

Probabilistic Inference and Learning in a Connectionist Causal Network

Carter Wendelken and Lokendra Shastri
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704
{carterw,shastri}@icsi.berkeley.edu

Abstract

The SHRUTI model demonstrates how a structured connectionist network can be used to encode relational causal knowledge and provide a basis for rapid inference. This paper explores the extent to which the evidential reasoning in SHRUTI can be viewed as probabilistic. An interpretation of the link weights is provided with which the results of spreading activation in the model accord well with probability theory. In addition, a simple local (Hebbian-like) learning mechanism is provided with which the described network structure and probabilistic link weights can be learned.

1 Introduction

The SHRUTI model demonstrates how a structured connectionist network can be used to encode relational causal knowledge and provide a basis for rapid inference. This paper explores the extent to which the evidential reasoning in SHRUTI can be viewed as probabilistic. Further, it provides a simple local learning mechanism which can account for the basic structure of the model.

1.1 Overview of SHRUTI

First we present the basic elements of SHRUTI relevant to the following discussion. The model is described in considerably more detail in [11] and [14]. SHRUTI is a neurally plausible (connectionist) model that demonstrates how a network of neuron-like elements could encode a large body of structured knowledge and perform a variety of inferences within a few hundred milliseconds.

SHRUTI suggests that the encoding of relational information (frames, predicates, etc.) is mediated by neural circuits composed of *focal clusters* and that the dynamic representation and communication of relational

instances involves the transient propagation of *rhythmic* activity across these clusters. A role-entity binding is represented in this rhythmic activity by the *synchronous* firing of appropriate cells. Rules are encoded by links that enable the propagation of rhythmic activity across focal clusters, and a fact in long-term memory is a temporal pattern matching circuit.

A focal cluster (also relation or predicate) in SHRUTI is a collection of related nodes of several different types. Activity of the collector (+) node reflects the amount of evidence collected in support of the given predicate. Activity of the enabler (?) node reflects the strength with which information about the predicate is being sought. Both of these are τ_{AND} nodes, meaning that they require more or less continuous pulse trains across some interval in order to fire and that they produce such pulse trains when they do. A focal cluster typically also includes role nodes, which only fire in certain phases in synchrony with role bindings; these are termed ρ or phasic nodes. Role bindings are represented by activity of phasic nodes in a connectionist type hierarchy.

Rules are formed by linking together predicate focal clusters, with the antecedent collector linked to the consequent collector, the consequent enabler linked to the antecedent enabler, and matching role nodes linked in both directions. Type restriction and instantiation of unbound variables are handled via connections between the rule structure and the type hierarchy.

1.2 Models of Inference

Any satisfactory model of human inference should demonstrate several characteristics. First, it should be capable of dealing with uncertainty. Probability is clearly the most accurate way to model uncertainty, since it can be shown that any system which does not utilize probabilistic models will behave irrationally in certain situations [8]. Although we humans are irrational at times, making it safe to say that we don't maintain

perfect probabilistic models of the world, the fact that we are able to reason effectively in most circumstances suggests that to a great extent the underlying mechanism is probabilistic in nature. While some experiments have purported to show that people’s inference is dramatically nonprobabilistic [17], more convincing work [2] indicates that in fact inferences do appear to rely on probabilistic models. Thus, any model of human inference should make contact with probability theory.

Probability, however, can provide only an incomplete model of the relationships between various phenomena. A notion of causality also appears necessary for any model of human inference. Correlational probabilities can suggest causal relations, but they don’t determine them. Yet an understanding of causality can be necessary for accurate prediction (e.g., Simpson’s Paradox, [9]). Moreover, it is fairly clear that people do tend to reason from cause to effect most readily, and in fact it has been shown that they are most comfortable assigning probabilities in this direction [8]. Therefore, it is reasonable to expect that any model of human inference would include causality as a key element.

As a formal system for computing probabilities of propositions, it is unlikely that one can do any better than a causal belief net [8]. Any gains in ease of computation must be accompanied by losses of accuracy or generality. However, the belief net is inadequate as a model of human inference, since any realistic model must allow for reasoning with relations. Much effort has been undertaken to combine relational and probabilistic reasoning (e.g., [6], [10], [4]); these efforts generally provide good solutions for artificial intelligence tasks but do not represent efforts to understand human reasoning and are in general even farther removed from any plausible neural interpretation than a belief net. SHRUTI was designed to handle relational causal inference; to the extent that it accurately handles probabilities it can be seen as an alternative to these approaches. More importantly though, SHRUTI does represent a plausible story of how human inference might operate.

