Journal of Statistical Physics, Vol. 57, Nos. 1/2, 1989

Optimal Performance of a Feed-Forward Network at Statistical Discrimination Tasks

W. Bialek,[1] R. Scalettar,[2] and A. Zee[3]

Received March 13, 1989; revision received June 20, 1989

Within the context of a particular statistical discrimination task, we make a quantitative comparison between the performance of a feed-forward neural network and the information-theoretic optimal performance. We also address the ability of such networks to generalize and the effect of network architecture on performance.

KEY WORDS: Neural network; information theory; optimal performance; statistical discrimination; visual system.

1. INTRODUCTION

In our efforts to understand the visual system, we must answer a number of basic questions before launching into speculation and theory. Recently, Bialek and Zee[1,2] emphasized that one such basic question concerns the computational abilities of the visual system. We are apt to feel that our visual systems perform extremely well. But a statement of this sort, which occurs frequently in the vision literature, begs the obvious question of what this alleged high performance is to be compared to. Bialek and Zee suggested that the actual performance of the visual system is to be compared to an optimal performance defined by information theory, in other words, the best performance achievable by a machine which has complete information on how the seen image is generated. They analyzed this optimal performance and showed that it can be characterized by concepts from statistical mechanics and (in the continuum case) field theory. They

[1] Physics Department, University of California, Berkeley, California 94720.
[2] Physics Department, University of California, Davis, California, 95616.
[3] Institute for Theoretical Physics, University of California, Santa Barbara, California 93106.

0022-4715/89/1000-0141$06.00/0

outlined how psychophysical experiments can be done on human subjects and the measured performance compared to the optimal performance.

In the last few years, neural networks have attracted a great deal of attention (see, e.g., ref. 3). In particular, with backpropagation, feed-forward networks can learn to perform a variety of tasks,[4] including perceptual tasks.[5] For instance, Scalettar and Zee[6] have showed how a network can learn to decide whether a given object is to the left or to the right of another object.

In this paper we make contact between the two studies by comparing the performance of networks with the optimal performance determined by information theory. The task assigned is statistical discrimination, of the type discussed in ref. 1. The network is presented with a configuration of a one-dimensional chain of Ising spins. The configuration belongs to the equilibrium ensemble at temperature β^{-1} of either a ferromagnet or an anti-ferromagnet. The network is to decide whether the configuration is ferromagnetic or antiferromagnetic. We also study the ability of the net to discriminate between ferromagnetic configurations at two different temperatures and to determine whether next-nearest-neighbor interactions were present in generating a particular spin configuration. For example, we can ask the net to discriminate between configurations generated when the inverse temperature $\beta = 1/T$ is zero, that is, totally random configurations, and configurations generated when β is small but nonzero, that is, configurations with incipient order. We chose this problem partly because humans appear to do quite well at detecting patterns and order and partly because it provides a simple model where optimal performance is analytically calculable and hence the performance of a neural net can be measured meaningfully. This problem also provides an example of learning with noise. In standard examples of learning, the net is told what the unique correct answer is after each trial. Here the net is confused by fluctuations: thus, a configuration generated at $\beta = 0$ could happen to have a high degree of order by chance.

The use of neural networks is potentially promising as a technique to study correlations in complicated statistical mechanical systems. In a sense the net learns to do a primitive sort of data analysis. After looking at the input configurations ("the data") it can tell us about the Hamiltonian that generated these configurations (by telling us whether next-nearest-neighbor interactions were present, for example.)

With this example we can also study how well the net generalizes from one task to another. Thus, we can test a net trained to discriminate between ferromagnetic order and antiferromagnetic order to see how well it will discriminate between ferromagetic orders at different temperatures.

An important issue is how the performance depends on the architecture of the net. Looking at the performance of a net, can we learn something about its architecture? It is quite a leap from simple nets to the human visual system, but we envisage the present work as something of a prototype of how one might one day explore the visual system quantitatively.[1,2] By measuring the performance of the human visual system and comparing the actual performance to the optimal performance for a variety of tasks, we may be able to learn something about its architecture.

2. OPTIMAL PERFORMANCE

An Ising ferromagnet in one dimension is described by the Hamiltonian

$$H = -\sum_i S_i S_{i+1} \tag{1}$$

S_i is a classical variable at site i which can take the values ± 1. If we choose periodic boundary conditions, the spin configuration $\{S_1, S_2, \ldots, S_N = S_1\}$ appears with the probability

$$P(\{S\} \mid \beta) = Z^{-1}(\beta) \exp\left(\beta \sum_i S_i S_{i+1}\right) \tag{2}$$

with the partition function

$$Z(\beta) = \sum_S \exp\left(\beta \sum_i S_i S_{i+1}\right) = 2^N[(\text{ch } \beta)^N + (\text{sh } \beta)^N] \tag{3}$$

The energy is minimized by configurations of parallel spins. Thus, at low temperatures the probability is strongly peaked in favor of configurations that are mostly all $S_i = +1$ or all $S_i = -1$. Indeed, in dimensions greater than one, below a certain critical temperature T_c the system locks into a phase where the magnetization m, the difference between the number of $+1$ and -1 spins, is nonzero.

