Introduction to Neural Networks
U. Minn. Psy 5038
Spring, 1999

Energy, Probability & the Boltzmann machine

Statistical physics, computation, and statistical inference

At the beginning of this course, we noted that John von Neumann, one of the principal minds behind the architecture of the modern digital computer, wrote that brain theory and theories of computation would eventually come to more resemble statistical mechanics or thermodynamics than formal logic. We have already seen in the Hopfield net, the development of the analogy between statistical physics systems and neural networks. In the past 15 years, the relationship between computation and statistical physics has received considerable study (cf. Hertz et al., 1991). We are going to look at a neural network model that exploits the relationship between thermodynamics and computation both to find global minima and to modify weights. Further, we will see how relating energy to probability leads naturally to statistical inference theory. Much of the current research in neural network theory is done in the context of statistical inference (Bishop, 1995; Ridley, 1996).

Probability and energy

For a more detailed review of basic probability, see **ProbabilitySupplement.nb** in lecture 9.

■ Conditional probabilities

Two events, a and b, are said to be independent if the probability of their occurring together (i.e. their "joint probability") is equal to the product of their probabilities:

$$p(a\&b) = p(a)p(b)$$

By definition, the conditional probability of a given b (or "the conditional probability of a on b") is:

$$p(a|b) = \frac{p(a\&b)}{p(b)}$$

If a and b are independent, what is the conditional probability of a given b? The intuition is that knowledge of b provides no help with making statistical decisions about a.

■ Probabilities of hypotheses contingent on data, Bayes' rule

Conditional probabilities are central to modeling statistical decisions about hypotheses that depend on data. For example, the posterior probability of H, given data d is:

$$p(H|d)$$

It is called "posterior", because it is the probability *after* one knows the data. It is more constrained than the *prior* probability, p(H). If one has a formula for the posterior probability, then it is possible to devise optimal strategies to achieve well-defined goals of inference. For example, imagine the data are fixed. A device that picks the hypothesis, H, that makes the posterior probability biggest is optimal in the sense that it makes the fewest errors on average. The well-defined goal is to achieve the least average error rate. This kind of decision maker is called a *maximum a posteriori* (or MAP) estimator.

Often it is easier to find a formula for p(d|H), then the other way around. If the prior is known, one can still do MAP estimation because of Bayes' rule:

$$p(H|d) = \frac{p(d|H)p(H)}{p(d)}$$

There are two assumptions that can simplify MAP estimation. First, p(d) is often assumed fixed (we have the data and it isn't changing while we try to decide on the best hypothesis to explain the data). Further, we often don't have reason to prefer one hypothesis H=H', over any other, say H=H''. So p(H) is constant. If these two conditions hold, then MAP estimation is equivalent to finding the H that makes p(d|H) biggest. This latter strategy is called *maximum likelihood* estimation.

■ Putting probability and energy together: The Gibbs distribution

Let $E(V_1,..,V_n)$ be an energy function for a network, as in a Hopfield model. Then we can write a probability function, called the Gibb's distribution, for the network as:

$$p(V_1,\ldots,V_n) = \kappa \exp\left(\frac{-E\left(V_1,\ldots,V_n\right)}{T}\right)$$

T is a parameter that controls the "peakedness" of the probability distribution (e.g. if the energy is a quadratic function of the V's, the Gibbs distribution becomes a Gaussian, and if each unit has the same variance, and they are independent, we have $T = 2\sigma^2$). From the physicist's point of view, T is temperature--e.g. the hotter the matter, the more variance there is in the particle velocities. For a magnetic material, an increase in thermal fluctuations makes it more likely for little atomic magnets to flip out of their otherwise regular arrangement. κ is a normalization constant determined by the constraint that the total probability over all possible states must equal one.

Question: So if the Hopfield net seeks states that minimize energy, what kind of statistical estimator is the Hopfield net?

Now imagine that some of the values of our units are given. In other words a subset of the units are declared to be the input, and are "clamped" at specific levels. Call these V_i^s. These values are fixed, and the others vary. The conditional probability is written as:

$$p(V_1,\ldots,V_m|V_1^s,\ldots,V_k^s) = \kappa' \exp\left(\frac{-E\left(V_1,\ldots,V_m;V_1^s,\ldots,V_k^s\right.}{T}\right.$$

So from a statistician's point of view, a network that is evolving to minimize an energy function, is doing a particular form of Bayesian estimation.

■ **Sidenote on terminology**

We've already noted that energy is equivalent to the Lyapunov function of dynamical systems. Other analogous terms you may run into are: Hamiltonian (in statistical mechanics), and cost or objective functions in optimization theory.