2 Probabilistic Reasoning in SHRUTI

The interpretation of link weights and activation values is intentionally underspecified in the SHRUTI model. The goal has been to provide a flexible and expressive representational structure which can be fine-tuned according to specific modelling and task requirements. In this section, we examine the extent to which the link weights can be interpreted as probabilities. The propositional case is considered first.

2.1 A single proposition

For a single proposition A, represented as a standard SHRUTI focal cluster with no role nodes, there must be some weight on the connection from ?A to +A and also a weight on the connection from +A to ?A. The former (called the taxon fact in SHRUTI) is the appropriate place to put the prior probability $P(A)$. The other link reflects the propensity of the system to seek information about predicates that have been asserted. If something is asserted which is regarded as unlikely (low prior), then this should lead to a stronger search for explanation; this unlikely predicate provides greater evidence for its potential causes than it would were it considered more likely [11]. This + to ? weight is therefore assigned the inverse of the prior.

For any predicate focal cluster, the activity level of the + node is taken as the probability that the predicate is true, and so this must range between 0 and 1. Asserting a predicate means activating its + node with value 1.0. Activation of the ? node with value γ (for simplicity, we assume $\gamma = 1.0$ for now) indicates that the predicate is being queried but provides no evidence to its truth. Larger activation values indicate positive evidence from below (diagnostic evidence), while activation levels less than γ indicate negative evidence.

If this single proposition is queried (by activating its enabler with value 1.0), then 1.0 is multiplied by the prior value while traversing the ? to + link and the resulting value at the + node is equal to the prior. If additional activity arrives from above (from possible causes), then this evidence combines with that on the prior link to determine the activation value.

2.2 A simple rule

Consider two predicates A and B. If A is a cause of B, then we can have a simple rule $A \rightarrow B [w_+, w_?]$, where w_+ represents the weights on the link from +A to +B and $w_?$ represents the weight on the link from ?A to ?B (see figure 1(a)). Let w_+ equal the the probability of B given only A (the causal strength of A and not simply $p(B|A)$; this is essentially the independent component of a noisy-or). If A is asserted with strength 1.0, then the system will initially conclude B with strength w_+ . This is not the end of the story, though. If there are other possible causes of B, then its activation should in fact be somewhat larger than w_+ , if it is to match B’s actual probability of $P(B|A)$. In this case, the initial activation of +B will lead to activation of ?B through the inverse prior link and ultimately to a search for these other causes.

Suppose now that B is asserted instead of A. Then +B is active at 1.0 and thus ?B takes on the value $1/p(B)$. If

we let w_i equal $p(B|A)$, then it is easily seen that the activity that arrives at +A along the path from +B through ?B and ?A has value $1.0 \times (1/P(B)) \times P(B|A) \times P(A) = P(A|B)$.

This very simple example gives a basic probability model that fits with the structure of SHRUTI. To recapitulate, the weights (for the proposition A and rule $A \rightarrow B$) are as follows:

Link	Weight
?A \rightarrow +A	p(A)
+A \rightarrow ?A	1/p(A)
+A \rightarrow +B	p(B only A)
?B \rightarrow ?A	p(B A)

2.3 Evidence combination

When there are multiple sources of evidence for some predicate, then we must have a way to combine them. Ideally, this combination would be via a conditional probability table wherein every fine-grained dependence between causes and effect could be modeled, as in a standard belief net. However, it is a requirement of the model that each cause communicate independently, utilizing a single link-weight parameter. The approach taken follows that of the noisy-or used in belief nets. However, to allow for more flexible evidence combination within this framework than what a single function can provide, a set of evidence combination functions was developed, based on notions of sufficiency or necessity of factors, and also on degrees of correlation. Interestingly, these functions suggest several different interpretations of the link weights. At one end of this range is the familiar *noisy-or* function $1 - \prod_i (1 - x_i * w_i)$, where each weight w_i is essentially a measure of the sufficiency of each potential cause for bringing about the effect. At the other end of the spectrum, a sort of *noisy-and* function $\prod_i (1 - (1 - x_i) * w_i)$ is used where the weight is interpreted as a degree of necessity, the probability that the consequent is false given that the particular antecedent is false (but all other necessary antecedents are true.) In between these are a *soft-or* where positive correlation is assumed, a set of power averages $((\sum_i X_i^k W_i) / (\sum_i W_i))^{1/k}$ ranging from max down to min depending on the parameter k, and a *soft-and* analogous to the *soft-or*. The evidence combination functions are described in greater detail in [15].

