Models of delay activity

Sebastian Seung

9.641 Lecture 16: November 12, 2002

1   Miyashita experiment

The monkey was trained on a task involving visual short-term memory. The sample and match stimuli were one of 97 randomly generated "fractal" color patterns. Both sample and match were presented for 200 ms. During the 16 second delay between sample and match, no visual stimulus was presented. During training, the sample cycled through a fixed sequence, while the match was chosen at random.

During the test phase, single unit recordings were made in anterior ventral temporal cortex. Now the sample stimulus was presented in random sequence. For the learned stimuli, there was strong delay activity. Neural response was selective for only a few learned stimuli. The delay activity patterns for stimuli that were adjacent in time during training were correlated with each other.

Caveat: The delay activity was weak for the 97 new stimuli. However, short-term memory performance on new patterns was as good as on old.

The delay activity patterns could be the fixed point attractors of the Hopfield model. But the attractors in the Hopfield model are uncorrelated with each other, so some modification of the model is necessary.

2   The GTA model

To model the Miyashita experiment, Griniasty, Tsodyks, and Amit proposed a modification of the Hopfield model with the following weight matrix:

$$W_{ij} = \frac{1}{N} \sum_{\mu=1}^{P} (\xi_i^\mu \xi_j^\mu + a\xi_i^{\mu+1}\xi_j^\mu + a\xi_i^{\mu-1}\xi_j^\mu)$$

The ordering of the patterns is supposed to be the order of presentation during training. And for simplicity, the ordering is assumed to be cyclic, so that $\xi_i^{N+1} = \xi_i^1$.

The natural order parameters are the overlaps

$$m^\mu = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu s_i$$

1

Since

$$\sum_j W_{ij}s_j = \frac{1}{N}\sum_{\mu=1}^{P}(\xi_i^\mu + a\xi_i^{\mu+1} + a\xi_i^{\mu-1})\sum_j \xi_j^\mu s_j$$

It follows from the update rule

$$s_i' = \text{sgn}\left(\sum_j W_{ij}s_j\right)$$

that the stationary states satisfy

$$m^\mu = \frac{1}{N}\sum_i \xi_i^\mu \,\text{sgn}\left(\sum_\nu m^\nu(\xi_i^\nu + a\xi_i^{\nu+1} + a\xi_i^{\nu-1})\right)$$

If we now average this by randomizing the patterns, then the average at each neuron is the same, so that we can drop the index i:

$$m^\mu = 2^{-P}\sum_\xi \xi^\mu \,\text{sgn}\left(\sum_\nu m^\nu(\xi^\nu + a\xi^{\nu+1} + a\xi^{\nu-1})\right)$$

Let's consider the stability of a pure pattern solution, for example $m^2 = 1$ with all other overlaps zero. If $a < 0.5$, then the right hand side of the above equation is

$$\frac{1}{N}\sum_i \xi_i^\mu \,\text{sgn}(\xi_i^2 + a\xi_i^3 + a\xi_i^1) = \delta^{\mu 2}$$

in the $N \to \infty$ limit, since $\text{sgn}(\xi_i^2 + a\xi_i^3 + a\xi_i^1) = \xi_i^2$. If $a > 0.5$, then the situation is different. On average, $\xi_i^3 = \xi_i^1 = -\xi_i^2$ for one quarter of the neurons, so that $\text{sgn}(\xi_i^2 + a\xi_i^3 + a\xi_i^1) = -\xi_i^2$. Therefore the right hand side is $0.5(\delta^{\mu 1} + \delta^{\mu 2} + \delta^{\mu 3})$.

We can numerically solve by iterating the fixed point equations to a steady state. For $0.5 < a < 1$, this converges to a state with a radius of five nonzero overlaps. If $a > 1$, the stable states have overlap with all stored patterns.

