A.N.

Submitted to the Fifth annual conference of the
Cognitive Science Society.

ANALYZING COOPERATIVE COMPUTATION

by
G. E. Hinton
Computer Science Department
Carnegie-Mellon University

and
T. J. Sejnowski
Biophysics Department
Johns Hopkins University

14 February 1983

Please send all communications to:

Prof. Geoffrey Hinton
Computer Science Dept.
Carnegie-Mellon University
Schenley Park
Pittsburgh PA 15213

ABSTRACT

Making a perceptual interpretation can be viewed as a computational process in which a plausible combination is chosen from among a large set of interdependent hypotheses. In a cooperative computation the hypotheses are implemented by units that interact non-linearly and in parallel via excitatory and inhibitory links (Julesz, 1971; Marr & Poggio, 1976; Sejnowski, 1976). A particular perceptual task is specified by external inputs to some of the units and the whole system must then discover a stable state of activity in which the active units represent the hypotheses that are taken as true. We describe a search procedure based on statistical mechanics that finds a plausible combination of hypotheses quickly, and we show that the hardware units required for its efficient implementation are remarkably similar to neurons. We also show that even though the individual units are non-linear, there is a linear relationship between the synaptic weights and the logs of the probabilities of global states into which the system settles. This linear relationship makes it possible to implement a convergent learning procedure which specifies just how the synaptic weights need to be changed in order to increase or decrease the probabilities of global states.

Introduction

Consider the problem of making a 3-D interpretation of a 2-D line drawing. Each line in the picture, considered in isolation, could depict any one of a large set of 3-D edges. We resolve this local ambiguity by using assumptions about the ways in which edges go together in the 3-D world. These assumptions make some combinations of edges far more plausible than others. There are two roughly separable problems in understanding the use of assumptions in perception. The first is to specify clearly what the assumptions are, and the second is to find a mechanism that can discover interpretations which optimally fit the input data and the assumptions, even when some of the assumptions conflict with one another (Attneave 1982).

In interpreting line drawings, our perceptual systems tend to interpret lines that are parallel in the picture as edges that are parallel in 3-D, and lines that are colinear in the picture as edges that are colinear in 3-D. If three lines meet in a junction that could be the projection of a rectangular corner we tend to interpret them that way. Other things being equal, we tend to make interpretations which are symmetrical. The precise definition of these various rules is complex (Hinton 1979), and will not be pursued further here. Our main concern is the second problem: How can we discover interpretations that optimally fit a large set of plausible assumptions?

Attneave (1982) and others (Hinton 1977) have proposed cooperative mechanisms in which neuron-like hardware units represent particular 3-D edges and the rules are implemented by

excitatory and inhibitory interactions between these units. Each line in the drawing provides input to the whole set of 3-D edges which it could depict, and under the influence of this input the whole system settles into a stable state of activity which represents the interpretation. It is not obvious that such a search process can be made to work. The apparent difficulty of analysing the behaviour of cross-coupled, non-linear systems makes it tempting to believe that the *only* way to make progress is through computer simulation. In this paper we attempt to show that mathematical analysis is possible and illuminating.

Most of the existing proposals for cooperative search mechanisms assume that there are real-valued activity levels which change smoothly during the search (Rosenfeld, Hummel & Zucker, 1976). These levels are often associated with the firing rates of neurons, and they are normally used to represent the value of a physical parameter such as slope in depth, or the current probability that a hypothesis is correct. The method we shall describe uses a very different representation. The units that stand for hypotheses only have two states, true and false. However, the decision rule which determines which state they enter is probabilistic, so they can change their state even if they are receiving constant input. The use of a probabilistic decision rule makes the cooperative search *easier* to analyse than with a deterministic rule because it makes it possible to apply the well-developed methods of statistical mechanics. Instead of being a problem, the non-determinism allows the system to escape from sub-optimal states. We start by describing a deterministic system which we shall then generalize.

Cooperative search with deterministic binary units

Hopfield (1982) postulates a system with a large number of binary units. The units are *symmetrically* connected, with the strength of the connection being the same in both directions. Hopfield has shown that with the right assumptions, the behaviour of a system of interacting binary threshold units can be analysed in an interesting way. Given the current inputs from outside the system, any particular state of the system has an associated "energy" and the whole system behaves in such a way as to minimize its energy. The energy of a state can be interpreted as the extent to which it violates a set of plausible constraints, so in minimizing its energy it is maximizing the extent to which it satisfies the constraints.