Boltzmann machine

Introduction

We've seen how local minima in an energy function can be useful stable points for storing memories. However, for constraint satisfaction problems such as the stereo example, local minima can be a real annoyance--one would like to find the *global* minimum, because this corresponds to the state-vector that should best satisfy the constraints inherent in the weights and the data input.

One of the early contributions to improving the odds of finding the global minimum was an algorithm called the Boltzmann machine by Ackley et al. (1985). Like the Hopfield network, the Boltzmann machine is a recurrent network with units connected to each other with symmetric weights. The units are binary threshold logic units. Unlike the 1982 Hopfield net, the Boltzmann machine uses a stochastic update rule that allows occasional increases in energy. First we take a look at the update rule, and then the learning rule. The learning rule will lead us to a different view of supervised learning, in which the goal is to model the state of the environment.

Finding the global minimum: Theory

■ TLUs and energy again

The starting point is the discrete Hopfield net (1982), with a view towards solving constraint satisfaction, rather than memory problems. Energy is then a measure of the extent to which a possible combination of hypotheses violates the constraints of the problem. We've seen this with the stereo problem. Let V1 and V2 be the outputs of neural elements 1 and 2. These two outputs can be thought to represent *local* "hypotheses". A positive connection weight (e.g. T_{12}) means that local hypotheses, V1 and V2 support one another. A negative weight would mean that the two hypotheses should not both be accepted.

Some of the inputs can be clamped, and the rest allowed to evolve. In this way the network finds the conditional local minimum (or equivalently, the maximum of the corresponding conditional probability).

$$V_i = \begin{cases} V_i^c & \text{if } i \text{ is a clamped unit} \\ 1 & \text{if } \Sigma\ T_{ij}\ V_j > U_i \\ 0 & \text{if } \Sigma\ T_{ij}\ V_j < U_i \end{cases} \tag{1}$$

$$E = -\sum_{i<j} T_{ij} V_i V_j + \sum_i U_i V_i$$

(Notation: the sum for i<j, is the same as the 1/2 the sum for i not equal to j, because the weight matrix is assumed to be symmetric, and the diagonals are not included. So this may look different than the Hopfield energy, but it isn't.).

As we have done several times before, we can remove the explicit dependence on the threshold U_i, by including weights $-U_i$, and a clamped input of 1 that effectively biases the unit. Then, as we saw for the Hopfield net:

$$E = -\sum_{i<j} T_{ij} V_i V_j$$

If V_i goes from 1 to 0, the energy gap between two states corresponding to Vi being off (hypothesis i rejected) or on (hypothesis i accepted) is:

$$\Delta E_i = \sum_j T_{ij} V_j$$

■ Boltzmann update rule

To allow escapes from local minima, the idea is to allow occasional *increases* in energy, an idea that goes back to 1953 (Metropolis et al., 1953). Let's see how it works.

$$V_i = 1 \text{ with probability } p_i$$

$$p_i = p(\Delta E_i) = \frac{1}{1 + e^{-\Delta E_i / T}}$$

T is a free parameter that plays the role of temperature in thermodynamics. When we implement the algorithm below, we draw a number between 0 and 1 out of a hat, and if that number is less than p_i, we set V_i to 1. Otherwise, it is set to zero. Specifically, the update rule is:

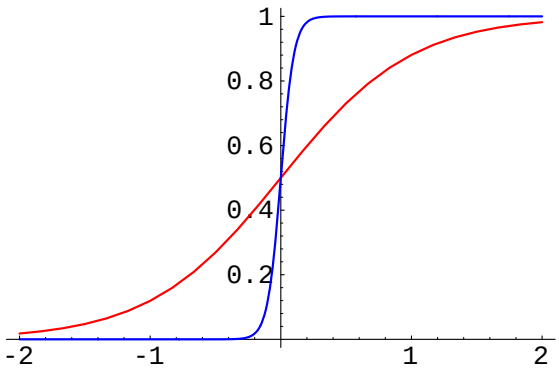
$$V_i = \begin{cases} 1 & \text{if Random[Real]} < p(\Delta E_i) = p(\sum_j T_{ij} V_j) \\ 0 & \text{otherwise} \end{cases}$$

If the temperature is very low (T about zero), the update rule is the same as that for a deterministic TLU. This is because if the weighted sum of inputs is bigger than zero, p is virtually at 1. The probability of setting V to 1 is then guaranteed. If the weighted sum is less than zero, p is zero, and the probability of setting V to 1 is nil.

```
boltz[x_,T_] := 1/(1 + Exp[-x/T]);
```

