ON THE PROBABILISTIC SEMANTICS OF CONNECTIONIST NETWORKS

Hector Geffner
Judea Pearl

July 1987
CSD-870033

To Appear in Proceedings of IEEE 1st International Conference on Neural Networks, June 1987, SAN DIEGO, CA.

TECHNICAL REPORT
R-~~93~~ 84
June 1987

ON THE PROBABILISTIC SEMANTICS OF CONNECTIONIST NETWORKS * †

Hector Geffner & Judea Pearl

Cognitive Systems Laboratory
UCLA Computer Science Department, L.A., CA. 90024-1600

ABSTRACT

The goodness/energy paradigm [Hopfield 82] has recently emerged as a useful framework for the construction and analysis of connectionist models. Its lack of a clear semantics however, makes the framework unsuitable as an specification language for the declarative content of those models. This paper establishes a correspondence between connectionist networks and a well known family of probabilistic networks, thus, endowing connectionist models with a well understood probabilistic semantics. Additionally we show how a natural extension of the energy formulation presented in [Hopfield 82] leads to models capable of expressing arbitrary probability distributions.

* This work was supported in part by the National Science Foundation, Grant #DCR 83-13875.
† To appear in Proceedings of IEEE First International Conference on Neural Networks, June 1987, San Diego, CA.

On the Probabilistic Semantics of Connectionist Networks

Hector Geffner & Judea Pearl

Cognitive Systems Laboratory, UCLA Computer Science Department, L.A., CA. 90024

ABSTRACT: The goodness/energy paradigm [Hopfield 82] has recently emerged as a useful framework for the construction and analysis of connectionist models. Its lack of a clear semantics however, makes the framework unsuitable as an specification language for the declarative content of those models. This paper establishes a correspondence between connectionist networks and a well known family of probabilistic networks, thus, endowing connectionist models with a well understood probabilistic semantics. Additionally we show how a natural extension of the energy formulation presented in [Hopfield 82] leads to models capable of expressing arbitrary probability distributions.

I.Introduction

Connectionist models appear to play an increasingly important role as mechanisms capable of displaying intelligent behavior. However, while carefully designed systems have attained impressive performance [Sejnowski 86a], it is still not well understood what makes these models achieve or fail to meet the desired specifications [Feldman 85]. Part of the difficulty has been the lack of an appropriate language to talk about the behavior of these models independently of implementation details.

The goodness/energy paradigm [Hopfield 82] has been advocated as providing a useful framework for the analysis of connectionist networks [Feldman 85]. The idea is essentially to attach to each state of the network a goodness measure which corresponds to a sum of local compatibility measures. Units can then compute how the global measure changes with local changes of state. Under certain conditions, if each unit changes state only if the change produces an increase in the overall goodness measure, it can be easily shown that that the network will be driven to a state of local maximum goodness. The thrust of this paradigm has been that it allows the decomposition of the specification of behavior in connectionist models into two different aspects : the specification of desired states for each input configuration as states which maximum goodness, and the characterization of distributed algorithms that given an input configuration will drive the network to a state of maximum goodness. These two subtasks can usually be dealt separately: the first is concerned with the static, declarative content of the model; the second is concerned with its dynamics.

This partition however, generates two subproblems which are far from being trivial. The goodness/energy formulation does not seem to provide any ties between the local compatibility measures and empirically observable relationships. In this paper we show how a well developed formalism, probability theory, can be brought to bear on this task: the specification of the declarative content of connectionist models. We show that a correspondence can be established between connectionist networks and a well known family of probabilistic networks, allowing the former to inherit the better understood semantics of the latter. We also show how a natural extension of the energy formulation as presented in [Hopfield 82] can endow connectionist networks with the expressive power necessary to express arbitrary probability distributions.

A probabilistic account of connectionist models can potentially offer a high level language to describe the

relationships embedded in a network. It might also help to determine whether a given architecture can be made to express (through learning or some other means) a given set of objects and relations. Additionally , for those models involving objects and relations "at a conceptual level", the probabilistic framework might offer a high level specification language from which the connectionist network parameters could be synthesized.

