Nonmonotonic Logic and Neural Networks*

Reinhard Blutner with Paul David Doherty
Humboldt University Berlin

Abstract

A puzzle in the philosophy of mind concerns the gap between symbolic and subsymbolic (neuron-like) modes of processing (e.g. Smolensky 1988). The aim of this paper is to overcome this gap by viewing symbolism as a high-level description of the properties of (a class of) neural networks. Combining methods of algebraic semantics and nonmonotonic logic, the possibility of *integrating* both modes of viewing cognition is demonstrated. The main results are (I) that certain activities of connectionist networks can be interpreted as *nonmonotonic inferences*, and (II) that there is a strict correspondence between the coding of knowledge in Hopfield networks and the knowledge representation in weight-annotated Poole systems. These results (a) show the usefulness of nonmonotonic logic as a descriptive and analytic tool for analyzing emerging properties of connectionist networks, (b) single out certain logical systems by giving them a "deeper justification", and (c) pave the way for using connectionist methods (e.g. "simulated annealing") in order to perform nonmonotonic inferences.

1 Introduction

There is a gap between two different modes of computation: the symbolic mode and the subsymbolic (neuron-like) mode. Complex symbolic systems like those of grammar and logic are essential when we try to understand the general features and the peculiarities of natural language, reasoning and other cognitive domains. On the other hand, most of us believe that cognition resides in the brain and that neuronal activity forms its basis. Yet neuronal computation appears to be numerical, not symbolic; parallel, not serial; distributed over a gigantic number of different elements, not as highly localized as in symbolic systems. Another aspect is that the brain is an adaptive system that is very sensitive to the statistical character of experience. Hard-edged rule systems are not suitable to deal with this side of behavior. A unified theory of cognition must overcome these gaps and must assign the proper roles to symbolic, neural and statistical computation (e.g. Smolensky 1988, 1996; Balkenius & Gärdenfors 1991).

The aim of this paper is to demonstrate that the gap between symbolic and neuronal computation can be overcome when we view symbolism as a high-level description of the properties of (a class of) neural networks. The important methodological point is to illustrate that the instruments of model-theoretic (algebraic) semantics and nonmonotonic logic may be very useful in realizing this goal. In this connection it is important to stress that the algebraic perspective is entirely neutral with respect to fundamental questions such as whether a "content" is in the head or is a platonic abstract entity (cf. Partee with Hendriks 1997, p. 18). Consequently, the kind of "psychologic" we pursue here isn't necessarily in conflict with the general setting of model-theoretic semantics.

Information states are the fundamental entities in the construction of propositions. In the next section we interpret information states as representing states of activations in a connectionist network. In section 3 we consider how activation spreads out and how it reaches, at least for certain types of networks, asymptotically stable output states. Following and extending ideas of Balkenius & Gärdenfors

*The research leading to this paper was funded in part by the Max-Planck-Gesellschaft. Many thanks to Birgit Ahlemeyer, Markus Steinbach and all our colleagues who provided valuable input.

1991, we show that the *fast dynamics* of the system can be described asymptotically as a nonmonotonic inferential relation between information states. Section 4 introduces the notion of weight-annotated Poole systems, and section 5 explains in which way these systems bring about the correspondence between connectionist and symbolic knowledge bases.

2 Information states in Hopfield networks

A neural network is a system of connected units ("neurons"). Each unit has a certain *working range of activity*. Let this be the set $\{-1, 0, +1\}$ (+1: maximal firing rate, 0: resting, −1: minimal firing rate). A possible state s of the system describes the activities of each neuron: $s \in \{-1, 0, +1\}^n$, with n = number of units. A possible *configuration* of the network is characterized by a *connection matrix* w. Hopfield networks are characterized by symmetric configurations and zero diagonals ($w_{ij} = w_{ji}$, $w_{ii} = 0$). The *fast dynamics* describes how neuron activities spread through that network. In the simplest case this is described by the following *update function*:

$$f(s)_i = \Theta \sum\nolimits_j w_{ij} \cdot s_j \qquad (\Theta \text{ a nonlinear function}). \tag{1}$$

