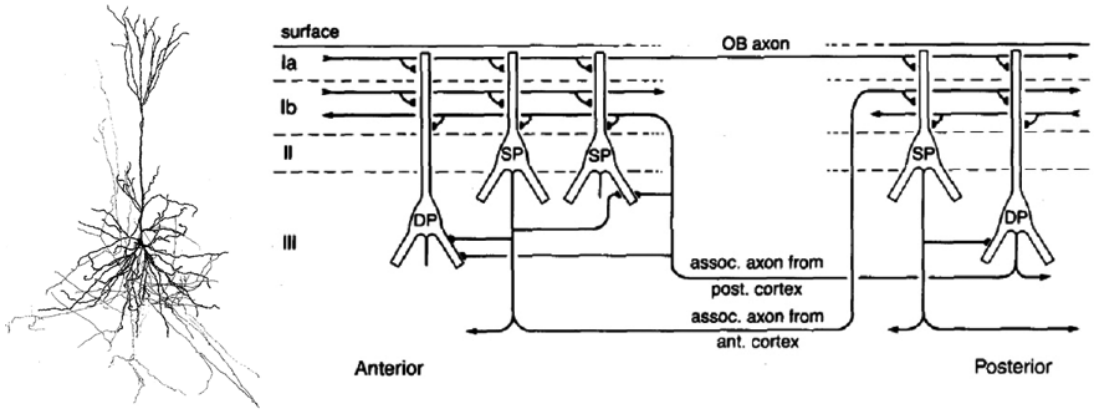
Revised 20 Jan 2024

2 Recurrent neuronal networks: Brain states and discrete attractors

2.1 Recurrent connections and memories

The storage of memories in the brain is an old and central issue in neuroscience. As we last discussed, it was known since the 1930's that bistable devices formed from threshold elements, like a digital flip-flop, could be built using feedback to hold electronic summing junctions in a particular state after their inputs had decayed away. By the 1970's, it was conjectured that networks with many summing junctions, or neurons, might be able to store a multitude of states if the feedback was extended across all pairs of cells, *i.e.*, order N^2 connections across N neurons. What are the expected motifs for such circuits? By extension of the idea of flip-flops, we might expect to find regions of the brain with neurons whose axon collaterals feed back onto other neurons. This anatomical arrangement was highlighted for the perform cortex of the olfactory system in a ca 1980's paper by Haberly (Figure 1) and by other researchers for the CA3 region of hippocampus. It was explored theoretically starting in the 1970s and culminated with a pivotal contribution by Hopfield in 1982 and an analysis of Hopfield's model by Amit, Gutfriend and Sompolinsky, by Gardner, and by others, in the mid 1980s.

Figure 1: Summary of major excitatory connections in piriform cortex. Each cell represents a population. From Haberly 1985.



The hippocampus seemed a particularly valuable region to consider feedback, as it is known for the occurrence of place cells. In their simplest substantiation, these are neurons that fire only when the animal reaches a particular location in the local environment. Different cells prefer to spike in different locations. Thus the animals builds up a map

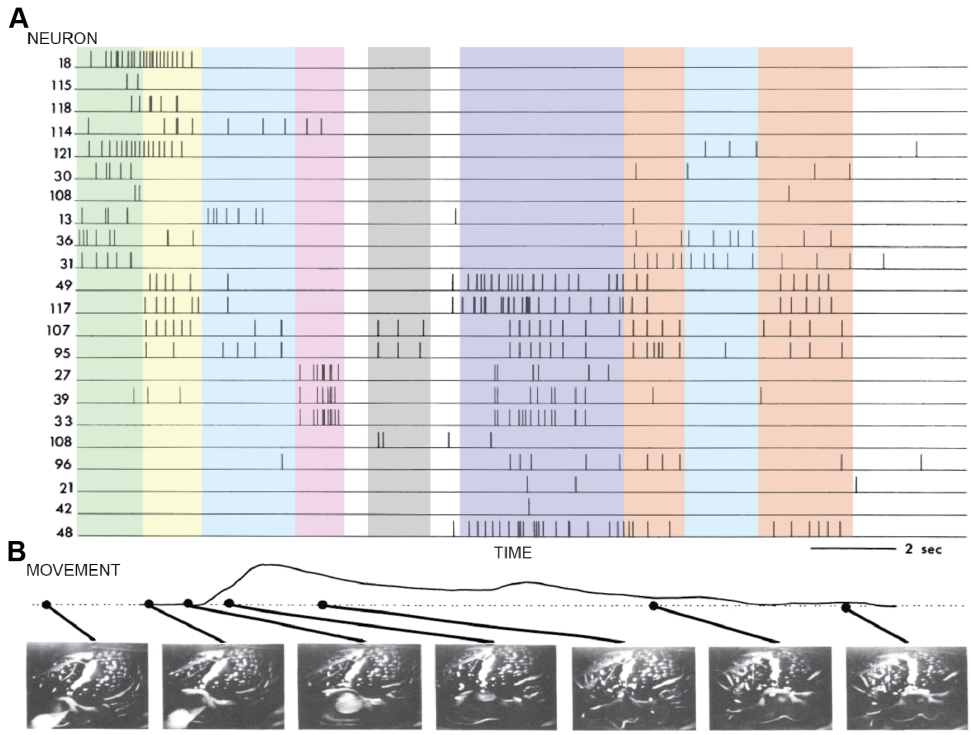
of the space, and in principle can use this map to determine a path to move from one location to another.

So we have an idea - the use of feedback connections to form memories of many places, or of anything by extrapolation, and we have biological motivation, in terms of the anatomical evidence, to understand the dynamics of such networks as well as search for them in real nervous systems.

2.2 What is a state?

We previously considered the output from neuronal networks with only two cells, so the notion of a state was pretty obvious. In general, the state is simply the arrangement of ON or active neurons (+1) and OFF or quiescent neurons (-1) under observation. Ideally this is every neuron in the circuit, which is possible in some preparations, like the invertebrate preparations that we considered in week 1. In fact, back in 1987 Larry Cohen and colleagues measured from about 100 neurons in the mollusc Navanax and showed that the neuronal activity that underlies feeding events appears to cluster into states (Figure 2).

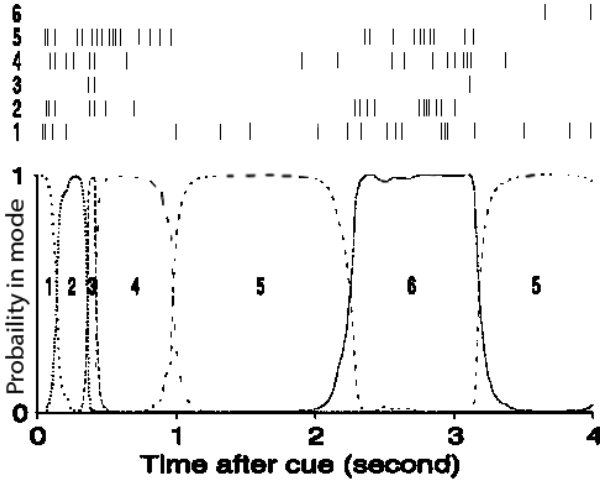
Figure 2: Navanax feeding. (A) Voltage sensitive dye recording of 115 neurons yields 22 neurons whose activity that contributes to 8 states. (B) The movement of the siphon during feeding. From London, Zecevic and Cohen, 1987.