In order to arrive to a probabilistic interpretation of connectionist models, we shall appeal to the language developed in the context of Bayesian Networks [Pearl 86a]; a class of directed acyclic graphs devised to represent the dependency structure that underlies a set of uncertain propositions. Bayesian networks and connectionist models are very akin to each other. Research on the former approach started with a semantically clear specification language (i.e., probability theory) defining the knowledge available, the queries to be asked and the answers desired. It then proceeded to search for a suitable implementation architecture and has found that many of the tasks could be accomplished in a parallel and distributed fashion characteristic of connectionist systems [Pearl 86a,b,c]. Connectionist models have evolved in the opposite way. First, the architecture was identified both as desirable and biologically feasible. Later, that architecture was shown capable of exhibiting some interesting behavior and, finally, a search is under way to find clear semantics for the system's components (e.g., units, activity, weights, topology etc.) in order to facilitate the synthesis of such systems directly, from conceptually meaningful packets of knowledge [Feldman 85].

The paper is structured as follows. Section II shows how connectionist networks can be constructed to express arbitrary second order probability distributions. Section III describes how the addition of multiplicative connections allows the networks to represent arbitrary distributions over binary variables. In section IV we illustrate how the previous result can be extended to distributions over multiple-valued variables. We conclude in section V summarizing some of the features that connectionist networks share with other well known probabilistic networks as well as discussing some of the features in which they differ.

II. A Probabilistic Interpretation of the Energy Coefficients

In [Hopfield 82] Hopfield has shown how some interesting computational abilities can emerge from networks of simple binary units adjusting their state as to minimize a global energy (negative goodness) measure. The slightly modified energy functional proposed in [Hinton 83] is given by :

$$E = -\frac{1}{2}\sum_{ij} w_{ij}\, s_i\, s_j - \sum_i (\eta_i - \theta_i)\, s_i \quad , \tag{1}$$

where η_i is the external input to the $i-th$ unit, w_{ij} is the strength of the connection from the $j-th$ unit to the $i-th$ unit, s_i and s_j are booleans truth values (1 = active , 0 = inactive), and θ_i is a fixed threshold. If connection strengths are symmetric, i.e. $w_{ij} = w_{ji}$, we can derive from (1) the change in energy ΔE_k due to the activation of unit k to be :

$$\Delta E_k = -\sum_i w_{ki}\, s_i - \eta_k + \theta_k \quad . \tag{2}$$

Moreover it can be easily shown that if units get activated only if the change is negative, and transmission delays are negligible, these systems will always settle into a local energy minimum. Furthermore, since the decision rule can be computed using information available in the local neighborhood, the algorithm is truly distributed.

The questions we shall address next are : a) what is the probabilistic interpretation of the terms in (1) and (2), and b) what kind of probability distributions can be captured by these models. These questions have partially been addressed by Hinton and Sejnowski in [Hinton 83]. In the following section we extend and correct some of their conclusions.

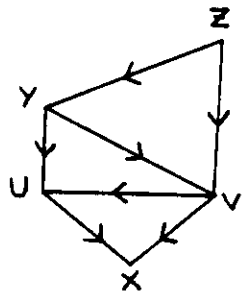
Formulation

Let $S = \{X, U, V, Y, \ldots, Z\}$ be a set of binary variables. We will denote their states in small italic letters, and use small bold letters to denote state variables. So while x and $\bar{x}$ stand for the active and inactive state of X, $\mathbf{x}$ will serve as a generic symbol for the values, x and $\bar{x}$, that X might attain. We will use the letters $A, B, \ldots$ as typical variables from $X, U, \ldots, Z$.