Let us interpret activations as indicating information specification: the activations +1 and −1 indicate maximal specification, the resting activation 0 indicates underspecification. Generalizing a notion introduced by Balkenius & Gärdenfors 1991, the set $S = \{-1, 0, +1\}^n$ of activation states can be partially ordered in accordance with their informational content:

$$s \geq t \text{ iff } s_i \geq t_i \geq 0 \text{ or } s_i \leq t_i \leq 0, \quad \text{for all } 1 \leq i \leq n \tag{2}$$

$s \geq t$ can be read as *s is at least as specific as t*. The poset $\langle S, \geq \rangle$ doesn't form a lattice. However, it can be extended to a lattice by introducing a set $\perp$ of *impossible activation states*: $\perp = \{-1, 0, +1, nil\}^n - S$, where *nil* designates the "impossible" activation of a unit. It can be shown that the extended poset of activation states $\langle S \cup \perp, \geq \rangle$ forms a DeMorgan lattice: Replace the former definition (2) of the informational ordering by the following:

$$s \geq t \text{ iff } s_i = nil \text{ or } s_i \geq t_i \geq 0 \text{ or } s_i \leq t_i \leq 0, \quad \text{for all } 1 \leq i \leq n \tag{3}$$

Conjunction $\circ$ can be interpreted as *simultaneous realization* of two activation states, disjunction $\oplus$ as some kind of generalization. This fact enables us to interpret activation states as propositional objects ("information states").

3 Asymptotic updates and nonmonotonic inference

In general, updating an information state s may result in a information state $f \ldots f(s)$ that doesn't include the information of s. However, for the following it is important to interpret updating as specification. If we want s to be informationally included in the resulting update, we have to "clamp" s somehow in the network. A technical way to do that has been proposed by Balkenius & Gärdenfors 1991. Let f designate the original update function (1) and $\underline{f}$ the clamped one, which can be defined as follows (including iterations):

$$\begin{aligned} \underline{f}(s) &= f(s) \circ s \\ \underline{f^{n+1}}(s) &= f\left(\underline{f^n}(s)\right) \circ s \end{aligned} \tag{4}$$

Hopfield networks (and other so-called *resonance systems*) exhibit a desirable property: when given an input state s the system stabilizes in a certain state (it is of no importance here whether the dynamics is clamped or not). Thus, the following set of *asymptotic updates* of s is well-defined:

$$ASUP_w(s) = \left\{t : t = \lim_{n\to\infty} \underline{f^n}(s)\right\} \tag{5}$$

Under certain conditions (asynchronous, stochastic updates) the function

$$E(s) = -\sum_{i>j} w_{ij} \cdot s_i \cdot s_j \tag{6}$$

is a Ljapunov function (energy function) of the dynamic system (Hopfield 1982). This enables us to characterize the asymptotic updates of s as those specifications of s that minimize E:

$$ASUP_w(s) = \min_E(s) \tag{7}$$

The notion of asymptotic updates naturally leads to a nonmonotonic inferential relation (between information states):

$$s \mid\!\sim_w t \qquad \text{iff } s' \geq t \text{ for each } s' \in ASUP_w(s) \tag{8}$$

With the help of the equivalence (7), the usual traits of nonmonotonic consequence relations can be shown:

$$\begin{array}{rl} \text{Supraclassicality:} & \text{if } s \geq t, \text{ then } s \mid\!\sim_w t \\ \text{Reflexivity:} & s \mid\!\sim_w s \\ \text{Cut:} & \text{if } s \mid\!\sim_w t \text{ and } s \circ t \mid\!\sim_w u, \text{ then } s \mid\!\sim_w u \\ \text{Cautious Monotonicity:} & \text{if } s \mid\!\sim_w t \text{ and } s \mid\!\sim_w u, \text{ then } s \circ t \mid\!\sim_w u \end{array} \tag{9}$$

This corresponds to results found by Balkenius & Gärdenfors 1991, who have considered information states for the case that they form a Boolean algebra.

4 Weight-annotated Poole systems

In connectionist systems knowledge is encoded in the connection matrix w (or, alternatively, the energy function E). Symbolic systems usually take a default logic and represent knowledge as a database consisting of expressions having default status. A prominent example of such a framework has been proposed by Poole (e.g. Poole 1988, 1996). In this section, we introduce a variant of Poole's systems, which we will call *weight-annotated Poole systems*. This variant will be proven to be useful for relating the different types of coding knowledge (see section 5).