2.4 Explaining away

SHRUTI allows for both deductive and diagnostic reasoning. It is well known that in rule-based systems the combination of these two can lead to serious problems; for example a rule-based system which concludes wetGrass (W) from isRaining (R), and sprinklerOn (S)

from wetGrass can have the unfortunate tendency to chain these together and conclude sprinklerOn based on isRaining (example from [8]). This problem is avoided in SHRUTI by constructing inhibitory links from the + node of any cause to the ? \rightarrow ? link leading to any other cause from a common effect (see figure 1(b)). For the example shown, belief in isRaining (activity of +R) modulates the activity along the ?W \rightarrow ?S link by an appropriate amount. A probabilistic interpretation of the weight on this inhibitory link is given by $\phi_S(R) = [P(W)P(W|S, R)P(S|R)] / [P(W|S)P(W|R)P(S)]$. It is easily seen that this leads to correct activation of R based on values of the other two nodes. The actual value of this modulatory weight of course depends on the evidence combination function; while a significant explaining away effect occurs for a noisy-or, none occurs for a noisy-and.

2.5 Relational rules

A key feature of SHRUTI is that it handles relations, and this has a number of ramifications when it comes to dealing with evidential reasoning. The probabilistic rules, for example, have type restrictions on the variables; it is thus possible and reasonable to have multiple rules connecting two relations, each accepting differently typed arguments. For relational predicates, a single weight for the ? to + prior link, and also for the + to ? inverse prior link, is generally inadequate. The probability of a relation being true depends very much on the arguments of the relation. For this reason, multiple links of this kind are allowed, and type restrictions are placed on these links. The connectionist form of these type restrictions, for both the prior and inverse prior links, is the same as that for episodic facts in SHRUTI (see [11]). A link from source ? or + to a mediating node is inhibited by the role activity unless that inhibition is in turn inhibited by synchronous activity of the appropriate type. In the example (figure 2(a)), the predicate *falls(x)* is constructed such that $P(\text{falls}(\text{Child})) > P(\text{falls}(\text{Adult}))$. There is no attempt to assign a distinct probability to a relation for every possible set of arguments to which it might apply; instead, specific priors might be learned for salient combinations (such as *falls(ChevyChase, SNLsketch)*) while new or less salient combinations would fall back on more general prior knowledge (such as *falls(Actor, Stage)* or *falls(Adult, Location)*).

2.6 Patterns of inference

A number of things should be noted about the patterns of reasoning induced by this model. Inference in SHRUTI is essentially an anytime algorithm. Unlike in a belief net,

