

Maximum Entropy Prediction in Neural Networks

William B Levy, Ph.D.
Department of Neurological Surgery
University of Virginia School of Medicine
Charlottesville, Virginia 22908

Neural networks can generate predictive representations. Particularly interesting are the networks that produce a predictive representation that is a vector of probabilities associated with a future representation. These predictive representations are interesting because, among other reasons, such vectors imply the predictive representation of maximum likelihood.

Such a vector of probabilities, called a type II predictive representation to distinguish it from a maximum likelihood predictive representation, consists of elements each of which is the probability of one particular neuron firing. More specifically, each neuron produces the probability of its own future state. In our investigations each neuron generates this probability by a local computation that uses maximum entropy (M.E.) inference, stored averages, and Bayes's equation.

The motivation for such investigations⁸ comes from the importance of prediction in the life of an animal^{3,17}, from various observations that point out how synaptic modification can lead to the encoding of a statistic^{1,4,14}, and from the existence of a unique, optimal procedure to produce probabilities based on statistics, i.e., M.E. inference^{6,15}.

In order for a neural network to produce this type II predictive representation, there is a small set of absolute axioms and requirements: a definition of the prediction problem, the requirements of M.E. inference, and complexity considerations. In turn the implementation of these requirements leads to a set of network characteristics as natural outcomes of the usual, classic characteristics of neurons and synapses. The purpose of this communication is to point out the implications these requirements have for neural networks which mediate predictive representations.

THE PREDICTION PROBLEM

It is first necessary to define the type of prediction we are studying. We want a network to generate predictions which are useful (i.e. usable by another network or the organism itself); moreover, we want a network to base such predictions on appropriate correlational information which is adaptively encoded at neurons and synapses.

There are three requirements of a predictive representation. (To emphasize that a prediction is also a representation, we call it a predictive representation as distinct from a "standard representation.") The first two requirements stem from our interest only in predictions which will be usable.

(1) Timeliness: A predictive representation must precede in time the standard representation being predicted (because the whole point of creating a prediction is to improve some outcome in the future).

(2) Meaningfulness: A predictive representation space must map into the standard representation space being predicted about. This requirement is necessary if a prediction is to be used for the benefit (survival, propagation) of the organism containing this network. A specific example might make the motivation for this requirement clearer. Consider a prediction generating network which is embedded in a larger network that alters W in the external world to suit some homeostatic purpose. This larger network needs to relate, or map, a predictive representation onto the standard representation of W in order to use the prediction to control W in a sensible way that anticipates homeostatic needs. Without this map, the homeostatic part of the network would be unable to take advantage of the predictions being generated.

The third requirement stems from our desire that the network encode and use appropriate associative (correlational) information.

(3) Aptness: Appropriate correlations must be used to generate predictions. Among the many constraints on a neural network there is the local principle which sensibly limits the information available to a neuron; e.g., of all the synaptic weights and neuronal activities in a network, only those which are inputs to a neuron are local to that neuron. Then suppose, because of the local principle and extant circuitry, a neuron can only learn an association between a representation in the space A and another representation in the space B where any one A representation precedes any one B representation. Then it is appropriate for this neuron to use an A representation to predict about B . However, it is not appropriate for the neuron to use a B representation to predict about A in the future or to use an A representation to predict about some nonlocal event C (except in

the sense that C can be represented as B). Furthermore, even though the local principle requires convergence of an input to a cell such convergence is not enough for aptness. Just because the inputs to a neuron are, by definition, local to that neuron, does not imply that it is appropriate for such a neuron to associate just any set of these inputs. Specifically, a neuron would be living in a fantasy world of inappropriate correlational encodings if it were to use its own predictive representation rather than a standard representation for synaptic reinforcement.