Let us consider the language L_{At} of propositional logic (referring to the alphabet At of atomic symbols). A triple $T = \langle At, \Delta, g\rangle$ is called a *weight-annotated Poole system* iff (i) Δ is a set of consistent sentences built on the basis of At (the possible hypotheses); (ii) $g : \Delta \mapsto [0, 1]$ (the weight function). A *scenario of a formula* α *in* T is a subset Δ' of Δ such that $\Delta' \cup \{\alpha\}$ is consistent. The *weight of a scenario* Δ' is

$$G(\Delta') = \sum_{\delta\in\Delta'} g(\delta) \;\; - \sum_{\delta\in(\Delta-\Delta')} g(\delta) \tag{10}$$

A *maximal scenario of* α *in* T is a scenario the weight of which is not exceeded by any other scenario (of α in T). With regard to a weight-annotated Poole system T, the following cumulative consequence relation can be defined:

$$\alpha \sim\!\!\!\!\supset_T \beta \qquad \text{iff } \beta \text{ is an ordinary consequence of each maximal scenario of } \alpha \text{ in } T \tag{11}$$

It is important to give a preference semantics for weight-annotated Poole systems. This preference semantics may be seen as the decisive link for establishing the correspondence between connectionist and symbolic systems.

Let υ denote an ordinary (total) interpretation for the language L_{At} (υ: $At \mapsto \{-1, 1\}$). The usual clauses apply for the evaluation of the formulas of L_{At} relative to υ. The following function indicates how strong an interpretation υ conflicts with the space of hypotheses Δ:

$$\mathfrak{E}(\upsilon) = -\sum_{\delta \in \Delta} g(\delta) \cdot [\delta]_\upsilon \qquad \text{(call it the "energy" of the interpretation)} \tag{12}$$

An interpretation υ is called a *model* of α just in case $[\alpha]_\upsilon = 1$. A *preferred model* of α is a model of α with minimal energy $\mathfrak{E}$ (with regard to the other models of α). As a semantic counterpart to the syntactic notion $\alpha \sim\!\!\!\!\supset_T \beta$, let us take the following relation:

$$\alpha \sim\!\!\!\!\supset\!\!\!\!=_T \beta \qquad \text{iff each preferred model of } \alpha \text{ is a model of } \beta \tag{13}$$

As a matter of fact, the syntactic and the semantic notions coincide. (A proof can be found in Blutner 1997).

5 The correspondence between Hopfield networks and weight-annotated Poole systems

Bringing about the correspondence between connectionist and symbolic knowledge bases, we have first to look for a symbolic representation of information states. Let us again consider the propositional language L_{At}, but let us now take this language as a symbolic means to speak about information states. Following usual practice in algebraic semantics, we can do this formally by interpreting (some subset of the) expressions of the propositional language by the corresponding elements of the DeMorgan algebra $\langle S \cup \bot, \geq \rangle$. More precisely, let us call the triple $\langle S \cup \bot, \geq, \uparrow \downarrow \rangle$ a *Hopfield model* (for L_{At}) iff $\uparrow \downarrow$ is a function assigning some element of $S \cup \bot$ to each atomic symbol and obtaining the following conditions: $\uparrow \alpha \wedge \beta \downarrow = \uparrow \alpha \downarrow \circ \uparrow \beta \downarrow$; $\uparrow \sim \alpha \downarrow = -\uparrow \alpha \downarrow$ ("$-$" converts positive into negative activation and vice versa).

A Hopfield model is called *local* (for L_{At}) iff it realizes the following assignments: $\uparrow p_1 \downarrow = \langle 1\ 0 \ldots 0 \rangle$, $\uparrow p_2 \downarrow = \langle 0\ 1 \ldots 0 \rangle$, $\ldots$, $\uparrow p_n \downarrow = \langle 0\ 0 \ldots 1 \rangle$. With regard to local Hopfield models each state can be represented by a conjunction of literals (atoms or their inner negation); e.g. $\langle 1\ 1\ 0 \rangle = \uparrow p_1 \wedge p_2 \downarrow$, $\langle 1\ 1\ -1 \rangle = \uparrow p_1 \wedge p_2 \wedge \sim p_3 \downarrow$. In other words, in the case of local models each information state can be considered as symbolic.

