3 Recurrent neuronal networks: Associative memory I

3.1

We saw that an associate memory based on a recurrent Hopfield network stores a number of memories that scales more weakly than the number of neurons if one cannot tolerate any errors upon recall. We now review the situation when a fixed, nonzero error rate is tolerated.

Figure 1: The Hopfield recurrent network. From Hertz, Krogh and Palmer 1991 following Hopfield 1082

3.2 Energy description and convergence

The following notes were abstracted from Chapter 2 of "Introduction to the Theory of Neural Computation" (Addison Wesley, 1991) by Hertz, Krogh and Palmer.

One of the most important contributions of Hopfield was to introduce the idea of an energy function into neural network theory. For the networks we are considering, the energy function $E$ is

$$E = -\frac{1}{2} \sum_{ij:i\neq j}^N W_{ij} S_i S_j.$$  (3.3)
The double sum is over all $i$ and all $j$. The $i = j$ terms are of no consequence because $S_i^2 = 1$; they just contribute a constant to $E$, and in any case we could choose $W_{ii} = 0$. The energy function is a function of the configuration $S_i$ of the system, where $S_i$ means the set of all the $S_i$’s. Typically this surface is quite hilly.

The central property of an energy function is that it always decreases (or remains constant) as the system evolves according to its dynamical rule. Thus the attractors (memorized patterns) are at local minima of the energy surface. For neural networks in general an energy function exists if the connection strengths are symmetric, i.e., $W_{ij} = W_{ji}$. In real networks of neurons this is an unreasonable assumption, but it is useful to study the symmetric case because of the extra insight that the existence of an energy function affords us. The Hebb prescription that we are now studying automatically yields symmetric $W_{ij}$’s.

For symmetric connections we can write the energy in the alternative form

$$E = -\sum_{(ij)} W_{ij} S_i S_j + \text{constant} \quad (3.4)$$

where $(ij)$ means all the distinct pairs of $ij$, counting for example ”1,2” as the same pair as ”2,1”. We exclude the $ii$ terms from $(ij)$; they give the constant. It now is easy to show that the dynamical rule can only decrease the energy. Let $S_i'$ be the new value of $S_i$ for some particular unit $i$:

$$S_i' = \text{sgn} \left( \sum_{j \neq 1} W_{ij} S_j \right). \quad (3.5)$$

Obviously if $S_i' = S_i$ the energy is unchanged. In the other case $S_i' = -S_i$ so, picking out the terms that involve $S_i$:

$$E' - E = -\sum_{j \neq i} W_{ij} S_i' S_j + \sum_{j \neq i} W_{ij} S_i S_j$$

$$= 2S_i \sum_{j \neq i} W_{ij} S_j.$$  

This term is negative from the update rule. Thus the energy decreases every time an $S_i$ changes, as claimed.

The idea of the energy function as something to be minimized in the stable states gives us an alternate way to derive the Hebb prescription. Let us start again with the single-pattern case. We want the energy to be minimized when the overlap between the network configuration and the stored pattern $\xi_i$ is largest. So we choose

$$E = -\frac{1}{2N} \sum_{k=1}^{P} \left( \sum_{i=1}^{N} S_i \xi_i^k \right)^2. \quad (3.7)$$

Multiplying this out gives

$$E = -\frac{1}{2N} \sum_{k=1}^{P} \left( \sum_{i=1}^{N} S_i \xi_i^k \right) \left( \sum_{j=1}^{N} S_j \xi_j^k \right) \quad (3.8)$$
Figure 2: A and B are the energy landscape for a model with symmetric W. C corresponds to an asymmetric W, for which the stem can drift or have limit cycles. From Hertz, Krogh and Palmer 1991.

\[
S_i S_j = -\frac{1}{2} \sum_{i \neq j} \left( \frac{1}{N} \sum_{k=1}^{P} \xi_i^k \xi_j^k \right) S_i S_j
\]

which is exactly the same as our original energy function if \( W_{ij} \) is given by the Hebb rule. This approach to finding appropriate \( W_{ij} \)'s is generally useful. If we can write down an energy function whose minimum satisfies a problem of interest, then we can multiply it out and identify the appropriate strength \( W_{ij} \) from the coefficient of \( S_i S_j \).

### 3.3 The issue of spurious attractors

The following notes were abstracted from Chapter 2 of ”Introduction to the Theory of Neural Computation” (Addison Wesley, 1991) by Hertz, Krogh and Palmer.

We have shown that the Hebb prescription gives us (for small enough \( P \)) a dynamical system that has attractors – local minima of the energy function – for the desired states \( \xi^k \). These are sometimes called the retrieval states. But we have not shown that these are the only attractors. And indeed there are others, as discovered by Dani Amit, Hannuck Gottfried and Hiam Sompolinsky in 1985.