Suppose that for a given configuration we are asked to decide whether the temperature is β_1^{-1} or β_2^{-1}. Define the discriminant

$$\lambda(\{S\}; \beta_1 \text{ vs } \beta_2) = \log \frac{P(\{S\} \mid \beta_1)}{P(\{S\} \mid \beta_2)} \tag{4}$$

According to signal detection theory,[7] the optimal *unbiased* discrimination is reached by maximum likelihood: we say the inverse temperature is β_1 if $\lambda > 0$ and β_2 if $\lambda < 0$, precisely what any sensible person would do. The probability of correctly identifying the ensemble at temperature β_1^{-1} is then

$$P_c(\text{if } \beta_1) = \sum_S P(\{S\} \mid \beta_1)\, \Theta[\lambda(\{S\}; \beta_1 \text{ vs } \beta_2)] \tag{5}$$

where Θ denotes the Heaviside step function. More generally, one can define the probability distribution of the discriminant

$$\mathscr{P}(\lambda; \beta_1 \text{ vs } \beta_2; \beta_1) = \sum_S P(\{S\} \mid \beta_1)\, \delta(\lambda - \lambda(\{S\}; \beta_1 \text{ vs } \beta_2)) \tag{6}$$

from which

$$P_c(\text{if } \beta_1) = \int_0^\infty d\lambda\, \mathscr{P}(\lambda; \beta_1 \text{ vs } \beta_2; \beta_1)$$

can be computed; and similarly for $\mathscr{P}(\lambda; \beta_1 \text{ vs } \beta_2; \beta_2)$ and

$$P_c(\text{if } \beta_2) = \int_{-\infty}^0 d\lambda\, \mathscr{P}(\lambda; \beta_1 \text{ vs } \beta_2; \beta_2)$$

If in the presentation the ensemble at β_a is chosen with probability $P(\beta_a)$ ($a = 1, 2$), then the probability of correct identification is

$$\begin{aligned} P_c &= P_c(\text{if } \beta_1)\, P(\beta_1) + P_c(\text{if } \beta_2)\, P(\beta_2) \\ &= \sum_S P(\{S\} \mid \beta_1)\, P(\beta_1)\, \Theta(\lambda(\{S\}; \beta_1 \text{ vs } \beta_2)) \\ &\quad + P(\{S\} \mid \beta_2)\, P(\beta_2)\, \Theta(-\lambda(\{S\}; \beta_1 \text{ vs } \beta_2)) \end{aligned}$$

In our simulation below, we always set $P(\beta_1) = P(\beta_2) = 1/2$.

Discrimination between ferromagnet and antiferromagnet is a special case of discrimination between two temperatures with $\beta_1 = \beta$ and $\beta_2 = -\beta$.

In general, with $P(\{S\} \mid \beta) = Z^{-1}(\beta)\, e^{-\beta H(S)}$ we have

$$\begin{aligned} \lambda(\{S\}; \beta_1 \text{ vs } \beta_2) &= -(\beta_1 - \beta_2)\, H(S) - \log \frac{Z(\beta_1)}{Z(\beta_2)} \\ &= \beta_1(F(\beta_1) - H(S)) - (\beta_1 \to \beta_2) \end{aligned} \tag{7}$$

where $F(\beta)$ is the free energy.

The probability distribution $\mathscr{P}(\lambda)$ is readily calculated to be

$$\begin{aligned} \mathscr{P}(\lambda) &= \int_{-\infty}^{\infty} \frac{d\alpha}{2\pi} \left\{ \exp\left[i\alpha \left(\lambda + \log \frac{Z(\beta_1)}{Z(\beta_2)} \right) \right] \right\} \sum_S \exp[+i\alpha(\beta_1 - \beta_2)\, H(S)] \\ &\quad \times P(\{S\} \mid \beta_1) \\ &= \int_{-\infty}^{\infty} \frac{d\alpha}{2\pi} \left\{ \exp\left[i\alpha \left(\lambda + \log \frac{Z(\beta_1)}{Z(\beta_2)} \right) \right] \right\} \frac{Z(\beta_1 - i\alpha(\beta_1 - \beta_2))}{Z(\beta_1)} \end{aligned} \tag{8}$$

Note that the partition function at complex temperature enters.