It is well known that any probability distribution on a set S of variables $X, U, V, \ldots, Z$ can be decomposed as a product of conditional probabilities of the form :

$$P(\mathbf{x}, \mathbf{u}, \mathbf{v}, \ldots, \mathbf{z}) = P(\mathbf{x} \mid \mathbf{f}_X) P(\mathbf{u} \mid \mathbf{f}_U) P(\mathbf{v} \mid \mathbf{f}_V) \cdots P(\mathbf{z} \mid \mathbf{f}_Z) \ , \tag{3}$$

where $\mathbf{f}_A$ is the value of a (possibly empty) subset F_A of variables in S. Moreover the right hand side of the equation determines a set of dependencies among the variables of S which can be captured by a directed acyclic graph, where each vertex A has the variables in F_A as its parents. This graphic representation of probability distributions has been called called Bayesian Networks [Pearl 86a,b], and will be used throughout the paper to display the decomposition of arbitrary probability distribution (Fig.1). We will also use the term ports of a variable A to refer to the set of factors in the right hand side of (3) in which the variable A appears.



$$P(\mathbf{x}, \mathbf{u}, \mathbf{v}, \mathbf{y}, \mathbf{z}) = P(\mathbf{x} \mid \mathbf{u}, \mathbf{v}) P(\mathbf{u} \mid \mathbf{y}, \mathbf{v}) P(\mathbf{v} \mid \mathbf{y}, \mathbf{z}) P(\mathbf{y} \mid \mathbf{z}) P(\mathbf{z})$$

Fig.1 A probability distribution and its Bayesian Network

Let us now define the energy measure associated with the state $\mathbf{s} = \mathbf{x}\,\mathbf{u}\,\mathbf{v} \cdots \mathbf{z}$ under the probability distribution P over S as [†] :

$$E(\mathbf{x}, \mathbf{u}, \mathbf{v}, \ldots, \mathbf{z}) = -\ln \frac{P(\mathbf{x}, \mathbf{u}, \mathbf{v}, \ldots, \mathbf{z})}{P(\bar{x}, \bar{u}, \bar{v}, \ldots, \bar{z})} \tag{4}$$

$$= -\ln \left[\frac{P(\mathbf{x} \mid \mathbf{f}_X)}{P(\bar{x} \mid \bar{f}_X)} \frac{P(\mathbf{u} \mid \mathbf{f}_U)}{P(\bar{u} \mid \bar{f}_U)} \frac{P(\mathbf{v} \mid \mathbf{f}_V)}{P(\bar{v} \mid \bar{f}_V)} \cdots \frac{P(\mathbf{z} \mid \mathbf{f}_Z)}{P(\bar{z} \mid \bar{f}_Z)} \right] .$$

Then, if we denote the set of children of X by C_X and abbreviate the set difference $F_A - X$ by F_A^X, we can write the change in energy $\Delta E(X)$ due to the activation of X in the current state $\mathbf{s}$ as :

$$\Delta E(X) = -\ln \frac{P(x, \mathbf{u}, \mathbf{v}, \ldots, \mathbf{z})}{P(\bar{x}, \mathbf{u}, \mathbf{v}, \ldots, \mathbf{z})} \tag{5}$$

[†] It is well known that every Markov Field P over an undirected graph G has an energy made up of the sum of local energies over the maximal cliques of G [Geman 84]. However, in contrast to the conditional probabilities appearing in (3), the terms that make up P do not have a clear, experiential content [Pearl 86a], making it less suitable as an specification language for connectionist models.

$$= -\ln \left[\frac{P(x \mid \mathbf{f}_X)}{P(\bar{x} \mid \mathbf{f}_X)} \prod_{A \in C_x} \frac{P(\mathbf{a} \mid x, \mathbf{f}_A^X)}{P(\mathbf{a} \mid \bar{x}, \mathbf{f}_A^X)} \right]$$

$$= -\left[LP(x \mid \mathbf{f}_X) - LP(\bar{x} \mid \mathbf{f}_X)\right] - \sum_{A \in C_x} \left[LP(\mathbf{a} \mid x, \mathbf{f}_A^X) - LP(\mathbf{a} \mid \bar{x}, \mathbf{f}_A^X)\right] ,$$