First of all, the reversed states \(-\xi^k\) are minima and have the same energy as the original patterns. The dynamics and the energy function both have a perfect symmetry, \( S_i \leftrightarrow -S_i \forall i \). This is not too troublesome for the retrieved patterns; we could agree to reverse all the remaining bits when a particular “sign bit” is \(-1\) for example.

Second, there are stable mixture states \( \xi_i^{mix} \), which are not equal to any single pattern, but instead correspond to linear combinations of an odd number of patterns. The simplest of these are symmetric combinations of three stored patterns with components:

\[
\xi_i^{mix} = sgn(\pm \xi_i^1 \pm \xi_i^2 \pm \xi_i^3) .
\]

(3.9)

All \( 2^3 = 8 \) sign combinations are possible, but we consider for definiteness the case where all the signs are chosen as +’s. The other cases are similar. Observe that on average \( \xi_i^{mix} \) has the same sign at \( \xi_i^1 \) three times out of four; only if \( \xi_i^2 \) and \( \xi_i^3 \)
both have the opposite sign can the overall sign be reversed? So $\xi_{i}^{\text{mix}}$ is Hamming distance $N/4$ from $\xi_{i}^{1}$, and of course from $\xi_{i}^{2}$ and $\xi_{i}^{3}$ too; the mixture states lie at points equidistant from their components. This also implies that $\sum_{i} \xi_{i}^{1} \xi_{i}^{\text{mix}} = N/2$ on average. To check the stability pick out the three special states with $k = 1$, 2, and 3, still with all $+$ signs, to find:

$$\mu_{i}^{\text{mix}} = \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{3} \xi_{i}^{k} \xi_{j}^{k} \xi_{j}^{\text{mix}}$$

$$= \frac{1}{2} \xi_{i}^{1} + \frac{1}{2} \xi_{i}^{2} + \frac{1}{2} \xi_{i}^{3} + \text{cross terms} .$$

Thus the stability condition is satisfied for the mixture state. Similarly 5, 7, ... patterns may be combined. The system does not choose an even number of patterns because they can add up to zero on some sites, whereas the units have to have nonzero inputs to have defined outputs of $\pm 1$.

Third, for large $P$ there are local minima that are not correlated with any finite number of the original patterns $\xi^{k}$.

### 3.4 The phase diagram of the Hopfield model

The following notes were abstracted from Chapter 2 of ”Introduction to the Theory of Neural Computation” (Addison Wesley, 1991) by Hertz, Krogh and Palmer.

A statistical mechanical analysis by Amit, Gottfried and Sompolinsky (1985) shows that there is a crucial value of $P/N$ where memory states no longer exist. A numerical evaluation gives

$$\alpha_{C} \equiv \frac{P}{N}\text{critical} \approx 0.138 .$$

The jump in the number of memory states is considerable: from near-perfect recall to zero. This tells us that with no internal noise we go discontinuously from a very good working memory with only a few bits in error for $\alpha < \alpha_{C}$ to a ”useless” memory system for $\alpha > \alpha_{C}$.

The phase diagram for the Hopfield model delineates different regimes of behavior in the Variance -- $\alpha$ plane (variance is $\sigma^{2}$ in our notation, but the statistical
mechanics literature uses $T$ for temperature). There is a roughly triangular region where the network is a good memory device, as indicated by regions A and B of the embedded figure. The result corresponds to the upper limit of $\alpha_C$ on the $\alpha$ axis, while the critical variance $T_C = 1$ for the $P \ll N$ case sets the limit on the variance axis. Between these limits there is a maximum variance or maximum load defined by a phase-transition line. As $\textit{Variance} \rightarrow 1$, $\alpha_C(T)$ goes to zero like $(1 - T)^2$.

Figure 4: The error rate upon retrieval for variance, $T = 0$. From Hertz, Krogh and Palmer 1991, following Amit, Gutfreund and Sompolinsky 1985.

In region C of the phase diagram the network still turns out to have many stable states, called spin glass states, but these are not correlated with any of the patterns $\xi^k_i$. However, if $T$ is raised to a sufficiently high value, into region D, the output of the network continuously fluctuates with $\langle S_i \rangle = 0$.

Regions A and B of the phase diagram both have the desired retrieval states, beside some percentage of wrong bits, but also have spin glass states. The spin states are the most stable states in region B, lower in energy than the desired states, whereas in region A the desired states are the global minima. For small enough $\alpha$ and $\textit{Variance}$ there are also mixture states that are correlated with an odd number of the patterns as discussed earlier. These always have higher free energy than the desired states. Each type of mixture state is stable in a triangular region (A and B), but with smaller intercepts on both axes. The most stable mixture states extend to 0.46 on the $\textit{Variance}$ axis and 0.03 on the $\alpha$ axis (a subregion of A).