Normally, as in the case discussed in refs. 1 and 2, the evaluation of P_c and $\mathscr{P}(\lambda)$ is intractable and essentially amounts to solving a statistical mechanics system or a field theory. Here we chose a particularly simple system so that the relevant quantities can all be evaluated analytically. The discriminant is

$$\lambda(\{S\}; \beta_1 \text{ vs } \beta_2) = (\beta_1 - \beta_2) \sum S_i S_{i+1} - \log \frac{Z(\beta_1)}{Z(\beta_2)}$$

For discrimination between ferromagnet and antiferromagnet and for N even, it simplifies to an expression

$$\lambda(\{S\}; \beta \text{ vs } -\beta) = 2\beta \sum S_i S_{i+1} = 2\beta H(S) \tag{9}$$

proportional to the Hamiltonian. $H(S)$ is just the number of pairs of parallel spins minus the number of pairs of antiparallel spins. Thus, the optimal unbiased strategy mentioned earlier is precisely what any sensible person would adopt: if there are lots of parallel spins, call it a ferromagnet; if not, call it an antiferromagnet. [The case N odd is different because the antiferromagnetic chain is then "frustrated." The distinction between N even and odd vanishes for large N when $(\operatorname{th} \beta)^N \ll 1$.]

For N even,

$$Z(\beta) = 2e^{-\beta N} \sum_{\substack{k=0 \\ \text{even}}}^{N} \binom{N}{k} e^{2\beta k}$$

and the integration over α may be performed term by term to give

$$\mathscr{P}(\lambda) = \frac{2e^{-\beta_1 N}}{Z(\beta_1)} \sum_{\substack{k=0 \\ \text{even}}}^{N} \binom{N}{k} e^{2k\beta_1} \delta\left(\lambda + \log \frac{Z(\beta_1)}{Z(\beta_2)} + (\beta_1 - \beta_2)(N - 2k)\right) \tag{10}$$

with

$$P_c = \int_0^\infty d\lambda\, \mathscr{P}(\lambda) = \frac{2e^{-\beta_1 N}}{Z(\beta_1)} \sum_{\substack{k=0 \\ \text{even}}}^{N}{}' \binom{N}{k} e^{2k\beta_1} \tag{11}$$

where the prime on the summation sign restricts the sum to

$$(2k - N)(\beta_1 - \beta_2) > \log \frac{Z(\beta_1)}{Z(\beta_2)}$$

The same result can also be obtained by carefully enumerating all possible configurations and weighting them by their probabilities of occurrence. These formulas simplify for $\beta_1 = -\beta_2 = \beta$, in which case we write $\lambda = 2\beta N x$ with the variable x ranging between -1 and $+1$. Using Stirling's formula for large N, we find

$$\mathscr{P}(x) = 2[(1+e^{2\beta})^N + (1-e^{2\beta})^N]\, e^{f(x)} \tag{12}$$

with

$$f(x) = -N \log \frac{N}{2} - \frac{N}{2}\,[(1+x)\log(1+x) + (1-x)\log(1+x) - 2\beta x] \tag{13}$$

The maximum of $f(x)$ occurs at $x = \operatorname{th}\beta$ at which $d^2f/dx^2 = -N \operatorname{ch}^2 \beta$.

3. NEURAL NET PERFORMANCE

We now proceed to compare these expressions for the optimal performance with the performance attained by a feed-forward neural network. Such a network consists of a set of "input neurons" which are given some values I_i from an external source. In our case these values are the configuration of an Ising spin system equilibrated to inverse temperature β, i.e.,

$$I_i = S_i \tag{14}$$

This input information is processed through intermediate layers of "hidden" neurons. We will consider only networks with one hidden layer. The values of the hidden neurons are obtained by

$$H_i = \tanh\left(\sum_j T_{ij} I_j - \theta_i\right) \tag{15}$$

Finally, the "output" neurons have activities given by

$$O_i = \tanh\left(\sum_j W_{ij} H_j - \phi_i\right) \tag{16}$$

In our case we have a single output neuron whose target value is chosen to be $T = -1$ if the spin configuration came from the ensemble at β_1, and $T = +1$ otherwise.

The simulation proceeds as follows: We first equilibrate two distinct Ising spin chains (with periodic boundary conditions) to temperatures β_1 and β_2 and set up the weights in our neural net at random. (We call $T_{i,j}$, $W_{i,j}$, θ_i, and ϕ_i collectively the weights.) We next select, with probability 1/2, a spin configuration from one of the two ensembles. We also sweep through the lattice a few times doing Monte Carlo updates to ensure that the next sampling of that ensemble yields an independent configuration. We then present the Ising spins as input to the network and calculate the values of the hidden neurons and the output neuron. Naturally, with the original random weights the output of the net will bear no relationship to the target. However, the net can "learn" the task of distinguishing the two ensembles if the weights and thresholds are updated by gradient descent to minimize the "energy"

$$E = \tfrac{1}{2}(T - O)^2 \tag{17}$$

That is, denoting the weights generically by w, we change the weights by an amount

$$\delta w = -\eta\, \partial E/\partial w \tag{18}$$

This procedure has been shown to give rise to weights which make the output O of the net match the target T in a variety of problems.[4] Recently, log likelihood criteria[9] have been discussed as replacements for the squared error, although we have restricted ourselves to the form (17). We are interested in how well the net can learn our particular task. There are two possible sources of error: As stated earlier, there is an unavoidable error rate associated with the fact that any spin configuration could have come from either ensemble. This feature of our problem is not present in the usual tasks where a well-defined "answer" is usually available. There is also an error rate associated with the failure of the net itself to learn the task perfectly.