The total energy of the system is defined as

$$E = -1/2 \sum_{ij} \omega_{ij} s_i s_j - \sum_i (\eta_i - \theta_i) s_i \qquad (1)$$

where η_i is the external input to the i^{th} unit, w_{ij} is the strength of connection from the j^{th} to the i^{th} unit, s_i is a boolean truth value (0 or 1), and θ_i is a threshold.

A simple algorithm for finding a combination of truth values that is a *local* minimum is to switch each hypothesis into whichever of its two states yields the lower total energy given the current states of the other hypotheses. If hardware units change their states asynchronously, and if transmission times are negligible, then the system always settles into a local energy minimum. Because the connections are symmetrical, the difference between the energy of the whole system with the k^{th} hypothesis true and its energy with the k^{th} hypothesis false can be determined locally by the k^{th} unit (Hopfield, 1982), and is just

$$\Delta E_k = -\sum_i (s_i w_{ki}) - \eta_k + \theta_k \qquad (2)$$

Therefore, the rule for minimizing the energy contributed by a unit is to adopt the true state if its total input exceeds its threshold, which is the familiar rule for binary threshold units (Minsky & Papert, 1968).

Using probabilistic decisions to escape from local minima

The deterministic algorithm suffers from the standard weakness of all hill-climbing (or gradient descent) methods. It gets stuck at *local* minima that are not globally optimal. This is an inevitable consequence of only allowing jumps to states of lower energy. If, however, jumps to higher energy states occasionally occur, it is possible to break out of local minima. An algorithm with this property was introduced by Metropolis *et. al.* (1953) to study average properties of thermodynamic systems (Binder, 1978) and has recently been applied to problems of constraint satisfaction (Kirkpatrick, Gelatt & Vecci, in press). We adopt a form of the Metropolis algorithm that is suitable for parallel computation: If the energy gap between the true and false states of the k^{th} unit is ΔE_k then regardless of the previous state set $s_k = 1$ with probability

$$P_k = \frac{1}{1 + e^{\Delta E_k / T}} \qquad (3)$$

which is graphed in Fig. 1. It can be shown that on average this parallel algorithm chooses global states with a Boltzman distribution, so that the relative probability of two global states is determined solely by their energy difference

$$P_\alpha / P_\beta = e^{-(E_\alpha - E_\beta)/T} \qquad (4)$$

This distribution ensures that there is a bias in favour of *globally* good states. At low temperatures this bias is strong, but the time required to reach equilibrium may be long. At higher temperatures the bias is not so favorable but equilibrium is reached faster.

Reducing the time to reach equilibrium

One technique that can be used to reach a good equilibrium distribution quickly is to start at a high temperature and then to cool down (Kirkpatrick *et. al.*, in press). This type of search by "simulated annealing" initially finds a large-scale global minimum or near minimum but rattles around inside this minimum because of the high temperature. At the next temperature down, a good minimum will be found within the large-scale minimum, and so on. In general, the search can be helped by changing the temperature appropriately, though it is impossible to *guarantee* that a global minimum will be found.

We are investigating two additional techniques which we shall only mention here. The first is to use special units which are active during the search process but are quiescent in the final state. When one of these special units is active it lowers the energy of a state that would have been an energy barrier between two local minima. Energy barriers are what prevent a system from reaching equilibrium rapidly at low temperature, and if they can be temporarily suppressed, equilibrium can be achieved rapidly at a temperature at which the distribution strongly favors the lower minima. The energy barriers cannot be permanently removed, because they correspond to states that violate the constraints, and the energies of these states must be kept high to prevent the system from settling into them. The special units are a way of implementing heuristic knowledge about how to search the space. They have no effect on the energies of final states, and in this respect they are like catalysts.

Another method is to use several different cooperative modules in parallel. The modules need to be coupled loosely enough so that they can explore different parts of the space, but tightly enough so that when one discovers a good region of the search space it can "siphon" the other modules into the same region.

Learning

So far, we have assumed that the interactions between the units implement the correct constraints, and we have focussed on the search problem. However, in a system where the weights represent many plausible assumptions that interact, it is not obvious how to choose the weights so as to produce the desired behavior; we need a learning algorithm. An important consequence of the probabilistic decision rule is that it makes it possible for a cooperative module to internalize the constraints in any domain simply by being told whether the solutions it comes up with are right or wrong. When the module settles to the wrong solution, it modifies the weights so as to raise the energy of that state and thus make it less likely to be found in future. Similarly, good solutions that are not found often enough have their energies lowered when they are found. This simple procedure is effective because of the *linear* relationship between the individual weights and the logs of probabilities of whole states at thermal equilibrium.