The additional ideal of repeating patterns of states emerged about the same time through the cortical studies of Moshe Abeles and colleagues. They recorded from frontal

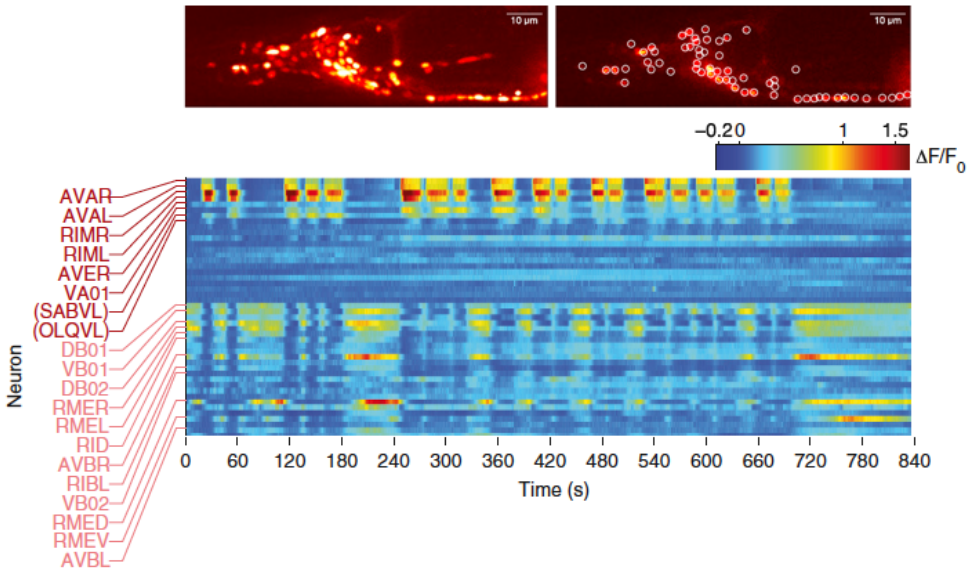
areas of monkey cortex and tended to see repeated patterns of spikes, even though they recorded from relatively few cells. Judge for yourself (Figure 3)!

Figure 3: Firing times of six neurons in monkey frontal cortex over a total of 93 trials were used to construct an hidden Markov model. Six states were identified. From Abeles, Bergman, Gati, Meilijson, Seidemann, Tishby and Vaadia 1995.



Zooming up to modern times, the technology of recording spikes from many cells at the same time has vastly improved to get a much better view of concurrent neuronal activity. States appear to occur in preparations that contain tens to hundreds of neurons in which every cell can be observed at effectively the same time. This is shown from recordings by Manuel Zimmer and colleagues of neurons in the worm c. Elegans (Figure 4).

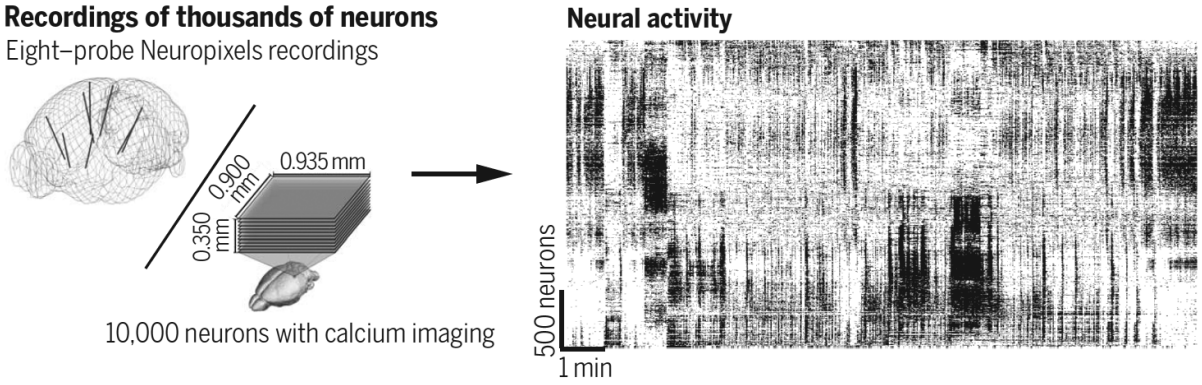
Figure 4: Calcium imaging from c. Elegan neurons during movement. Kato, Kaplan, Schrodel, Skora, Lindsay, Yemini, Lockery and Zimmer 2015



Moving to mammals, high density electrodes - Neuropixels - by Timothy Harris permits cylindrical volumes of the mouse brain to be probed, as seen in the data of Mateo

Carandini and Kenneth Harris and colleagues (Figure 5). We see many repeating or near repeating patterns among what is really a very sparse sample, i.e, 10^4 neurons among the 10^8 neurons in the mouse brain. The same neurons can be active or quiescent across a multitude of states.

Figure 5: Sorted output from Neuropixels probes in the brain of mouse, From Stringer, Pachitariu, Steinmetz, Reddy, Carandini and Harris 2019



One special aspect of all of these and related data is that stable firing patterns exist. In the last two case one could see patterns without special statistical tools - just recording of the presentation. A second aspect is that the number of states are few, i.e.,less than the number of cells, denoted N, and far, far less than the number of possible states, i.e., 2^N.

2.3 Are real networks highly interconnected?

The connectome, or wiring diagram of the brain, has been brain completed only for most of the fly and just for parts of other animals. We consider for the moment the connections among the neural integrator for horizontal eye position position in the juvenile zebrafish, which has been reconstructed over a large enough region to draw some conclusions (Figures 6 and 7). Here, about 0.1 of the neurons make recurrent connections on each other; this should be taken as a lower bound on connections (Figure 8). In any case all this means is that we need $0.1 * N \gg logN$ or $N \gg 35$, which is consistent with about 500 neurons in the integrator.

2.4 The network

We consider the dynamics of a fully connected recurrent neuronal network. We will begin our analysis guided by this task:

> Store a set of P patterns $\vec{\xi}^k$ in such a way that when presented with a new pattern that has partial overlap with an existing pattern $\vec{\mathbf{S}}^{test}$, the network responds by producing whichever one of the stored patterns most closely resembles $\vec{\mathbf{S}}^{test}$ (Figure 9). Close is defined by the Hamming distance, the number of different "bits" in the pattern.

Figure 6: Velocity-to-position neural integrator. Schematic showing the proposed wiring of modO, cells that project to the periphery, along with the two submodules modOI and modOM, and DO neurons that synapses onto ABDM and ABDI. From Vishwanathan, Ramirez, Wu, Sood, Yang, Kemnitz, Ih, Turner, Lee, Tartavull, Silversmith, Jordan, David, Bland, Goldman, Aksay and Seung, unpublished

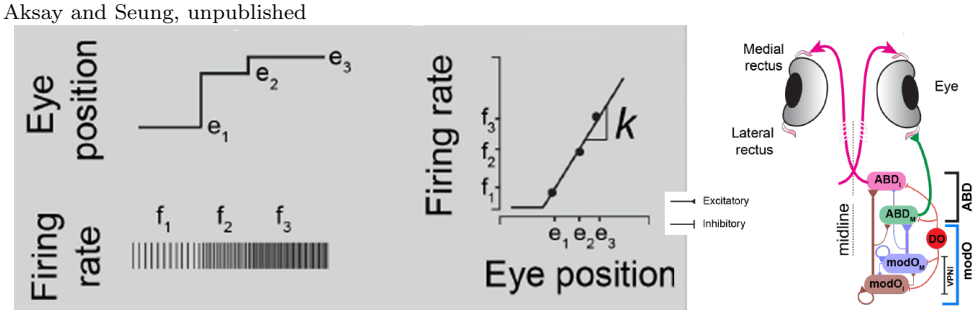
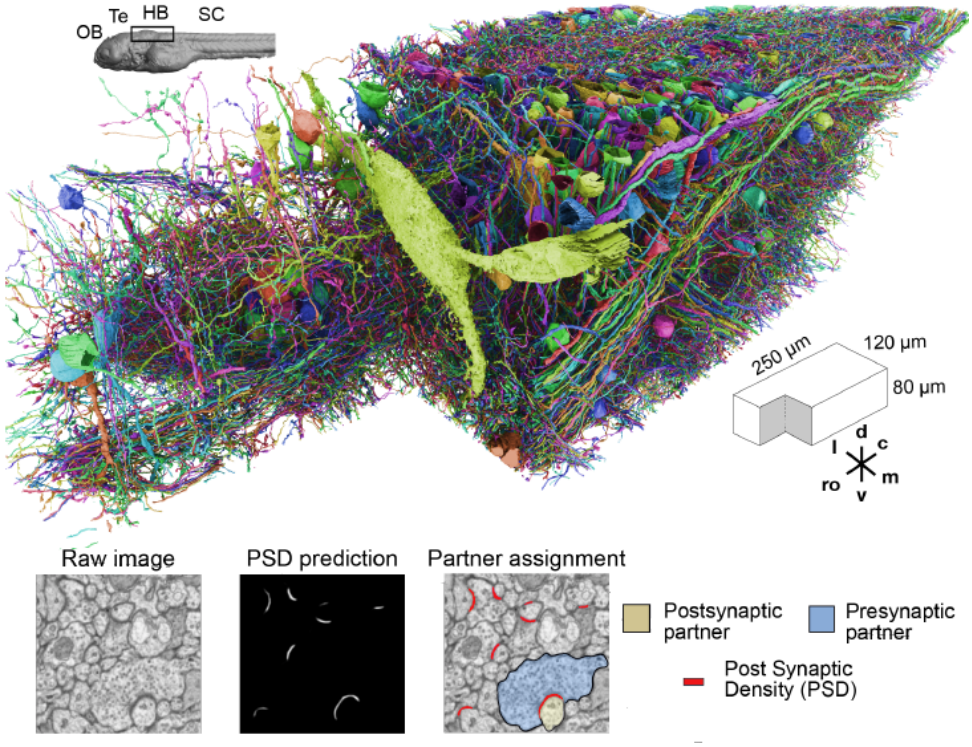


Figure 7: Cut-section view of the reconstructed volume and labeling of a synapse. From Vishwanathan, Ramirez, Wu, Sood, Yang, Kemnitz, Ih, Turner, Lee, Tartavull, Silversmith, Jordan, David, Bland, Goldman, Aksay and Seung, unpublished



The neurons are labelled by $i = 1, 2, \ldots, N$ and the individual stable patterns are labeled by $k = 1, 2, \ldots, P$.