We first studied a network without a hidden layer and with sparse connections: each hidden unit is connected only to three adjacent input neurons. Further, we required that the weights be translationally invariant. Thus, there are only six variables in this simple problem, four weights and two thresholds. We studied lattices which ranged in size from 8 to 32 spins. Learning rates and optimal performances were insensitive to lattice size.

We show in Figure 1 how the net learns to discriminate between ferromagnetic ordering at $\beta_1 = 0$ and at $\beta_2 = 1/2$. The performance P_c is shown as a function of the number of configurations presented. We have defined a correct response to be one in which the output is within 0.2 of its target value. It is found that the results are not dependent on this choice. Indeed, the neural net's output after training was strongly peaked at ± 1, i.e., if the net gave the wrong output, it was typically completely wrong, not some intermediate value near 0. Since the initial weights were random, P_c is initially 1/2. The net then subsequently "learns" and increases P_c. We expect that the net has an easier time for β_2 large than

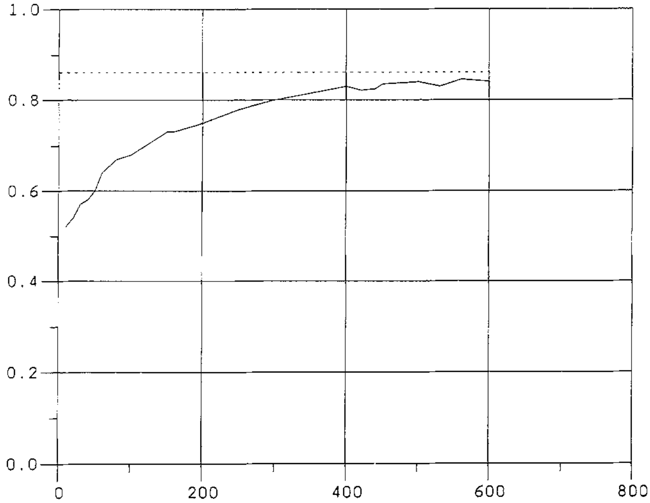

Fig. 1. The performance P_c of a net whose task is discriminating configurations generated from ensembles at inverse temperature $\beta_1 = 0$ and $\beta_2 = 1/2$ shown as a function of the number of input configurations presented. The dashed horizontal line is the optimal performance.

for $\beta_2 \approx 0$. This is indeed borne out in the simulation. We say that the net has learned if the fraction of correct choice P_c has flattened out more or less to some asymptotic value. In our simulations, we allow the net to "graduate" if ΔP_c from one trial to the next is less than 0.01. (Note that graduation is not based on the net having mastered the task, but by its inability to learn further: be all it can be, as the slogan goes.) The result for this asymptotic value is plotted in Fig. 2. We note that the net's performance is remarkably close to the optimal performance. In Fig. 3, we plot the efficiency, defined as $P_c(\text{net})/P_c(\text{optimal})$, as a function of β_2. We see that this increases with increasing β_2, as one might expect. Clearly, if we allow the net to graduate after some fixed N_G of trials with N_G independent of β_2, the corresponding plots of Figs. 2 and 3 will be slightly different. The net underperforms by a wider margin at small β_2.

We can now address the issue of generalization by asking whether a net trained to distinguish ferromagnetic and antiferromagnetic ensembles will also distinguish high and low temperature. We train a net at the ferromagnetic–antiferromagnetic task. Its performance, as shown in Fig. 4, is nearly optimal. We then take the net and the weights it developed and compare its output with the target for the two-temperature task *without* any further evolution of its weights at the new task. These results are shown in Fig. 5. We see the performance is nearly as good as that given in Fig. 2. This is a real test of generalization, since the configurations realized

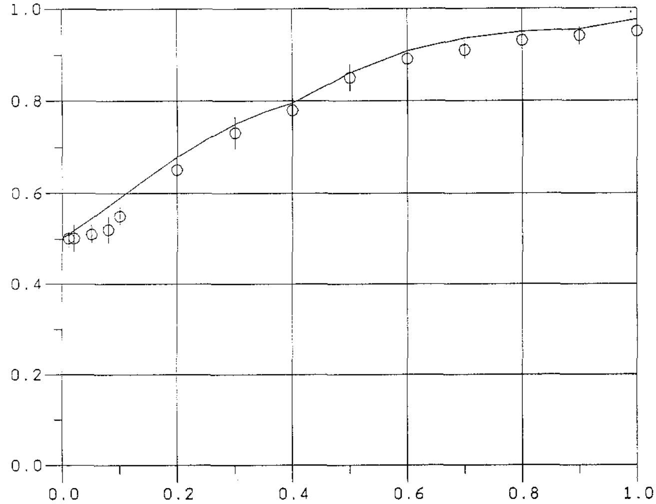

Fig. 2. The asymptotic value for P_c for distinguishing $\beta_1 = 0$ from β_2 shown as a function of β_2. The smooth curve is the optimal performance and the circles are the performance of the net.