NETWORK CHARACTERISTICS DERIVATIVE OF THIS DEFINITION

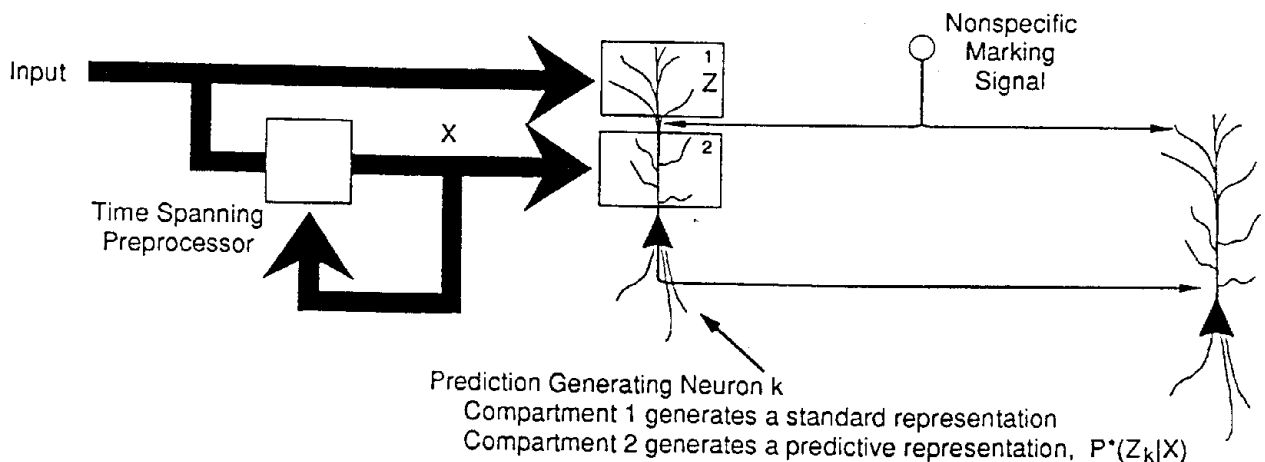
The requirement for a timely prediction is, in part, implemented by the temporal characteristics of associative modification rules^{5,12}. These characteristics allow time spanning associations between two standard representations. However, the allowable time span of these rules is rather short, certainly less than 125 ms. This span is hardly large enough for most practical purposes requiring prediction. An arbitrarily large time gap, however, can be spanned by using feedback circuitry for preprocessing the representation that will be used as the conditioning variable in prediction generation⁸.

Requirements (2) and (3) produce the need for a bidirectional mapping between the standard representation space and the corresponding predictive representation space. The obvious implementation of this bidirectional mapping is that the same neuron mediates both a standard representation and its own predictive representation. There are advantages to such a mapping including simplicity, immediate interpretability, and constancy over changes in circuitry due to synaptic modification. However, although this mapping prevents confusion as to the applicable space of a prediction, it does not prevent, rather it can cause, confusion between the two types of representations themselves.

Such a dual representational implementation within a single neuron leads to the possibility that the two different types of representations will be confused. Such confusion could happen in two places. One place is at the associatively modifiable synapses of the prediction generating neuron. Associative synaptic modification could be reinforced by the unconfirmed state of a predictive representation instead of the reality of a standard representation. The other place of possible confusion is at the postsynaptic cells receiving inputs from such a prediction generating neuron. These postsynaptic neurons must distinguish which input signals are predictive representations and which are standard representations.

The distinction which solves the first problem can be accomplished by a nonspecific, low dimension, marking signal, a two-compartment neuron, and non-simultaneity of the two types of representations. The second problem needs only a marking signal and the non-simultaneity of the two representations to produce the necessary distinction. Such marking signals would alter the interaction of postsynaptic cells according to the marking signal's temporal relationship with one of the two types of representations.

Combined with the network characteristics just noted, aptness is satisfied by an associative modification rule in which reinforcement of synaptic modification is a rectified affair. Consider the event being predicted as the voltage of compartment 1, Z, in the figure. Then Z in compartment 1 can reinforce compartment 2 synaptic modification for a state of each input X_i in compartment 2 that precedes the state Z but cannot reinforce modification of states which occur after state Z in time. Furthermore, the prediction generating compartment (2 in the figure) cannot reinforce synapses in compartment 1.



Although the definitional requirements constrain the characteristics of suitable networks, our implementations are probably not unique in satisfying these requirements because so many different types of neurons exist. Still, the suggested network characteristics seem rather natural, straightforward constructions, particularly when considered in the context provided by the hippocampus of the mammalian brain⁸.

Let us now consider computational constraints which limit the characteristics of a neural network. For simplicity of exposition, we consider a standard representation space in which each neuron takes on a state in the set {0,1} and a predictive representation space in which each neuron takes on some monotonic function of a probability; for expository purposes, let it take on values in [0,1] as the probability itself.