where $LP(\cdot)$ is an abbreviation for $\ln P(\cdot)$ *. Thus, $\Delta E(X)$ appears as the sum of the energies contributed by the ports associated with the parents and children of X. If we denote by P^X the set of ports associated with X, and by P_i^X, its $i-th$ port, we can rewrite (5) as :

$$\Delta E(X) = \sum_{i : P_i^X \in P^X} \Delta E_i(X) \quad , \tag{6}$$

where $\Delta E_i(X)$ is given by :

$$\Delta E_i(X) = \begin{cases} -\left[LP(x \mid \mathbf{f}_X) - LP(\bar{x} \mid \mathbf{f}_X)\right] & \text{if } P_i^X \text{ is associated with } X\text{'s parents } F_X \\ -\left[LP(\mathbf{a} \mid x, \mathbf{f}_A^X) - LP(\mathbf{a} \mid \bar{x}, \mathbf{f}_A^X)\right] & \text{if } P_i^X \text{ associated with } X\text{'s child } A \end{cases} \tag{7}$$

and stands for the contribution to $\Delta E(X)$ coming from the $i-th$ port of X.

Second Order Constraints

We will assume in this section that the probability distribution P is of second order, i.e. there exists a decomposition in which for every variable A, F_A is either empty or is such that : (8)

$$P(\mathbf{a} \mid \mathbf{f}_A) = k\, P(\mathbf{a} \mid \mathbf{f}_{A,1}) P(\mathbf{a} \mid \mathbf{f}_{A,2}) \cdots P(\mathbf{a} \mid \mathbf{f}_{A,n}) \ ,$$

where k is a normalizing constant, and the $F_{A,i}$'s stand for the individual variables included in F_A. A special case of (8) is a tree, where each node has a single parent.

We can now rewrite (5) as :

$$\Delta E(X) = -\ln \prod_i \frac{P(x \mid \mathbf{f}_{X,1})}{P(\bar{x} \mid \mathbf{f}_{X,1})} \prod_{A \in C_x} \frac{P(\mathbf{a} \mid x) P(\mathbf{a} \mid \mathbf{f}_A^X)}{P(\mathbf{a} \mid \bar{x}) P(\mathbf{a} \mid \mathbf{f}_A^X)} \tag{9}$$

$$= -\sum_i \left[LP(x \mid \mathbf{f}_{X,1}) - LP(\bar{x} \mid \mathbf{f}_{X,1})\right] - \sum_{A \in C_x} \left[LP(\mathbf{a} \mid x) - LP(\mathbf{a} \mid \bar{x})\right] .$$

Note that, due to the assumptions implicit in (8), and in contrast with (5), the terms in the last expression reflect only pairwise interactions among variables .

We are interested in expressing $\Delta E(X)$ in the form :

$$\Delta E(X) = -\sum_l w_{Xl} s_l + \theta_X \quad , \tag{10}$$

where w_{Xl} and θ_X are real coefficients and s_l is a Boolean value. Since we have found in (6) that $\Delta E(X)$ can be expressed as a linear combination of the $\Delta E_i(X)$, it would suffice to find the coefficients of the expansion of the

* Note that the probability of $X = x$ can be recovered as : $P(x \mid \mathbf{s}) = (1 + e^{-\Delta E(X)})^{-1}$. This measure turns out to be essential for stochastic relaxation algorithms (see for instance [Pearl 87]).

latter, which under the current second order restriction would look like :

$$\Delta E_i(X) = -w^i_{XY}\, s_Y + \theta^i_X \quad , \tag{11}$$

where Y denotes the only variable within P^X_i. Note that since each variable identifies a port and viceversa, we can safely drop the port index from the weights w. The other coefficient of $\Delta E(X)$ can then be computed from :

$$\theta_X = \sum_i \theta^i_X \quad .$$

We have from (7) that, if Y is a parent of X :

$$\Delta E_i(X) = -[LP(x|y) - LP(\bar{x}|y)] \quad , \tag{12}$$

so after equating (11) and (12) for both possible values of Y we obtain :

$$\theta^i_X = -[LP(x|\bar{y}) - LP(\bar{x}|\bar{y})]$$

and

$$w_{XY} = [LP(x|y) - LP(\bar{x}|y)] - [LP(x|\bar{y}) - LP(\bar{x}|\bar{y})] \quad .$$