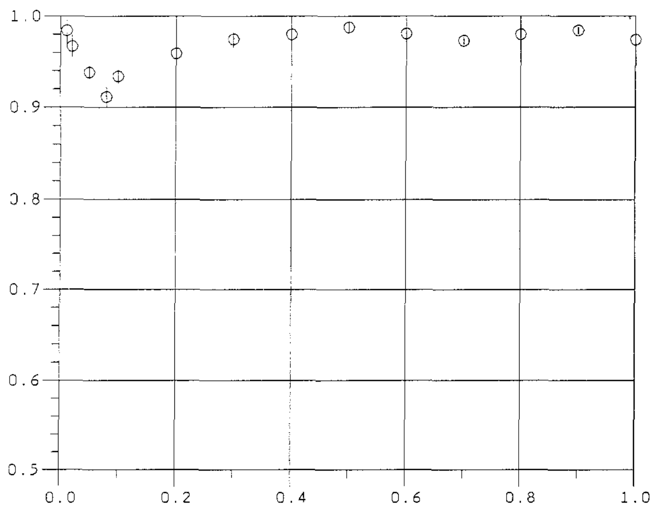

Fig. 3. The ratio of the net's performance to the optimal performance (data from Fig. 2), as a function of β_2.

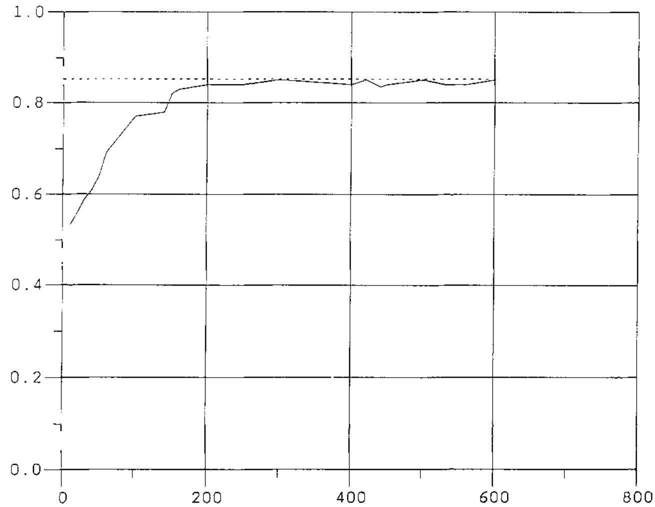

Fig. 4. The performance of a net trained at the task of distinguishing ferromagnetic and antiferromagnetic configurations. $\beta = 1/4$. The dashed horizontal line is the optimal performance.

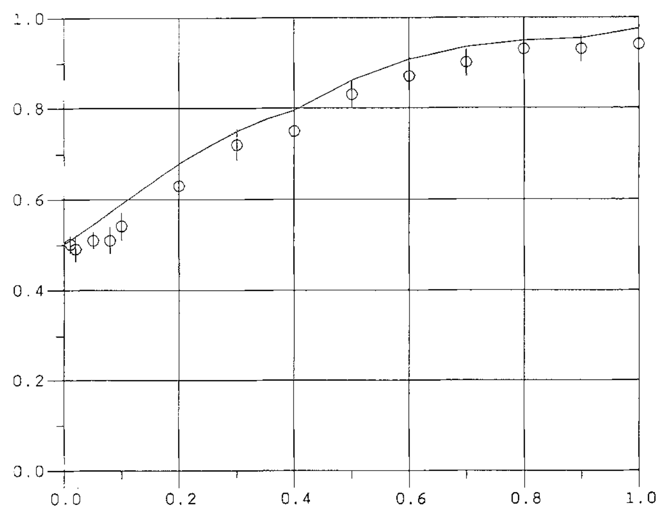

Fig. 5. Same as for fig. 2, except that the net was one trained on the ferromagnetic–antiferromagnetic task and then tested on the two-temperature task without any further evolution of its weights.

for an antiferromagnet are concentrated about the "staggered" ones where ± 1 alternate. While these do share the characteristic $m \approx 0$ of high temperature, they clearly are a very restricted subset of the set that the high-T ensemble samples. This remarkable performance can, however, be readily understood if the net is using a means of distinguishing the ensembles, like measuring the mean magnetization m, which works for either task. Finally, Fig. 6 shows asymptotic values for P_c for the ferromagnetic–antiferromagnetic task as a function of β.