### 3.5 Can we relate stable states to a task?

The previous data by Carandini and Harris implies states, but the states were tied to ongoing sensory input. Can we tie states to a task that is is ongoing, such as a memory task, where the external cues were removed? This is captured by the delay-to-match task of Joaquin Fuster; we show a more recent incarnation by Yasushi Miyashita. The monkey is asked to remember a picture and then, after a delay without visual input, compare a new picture with the old picture. The monkey signals if the two are part of a matched set.
The spike rate of different neurons in inferior temporal cortex are measured while the monkey is performing this task. Critically, some neurons go up in their firing rate while others go down in rate. An interesting observation is that activity continues throughout the period of the delay, for which there is no stimulus. This can occur for 20 seconds or more, i.e., one to two orders of magnitudes longer than the integration time of neurons. We take this as evidence for sustained activity based on neuronal interactions as the recordings are in regions that appear heavily interconnected. Further, with one exceptional case found so far, individual neurons do not show multistability.

3.6 Noise and spontaneous excitatory states as a model for epilepsy

It is worth asking if, by connection with ferromagnetic systems, rate equations of the form used for the Hopfield model naturally go into an epileptic state of continuous firing, but not necessarily with every cell firing. This exercise also allows us to bring up the issue of fast noise (variance) that is uncorrelated from neuron to neuron.

We consider \( N \) binary neurons, with \( N >> 1 \), each of which is connected to all other neighboring neurons. For simplicity, we assume that the synaptic weights \( W_{ij} \) are the same for each connection, i.e., \( W_{ij} = W_0 \). Then there is no spatial structure in the network and the total input to a given cell has two contributions. One term from the neighboring cells and one from an external input, which we also take to be the same for all cells and denote \( I^{ext} \). Then the input is

\[
\mu_i = W_0 \sum_{j=1}^{N} S_j + I^{ext}.
\]  

(3.12)

The energy per neuron, denoted \( \epsilon_i \), is then

\[
\epsilon_i = -S_i \mu_i
\]  

(3.13)

\[
= -S_i W_0 \sum_{j=1}^{N} S_j - S_i I^{ext}
\]
The insight for solving this system is the mean-field approach. We replace the sum of all neurons by the mean value of $S_i$, denoted $<S>$, where

$$< S > = \frac{1}{N} \sum_{j=1}^{N} S_j,$$

so that

$$\epsilon_i = -S_i \left( W_0 N < S > + I^{ext} \right).$$

We can now use the expression for the value of the energy in term of the average spike rate, $< S >$, to solve self consistently for $< S >$. We know that the average rate is given by a Boltzmann factor over all of the $S_i$. Thus

$$< S > = \frac{\sum_{S_i=\pm 1} S_i e^{-\epsilon_i/k_B T}}{\sum_{S_i=\pm 1} e^{-\epsilon_i/k_B T}}$$

$$= \frac{\sum_{S_i=\pm 1} S_i e^{S_i(W_0 N < S > + I^{ext})/k_B T}}{\sum_{S_i=\pm 1} e^{S_i(W_0 N < S > + I^{ext})/k_B T}}$$

$$= e^{-W_0 N < S > + I^{ext})/k_B T} - e^{W_0 N < S > + I^{ext})/k_B T}$$

$$= e^{-(W_0 N < S > + I^{ext})/k_B T} + e^{(W_0 N < S > + I^{ext})/k_B T}$$

$$= \tanh \left( \frac{W_0 N < S > + I^{ext}}{k_B T} \right).$$

where we made of the fact that $S_i = \pm 1$. This is the neuronal equivalent of the famous Weiss equation for ferromagnetism. The properties of the solution clearly depend on the ratio $W_0 N k_B T$, which pits the connection strength $W_0$ against the noise level $T/N$. We also see how the input-output function $\tanh \{x\}$ naturally arises.

- For $\frac{W_0 N}{k_B T} < 1$, the high noise limit, there is only the solution $< S > = 0$ in the absence of an external input $h_0$.

- For $\frac{W_0 N}{k_B T} > 1$, the low noise limit, there are three solutions in the absence of an external input $h_0$. One has $< S > = 0$ but is unstable. The other two solutions have $< S > \neq 0$ and must be found graphically or numerically.

- For sufficiently large $|I^{ext}|$ the network is pushed to a state with $< S > = \text{sgn}(I^{ext}/k_B T)$ independent of the interactions.

We see that there is a critical noise level for the onset of an active state and that this level depends on the strength of the connections and the number of cells. We also see that an active state can occur spontaneously for $\frac{W_0 N}{k_B T} > 1$ or $T < \frac{W_0 N}{k_B}$. This is a metaphor for epilepsy, in which recurrent excitatory connections maintain a spiking output (although a lack of inhibition appears to be required as a seed).