Likewise, if we associate X with the $k-th$ port of its parent Y, we have from (7) that :

$$\Delta E_k(Y) = -[LP(x|y) - LP(x|\bar{y})]$$

and hence, we obtain a threshold and a symmetric link weight given by :

$$\theta^k_Y = -[LP(\bar{x}|y) - LP(\bar{x}|\bar{y})]$$

and

$$w_{YX} = w_{XY} \quad .$$

These equations are sufficient to derive the weights and thresholds of a connectionist network that captures the behavior of an arbitrary second order probability distribution over a finite set of binary variables [*]. There is however a term missing. For those variables R having no parents, F_R is empty and there is a term contributing to ΔE due to the priors of the form :

$$\Delta E_p(R) = -[LP(r) - LP(\bar{r})] \quad ,$$

where p identifies R's empty parents port. This term appears as a sustained input which is captured by the the input term η of (1), i.e. :

$$\eta_R = -[LP(r) - LP(\bar{r})] \quad .$$

We have shown so far how to synthesize the coefficients of (1) in such a way as to capture any second order probability distribution. The interpretation of connection strengths is the same as the one arrived by Hinton and Sejnowski [Hinton 83]. The expressions for thresholds and input coefficients is however slightly different, and weaker independence assumptions are needed. We proceed next to show how higher order probability distributions can be captured in connectionist networks with symmetric weights.

II. Capturing Higher Order Constraints

We have shown above how, for any variable X, the coefficients of $\Delta E(X)$ can be computed from the coefficients of the $\Delta E_i(X)$'s contributed by each of its ports. Taking advantage of this result, we will simplify the forthcoming discussion by considering a *single port net*, as the one depicted in Fig.1, in which variables U and V are the parents of variable X, both with priors equal to 1/2 (i.e. the energy contribution due to the priors of U and V is assumed to be 0). As we shall see the results will still hold for any number of X's parents.

---

[*] We are not considering however the limitations imposed by bounded connection strengths in capturing "extreme" probability relations (with 0's and 1's). This limitations could in principle be overcome if null probability entries are replaced by small ε's.

An arbitrary probability $P(\mathbf{x} \mid \mathbf{u},\mathbf{v})$ will not be in general expressible as the product of $P(\mathbf{x} \mid \mathbf{u})$ and $P(\mathbf{x} \mid \mathbf{v})$. Instead, it constitutes an irreducible third order constraint among the variables X,U and V. To capture this constraint we shall extend the form of $\Delta E(A)$, for $A \in \{U,V,X\}$, to be :

$$\Delta E(A) = -\sum_{ij} w_{A\,i\,j}\, s_i\, s_j - \sum_k w_{A\,k}\, s_k - w_A \quad , \tag{13}$$

where we have introduced a new second order term and changed the notation used to denote the threshold from θ_i to $-w_i$. Note that the local computation of ΔE according to (13) will now require an architecture admitting "multiplicative connections".

Since we will need to make the current network state explicit, we will use ΔE_S^A to refer to the change in global energy due to activating A when only the neighbor variables appearing in S are already on.

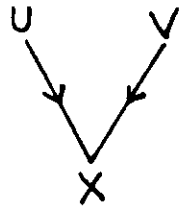


Fig.2 Small Bayesian Network : X has two parents U and V.