Here is a plot of p_i, with a high and low temperatures:

```
Plot[{boltz[x,.5], boltz[x,.05]}, {x,-2,2},
    PlotStyle->{RGBColor[1,0,0],RGBColor[0,0,1]}];
```



■ Simulated annealing

Now if we just let the Boltzmann rule update at a high temperature, the network will never settle to a stable point in state space. Conversely, if we set the temperature low, the network is likely to get stuck in a local minimum. The key idea behind introducing the notion of temperature is to start off with a high temperature, and then gradually cool the network. This simulates the physical process of annealing. If one heats metal, and then cools it rapidly, there is less crystalline structure or alignment of the atoms. This is a high energy state, and is desirable for making strong metal. The steel has been tempered. Slower cooling allows the substance to achieve a lower energy state with more alignment, with correspondingly more potential fractures. Although bad for metal strength, slow annealing is good for constraint satisfaction problems.

It has been shown that a suitably slow annealing schedule will guarantee convergence (Geman and Geman, 1984):

$$T(n) > \frac{c}{\log(1+n)}$$

This annealing schedule, however, can be VERY slow, in fact too slow to be usually practical, except for small scale problems.

The "Gibbs Sampler" is the more general form of updating for n-valued nodes making up a Markov Random Field (Geman and Geman, 1984).

Local minimum demonstration

Let's look at an example where the standard discrete Hopfield net gets stuck in a local minimum, but the Boltzmann machine with annealing gets out of it.

■ Initialization

We will use a toroidal geometry to keep the programming simple.

```
Mod2[x_,n_] := Mod[x-1,n] + 1;
threshold[x_] := N[If[x>=0,1,-1]];
```

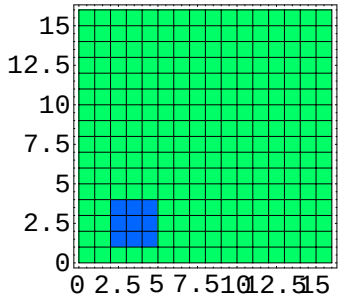
```
weight = 1;
size = 16;
numiterations = 10;
```

Below, we will deliberately construct a weight matrix so that the energy function has a local minimum at the following state vector:

```
V = Table[-1,{i,size},{j,size}];

V[[2,3]] = 1; V[[2,4]]= 1;  V[[2,5]] = 1;
V[[3,3]] = 1; V[[3,4]] = 1; V[[3,5]] = 1;
V[[4,3]] = 1; V[[4,4]] = 1; V[[4,5]] = 1;
```

Here is a picture of the state vector with the local minimum:

```
ListDensityPlot[V,ColorFunction->Hue,
    PlotRange->{-5,5}];
```



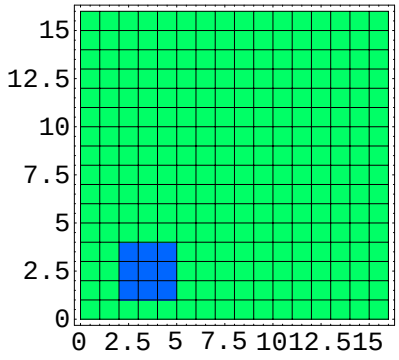
■ Asynchronous updating without annealing: Getting stuck

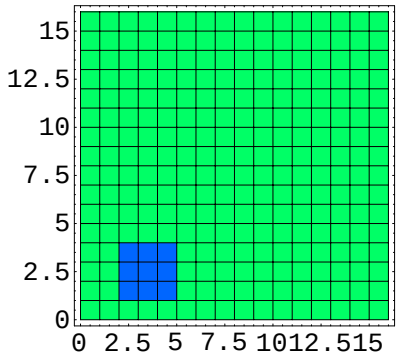
Each unit is connected to its four nearest neighbors with weights given by **weight** (=1). The rest of the weights are zero. So the update rule is:

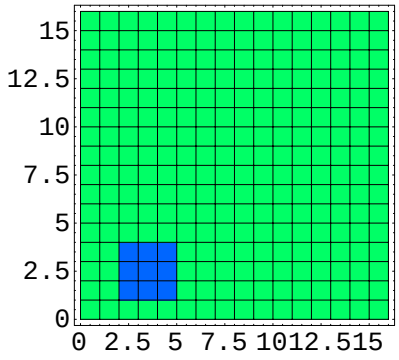
```
update[Vv_,ii_,jj_] :=
           threshold[weight (Vv[[ Mod2[ii + 1,size],
            jj  ]] +
           Vv[[  Mod2[ii - 1,size], jj  ]] +
           Vv[[  ii, Mod2[jj - 1,size]   ]] +
           Vv[[  ii,  Mod2[jj + 1,size]   ]])];
```

```
numiterations = 3;
```

```
For[iter=1,iter<=numiterations,iter++,
    For[i=1,i<=size*size,i++,
        iindex = Random[Integer,size-1]+1;
        jindex = Random[Integer,size-1]+1;
        V[[iindex,jindex]] = update[V,iindex,jindex];
    ];
    ListDensityPlot[V,ColorFunction->Hue,
        PlotRange->{-5,5}];
];
```