We denote the activity of the $i - th$ neuron by S_i. The input to neuron i is denoted by μ_i and is given by

$$\mu_i = \sum_{j=1;\, j\neq i}^{N} W_{ij} S_j + I_i^{ext} \tag{2.1}$$

where the W_{ij} are analog-valued synaptic weights and I_i^{ext} is an external input. The

Figure 8: Connectivity matrix of center neurons organized into two modules (modA, modO). Neurons in the center were clustered whereas neurons in the periphery were not. Neurons in the periphery were organized by known cell types, vSPNs and ABD neurons. Colored dots represent the number of synapses. From Vishwanathan, Ramirez, Wu, Sood, Yang, Kemnitz, Ih, Turner, Lee, Tartavull, Silversmith, Jordan, David, Bland, Goldman, Aksay and Seung, unpublished

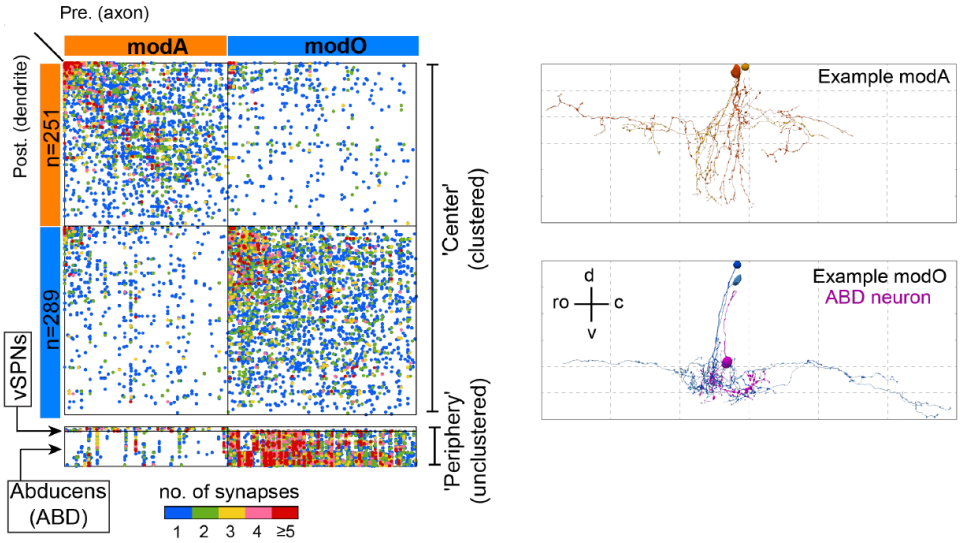
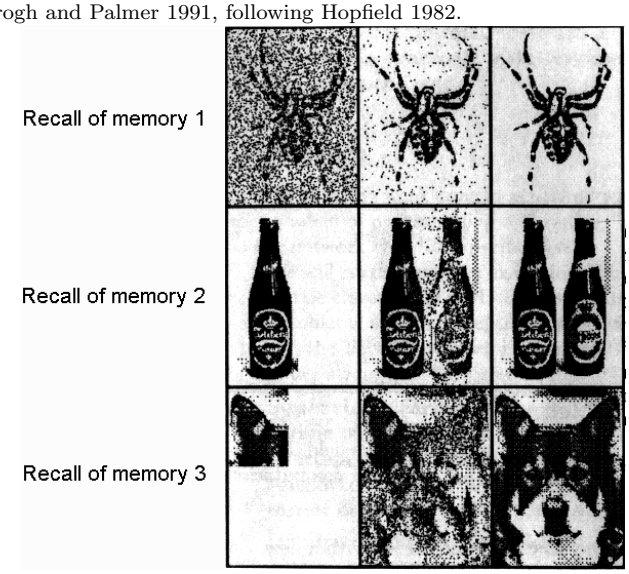


Figure 9: Function of the network as a content addressable memory in the recovery of a full memory from partial initial information. from Hertz, Krogh and Palmer 1991, following Hopfield 1982.



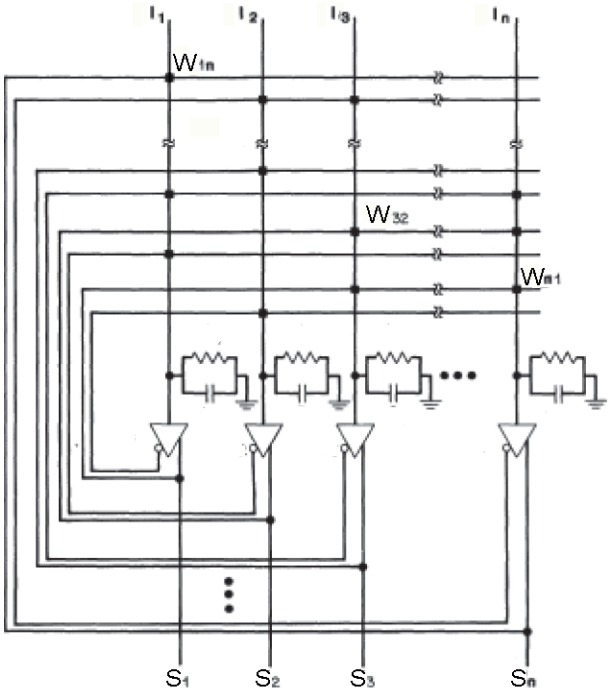
dynamics of the network are (Figure 10):

$$S_i \equiv sgn\left(\mu_i - \theta_i\right) \tag{2.2}$$

Clearly the output S_I is driven by the external input when I_i^{ext} is sufficiently large.

Going forward, we may take $\theta_i = 0\ \forall i$ as befits the case of random patterns on which neuronal outputs take on the values $+1$ and -1 with equal probability. In the further

Figure 10: Basic associative or "Hopfield" network. From Hertz, Krogh and Palmer 1991, following Hopfield 1982.



absence of external input, we have the minimal description

$$S_i \equiv sgn\left(\sum_{j\neq i}^{N} W_{ij}S_j\right). \tag{2.3}$$

There are at least two ways in which we might carry out the updating specified by the above equation. We could do it synchronously, updating all units simultaneously at each time step. Or we could do it asynchronously, updating them one at a time. Both kinds of models are interesting, but the asynchronous choice is more natural for both brains and artificial networks. The synchronous choice requires a central clock or pacemaker, and is potentially sensitive to timing errors, as is the case of sequential updating. In the asynchronous case, which we adopt henceforth, we can proceed in either of two ways:

- At each time step, select at random a unit i to be updated, and apply the update rule.
- Let each unit independently choose to update itself according to the update rule, with some constant probability per unit time.

These choices are equivalent, except for the distribution of update intervals. For the second case there is vanishing small probability of two units choosing to update at exactly the same moment.

Rather than study a specific problem such as memorizing a particular set of pictures, we examine the more generic problem of a random set of patterns drawn from a distri-

bution. For convenience, we will usually take the patterns to be made up of independent bits ξ_i that can each take on the values +1 and -1 with equal probability.

Our procedure for testing whether a proposed form of W_{ij} is acceptable is first to see whether the patterns to be memorized are themselves stable, and then to check whether small deviations from these patterns are corrected as the network evolves.