To explain the learning procedure, we invent a fictional ideal system which settles into global states with exactly the probabilities required by an evaluation function. We then show that if the same evaluation function is used to tell an actual system whether its current probabilities for particular states are too high or too low, the actual system can modify its weights so that they more closely resemble the weights in the fictional ideal system.

Suppose that under the influence of a constant external input vector, the actual system settles into two different states, S_1 , S_2 with probability ratio P_1/P_2. Suppose that the probability ratio demanded by the evaluation function (and achieved by the ideal system) is P'_1/P'_2 which is higher. The actual system can increase its probability ratio by increasing the energy difference, $E_2 - E_1$. This can be done by adding δ to each weight between a pair of active units in S_1 and subtracting δ from each weight between a pair of active units in S_2. The net change in a weight is then $\delta.h_{ij}$ where h_{ij} takes the following values:

1 if the units s_i and s_j are both on in S_1 and not both on in S_2
-1 if the units s_i and s_j are both on in S_2 and not both on in S_1
0 otherwise.

Provided δ is sufficiently small, each application of this learning procedure is guaranteed to reduce the Euclidean distance, D, between the current set of weights, w_{ij}, and the ideal ones, w'_{ij}.

If the actual value of $\ln(P_1/P_2)$ is r less than the ideal value, and if we assume that the actual and ideal systems have the same external inputs and thresholds, and that the temperature is 1, we have (from Eq. 4)

$$r = \left(E_2' - E_1'\right) - \left(E_2 - E_1\right)$$

$$= \sum_{ij} h_{ij} w_{ij}' - \sum_{ij} h_{ij} w_{ij}$$

Before learning we have

$$D^2_{Before} = \sum_{ij} \left(w_{ij} - w_{ij}'\right)^2$$

After learning we have

$$D^2_{After} = \sum_{ij} \left(w_{ij} + \delta . h_{ij} - w_{ij}'\right)^2$$

$$= D^2_{Before} - \delta \sum_{ij} \left(2 h_{ij} w_{ij}' - 2 h_{ij} w_{ij} - \delta (h_{ij})^2\right)$$

$$= D^2_{Before} - \delta \left(2r - \delta \sum_{ij} (h_{ij})^2\right)$$

So the distance is reduced iff $\delta < 2r/n$

where $n = \sum_{ij} (h_{ij})^2$ = the number of weights that are changed.

The discovery of a simple convergent learning procedure for a non-linear system is important because it allows the synaptic weights that implement the energy function to be determined by feedback from the correctness of the interpretation that the system settles into. Thus the constraints implicit in the task can be programmed into the system simply by telling it how well it is doing.

The proof assumes that the function which evaluates global states can "see" the whole state. If this is not the case, there will be sets of "equivalent" global states which differ only in the activities of units that the evaluation function cannot see. This causes serious complications. If a system with the ideal set of weights would probably settle into state S_1' under the influence of a particular external input vector, the evaluation function must reinforce the actual system for settling into a state which is indiscriminable from S_1' so far as the evaluation function is concerned. Unfortunately, the actual state may be very different from S_1', and so the changes in the weights that occur when the actual state is reinforced may cause them to get further away from the ideal weights. In general, convergence is only guaranteed if the actual weights are already similar enough to the ideal ones so that the states which the actual system adopts are similar to the ones the ideal system would adopt.

A more general learning procedure, which we describe only briefly here, is to minimize the difference between the required probabilities of states P'_α and the actual ones, P_α. The information theoretic measure of the difference between these two probability distributions is

$$G = \sum_\alpha P_\alpha \ln\left(\frac{P_\alpha}{P'_\alpha}\right)$$

For each set of weights, there is a value for G, and so we can reduce G by changing the w_{ij} so as to move in the direction of the gradient of G in weight space. The changes required are given by:

$$\begin{aligned} \partial w_{ij} &= -\nabla_{w_{ij}} G \\ &= -\frac{1}{T} \sum_\alpha \left(P_\alpha \ln\left(\frac{P_\alpha}{P'_\alpha}\right) \left[s_i^\alpha s_j^\alpha - \sum_\beta P_\beta s_i^\beta s_j^\beta \right] \right) \end{aligned} \qquad (5)$$

where s_i^α is the state of the i^{th} unit in the global state α.