For the small network depicted in Fig.2, we can obtain from (5), (7) and (13) the following equations characterizing the transitions due to a change in X :

$$-\Delta E^x = LP(x \mid \bar{u},\bar{v}) - LP(\bar{x} \mid \bar{u},\bar{v}) = w_X$$

$$-\Delta E_u^x = LP(x \mid u,\bar{v}) - LP(\bar{x} \mid u,\bar{v}) = w_{X\,U} + w_X$$

$$-\Delta E_v^x = LP(x \mid \bar{u},v) - LP(\bar{x} \mid \bar{u},v) = w_{X\,V} + w_X$$

$$-\Delta E_{u\,v}^x = LP(x \mid u,v) - LP(\bar{x} \mid u,v) = w_{X\,U\,V} + w_{X\,U} + w_{X\,V} + w_X \quad ,$$

and the following equations for transitions due to a change in U :

$$-\Delta E^u = LP(\bar{x} \mid u,\bar{v}) - LP(\bar{x} \mid \bar{u},\bar{v}) = w_U$$

$$-\Delta E_x^u = LP(x \mid u,\bar{v}) - LP(x \mid \bar{u},\bar{v}) = w_{U\,X} + w_U$$

$$-\Delta E_v^u = LP(\bar{x} \mid u,v) - LP(\bar{x} \mid \bar{u},v) = w_{U\,V} + w_U$$

$$-\Delta E_{x\,v}^u = LP(x \mid u,v) - LP(x \mid \bar{u},v) = w_{U\,X\,V} + w_{U\,X} + w_{U\,V} + w_U$$

A similar set of equations can be obtained for V. Notice that we can easily obtain the values of the coefficients from the probability relations : each new equations introduces a single new unknown. Moreover this fact will not depend on the degree of the relationship.

The second step is to show that the coefficients computed above are indeed symmetric. For that purpose let E^S stand for the energy of the net when only the variables in S are activated. Then from the probabilistic definitions of E and ΔE, it follows after simple manipulations that :

$$E^{x\,u} = \Delta E^x + \Delta E_x^u = \Delta E^u + \Delta E_u^x$$

and therefore from the equations above, it follows that : $w_{X\,U} = w_{U\,X}$.

Similarly :

$$E^{uv} = \Delta E^{u} + \Delta E^{v}_{u} = \Delta E^{v} + \Delta E^{u}_{v}$$

and therefore $w_{UV} = w_{VU}$. In the same fashion we can show, corresponding to the intersection points of Fig.3, that $w_{XV} = w_{VX}$ and $w_{XUV} = w_{UXV} = w_{VXU}$.

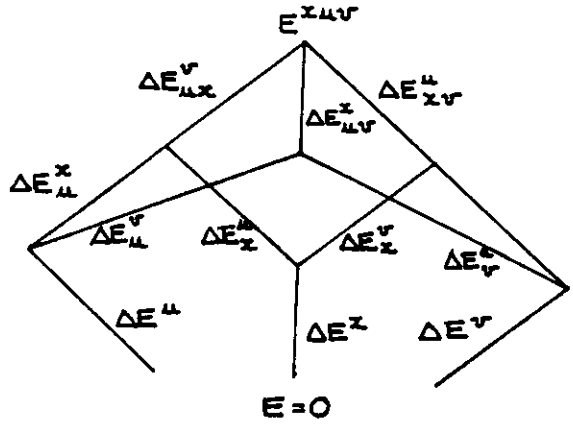


Fig.3 Energy transitions for the network relating U, V and X.

Example

Let us consider how the equations above can be used to construct a tiny net composed of three units U, V and X, where the desired relationship is a noisy-OR gate. This type of relationship is common among systems performing word sense disambiguation [Cotrell 84]. For instance, the variables X, U and V could take part in a larger network standing for a given word and its two alternative word senses. The relationship between X, U and V could be then defined as :

$$P(x \mid u,v) = \begin{cases} 0.9 & \text{if } u = u \text{ or } v = v \\ 0.1 & \text{otherwise} \end{cases}$$

Plugging these values into the equations above we obtain the net parameters :

$$w_X = w_U = w_V = -k \qquad w_{XU} = w_{XV} = 2k \qquad w_{XUV} = -2k$$

all symmetric, in which $k = \ln 9$ (Fig.4). We might also go in the opposite direction and ask how to modify the specification of the noisy-OR gate above in order to leave the net with pairwise interactions only, i.e., making $w_{XUV} = 0$. Again by simple manipulations it turns out that if we only change the entry $P(x \mid u,v)$ to 0.9878, we get $w_{XUV} = 0$ with $w_{UV} = -k$.