2.5 Storing one pattern

To motivate our choice for the connection weights, we consider first the simple case whether there is just one pattern ξ_i that we want to memorize. The condition for this pattern to be stable is just

$$sgn\left(\sum_{j\neq i}^{N} W_{ij}\xi_j\right) = \xi_i \qquad\qquad \forall i \tag{2.4}$$

since the update rule produces no changes. It is easy to verify this if we take

$$W_{ij} \propto \xi_i\xi_j \tag{2.5}$$

since $\xi_j^2 = 1$. We take the constant of proportionality to be $1/N$, where N is the number of units in the network, which yields

$$W_{ij} = \frac{1}{N}\xi_i\xi_j. \tag{2.6}$$

Furthermore, it is also obvious that even if a number (fewer than half) of the bits of the starting pattern S_i are wrong, i.e., not equal to ξ_i, they will be overwhelmed in the sum for the net input $\sum_{j\neq i}^{N} W_{ij}S_j$ by the majority that are correct, so that $\mathrm{sgn}[\sum_{j\neq i}^{N} W_{ij}S_j]$ will still give ξ_i.

An initial configuration near to ξ_i will therefore quickly relax to ξ_i. This means that the network will correct errors as desired, and we can say that the pattern ξ_i is an attractor. Actually, there are two attractors in this simple case; the other one is at $-\xi_i$. This is called a reversed state. All starting configurations with more than half the bits different from the original pattern will end up in the reversed state, $-\xi_i$.

2.6 Storing many patterns

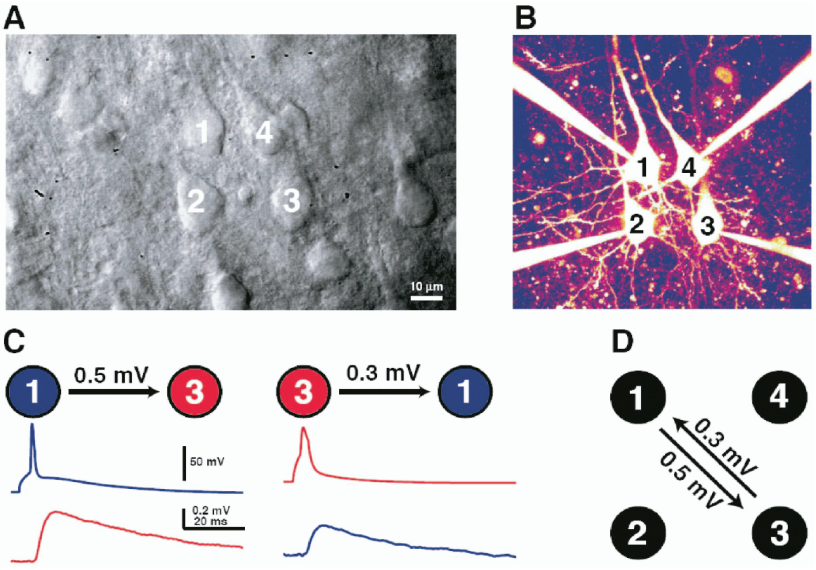
How do we get the system to recall the most similar of many patterns? The simplest answer is just to make the synaptic weights W_{ij} by an outer product rule for each of the P patterns, which corresponds to

$$W_{ij} = \frac{1}{N}\sum_{k=1}^{P} \xi_i^k\xi_j^k. \tag{2.7}$$

The above rule for synaptic weights is called the "Hebbian rule" because of the similarity with a hypothesis made by Hebb in 1949 about the way in which synaptic strengths in the brain change in response to experience: Hebb suggested changes are proportional

to the correlation between the firing of the pre- and post-synaptic neurons. The Hebb prescription automatically yields symmetric W_{ij}'s. This is an unreasonable assumption, although experimentally symmetric synapses occur more than expected by chance (Figure 11). Nonetheless, it is useful to study the symmetric case because of the extra insight that the existence of an energy function affords us.

Figure 11: Experimental evidence for symmetric synapses based on pairwise recordings form L5 pyramidal neurons in mouse brain slice. Only one pair of neurons are connected in these measurements. From Song, Sjostrom, Reigl, Nelson and Chklovskii, 2005



2.7 Scaling for error-free storage of many patterns

We consider a Hopfield network with the standard Hebb-like learning rule and ask how many memories we can imbed in a network of N neurons with the constraint that we will accept at most one bit of error, i.e., one neuron's output in only one of the memory states. The input is

$$\begin{aligned} \mu_i &= \sum_{j\neq i}^{N} W_{ij} S_j \\ &= \frac{1}{N} \sum_{k=1}^{P} \sum_{j\neq i}^{N} \xi_i^k \xi_j^k S_j. \end{aligned} \tag{2.8}$$

Let $S_j \equiv \xi_j^1$, one of the stored memory states, so that

$$\begin{aligned} \mu_i &= \frac{1}{N} \sum_{k=1}^{P} \sum_{j\neq i}^{N} \xi_i^k \xi_j^k \xi_j^1 \\ &= \frac{1}{N} \sum_{k=1}^{P} \xi_i^k \sum_{j\neq i}^{N} \xi_j^k \xi_j^1 \end{aligned} \tag{2.9}$$

$$= \frac{1}{N}\ \xi_i^1 \sum_{j\neq i}^{N} \xi_j^1 \xi_j^1 + \frac{1}{N} \sum_{k\neq 1}^{P} \xi_i^k \sum_{j\neq i}^{N} \xi_j^k \xi_j^1$$
$$= \frac{N-1}{N} \xi_j^1 + \frac{1}{N} \sum_{k\neq 1}^{P} \xi_i^k \sum_{j\neq i}^{N} \xi_j^k \xi_j^1$$

Thus, in the limit of large N, the first term leads to stability while the second term goes to zero, so that the average input is

$$< \mu_i > \ \simeq\ \xi_i^1 \tag{2.10}$$

Even when the second term for pattern 1 is not zero, the state $\vec{\xi^1}$ is stable if the magnitude of the second term is smaller than 1, i.e., if the second term cannot change the sign of the output S_i^l. It turns out that the second term is less than 1 in many cases of interest if P, the number of patterns, is sufficiently small. Then the stored patterns are all stable – if we start the system from one of these states the system will remain in that state. A small fraction of bits different from a stored pattern will be corrected in the same way as in the single-pattern case; they are overwhelmed in $\sum_{j\neq i}^{N} \sum_{k\neq l}^{P} W_{ij} S_j$ by the vast majority of correct bits. A configuration near to ξ_i^1 thus relaxes to ξ_i^1.

What is the variance, denoted σ^2, induced by the storage of many memories, the so-called structural noise? The second term consists of $(P-1)$ inner products of random vectors with $(N-1)$ terms. Each term is $+1$ or -1, i..e., binomially distributed, so that the fluctuation to the input is (see Box 1 for details):

$$\begin{aligned} \sigma \quad &= \frac{1}{N} \cdot \sqrt{P-1} \cdot \sqrt{N-1} \\ &\simeq \sqrt{\frac{P}{N}}. \end{aligned} \tag{2.11}$$

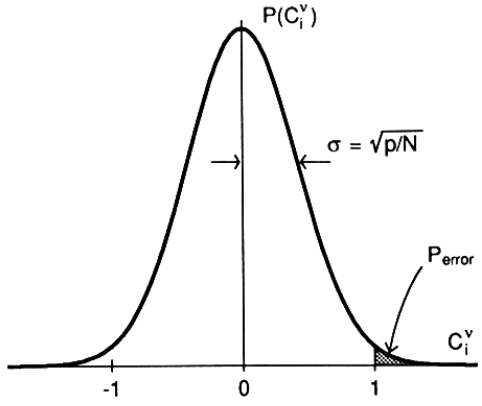
Noise hurts only if the magnitude of the noise term exceeds $\sigma = 1$. By the Central Limit Theorem, the noise becomes Gaussian for large P and N, but constant P/N (Figure 12). Thus the probability of an error in the recall of all stored states is

$$\begin{aligned} p_{error} \quad &= \frac{1}{\sqrt{2\pi}\ \sigma} \left[\int_{-\infty}^{-1} e^{-x^2/2\sigma^2}\ dx\ + \int_{+1}^{\infty} e^{-x^2/2\sigma^2}\ dx \right] \\ &= \frac{\sqrt{2}}{\sqrt{\pi}\ \sigma} \int_{+1}^{\infty} e^{-x^2/2\sigma^2}\ dx \\ &= \frac{2}{\sqrt{\pi}} \int_{\frac{1}{\sqrt{2}\sigma}}^{\infty} e^{-x^2}\ dx \\ &\equiv \text{erfc}\left(\frac{1}{\sqrt{2}\sigma}\right) \end{aligned} \tag{2.12}$$

where efrc(x) is the complementary error function and we again note that the average of the error term is zero. Thus

$$p_{\text{error}} = \text{erfc}\left(\sqrt{\frac{N}{2P}}\right). \tag{2.13}$$

Figure 12: We compute the probability in the tail of the Gaussian. From Hertz, Krogh and Palmer 1991.