The term in square brackets is the covariance, which is the correlation between the i^{th} and the j^{th} units in the α^{th} global state minus their average correlation at thermal equilibrium. The covariance is positive if both units are on but will be small if they are often true togehter. If at least one of the units is off, then the covariance is negative. The covariance is multiplied by a term which is positive iff P_α needs to be increased. Thus whether the weight w_{ij} is raised or lowered depemds on whether the state α needs to be helped or hurt to match P'_α, and on the correlation between the units in this state relative to their average correlation over all states.

This learning rule will minimize G even if the evaluation function cannot discriminate between a whole set of equivalent global states. However, it will only find a *local* minimum of G. Also, unless the parameters in equation (5) are estimated over a very long time scale, the estimate of δw_{ij} will fluctuate about its true value and so G may sometimes increase. Of course, this may be helpful if the problem is to escape from non-optimal local minima in G.

Relation to the brain

There are two different ways to interpret the input-output function that hardware units should have to implement the parallel search (Fig. 1). During a short interval the sigmoid curve describes the probability of a unit being in the true state as a function of the energy gap between the false and true states. For much longer time intervals the curve describes the proportion of time that the unit is in its true state. If we assume that a hypothesis which is true all the time is represented by a neuron firing at its maximum rate, then the curve in Fig. 1 can be interpreted as the firing rate of a neuron as a

function of its average input (Sejnowski, 1977). However, the way in which truth values are represented by action potentials is not the kind of simple encoding in which two different voltage levels stand for the two truth values. Instead, it appears that an action potential only provides a delta-function type of signal that drives integrative processes in the recipient neurons. This amounts to treating a hypothesis as "true" for a whole refractory period after an action potential has been emitted.

The probabilistic nature of electrical responses of single neurons is well known, but has generally been regarded as evidence of imprecision (Winograd & Cowan, 1963). Probability, however, may be a central design principle of the nervous system, particularly in cerebral cortex (Sejnowski, 1981). We suggest that the fluctuations may be deliberately added to neural signals to avoid locking the network into unwanted local optima that occur in the deterministic zero-temperature case, and to search for a global optimum.

The parallel algorithm for cooperative search depends on the computation of energy gaps ΔE_i. In the case of symmetrically connected units the global energy gaps can be computed locally by single units. With asymmetrical connections this is not possible, but if each unit receives many inputs, it is still possible for it to estimate what it would have received if all the connections had been symmetrical, and so the system can behave in a similar way (Hopfield, 1982). The crucial requirement is that the connection strengths needed to implement the task be symmetrical. If this is true, an asymmetrical network can simulate a symmetrical one. Fortunately, constraint satisfaction tasks always require symmetrical connections because when the states of two units violate a constraint, the violation can be removed by switching the state of either unit, and so the "pressure" on either unit due to a violated constraint needs to be the same.

The assumption that there are no time delays in transmission simplifies the analysis, but it is not necessary. Simple simulations show that the introduction of time delays has an effect very similar to raising the temperature.

SUMMARY

By using a particular stochastic function for the units and by allowing the system to reach thermal equilibrium, we achieve a linear relationship between the microscopic weights and the logs of the probabilities of the macroscopic states. This simple relationship makes it easy to modify the weights so as to change the probabilities of global states. The changes encode into the synaptic weights the implicit constraints that determine the correctness of a perceptual interpretation.

Acknowledgements

This work was supported by grants from the System Development Foundation and by earlier grants from the Sloan Foundation to Don Norman and to Jerry Feldman. We thank Scott Fahlman, David Rumelhart, Paul Smolensky, and Francis Crick for helpful discussions.

REFERENCES

Attneave, F. Prägnanz and soap-bubble systems: A theoretical exploration. In J. Beck (Ed.) *Organization and Representation in Perception*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1982.

Binder, K. (Ed.) *The Monte-Carlo Method in Statistical Physics* New York: Springer-Verlag, 1978.

Hinton, G. E. Relaxation and its role in vision. PhD Thesis, University of Edinburgh, 1977; Described in: *Computer Vision*, D. H. Ballard & C. M. Brown (Eds.) Englewood Cliffs, NJ: Prentice-Hall, 1982, pp. 408-430.

Hinton, G. E. Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, 1979, *3*, pp 231-250.

Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA*, 1982, *79* pp 2554-2558.

Julesz, B. *Foundations of Cyclopean Perception* Chicago: University of Chicago Press, 1971.

