E^c, W^c, A^c, V^c \rangle$$

- $N^c_{h,n} = \{\alpha \in \mathcal{C}_{h,n} \mid \alpha \text{ is consistent}\}$. Let's fix an order on these nodes: $\alpha_1, \alpha_2, \alpha_3, \ldots$
- $\beta E^c \alpha$ iff $\alpha \wedge \Diamond \beta$ is consistent.
- Suppose $\alpha_i \in N^c_{h,n}$ has predecessors $\beta_{i1}, \beta_{i2}, \ldots, \beta_{ik}$. For each $\beta_{ij} E^c \alpha_i$, let

  $$W^c(\beta_{ij}, \alpha_i) = (p_i)^j$$

  where p_i is the i^{th} prime number. **Intuition:** Each prime p_i uniquely codes for the node α_i, and the weight between α_i and its predecessor β_{ij} is a power of p_i that uniquely codes for β_{ij}. So any activation value x we care about is going to be a sum of powers of p_i, from which we can reconstruct precisely the α_i being activated and the predecessors β_{ij} that were used to activate it.
- Let $x \in \mathbb{Q}$, and suppose x uniquely identifies the subset of predecessors $\{\beta_{ij} \mid \beta_{ij} E^c \alpha_i \text{ and } j \in X\}$, for some $X \subseteq \{1, \ldots, k\}$. I.e., $x = \sum_{j \in X} (p_i)^j$ for some choice of i. Let the activation function $A^c(x)$ be defined as follows:

  $$A^c(x) = 1 \text{ iff } \textbf{[This is what I need help with!]}$$

  (If x does not code for any valid subset X, then simply set $A^c(x) = 0$.)
- $\alpha \in V^c(p)$ iff $\vdash \alpha \to p$

Some ideas: $A^c(x) = 1$ should be true exactly when the model *says* the β_{ij}'s activate α_i. We have access to a_i and the β_{ij}'s, as well as the set X. So we can state conditions such as:

$$A^c(x) = 1 \text{ iff } \alpha_i \wedge \langle \mathbf{C} \rangle \left(\bigwedge_{\beta_{ij} E^c \alpha_i \text{ and } j \in X} \beta_{ij} \right) \text{ is consistent}$$

Check that $\mathcal{N}^c_{h,n}$ is in Net. There is some $k \in \mathbb{N}$ such that $F_{S_0}(F^k_{S_0}(S_0)) = F^k_{S_0}(S_0)$, and for all other states S, if $F_{S_0}(S) = S$, then $S = F^k_{S_0}(S_0)$.

Proof. **[I'm having a lot of trouble starting, but I think a proof should use the (CM) rule...]** □

The Truth Lemma I need, ⟨C⟩ case. $\mathcal{N}^c_{h,n}, \alpha \Vdash \langle \mathbf{C} \rangle \varphi$ iff $\vdash \alpha \to \langle \mathbf{C} \rangle \varphi$

Proof Sketch. Observe that the claim is equivalent to:

$$\mathsf{Clos}_{\mathcal{N}^c_{h,n}}(\llbracket \varphi \rrbracket) = \{\alpha \mid \vdash \alpha \to \langle \mathbf{C} \rangle \varphi\}$$

By definition, $\mathsf{Clos}_{\mathcal{N}^c_{h,n}}(\llbracket \varphi \rrbracket)$ is the unique fixed point of the transition function $F_{\llbracket \varphi \rrbracket}$ under $\llbracket \varphi \rrbracket$. I will show here that the set $\{\alpha \mid \vdash \alpha \to \langle \mathbf{C} \rangle \varphi\}$ is *also* such a fixed point.

$$\textbf{[GOAL:]}\ F_{\llbracket \varphi \rrbracket}(\{\alpha \mid \vdash \alpha \to \langle \mathbf{C} \rangle \varphi\}) = \{\alpha \mid \vdash \alpha \to \langle \mathbf{C} \rangle \varphi\}$$

Since $\mathsf{Clos}_{\mathcal{N}^c_{h,n}}(\llbracket \varphi \rrbracket)$ is the *unique* fixed point under $\llbracket \varphi \rrbracket$, it will follow that these two states are the same. □

References

[1] Lawrence S Moss. Finite models constructed from canonical formulas. *Journal of Philosophical Logic*, 36:605–640, 2007.