For $N/P \gg 1$ the complementary error function may be approximated by an asymptotic closed form, whose leading term is given by

$$p_{\rm error} \simeq \frac{2}{\sqrt{\pi}} \frac{P}{N}\ e^{-N/2P} \tag{2.14}$$

so that to leading order

$$log\{p_{\rm error}\} \simeq -\frac{N}{2P} - \log\{\frac{N}{P}\}. \tag{2.15}$$

Now NP is total number of "bits" in the network. Suppose only less than one bit can be in error. Then we equate probabilities of correct to within a factor of one bit, or $1/(NP)$. Thus

$$1 - p_{\rm error} \geq 1 - \frac{1}{NP} \tag{2.16}$$

or

$$\log\{p_{\rm error}\} < -\log\{NP\}. \tag{2.17}$$

Thus

$$-\frac{N}{2P} - \log\{\frac{N}{P}\}\ < -\log\{NP\} \tag{2.18}$$

or

$$-\frac{N}{2P} < -2\log\{P\} \tag{2.19}$$

so

$$P\ <\ \frac{1}{4}\ \frac{N}{\log\{P\}}. \tag{2.20}$$

Since P scales sublinearly with N, we can iterate to write

$$P\ <\ \frac{1}{4}\ \frac{N}{\log\{N\}}. \tag{2.21}$$

Thus we see 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. Keep a mind that a linear network stores only one stable state, e.g., an integrator state. So things are looking good.

This is a worst case analysis that holds in the limit of $N \to \infty$. More typically we want to store states with a fixed, nonzero albeit small error rate. We will explore this possibility next and see if the scaling among P and N changes. This analysis makes us of statistical mechanics and starts with an energy description of the state-state of the Hopfield model (See Box 2 for this topic, which brings intuition as well as a path for calculations).

The essential ingredients for storing multiple stable output patterns in the same set of synaptic weights are

- Recursive dynamics.
- A nonlinear input-output relation. A linear input-output will support only a stable stable pattern independent of N (See Boxes 3 and 4)
- A bound on the number of stored states; we will nest see that this bound can be relaxed to $P \propto N$.

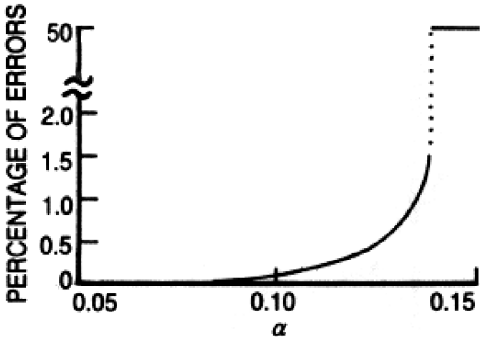
2.8 The phase diagram of the Hopfield model

This section was abstracted from Hertz, Krogh and Palmer (1991) 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

$$\alpha_C \equiv \frac{P}{N}|_{\text{critical}} \approx 0.138 \quad . \tag{2.22}$$

The jump in the number of memory states is considerable: from near-perfect recall to zero (Figure 13). This tells us that with no internal fast, or thermal, noise the system jumps discontinuously from a very good working memory with only a few bits in error for $\alpha < \alpha_C$ to a "useless" memory system for $\alpha > \alpha_C$.

Figure 13: The error rate upon retrieval for variance, T = 0. From Hertz, Krogh and Palmer 1991, following Amit, Gutfreund and Sompolinsky 1985.



The **phase diagram** for the Hopfield model delineates different regimes of behavior in the $Variance - \alpha$ plane (variance is σ^2 in our notation, but the statistical mechanics

literature uses T for temperature) (Figure 14). There is a roughly triangular region where the network functions as a memory device (Figure refphase1A,B). In region "A" the stored memory states form the absolute minima in the system. Their presence can be defined by the non-zero value of the order parameters

$$m^{\mu} \equiv \langle\langle \frac{1}{N} \sum_{i=1}^{N} \xi_i^{\mu} \langle S_i \rangle \rangle\rangle \tag{2.23}$$

where the averaging is over all configuration and tome (or noise). In region B the stored the memory states are still minima, but not absolute minima as "spin glass states" with zero overlap with the memory states are now the absolute minima.

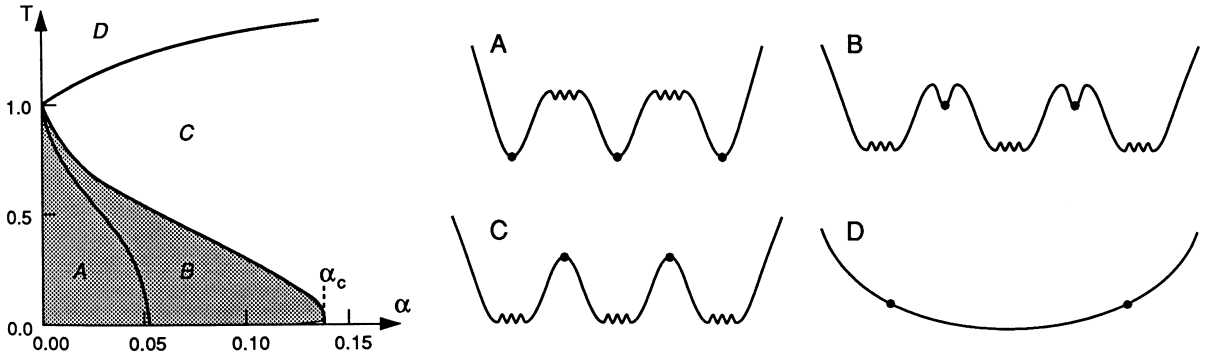
For small enough α and *variance* there are also mixture states that are correlated with an odd number of the patterns as discussed earlier. These always have higher free energy than the desired states. Each type of mixture state is stable in the triangular region defined by "A" and "B", but with smaller intercepts on both axes. The most stable mixture states, the triplets we discussed above, live within region "A" extend to 0.46 on the $Variance$ (T) axis and 0.03 on the α axis.

As we cross into region "C" the memory states are no longer attractors. There are only "spin glass states" and $m^{\mu} = 0 \forall \mu$. However, the network may be stuck network in a state, particularly as this phase encompasses $T = 0$. Thus an order parameter, q, that distinguished between a fixed or "frozen" system and one that perpetually drifts may not be zero, i.e.,

$$q \equiv \langle\langle \frac{1}{N} \sum_{i=1}^{N} \langle S_i \rangle^2 \rangle\rangle \tag{2.24}$$

In region D the network is completely ergodic, i.e., output of the network continuously fluctuates with $\langle S_i \rangle = 0$ and thus $q = 0$.

Figure 14: The phase diagram of the Hopfield model. From Hertz, Krogh and Palmer 1991, following Amit, Gutfreund and Sompolinsky 1985.