Kirkpatrick, S. Gelatt, C. D. & Vecci, M. P. Optimization by simulated annealing. *Science* (in press)

Marr, D. & Poggio, T. Cooperative computation of stereo disparity. *Science*, 1976 *194*, p 283-287.

Metropolis, N. Rosenbluth, A. W. Rosenbluth, M. N. Teller, A. H. Teller, E. *Journal of Chemical Physics*, 1953 6, p 1087.

Minsky, M. Pappert, S. *Perceptrons* Cambridge, MA: MIT Press, 1968.

Rosenfeld, A. Hummel, R. A. & Zucker, S. W. Scene labeling by relaxation operations. *IEEE Transactions on Systems, Man, & Cybernetics*. SMC-6, 1976, pp 420-433.

Sejnowski, T. J. On global properties of neuronal interaction. *Biological Cybernetics*, 1976, *22*, pp 85-95.

Sejnowski, T. J. Storing covariance with non-linearly interacting neurons. *Journal of Mathematical Biology*, 1977, *4*, pp 303-321.

Sejnowski, T. J. Skeleton filters in the brain. In G. E. Hinton & J. A. Anderson (Eds.) *Parallel Models of Associative Memory*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1981, pp 189-212.

Winograd, S. Cowan, J. D. *Reliable Computation in the Presence of Noise* Cambridge, MA: MIT Press, 1963.

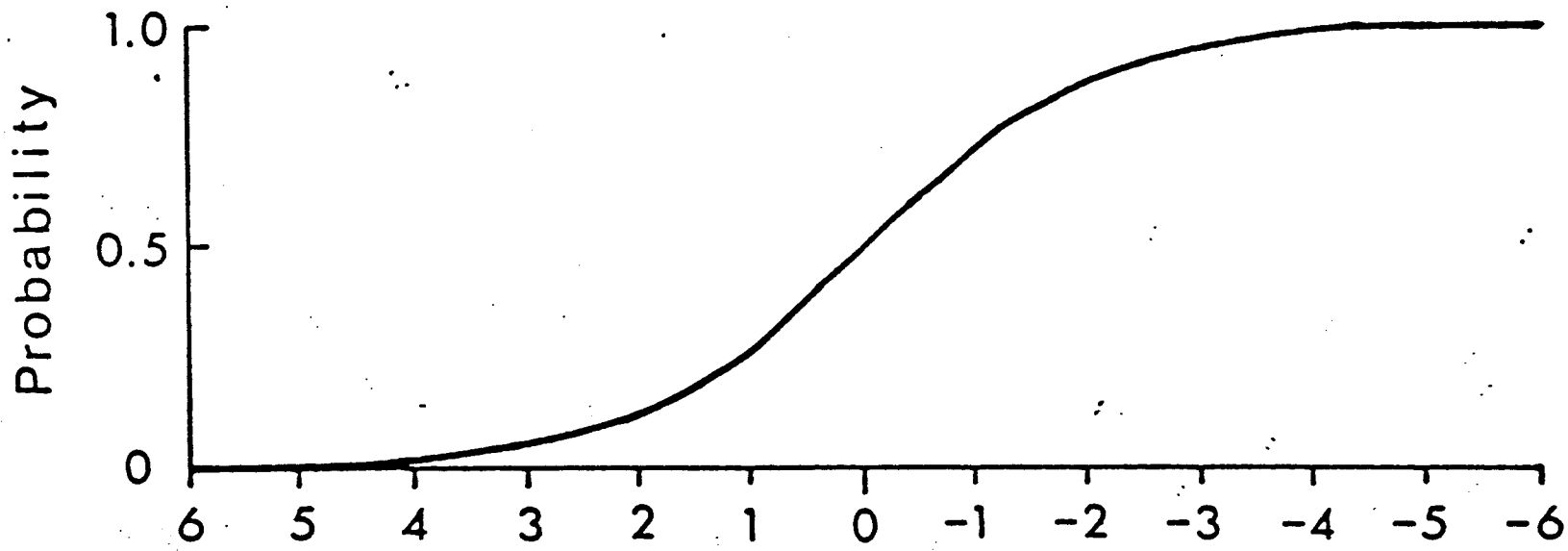

Fig. 1 Probability $p(\Delta E)$ that a unit is in its "true" state as a function of its energy gap ΔE plotted for $T = 1$ (Eq. 3). As the temperature is lowered to zero the sigmoid approaches a step function.