As a further test of generalization we train a net to discriminate between ferromagnetic ordering at β_1 and β_2. We then ask the net to discriminate between ferromagnetic order at β'_1 and β'_2. We choose $\beta_1 = 0$ and $\beta_2 = 0.5$ and then test generalization on $\beta_1 = \delta\beta$ and $\beta_2 = 0.5 + \delta\beta$. Figure 7 shows P/P_{opt} as a function of $\delta\beta$.

A further interesting question in the study of neural networks is to determine the way in which the problem is "represented" in the net, i.e., to look at the weights the net develops. In particular, one might ask whether the net will develop weights which can be identified as counting the number of parallel and antiparallel spins. These weights would clearly look like the weights developed in solving the XOR problem. We find that the weights do indeed correspond to measuring the number of parallel near-neighbor spins. A net with such sparse connections was actually somewhat better at learning the task at hand than a fully connected net. This should perhaps

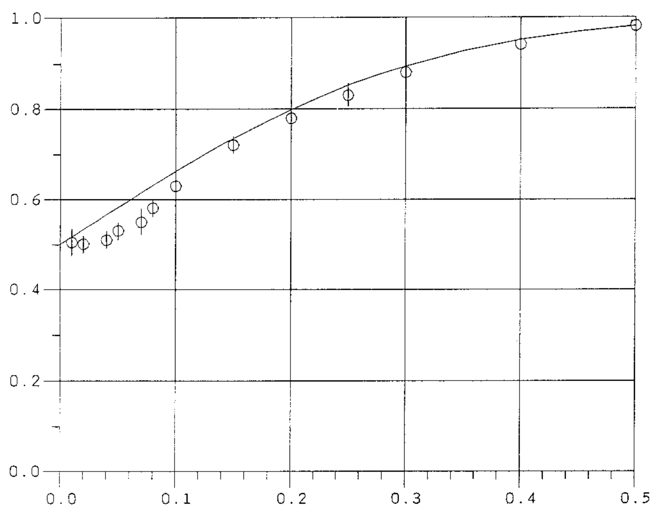

Fig. 6. The asymptotic value for P_c for distinguishing ferromagnetic and antiferromagnetic configurations shown as a function of β. The smooth curve is the optimal performance and the circles are the performance of the net.

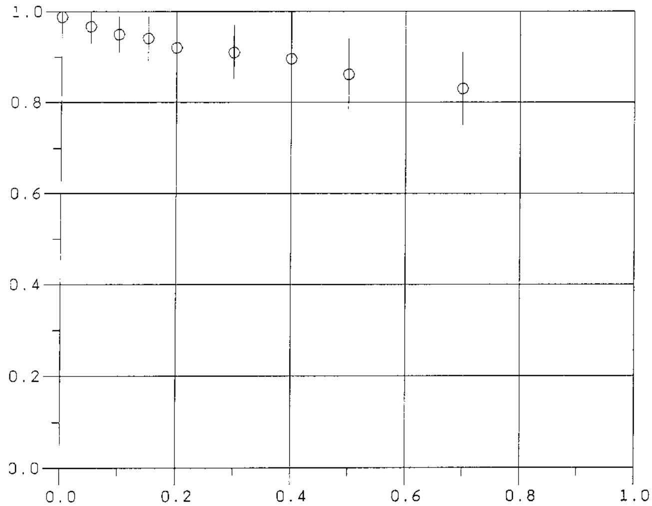

Fig. 7. Performance of a net trained at distinguishing $\beta_1=0$ and $\beta_2=1/2$ at the task of distinguishing $\beta_1=\delta\beta$ and $\beta_2=1/2+\delta\beta$, as a function of $\delta\beta$, without any further evolution of its weights.

be expected, since the relevant discriminant depends most crucially on S_iS_{i+1} and one is forcing the net to look at this function and since the fully connected net contains more weights that can be varied.

This naturally leads us to ask another task of our net. We modify the Ising Hamiltonian to include next-nearest interactions

$$H=-\sum_i S_iS_{i+1}+K\sum_i S_iS_{i+2} \tag{19}$$

We can ask the network to distinguish configurations at the same temperature, but with different values of K. In Fig. 8a we show the results of discriminating between $K=0$ and K positive, and in Fig. 8b between $K=0$ and K negative. In the former case, the interactions compete. The open circles represent results for a net with translationally invariant local connections, and the crosses a fully connected net with independent weights. (This generalized Ising model can also be solved explicitly, as all short-ranged one-dimensional Ising models can be. However, we have not bothered to work out the analytic expressions for the optimal performance, but merely contended ourselves with plotting the optimal performance numerically.) We see that the net can master this task and that nonlocal connections help as the range of the interactions increases with this additional next-near-neighbor term.