COMPUTATIONAL CONSIDERATIONS

In an idealized situation each prediction generating neuron, k , would, at time t , generate $P(Z_k(t+n)=1 | X(t)=x)$, the conditional probability that k will be in the one state at some n steps into the future given that its inputs, the vector variable X , is in state x now at time t .

Because the dimension of X is quite large, there will be many configurations X that have never been experienced before the prediction generation time t . Moreover, even with sufficient sampling, it is impossible to store the exponentially many statistics, e.g. the expectation $E[Z_k = 1 | X = x]$, that might be needed. The solution to this problem is to use Bayes's equation and M.E. inference on low-order moments. In accord with Bayes's equation and leaving implicit the time notations mentioned above, each neuron would compute

$$P^*(Z_k=1 | X=x) = P^*(X=x | Z_k=1) \cdot P^*(Z_k=1) / P^*(X=x) \quad (1)$$

where, on the right hand side, $P^*(\)$ is a M.E. inferred probability distribution computed from sample averages that have nearly converged to the population based expectation. A M.E. inferred probability based on the lowest order moments of interest is:

$$P^*(X=x | Z_k=1) = \prod_i P^*(X_i | Z_k=1) = \prod_i \bar{P}(X_i=1 | Z_k=1)^{x_i} \cdot (1 - \bar{P}(X_i=1 | Z_k=1))^{(1-x_i)} \quad (2)$$

$$= \exp\left\{ \sum_i x_i \cdot \log \bar{P}(X_i=1 | Z_k=1) + (1-x_i) \cdot \log (1 - \bar{P}(X_i=1 | Z_k=1)) \right\} \quad (3)$$

where the $\bar{P}(\)$ are sample based averages.

As in many neural networks, synapses would store statistics, e.g. $(\bar{P}(X_i | Z_k=1))$, and the readout of the appropriate statistics is just a natural result of the signal flow, x_i . The "learning" of such statistics could result from synaptic modification rules similar to those known to exist in the brain^{7,9,10,11}.

The existence of synaptic modification rules that encode averages provided much of the impetus for the approach described here. Equation (1) and computations like equation (2) imply the need for three different types of averages: $\bar{P}(X_i=1 | Z_k=1)$, $\bar{P}(Z_k=1)$, and $\bar{P}(X_i=1 | Z_k=0)$. Note that $\bar{P}(Z_k)$ requires a neuron to encode its own average activity or to guarantee a preset value. Note also that a synaptic encoding of $\bar{P}(X_i=1 | Z_k=1)/\bar{P}(X_i=1)$ might substitute for the pair $(\bar{P}(X_i=1 | Z_k=0), \bar{P}(X_i=1 | Z_k=1))$. In either case a neuron has the information to calculate the denominator of equation (1). Interestingly there is recent evidence for the existence of a synaptic modification rule that would encode $\bar{P}(X_i | Z_k=0)$ or the alternate statistic^{9,16}.

Regardless of the statistics in the constraint set (in our case the low order statistical correlations), M.E. inference will always use what is essentially a multiplicative form². This requirement leads to an obvious implementation in a neural network: add logarithms and exponentiate. Because synaptic currents vary with the logarithm of their conductance and because depolarization translates into cell firing in a nonlinear way, such hypothetical characteristics are plausible. (Of course a maximum likelihood predictive representation does not even require exponentiation before its formation from a monotonic function of a type II predictive representation formed with logarithms.)

ANOTHER COMPLEXITY CONSTRAINT

M.E. inference can make use of moment constraints, which are essentially correlations, of any order. For a postsynaptic cell k and inputs X_i, X_j, X_h , a lowest order constraint of interest would be $\bar{P}(X_i=1 | Z_k=1)$ and examples of higher order constraints are $\bar{P}(X_i \cdot X_h=1 | Z_k=1)$ and $\bar{P}(X_i \cdot X_j \cdot X_h=1 | Z_k=1)$.