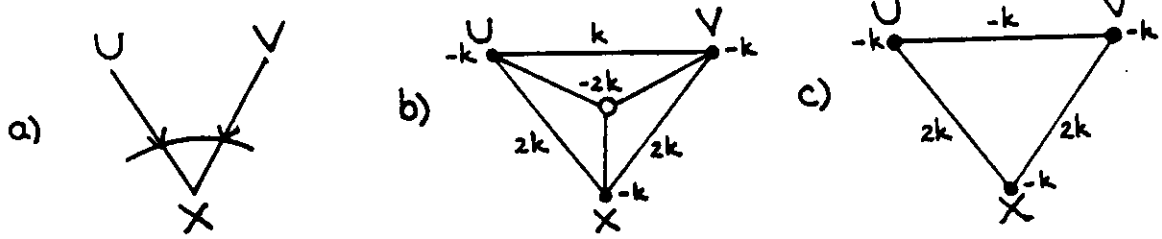


Fig.4 a) a noisy-OR b) the equivalent net c) net with $w_{XUV} = 0$

Note that the probability distribution unambiguously specifies the relationships among the variables that we want the network to express. It is more reasonable to specify the entries of $P(x \mid u,v)$ than directly specify the parameters of the equivalent net. The reason being, as we said above, that the former has a clear experiential meaning that is missing from the latter.

IV.Discussion

We have so far focused on translating probability distributions defined over binary variables to networks with of 2-state units standing in one-to-one correspondence to these variables. We shall now briefly address the case of non-binary variables, where a direct correspondence of variables to units and values to states no longer holds.

The mapping of a set of non-binary variables to a set of binary units can be achieved by associating with each variable a set of "private" units, so that a particular variable instantiation corresponds to particular state(s) of its set of units [†]. Let $X,Y,...,Z$ stand for the possibly non-binary variables in the set S and let $X^u,Y^u,\ldots,Z^u$ denote their corresponding set of units. Then the probability distribution $P(\cdot)$ over the set of variables S induces a probability $P_u(\cdot)$ over the set of units S^u which satisfies :

$$P_u(\mathbf{x}^u,\mathbf{y}^u,\ldots,\mathbf{z}^u) = \begin{cases} P(\mathbf{x},\mathbf{y},\ldots,\mathbf{z}) & \text{for } \mathbf{x}^u,\mathbf{y}^u,\ldots,\mathbf{z}^u \text{ encoding the values } \mathbf{x},\mathbf{y},\ldots,\mathbf{z} \\ \varepsilon \rightarrow 0 & \text{otherwise} \end{cases}$$

In other words, the probabilities associated with the states of S^u which encode instantiations of the variables in S are kept in the same ratio, while all other configurations are neglected. In this sense we can say that the probability distribution over S^u "represents" the relationships defined by $P(\cdot)$ over S.

Since the resulting probability distribution $P_u(\cdot)$ is defined over a set of binary variables (i.e. units), the method described in the previous section is applicable. We can then determine the network weights necessary to express $P_u(\cdot)$ over a set of units S^u which, as we argued above, captures the relationships embodied by $P(\cdot)$ over the multi-valued variables in S.

Note that the ability to express probability distributions over a set of multi-valued variables can be applied back to the binary case to *cluster* several binary-valued variables into a single multi-valued one. Clustering and other encodings of multiple-valued variables over sets of binary units, determine a whole range of representations with equivalent expressive power (i.e. they capture the same probability distribution among the variables), but an interesting spectrum of distinct features. While we have not investigated these features throughly, we tend to believe that these representations will differ mainly in the resulting *order* of the constraints induced among the units as in the *shape* of the resulting energy landscape. An interesting challenge will be to minimize the order of the constraints induced (to minimize the complexity of the architecture), while preserving the smoothness of the energy surface (to reduce the myopic effects of relaxation algorithms).