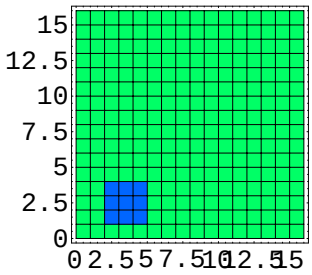


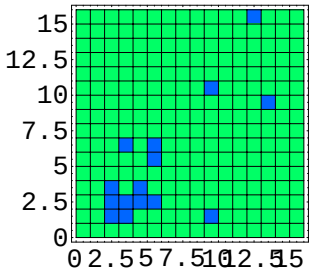
■ Asynchronous updating with annealing: Getting unstuck

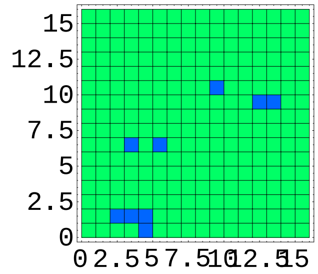
```
boltz[x_,T_] := 1/(1 + Exp[-x/T]);
```

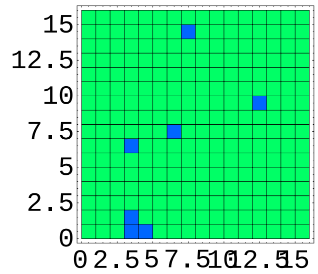
```
temp0=1;
numiterations = 10;
ListDensityPlot[V,ColorFunction->Hue,
    PlotRange->{-5,5}];
For[iter=1,iter<=numiterations,iter++,
    temp = temp0/Log[1 + iter];
    For[i=1,i<=size*size,i++,
        iindex = Random[Integer,size-1]+1;
        jindex = Random[Integer,size-1]+1;

        pdeltaE =
            N[boltz[weight (V[[ Mod2[iindex +
             1,size],jindex]] +
            V[[Mod2[iindex - 1,size],jindex]] +
            V[[iindex,Mod2[jindex - 1,size]   ]] +
            V[[iindex,Mod2[jindex + 1,size]   ]]),
             temp]];

        V[[iindex,jindex]] =
            If[ pdeltaE >= Random[],1,-1];
    ];
        ListDensityPlot[V,ColorFunction->Hue,
            PlotRange->{-5,5}];
];
```

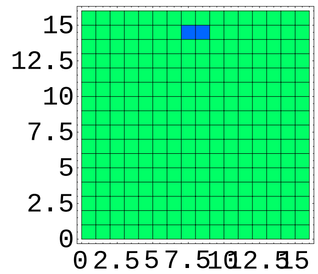
Here are the results of a simulation:

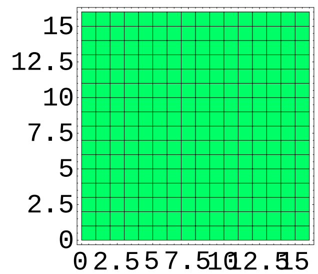


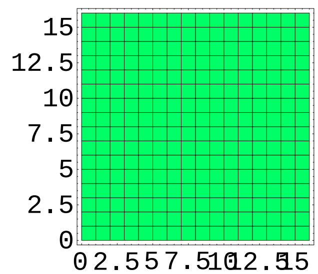



15
12.5
10
7.5
5
2.5
0
0 2.5 5 7.5 10 12.5 15


15
12.5
10
7.5
5
2.5
0
0 2.5 5 7.5 10 12.5 15


15
12.5
10
7.5
5
2.5
0
0 2.5 5 7.5 10 12.5 15


15
12.5
10
7.5
5
2.5
0
0 2.5 5 7.5 10 12.5 15


15
12.5
10
7.5
5
2.5
0
0 2.5 5 7.5 10 12.5 15

Optional exercise 1: Energy function

Write a function to calculate the energy function for the above network.

What is the energy of the ground state?

What is the energy of the local minimum we constructed above?

Optional Exercise 2 : Thermal equilbrium

Make a two versions of the above simulation in which 1) the temperature is fixed; 2) the temperature is gradually lowered. Start with a random initial setting of the network.