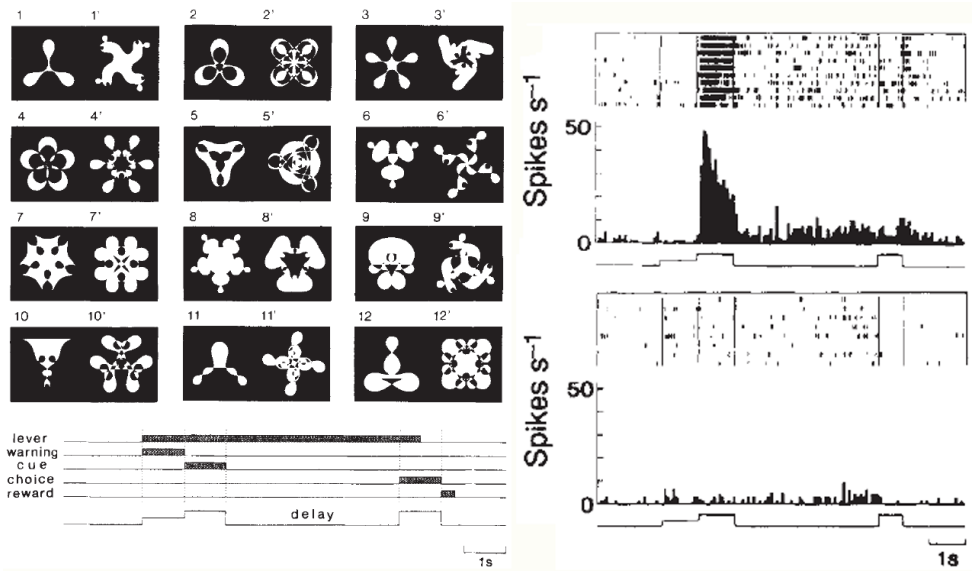


2.9 Can we relate stable states to a task?

The previous Neuropixels data by Carandini and Harris implies the presence of states (Figure 5), as did other data sets, 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. Here the monkey is asked to remember a picture and then, after a delay without visual input, compare a new picture with the old picture (Figure 15). The monkey signals if the two are part of a matched set.

Figure 15: Delayed match after sample task in monkey recording from IT cortex. From Sakai and Miyashita 1991



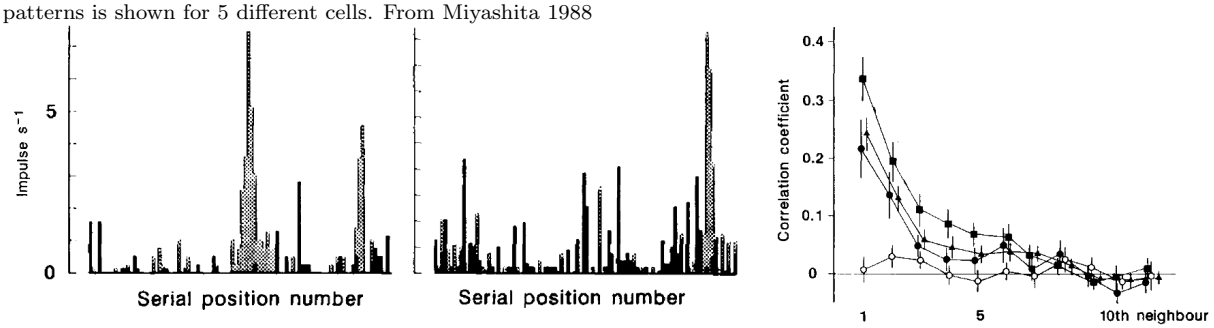
The spike rate of different neurons in inferiotemporal cortex cortex are measured while the monkey is performing this task. Critically, some neurons go up in their firing rate while others go done in rate. An interested observation is that activity continues throughout the period of the delay, for which there is no stimulus. This can occur for 20 seconds or more, i.e., one to two order 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.

These experiments also addressed an issue of coding. The visual patterns must be represented as a state, i.e., a pattern of activation across the neurons. Are these patterns statistically independent of each other, i.e., are their cross-correlations of order $1/\sqrt{N}$? Miyashita addressed this by looking at the likelihood of a neuron firing in response to different visual patterns. Interestingly, he found that the patterns of neuronal firing are related to the order of presentation of the visual images during training. Images next in sequence tend to have correlated firing patterns; the autocorrelation for five neurons decays to $1/e$ after three patterns (Figure 16),

The experimental correlation length of 3 to 4. In a theoretical work by Amit, Brunel and Tsodyks (1994) that followed this experimental work, a correlation length of 3 to 4was found adding a correlation term to the Hebbian learning rule, i.e.,

$$W_{ij} = \frac{1}{N}\sum_{k=1}^{P} \xi_i^k \xi_j^k + \frac{a}{N}\sum_{k=1}^{P} \xi_i^k \xi_j^{k+1} \quad . \tag{2.25}$$

Figure 16: Overlap of firing of two neurons for the fixed sequence for patterns used for training. The correlation across patterns is shown for 5 different cells. From Miyashita 1988



where $a < 1$; $a = 0.5$ was used in the published simulations. Let's just say that this is all very suggestive given the simplicity of the model.

2.10 Can we manipulate a stable state?

Recorded neuronal activity is not necessarily from brain regions that are part of the pathway that drives a task. While stimulation of one or a cluster of neighboring neurons has been shown to bias behavior in regions that map sensory stimuli to the cortical mantle, or map motor output, one can ask if manipulating a randomly represented state can lead to a change in behavior. Such an experiment was performed by Michael Hausser and colleagues, albeit attempted by others. They made us of two properties of hippocampus, a structure at the apex of internal loops in the brain (Figure 17). First, the circuitry in areas CA3 has heavy recurrence (Figure 18). Second, the neurons in this region respond to a specific location in space after they have run around to form a map of the regions (Figure 19), suggestive of the formation of an attractor. In fact, attractors models of the hippocampus are always in fashion.

Hausser recorded from neurons in hippocampus that responded to locations all along a virtual linear track (Figure 20). They selected on one location to focus their interest and stacked the deck by asking the mouse to lick at this location on the track, designated the reward location. Thus a readily observable behavior was linked to a place.

Hausser demonstrated that cells were excited at all phases along the virtual track (Figure 21). And that he could target cells for stimulation. Thus he could potentially initiate convergence to a state, more than less. During a test trial, Hausser stimulated a fraction (≈ 10) of the cells that responded at the place on the virtual track that the animal drank the reward, but during an earlier part of the run. He found that, indeed, stimulation led to licking (Figure 22. It was as though the animal thought it was at the reward location, although it was elsewhere.

All this is consistent with, but not a strong demonstration of, attractor networks. Yet we are still in need of experiment that probes the representation in the brain as it discriminated among a multitude of attractors.

Figure 17: Schematic of brain-wide circuitry centered on hippocampus.

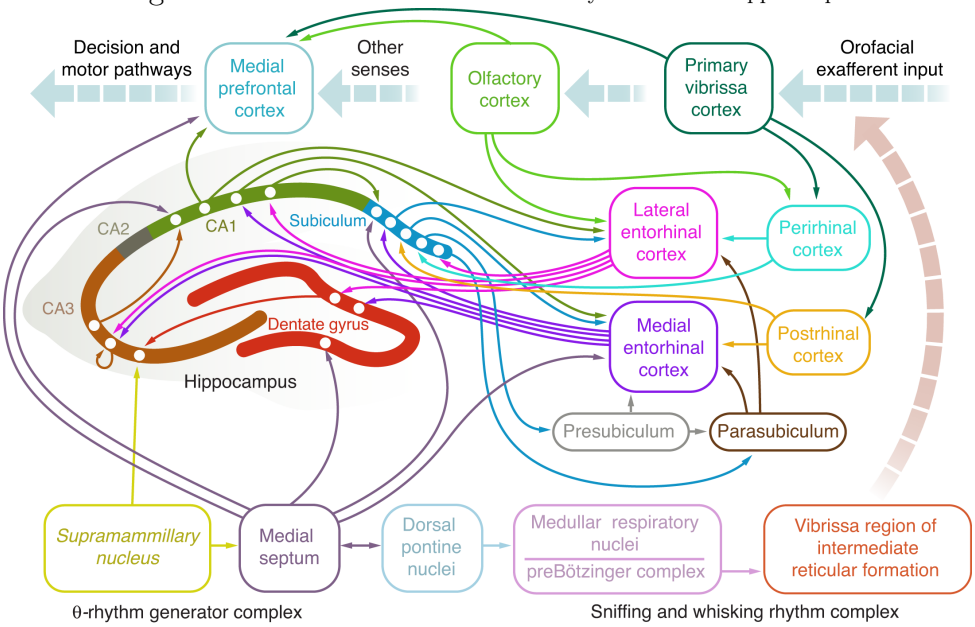
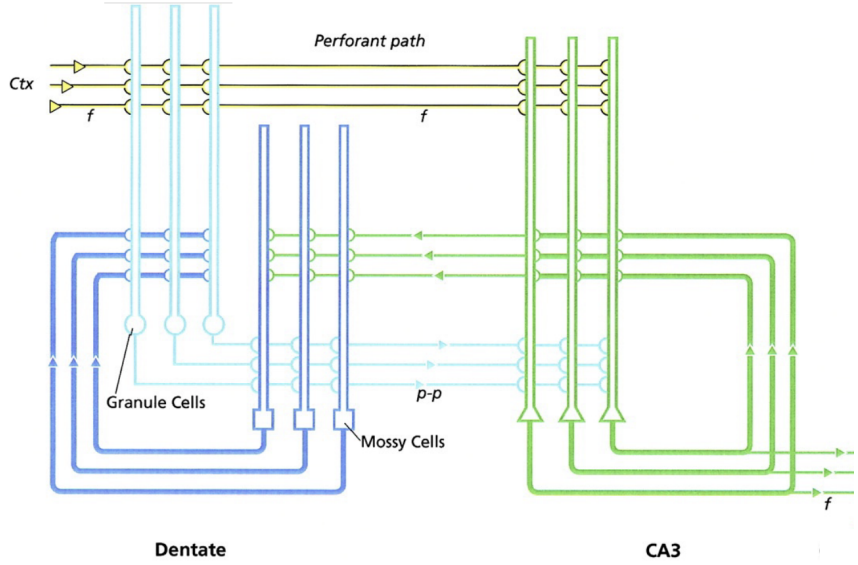