Suppose that m^μ is the vector of overlaps at a steady state. Then the steady state is of the form

$$s_i = \text{sgn}\left(\sum_\mu m^\mu(\xi_i^\mu + a\xi_i^{\mu+1} + a\xi_i^{\mu-1})\right)$$

So if each attractor has overlap with several patterns, then the attractors associated with different patterns must also overlap.

3   Associative memory with sparse patterns

The ± 1 symmetry in the Hopfield model is rather artificial. The activity patterns in the brain are generally sparse. That is, the number of active neurons is much smaller

than the total number of neurons. To treat the case of sparse patterns, it is convenient to switch to $s_i = 0$ or 1, and write the dynamics as

$$s_i(t+1) = H\left(\sum_j W_{ij}s_j(t) - \theta\right) \tag{1}$$

where θ is a threshold, H is the Heaviside step function, and W_{ij} are the synaptic weights.

Let P random patterns be given, ξ_i^μ with $\mu = 1$ to P. Each component is one with probability f and zero with probability $1 - f$. How should W be chosen so as to embed the patterns as attractors of the network dynamics? A number of Hebbian rules are considered below.

4   Covariance rule

The covariance rule is

$$W_{ij} = \frac{1}{Nf(1-f)} \sum_\mu (\xi_i^\mu - f)(\xi_j^\mu - f)$$

Before we go into a detailed analysis of the covariance rule, let's obtain some rough intuition as to why it works. The basic reason is that the formula

$$\sum_j W_{ij}\xi_j^\mu \approx \xi_i^\mu - f$$

is approximately true. If this formula were exactly true, ξ^μ would be a steady state of the dynamics (1), provided that the threshold θ were set somewhere between $1 - f$ and $-f$.

This approximate formula is true in turn, provided that the following "orthogonality" condition is approximately satisfied:

$$\frac{1}{Nf(1-f)} \sum_j (\xi_j^\mu - f)\xi_j^\nu \approx \delta^{\mu\nu}$$

The $\mu \neq \nu$ term is small because $\xi_j^\mu - f$ is a zero mean random variable, and uncorrelated with ξ_j^ν. The approximation becomes better as $N \to \infty$ with f held fixed.

Now let's do a calculation to see when the above approximations are accurate. In the following we'll take the limits $N \to \infty$ and $P \to \infty$ with their ratio $\alpha = P/N$ held fixed.

$$\sum_j W_{ij}\xi_j^\nu \approx \frac{1}{Nf(1-f)} \sum_\mu (\xi_i^\mu - f) \sum_{j,j\neq i} (\xi_j^\mu - f)\xi_j^\nu \tag{2}$$

$$\approx (\xi_i^\nu - f) + \frac{1}{Nf(1-f)} \sum_{\mu,\mu\neq\nu} (\xi_i^\mu - f) \sum_{j,j\neq i} (\xi_j^\mu - f)\xi_j^\nu \tag{3}$$

3

The first term is the "signal," while the second term is the "noise." If the noise term were zero, ξ^ν would be a steady state for any threshold θ between $1 - f$ and $-f$. This will still be true with high likelihood if the noise is much smaller than the signal. Therefore it is important to estimate the size of the noise.

The noise term has zero mean and variance

$$\sigma^2 = \alpha f$$

Since the signal is order unity, we expect that $\alpha f \ll 1$ should be a sufficient condition for faithful storage. In other words, we estimate the capacity of the system to be $\alpha_c \sim 1/f$.

A more complicated calculation gives the result

$$\alpha_c \approx \frac{1}{2f |\log f|}$$

This has a logarithmic correction relative to the result derived by our simpler argument.

The capacity increases as the patterns become more sparse ($f \rightarrow 0$). This makes sense, as each pattern contains less information as f decreases.

5   Separating excitation and inhibition

The covariance rule is mathematically nice, but does not implement the biological constraint that excitatory and inhibitory neurons are distinct. An alternative is

$$W_{ij} = \frac{1}{Nf(1-f)} \sum_\mu (\xi_i^\mu \xi_j^\mu - f^2)$$

This can be implemented by a network of N excitatory neurons and a single global inhibitory neuron.