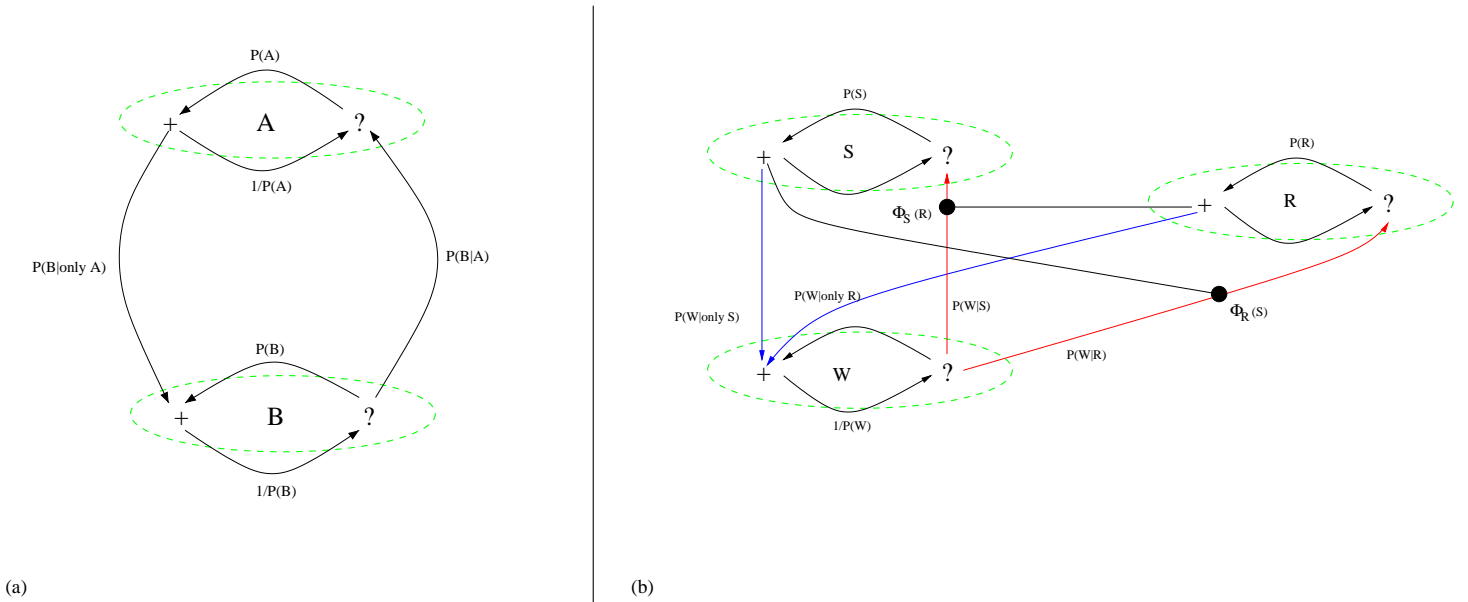


Figure 1: (a) A simple propositional rule $A \rightarrow B$ showing link weights. If $+A$ is activated the $+B$ takes on value $P(B|only\ A)$. If $+B$ is activated then $+A$ obtains the value $P(A|B)$. (b) An example showing the need for explaining away (from Pearl 1988). If $+W$ is activated in the absence of information about S , then clearly we obtain activity at $+R$ equivalent to $P(R|W)$, as before. If however $+S$ is activated, then this amount must be multiplied by $\phi_S(R)$ to yield $P(R|W,S)$.

responses to a query are generated almost immediately, based on the prior information stored for the queried predicate. As inference is allowed to progress, early estimates are repeatedly refined as more and more evidence is brought in from further up or down the causal chain. In a neural system, the depth to which this search for evidence occurs would be limited, such that only evidence within a certain distance (along any casual chain) would be considered. Presumably, this depth could be modulated somewhat by attention or other factors. Importantly, this is a model which scales up naturally to huge domains without any performance loss. The model performs predictive or diagnostic inference accurately, assuming that chosen combination functions correctly represent the desired probability model. Since probabilities are defined in terms of causal strengths, this is more easily achieved in the forward direction; the combination function chosen for the enabler in general can only approximate the actual function which is dependent on the top-down combination. Combining predictive and diagnostic inference can lead to some loss of accuracy, since sufficient information is not generally available at a collector node to determine precisely how to merge these two types of evidence. In figure 2(b), it can be seen that $P(B|A)$, $P(B|A,E)$, $P(B|C)$, and $P(B|C,F)$ can all be accurately computed by activating the appropri-

ate nodes. The precise probability $P(B|A,C)$ cannot be computed, since it requires the factor $P(C|A)$ which is unavailable at B . It is left to the combination function at B to appropriately combine the factors $P(B|A)$ and $P(B|C)$ which are available. Since it is known in each case whether the factor represents a positive or negative contribution (whether it is higher or lower, respectively, than the associated prior), this combination can produce a good approximation of the actual probability. Concerning human reasoning, this model predicts that inference involving evidence from both causes and effects will be the less reliable than inference in only one direction or the other.