Figure 18: Schematic of brain circuitry centered on hippocampal area CA3.



2.11 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 (Figure 23). Epilepsy typically followed a loss or reduction in inhibition, so that a particularly simple model is a network with only

Figure 19: Intracellular recording from CA1 in the behaving, free-ranging mouse. From Lee, Manns, Sakmann and Brecht, 2006

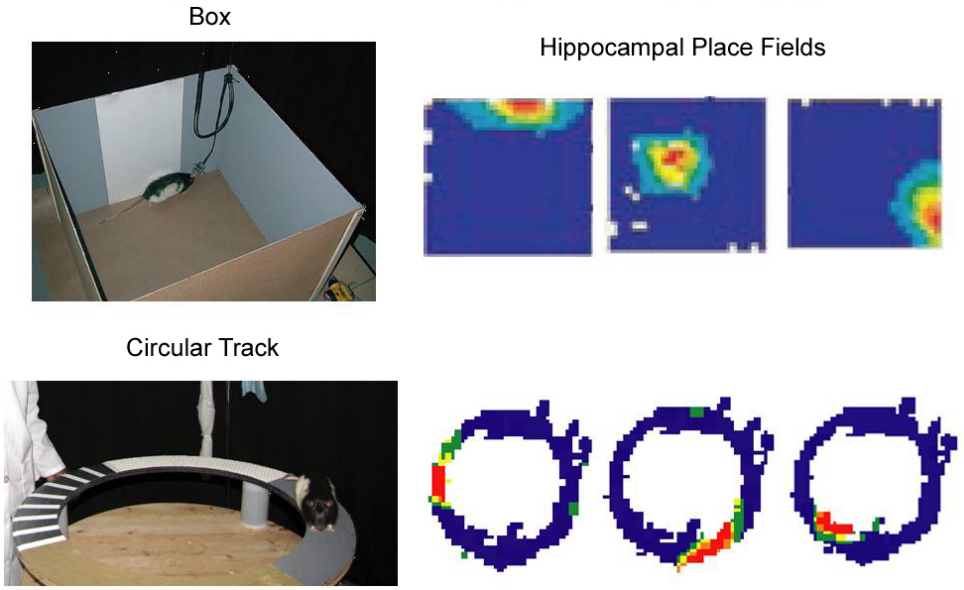
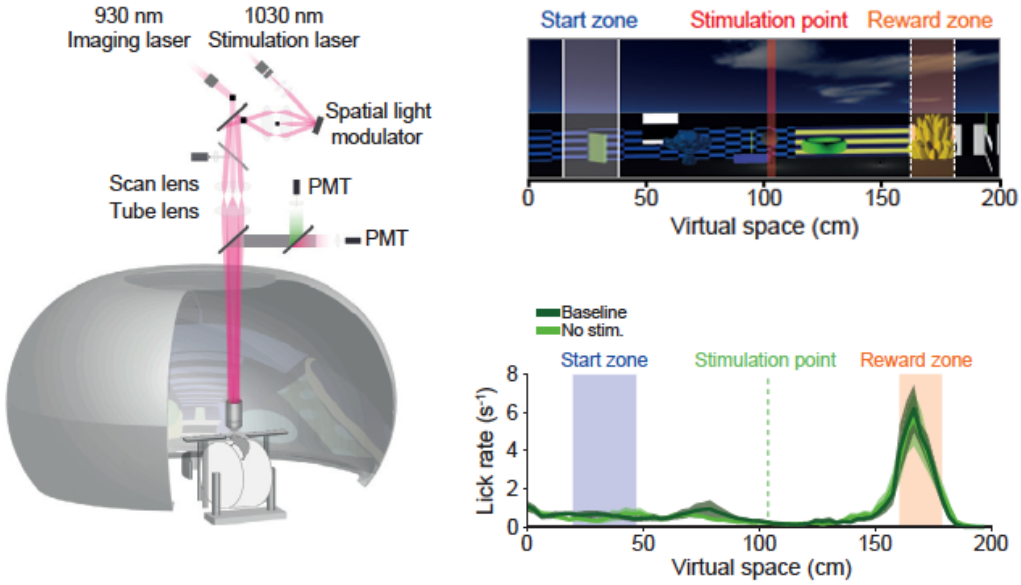


Figure 20: Set up of the virtual record and stimulation task. From Robinson, Descamps, Russell, Buchholz, Bicknell, Antonov, Lau, Nutbrown, Schmidt-Hieber and Hausser, 2020.



excitatory connections. 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 $N \gg 1$, 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 connections, *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

Figure 21: Imaging shows neurons that respond to all locations along the track. From Robinson, Descamps, Russell, Buchholz, Bicknell, Antonov, Lau, Nutbrown, Schmidt-Hieber and Hausser, 2020.

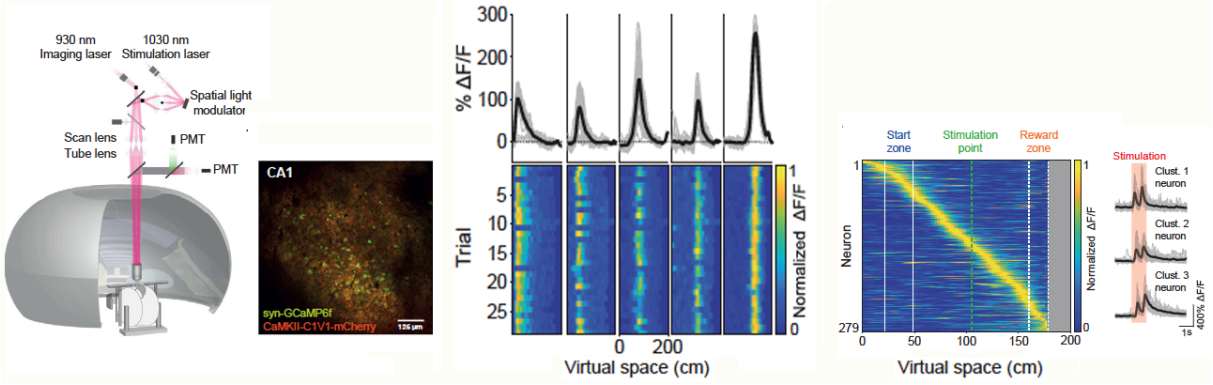


Figure 22: Stimulating about ten neurons normally active at the reward zone leads to enhanced licking at the time of stimulation. PC = place cell. From Robinson, Descamps, Russell, Buchholz, Bicknell, Antonov, Lau, Nutbrown, Schmidt-Hieber and Hausser, 2020.