Unfortunately computational complexity issues often make it impossible, in practice, to compute a M.E. derived probability from a constraint set of arbitrary moments of arbitrary order. A computationally intractable

problem can arise if the constraint set does not consist solely of lowest order correlates because there might be overlap among two constraints in the set. When such overlap exists, there can be a need for an exponential number of variables to calculate the M.E. inferred probability distribution. This exponential requirement renders the M.E. method computationally intractable in such cases.

Example of a set of nonoverlapping moment constraints:

$$\{\bar{P}(X_1=1 | Z_k=1), \bar{P}(X_2=1 | Z_k=1), \bar{P}(X_3 \cdot X_4=1 | Z_k=1)\}.$$

Example of a set with overlapping constraints:

$$\{\bar{P}(X_1=1 | Z_k=1), \bar{P}(X_1 \cdot X_4=1 | Z_k=1)\}.$$

On the other hand, intractability will always be avoided if the constraints in a set do not overlap. That is, if a conditioned variable X_i (i.e. the activity of input line i) appears in no more than one moment constraint in a set of constraints, then there will be no overlap, and this particular complexity problem is avoided.

We view this complexity problem and its solution as a constraint affecting the characteristic computation of a prediction generating network. More exactly the affect is on the preprocessor computation that produces the X inputs of the prediction generating neurons. It would be useful for the prediction generating neurons to receive their inputs from a network preprocessor that moves information out of higher-order moments into lower-order moments and that avoids overlapping constraints with high probability. Moreover, it seems possible that this preprocessor and the time spanning preprocessor are identical⁸.

Full connectivity from the X space to a prediction generating neuron is not a requirement because M.E. inference remains consistent even with missing moment constraints.

Thus the requirements stemming from this particular computation, equations (1) and (2) or (3), imply several network characteristics. These requirements include: a multiplicative combination of probabilities; computation of M.E. inference as if working with lowest order moments; and a requirement for averages over three different kinds of distributions. The combination of these computational requirements and the definitional requirements create an important set of restrictions on the class of acceptable neural networks that create predictions.

This research and WBL are supported in part by the NIH (NS14588) and by an NIMH RSDA (MH00622). The help of D. Adelsberger-Mangan, C. M. Colbert, and N. L. Desmond is greatly appreciated.

References

1. Amari, S.-I. (1977) *Biol. Cybern.*, 26, 175-185.
2. Csiszär, I. (1975) *Ann. Prob.* 3, 146-158.
3. Dawkins, R. (1976) *The selfish gene*. New York: Oxford University Press.
4. Geman, S. (1981) *SIAM AMS Proc.*, 13, 91-105.
5. Gustaffson, B., Wigström, H., Abraham, W. C., and Huang, Y.-Y. (1985) *J. Neurosci.* 7, 774-780.
6. Jaynes, E. T. (1978) In R. D. Levine & M. Tribus, Eds., *The maximum entropy formalism*. 15-118. Cambridge: MIT Press.
7. Levy, W. B. (1982) *Proc. Fourth Annual Conference of Cognitive Science Society*, 135-136.
8. Levy, W. B. (in press) In: R. D. Hawkins and G. H. Bower, Eds., *Computational models of learning in simple neuronal systems*. New York: Academic Press.
9. Levy, W. B., Colbert, C. M. and Desmond, N. L. (1989) In: M. A. Gluck and D. E. Rumelhart, Eds., *Neuroscience and connectionist models*. Hillsdale, NJ: Lawrence Erlbaum Assoc., Inc.
10. Levy, W. B. and Desmond, N. L. (1985) In G. Buzsáki and C. H. Vanderwolf, Eds., *Electrical activity of the archicortex*. Budapest, Hungary: Akademiai Kiado, 359-373.
11. Levy, W. B. and Steward, O. (1979) *Brain Res.* 175, 233-245.
12. Levy, W. B. and Steward, O. (1983) *Neurosci.* 8, 791-797.
13. Lorente de Nó, R. (1938) *J. Neurophysiol.* 1, 207-244.
14. Rosenblatt, F. (1962) *Principles of neurodynamics*. Washington, DC: Spartan Books.
15. Shore, J. E., & Johnson, R. W. (1980) *IEEE Trans. Information Theory*, IT-26, 26-37.
16. Stanton, P. K. and Sejnowski, T. J. (1989) *Nature* 339, 215-218.
17. Young, J. Z., Ed. (1970) *The life of mammals*. Oxford: Clarendon Press.