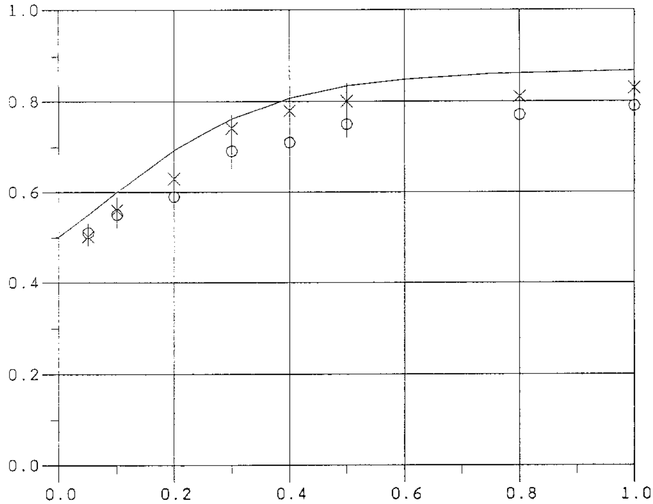

Fig. 8a. Performance of a net distinguishing the presence of a $K>0$ next-near-neighbor interaction, as a function of K. (See text.)

In a sense, the tasks we have chosen are too simple and the net learns these tasks quite readily. Essentially the net only needs to measure local correlations. It is perhaps more interesting to find some more difficult example, suppose we consider

$$H = -\sum S_i S_{i+1} + J \sum_i S_i S_{i+5}$$

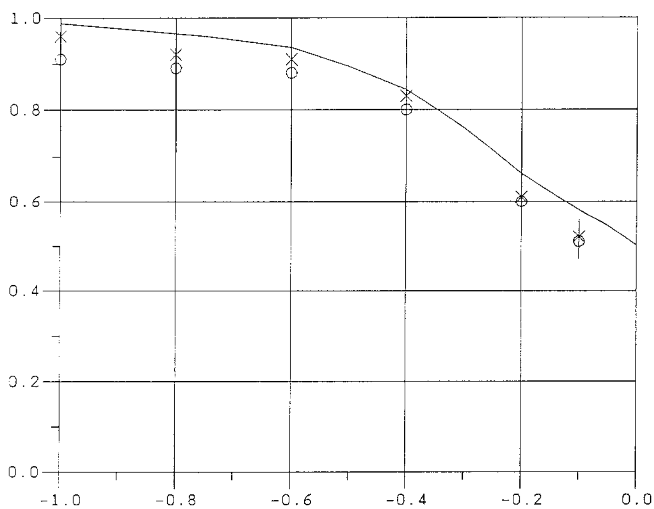

Fig. 8b. Same as Fig. 8a, except $K<0$.

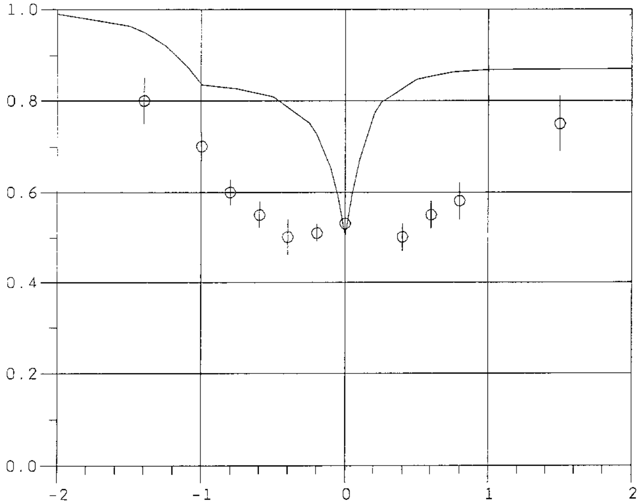

Fig. 9. Performance of a net (circles) at distinguishing the presence of a $J \sum S_i S_{i+5}$ interaction, as a function of J. (See text.) The curve is the optimal performance.

The J interaction has a range (arbitrarily chosen to be) outside the view of each individual hidden unit. We expect that when $J > 1$ the net will have a great deal of difficulty mastering this task. This is indeed shown to be the case in Fig. 9. To pick out the longer range correlation, we add a second hidden layer as shown in Fig. 10. We expect the performance to improve,

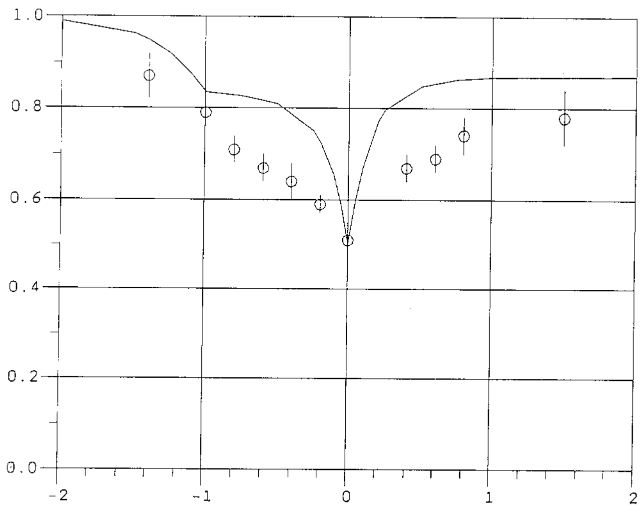

Fig. 10. Same as for Fig. 9, except for a net with a hidden layer.

as it indeed does. We expect in general that there will be a relation between the architectures of nets that perform well and the Hamiltonian that is generating the ensembles.