3 Learning

Having shown that a neurally plausible connectionist system, with weights properly assigned, can perform evidential reasoning which doesn't stray far from a probabilistic interpretation, it remains to be seen how these connection weights can be learned in a neurally plausible manner. The SHRUTI model is amenable to supervised learning via backpropagation, and this procedure has been used to train the network in a number of domains. Backpropagation, however, is not plausi-

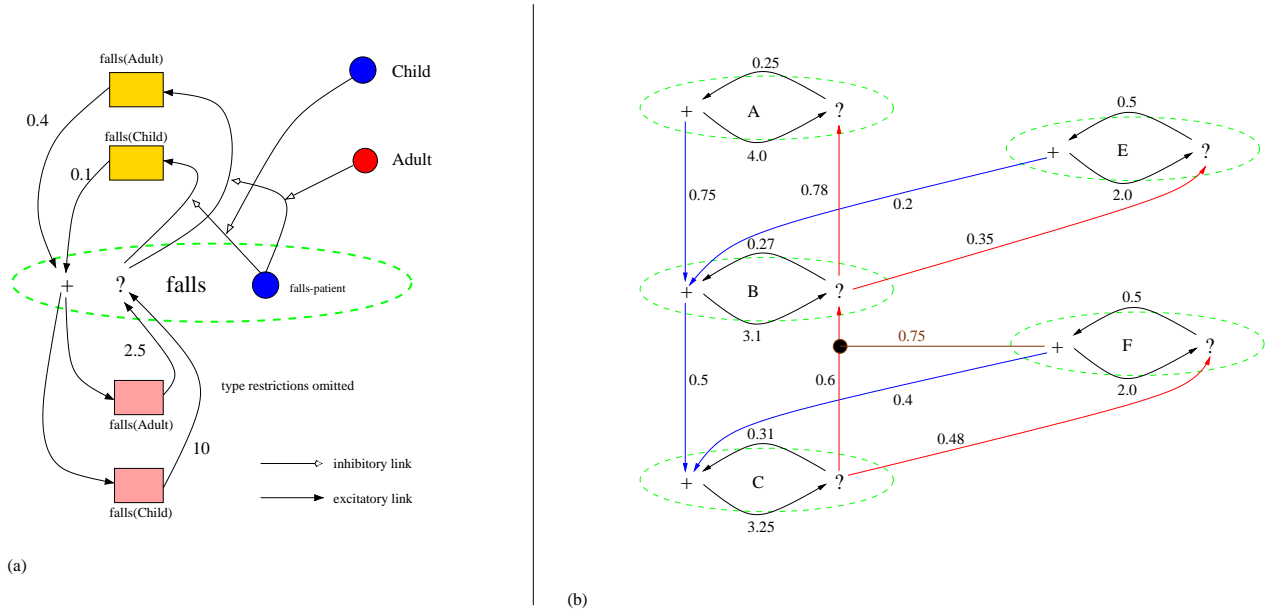


Figure 2: (a) A focal cluster for falls(x) with two typed priors and inverse priors. (b) A network with link weight values shown.

ble as a neural learning mechanism, and so other forms of learning must be explored. In this context, Hebbian learning is a much more desirable alternative. With Hebbian learning, a synapse is potentiated whenever activity in the target neuron is correlated with activity at the synapse, and possibly depressed when uncorrelated [5]. In SHRUTI terms, this would mean that a link weight increases whenever source and target nodes are both active, and decreases when only the source is active. We have examined the role of fast Hebbian learning (i.e. priming) during inference [16]. Below we discuss how a modified form of Hebbian learning can lead to learning of link weights that have the desired probabilistic interpretation.

3.1 A modified Hebbian learning rule

Hebbian learning is ideal for building associations. However, the causal model presented above encodes more than just associations between components of a rule; it also encodes a directionality which is vital for correct inference. We must have a learning method that allows these asymmetric connections to develop. One important feature of causation is that causes precede effects. It is therefore reasonable to assume that we would only learn a causal rule $A \rightarrow B$ when we observe A to occur before B. As a learning rule for SHRUTI, this can mean that a +A to+B link should be strengthened only when the activity at the source of the link (+A) precedes

activity at the target (+B). For a ? to ? link, we assume the opposite behavior, namely that such a link should be strengthened if and only if its source becomes active once its target is already firing. In this manner, we require that one observation precede another by some amount of time and also that both occur within a specified window.

With these learning rules, it can be shown that a structured causal network with appropriate probabilistic link weights can be constructed from a set of relations in response to observations. The starting point is a collection of relational focal clusters, differentiated into + nodes, ? nodes, and role nodes. For simplicity we assume that each type of node is connected to all others of its own type with low weight but to no others. At another level of detail, this can be thought of as the result of a recruitment learning process over a differentiated neural substrate [1] [18] [12] [13], wherein recruitment of a predicate focal cluster involves acquisition of neurons from several different functional regions.