V. Conclusion

We have argued that probability theory provides a framework in which to analyze the semantics of connectionist networks. We have also described how to construct networks which capture arbitrary probability distributions. The inverse process also holds; the theory of Markov fields permits one to uncover the probability distribution embedded in a network, once the local energy terms has been specified [Geman 84]. Thus, connectionist networks as proposed by Hopfield [Hopfield 82], with the addition of higher order terms, posses an expressive power equivalent to that of Bayes and Markov Networks [Pearl 86a].

The challenge remains to exploit the features that make these networks different. While we have been assuming the feasibility of multiplicative connections among units, workers in connectionist networks have avoided whenever possible its introduction (see [Sejnowski 86b] for a discussion of the issue). However, clustering as well as different encoding of variables over units determine a whole spectrum of representations that may render pairwise interactions among units and a smooth energy landscape. Moreover the addition of hidden units has been

[†] This idea, in a slightly different flavor, has been advocated in the context of "distributed representations" [Hinton 84].

shown to further increase the expressive power of second order nets [†]. While we have not considered hidden units here, we hope that the framework laid out in this paper can be extended to formally characterize the spectrum of representations that will emerge from the interaction of hidden units with the other two factors mentioned above. We also expect that this characterization might shed some light on the *functional* properties emergent from distributed representations [Hinton 84].

Acknowledgment: This work was supported in part by the National Science Foundation grant #DCR 83-13875.

References

[Cotrell 84] Cottrell G. (1984). A model of Lexical Access of Ambiguous Words, *Proceedings of AAAI-84*, Texas, Austin, pp 61-67.

[Feldman 85] Feldman J. (1985). *Energy and the Behavior of Connectionist Models*, CSD University of Rochester, TR-155.

[Geman 84] Geman S., Geman D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6, No 6, pp 721-741, November 1984.

[Hinton 83] Hinton G. & Sejnowski T. (1983). Optimal perceptual inference. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 448-443.

[Hinton 84a] Hinton G. (1984) *Distributed Representations* CSD Carnegie-Mellon University, CMU-CS-84-157 .

[Hopfield 82] Hopfield J. (1982). Neural networks and physical systems with emergent collective abilities. *Proceedings of the National Academy of Sciences, USA, 79*,2554-2558.

[Pearl 86a] Pearl J. (1986) Markov and Bayes Networks, UCLA, Cog. Systems Lab TR-46.

[Pearl 86b] Pearl J. (1986) Fusion, Propagation, and Structuring in Belief Networks, *Artificial Intelligence Journal*, Vol. 29, No 3, pp 241-288.

[Pearl 86c] Pearl J. (1986) Distributed Revision of Belief Commitment in Multi-Hypothesis Interpretation, *2nd AAAI Workshop on Uncertainty in AI*, Philadelphia, PA. Also in *Artificial Intelligence Journal* (forthcoming).

[Pearl 87] Pearl J. (1986) Evidential Reasoning Using Stochastic Simulation of Causal Models, *Artificial Intelligence Journal*, Vol. 32, pp 245-257.

[Sejnowski 86a] Sejnowski T., Rosenberg C. (1986) NETtalk : A Parallel Network that Learns to Read Aloud, John Hopkins University, Electrical Engineering and Computer Science, TR JHU/EECS-86/01

[Sejnowski 86b] Sejnowski T. (1986) Open questions about the computation in cerebral cortex in McClelland J. and Rumelhart D. (Eds) *Parallel Distributed Processing*, Vol 2, Brandford Books/MIT Press.

[†] See [Pearl 86b] for a treatment of hidden units within the context of Bayesian Networks. An algorithm is described that finds a set of pairwise probabilistic constraints that, together, capture a class of higher order probability distributions.