4. CONCLUSIONS

Here we summarize some of our results. First, we have studied a somewhat novel neural network task. In the literature, networks are typically given problems to which there is a definite answer on any given trial. Here, in contrast, a configuration may look entirely ferromagnetic, but could in fact be an unlikely fluctuation of an antiferromagnet. Thus, already in the learning phase, the network is subjected to confusing and possibly contradictory signals. Also, the network is clearly required to generalize, since for N large the network could have seen only a tiny fraction of the 2^N possible spin configurations. Clearly it can do this since the input configurations are generated according to some well-defined rules.

Second, we have shown that neural networks indeed perform at optimal in at least some cases. Suppose the psychophysics experiments outlined in the program proposed in refs. 1 and 2 were actually done. If similar studies were done to determine the optimal performance of various models of the human visual system, we could perhaps eliminate or advance some models according to whether their performance matched up reasonably with that of real biological systems. As remarked in the introduction, the hope is that by measuring the performance of a net or the human visual system as a black box we can learn something about its architecture. In our simple example, this can clearly be done. At a rather trivial level, we see readily that the net would fail rather drastically at these discrimination tasks were a hidden layer not present. In the problem, two configurations generated one from the other by flipping all the spins are regarded as the same, while the net (without any hidden layer) would clearly treat these configurations as different. Thus, the net would not be able to learn. As another example, by measuring the performance of the net at the task described in Fig. 10, we could have learned, in principle, that there must be more than one hidden layer.

Finally, while the problem we chose is particularly simple, the hope is that neural nets might eventually be trained to determine the phase of more complicated systems. It would be interesting, for example, if a net could determine whether a frustrated spin system were in a glassy phase, rather than using some more traditional measurement of the specific heat or susceptibility. Further, in cases where the order parameter might not be well established, the weights the net develops might help in determining the relevant correlation functions to study. This may be useful in Monte Carlo simulations where the symmetry of the order parameter is not immediately obvious.

With some exaggeration, we may regard the mastery of the type of tasks discussed here as a crude prototype of physics itself. By detailed examination of the world (the input configurations) can careful analysis of the correlations see, physicists try to determine the Hamiltonian responsible. To dramatize an obvious point, we have plotted in Fig. 13 the performance at distinguishing $\beta_1 = 0$ from $\beta_2 = \beta$ as a function of β for a net seeing ensembles generated with

$$H = -J \sum S_i S_{i+1} - K \sum S_i S_{i+2}$$

with $J = 1$ and $K = 1/2$. The optimal performances for different K are shown. We see that $K = 1/2$ does best fit the net's actual performance. We could in this case have roughly determined H from the neural net performance.

In general, if the actual performance falls far below the optimal, we can conclude that the Hamiltonian is incorrect or the architecture is inappropriate. We could then try some other Hamiltonian or some other architecture. On the other hand, if the actual performance exceeds the optimal performance, then clearly we have again guessed the wrong Hamiltonian.

In conclusion, we have illustrated a number of points with a particularly simple problem. We expect that the possibilities illustrated here will also hold in more complicated situations.

ACKNOWLEDGMENT

This work was supported in part by NSF grant PHY82-17853, supplemented by funds from NASA.

REFERENCES

1. W. Bialek and A. Zee, *Phys. Rev. Lett.* **58**:741 (1987).
2. W. Bialek and A. Zee, *Phys. Rev. Lett.* **61**:1512 (1988).
3. D. Waltz and J. A. Feldman, eds., *Connectionist Models and Their Implications* (Ablex, Norwood, New Jersey, 1988), and references therein.
4. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D. E. Rumelhart and J. L. McClelland, eds. (Bradford, Cambridge, 1986).
5. T. J. Sejnowski and C. R. Rosenberg, *Complex Systems* **1**:145–168 (1987).
6. R. T. Scalettar and A. Zee, *Biol. Cybernet.* **58**:193 (1988).
7. D. M. Green and J. A. Siwets, *Signal Detection Theory in Psychophysics* (Krieger, New York, 1966).
8. N. Qian and T. Sejnowski, *J. Mol. Biol.* **202**, in press.
9. E. Baum and F. Wilczek, in *Neural Information Processing Systems*, D. Z. Anderson, ed. (AIP Press, New York, 1988), pp. 52–61.