The learning rule for a + link is as follows: if the source has been active sufficiently long and the target then becomes active, the link weight is updated as $w_{t+1} = w_t + \alpha * (1 - w_t)$ where $\alpha = 1/\#updates$; otherwise if the source has been active sufficiently long and the target fails to fire, the weight is decremented $w_{t+1} = w_t - \alpha * w_t$. If the target becomes active before the source, then there is no change. It is easily seen that $w_{t+1} = (\#increases)/(\#updates)$ for $w_0 = 0$,

correctly encoding the probability that the target of the link follows the source within the specified time parameters. A modification of this rule including a normalization term, to account for the possibility of multiple sources, reduces the weight increase on a link by a factor proportional to the number of active links impinging on the same target [3]. This allows a link weight for $+A \rightarrow +B$ to encode $P(B|only\ A)$ and not just $P(B|A)$. It should be noted that this is the link weight corresponding to a noisy-or combination of evidence; learning of weights for other sorts of combination functions, and also learning of explaining away inhibition, are important issues not yet addressed.

For $? \rightarrow ?$ links, the learning rule is nearly the reverse of the above. In this case, a similar weight increase occurs whenever a link target has been active for sufficiently long and a source becomes active, and a weight decrease occurs when a target remains active for too long without activity at the source. If the source becomes active first, then there is no change. As above, it is easily seen that this link correctly records the probability that the source fires after the target (within designated time parameters), and for a link $?B \rightarrow ?A$ this can be reasonably interpreted as $P(B|A)$.

Role links are learned via standard Hebbian learning, since bidirectional associations are appropriate for the connections between role nodes. The weights on the links connecting the enabler to the collector and vice-versa, the prior and inverse prior, also each have their own learning rule, which reflects the number of times the $+$ node becomes active divided by the total number of times the $?$ becomes active. Since activation of the $?$ node can provide negative evidence or positive and is not biased toward either, it is reasonable to interpret this ratio as a prior probability.

This model predicts that learning of asymmetric causal links depends the timing of observations. The particular time requirements here are arbitrary and certainly oversimplified. It is to be expected that such a learning mechanism ought to operate over different time scales, to account for different levels of causal relation (see [7]). Learning of the explaining away inhibitory links, or of weights for any combination function other than noisy-or, is not addressed here. Also ignored here is the problem of type restrictions in the learning of causal rules.

3.2 Simulation results

A prototype version of the SHRUTI system was developed in which these learning mechanisms could be explored. As an initial demonstration, scenarios were presented to the system consisting of the propositions *buy*, *find*, and *own* such that observation of *buy* (e.g., activation

of *+buy*) was often followed by observation of *own* and observation of *find* was less frequently followed by the same. Other observation sequences were not presented. It can be seen in figure 3 that the propositional rules [*buy* \rightarrow *own* 0.75] and [*find* \rightarrow *own* 0.45] (or $P(own|onlybuy) = .75$ and $P(own|onlyfind) = .45$) were easily induced from the data, in that the values on the four relevant links each obtained a value close to that designated for it in the probabilistic model presented above, and other links finished with weights of zero. A similar experiment was performed with relational versions of the same predicates and rules, which resulted in appropriate connections being formed between role nodes of different relations in addition to those formed in the propositional case.

The next experiment involved combining learning and inference. Here the task was to learn a particular causal relation from observation and then examine how this affects what is subsequently learned. In particular, an underlying causal model of [*wetFloor* \rightarrow *slips* 0.7] and [*slips* \rightarrow *hurt* 0.9] was utilized. The system first repeatedly observed *wetFloor* often followed by *slips*, until it had established this causal link. Then the observations pattern changed to *wetFloor* followed by (in a somewhat larger time interval) *hurt*. Because of the activation of *+slips* due to forward propagation of activity from *wetFloor*, both *wetFloor* and *slips* were active observations when *hurt* arrives and so both were learned as possible causes. In a related test where the additional rule [*trips* \rightarrow *hurt* 0.8] is induced and *+trips* is observed alongside *+wetFloor* preceding observation of *+hurt* (with $P(hurt|wetFloor)$ unchanged), the learned causal strength of *wetFloor* and *slips* is diminished.