Local Hopfield models give the opportunity to relate connectionist and symbolic knowledge bases in a way that allows to represent nonmonotonic inferences in neural (Hopfield) networks by inferences in weight-annotated Poole systems. The crucial point is the translation of the connection matrix w into an associated Poole system T_w. Let us consider a Hopfield system (n neurons) with connection matrix w, and let $At = \{p_1, \ldots, p_n\}$ be a set of atomic symbols. Take the following formulae α_{ij} of L_{At}:

$$\alpha_{ij} =_{def} (p_i \leftrightarrow sign\,(w_{ij})\; p_j) \qquad \text{for } 1 \leq i < j \leq n \tag{14}$$

For each connection matrix w the *associated Poole system* is defined as $T_w = \langle At, \Delta_w, g_w \rangle$, where the following clauses apply:

$$\begin{aligned} &\text{a.} \quad \Delta_w = \{\alpha_{ij} : \; 1 \leq i < j \leq n\} \\ &\text{b.} \quad g_w(\alpha_{ij}) = |\, w_{ij} \,| \end{aligned} \tag{15}$$

Updating information states came out as a kind of specification in section 3. Under certain conditions (no isolated nodes) it can be shown that each (partial) information state is completed asymptotically. Consequently, $ASUP_w(s)$ contains only total information states. Together with the equivalence (7) and the definitions (12) & (13), this fact allows us to prove that nonmonotonic inferences based upon asymptotic updates can be represented by inferences in weight-annotated Poole systems:

Theorem

Assume any formulae α and β that are conjunctions of literals. Let the Poole system T be *associated* with the connection matrix w. Then

$$\lceil \alpha \rfloor \; \mid\!\sim_w \; \lceil \beta \rfloor \qquad \text{iff } \alpha \; \supset\!\!=_T \; \beta \quad (\text{iff } \alpha \; \supset\!\!-_T \; \beta).$$

This result shows that we can use nonmonotonic logic to characterize asymptotically how neuron activities spread through the connectionist network. In particular, a weighted variant of Poole's logical framework for default reasoning has proven to be useful. One possible application of the correspondence may be the use of connectionist techniques to perform nonmonotonic inferences ("simulated annealing", cf. Derthick 1990).

Finally, we should stress that our primary aim was a methodological one: the demonstration that model-theoretic semantics may be very useful for analyzing (emerging properties of) connectionist networks. Admittedly, the results found so far are much too simplistic to count as a real contribution to closing the gap between symbolism and connectionism. What is important, in our view, is to get an active dialog between the traditional symbolic approaches to logic, information and language and the connectionist paradigm. Perhaps, this dialog may shed new light on old notions like partiality, updates, underspecification, learning, genericity, homogeneity, salience, probabilistic logic, randomized computation, etc.

References

Balkenius, C. & Gärdenfors, P. (1991): "Nonmonotonic inferences in neural networks". In J.A. Allen, R. Fikes, & E. Sandewall (eds.), *Principles of knowledge representation and reasoning*. San Mateo, CA: Morgan Kaufmann.

Blutner, R. (1997): "Psychologic". Handout. Available as `http://www2.rz.hu-berlin.de/asg/blutner/psylogic.ps`.

Derthick, M. (1990): "Mundane reasoning by settling on a plausible model". *Artificial Intelligence* 46, 107–157.

Hopfield, J.J. (1982): "Neural networks and physical systems with emergent collective computational abilities". *Proceedings of the National Academy of Sciences* 79, 2554–2558.

Partee, B. with Hendriks, H.L.W. (1997): "Montague Grammar". In J. van Benthem & A. ter Meulen (eds.), *Handbook of Logic and Language*. Cambridge: MIT Press. 5–91.

Poole, D. (1988): "A logical framework for default reasoning". *Artificial Intelligence* 36, 27–47.

Poole, D. (1996): "Who chooses the assumptions?". In P. O'Rorke (ed.), *Abductive Reasoning*. Cambridge: MIT Press.

Smolensky, P. (1988): "On the proper treatment of connectionism". *Behavioral and Brain Sciences* 11, 1–23.

Smolensky, P. (1996): "Computational, dynamical, and statistical perspectives on the processing and learning problems in neural network theory". In P. Smolensky, M.C. Mozer, & D.E. Rumelhart (eds.), *Mathematical Perspectives on Neural Networks*. Mahwah, NJ: Lawrence Erlbaum Associates. 1–13.