Boltzmann learning

We've seen how a stochastic update rule improves the chances of a network evolving to a global minimum. Now let's see how learning weights can be formulated as a statistical problem.

■ The Gibbs distribution again

Suppose T is fixed at some value, say T=1. Then we could update the network and let the it settle to thermal equilibrium, a state characterized by some statistical stability, but with occasional jiggles. Let V_α represent the vector of neural activities. The probability of a particular state α is given by:

$$p(V_\alpha) = \kappa e^{-E_\alpha / T}$$

$$\kappa = \frac{1}{\sum_{\text{all states } k} e^{-E_k / T}}$$

Recall that the second equation is the normalization constant the ensures that the total probability (i.e. over all states) is 1.

We divide up the units into two classes: **hidden** and **visible** units. Values of the visible units are determined by the environment. Our goal is to have the hidden units discover the structure of the environment. Once learned, if the network were left to run freely, the visible units would take on values that reflect the structure of the environment they learned.

Consider two probabilities over the visible units:

P - probability of visible units taking on certain values determined by the environment.

P' - probability that the visible units take on certain values while the network is running at thermal equilibrium.

If the hidden units have actually "discovered" the structure of the environment, then the probability P should match P'. How can one achieve this goal? Recall that for the Widrow-Hoff and error backpropagation rules, we started from the constraint

that the network should minimize the error between the network's prediction of the output, and the actual target values supplied during training. We need some measure of the discrepancy between the desired and target states for the Boltzmann machine. The idea is to construct a measure of how far away P is from P'. One such function is the Kullback-Leibler measure or relative entropy (also known as the "Gibbs G measure").

$$G\left(T_{12}, T_{13}\ldots, T_{ij}, \ldots\right) = \sum_{\substack{\text{all states} \\ \text{over visible units}}} P(V_\alpha) \log\left(\frac{P(V_\alpha)}{P'(V_\alpha)}\right)$$

Then we need a rule to adjust the weights so as to bring P' -> P. Ackley et al. derived the following rule for updating the weights so as to bring the probabilities closer together:

$$\Delta T_{ij} = \varepsilon\left(p_{ij} - p'_{ij}\right)$$

where p_{ij} is the probability of V_i and V_j both being 1 when environment is clamping the states at thermal equilibrium averaged over many samples. p'_{ij} is the probability of V_i and V_j being 1 when the network is running freely without the environment at equilibrium.

Descendants of Boltzmann machines

As noted above, convergence through simulated annealing can be impractically slow. Mean field approximation is one technique used to improve convergence (cf. Ripley, 1996). Boltzmann machines can be considered a special case of belief networks (Ripley, 1996).

The Boltzmann machine learns to approximate the joint probability distribution on a set of binary random variables. Some of the variables are designated inputs, and others outputs. Learning large scale joint distributions is known to be a hard problem in statistics, and the success of the Boltzmann machine has been limited to small scale problems. One successor to the Boltzmann machine is the Helmholtz machine and its derivaties (Dayan et al., 1995; Hinton, 1997). A recent advance in learning pattern distributions is the Minimax theory (Zhu and Mumford, 1997).

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. Cognitive Science, 9, 147-169.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. I.E.E.E. Transactions Pattern Analysis and Machine Intelligence, PAMI-6, 721-741.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz Machine. Neural Computation, 7(5), 889-904.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). Introduction to the theory of neural computation (Santa Fe Institute Studies in the Sciences of Complexity ed.). Reading, MA: Addison-Wesley Publishing Company.

Hinton, G. E., & Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. The Philosophical Transactions of the Royal Society, 352(1358), 1177-1190.

Kersten, D. (1990). Statistical limits to image understanding. In C. Blakemore (Ed.), Vision: Coding and Efficiency, (pp. 32-44). Cambridge, UK: Cambridge University Press.

Knill, D. C., & Kersten, D. K. (1991). Ideal Perceptual Observers for Computation, Psychophysics, and Neural Networks. In R. J. Watt (Ed.), Pattern Recognition by Man and Machine, (Vol. 14,): MacMillan Press.

Knill, D. C., Kersten, D., & Yuille, A. (1996). A Bayesian Formulation of Visual Perception. In K. D.C. & R. W. (Eds.), Perception as Bayesian Inference, (pp. Chap. 1): Cambridge University Press.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. J. Phys. Chem, 21, 1087.

Ripley, B. D. (1996). Pattern Recognition and Neural Networks . Cambridge, UK: Cambridge University Press.

Zhu, S. C., Wu, Y., & Mumford, D. (1997). Minimax Entropy Principle and Its Applications to Texture Modeling. Neural Computation, 9(8), 1627-1660.