4 Conclusion

The version of SHRUTI presented here demonstrates that a neurally plausible connectionist system can handle evidential reasoning in a manner that accords well with the laws of probability. Moreover, this model is relational, and so provides a useful approach to the difficult problem of combining relational knowledge and probability. Finally, it has been shown that a simple local learning mechanism is capable of learning significant components of the structure of this probabilistic causal model. How a local learning mechanism can account for the range of combination functions and how the explaining-away inhibitory links can be learned remain important open questions.

Acknowledgements

This work was partially funded by NSF grants SBR-9720398 and ECS-9970890.

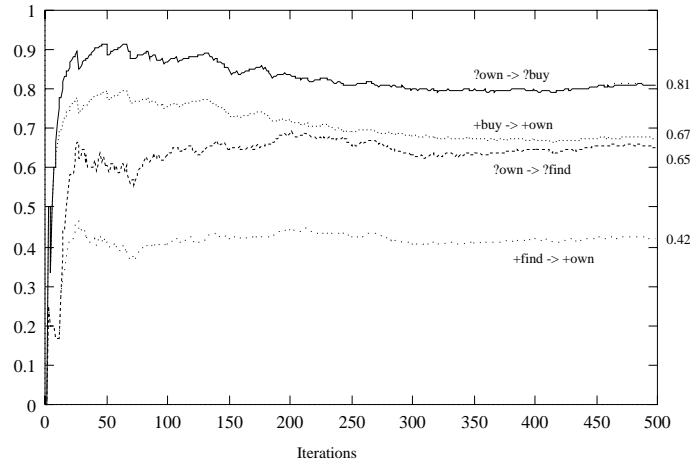


Figure 3: Link weights learned over a sequence of observations.

References

- [1] Feldman, J. (1982) Dynamic Connections in neural networks. *Bio-Cybernetics* 46. 27-39
- [2] Gigerenzer, G. and Hoffrage, U. (1995) How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* 102. 684-704.
- [3] Grossberg, S. (1987) Competitive learning: from interactive adaptation to adaptive resonance. *Cognitive Science* 11, p. 23-26.
- [4] Halpern, J. (1990) An analysis of first-order logics of probability. *Artificial Intelligence* 46. 311-350
- [5] Hebb, D. O. (1949) *The Organization of Behavior*. New York: Wiley.
- [6] Koller, D. and Pfeffer, A.. (1998) Probabilistic frame-based systems. *Proc. AAAI*
- [7] Mobus, G. (1994) Toward a theory of Learning and Representing Causal Inferences in Neural Networks., In: D.S Levine and M. Aparicio (Eds.) *Neural Networks for Knowledge Representation and Inference*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- [8] Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.
- [9] Pearl, J. (1999) *Simpson's Paradox: An Anatomy*
- [10] Poole, D. (1993) Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence* 64(1). 81-129.
- [11] Shastri, L. (1999) Advances in SHRUTI — A neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony., *Applied Intelligence* 11. 79-108.
- [12] Shastri, L. (1999) A Biological Grounding of Recruitment Learning and Vicinal Algorithms. Technical Report TR-99-009, International Computer Science Institute, Berkeley, CA, 94704. April, 1999.
- [13] Shastri, L. (1999) Recruitment of binding and binding-error detector circuits via long-term potentiation. *Neurocomputing* 26-27. 865-874.
- [14] Shastri, L. and Ajjanagadde V. (1993) From simple associations to systematic reasoning. *Behavioral and Brain Sciences*, 16:3. 417-494.
- [15] Shastri, L. and Wendelken, C. (1999) Soft Computing in SHRUTI. In *Proc. the Third International Symposium on Soft Computing Italy*. June, 1999. 741-747.
- [16] Shastri, L. and Wendelken, C. (1999) Knowledge fusion in the Large – taking a cue from the brain. *Proceedings of the 2nd Int'l Conference on Information Fusion* Sunnyvale, Ca.
- [17] Tversky, A. and Kahneman, D. (1974) Judgement under uncertainty: Heuristics and biases. *Science* 185. 1124-1137.
- [18] Valiant, L. (1994) *Circuits of the Mind*. New York: Oxford University Press.