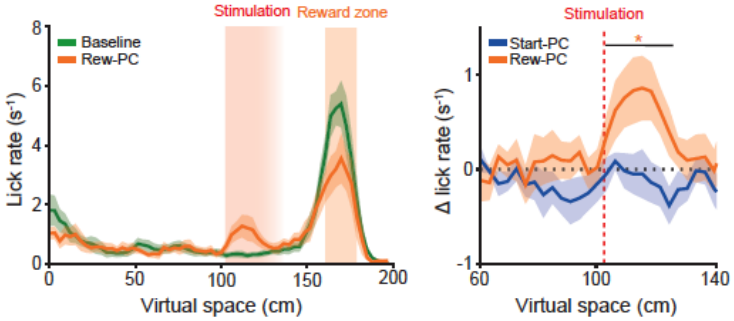
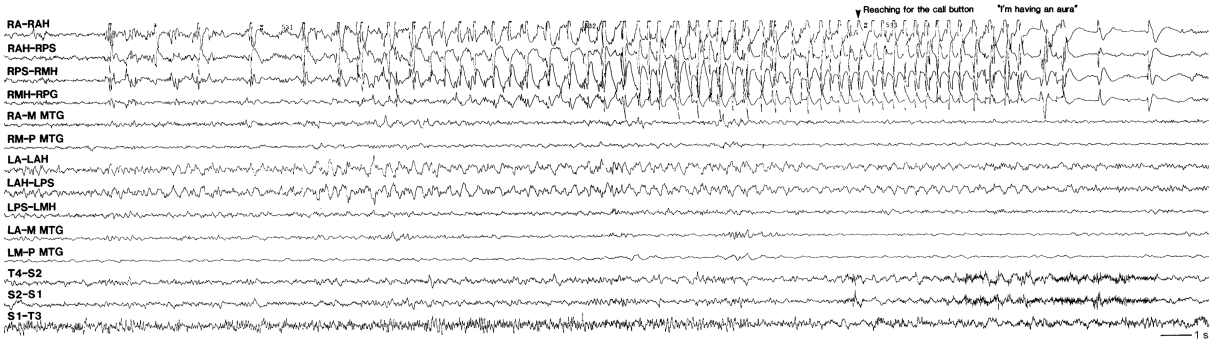


Figure 23: The onset of epilepsy recorded in the human brain with indwelling surface electrodes. From Engel 1993



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

$$\mu_i \;=\; W_0 \sum_{j=1}^{N} \; S_j \;+\; I^{ext}. \tag{2.26}$$

The energy per neuron, denoted ϵ_i, is then defined as

$$\begin{aligned} \epsilon_i \;&=\; -S_i \, \mu_i \\ &=\; -S_i \, W_0 \sum_{j=1}^{N} \; S_j \;-\; S_i \, I^{ext} \end{aligned} \tag{2.27}$$

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

$$< S > = \; \frac{1}{N} \sum_{j=1}^{N} \; S_j. \tag{2.28}$$

so that

$$\epsilon_i \;=\; -S_i \, (W_0 N < S > \; + I^{ext}). \tag{2.29}$$

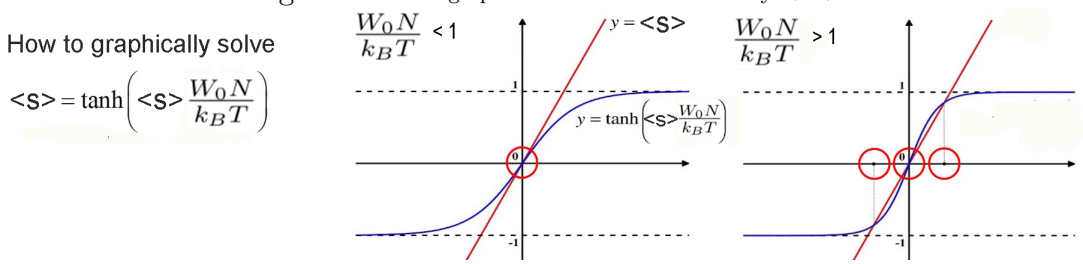
We can now use the expression for the value of the energy in term of the average spike rate, $< S >$, to solve self consistently for $< S >$. We know that the average rate is given by a Boltzman factor over all of the S_i. Thus

$$\begin{aligned} < S > \quad &= \frac{\sum_{S_i=\pm 1} \; S_i \; e^{-\epsilon_i / k_B T}}{\sum_{S_i=\pm 1} \; e^{-\epsilon_i / k_B T}} \\ &= \frac{\sum_{S_i=\pm 1} \; S_i \; e^{S_i (W_0 N < S > + I^{ext}) / k_B T}}{\sum_{S_i=\pm 1} \; e^{S_i (W_0 N < S > + I^{ext}) / k_B T}} \\ &= \frac{e^{-(W_0 N < S > + I^{ext}) / k_B T} \;-\; e^{(W_0 N < S > + I^{ext}) / k_B T}}{e^{-(W_0 N < S > + I^{ext}) / k_B T} \;+\; e^{(W_0 N < S > + I^{ext}) / k_B T}} \\ &= \tanh \left(\frac{W_0 N < S > + I^{ext}}{k_B T} \right). \end{aligned} \tag{2.30}$$

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 (Figure 24). We also see how the input-output function $tanh\{x\}$ naturally arises.

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

Figure 24: The graphical solution to the activity $< S >$.



- For sufficiently large $|I^{ext}|$ the network is pushed to a state with $< S >= sgn(I^{ext}/k_BT)$ 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 $W_0N/(k_BT) > 1$ or $k_BT < W_0N$. 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).

Box 1. A laboriously derivation of the variance

$$
\begin{aligned}
\sigma^2 \quad &= \frac{1}{N}\sum_{i=i}^{N}\left(\frac{1}{N}\sum_{k\neq 1}^{P}\xi_i^k\sum_{j\neq i}^{N}\xi_j^k\xi_j^1\right)\left(\frac{1}{N}\sum_{k'\neq 1}^{P}\xi_i^{k'}\sum_{j'\neq i}^{N}\xi_{j'}^{k'}\xi_{j'}^1\right) \qquad (2.31)\\
&= \frac{1}{N^3}\sum_{k\neq 1}^{P}\sum_{k'\neq 1}^{P}\left(\sum_{i=i}^{N}\xi_i^k\xi_i^{k'}\right)\sum_{j\neq i}^{N}\xi_j^k\xi_j^1\sum_{j'\neq i}^{N}\xi_{j'}^{k'}\xi_{j'}^1\\
&\xrightarrow{N\to\infty} \frac{1}{N^3}\sum_{k\neq 1}^{P}\sum_{k'\neq 1}^{P} N\ \delta(k-k')\ \sum_{j\neq i}^{N}\xi_j^k\xi_j^1\sum_{j'\neq i}^{N}\xi_{j'}^{k'}\xi_{j'}^1\\
&\xrightarrow{N\to\infty} \frac{1}{N^2}\sum_{k\neq 1}^{P}\ \sum_{j\neq i}^{N}\xi_j^k\xi_j^1\sum_{j'\neq i}^{N}\xi_{j'}^k\xi_{j'}^1\\
&\xrightarrow{N\to\infty} \frac{1}{N^2}\sum_{j\neq i}^{N}\xi_j^1\sum_{j'\neq i}^{N}\xi_{j'}^1\left(\sum_{k\neq 1}^{P}\ \xi_j^k\xi_{j'}^k\right)\\
&\xrightarrow{N\to\infty;\ P\to\infty} \frac{1}{N^2}\sum_{j\neq i}^{N}\xi_j^1\sum_{j'\neq i}^{N}\xi_{j'}^1\ (P-1)\ \delta(j-j')\\
&\xrightarrow{N\to\infty;\ P\to\infty} \frac{P-1}{N^2}\sum_{j\neq i}^{N}\left(\xi_j^1\right)^2\\
&\xrightarrow{N\to\infty;\ P\to\infty} \frac{(P-1)(N-1)}{N^2}\\
&\xrightarrow{N\to\infty;\ P\to\infty} \frac{P}{N}
\end{aligned}
$$

Box 2. Energy description and convergence

This advanced section was abstracted from Hertz, Krogh and Palmer (1991). 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, in which the connection strengths are *symmetric*, *i.e.*, $W_{ij} = W_{ji}$, the energy per neuron, ϵ_i, is

$$\epsilon_i = -\frac{1}{2}\, S_i \sum_{j;i\neq j}^{N} W_{ij} S_j \ . \tag{2.32}$$

and the total energy function E is

$$\begin{aligned} E &= \sum_{i}^{N} \epsilon_i \\ &= -\frac{1}{2} \sum_{ij;i\neq j}^{N} W_{ij} S_i S_j \ . \end{aligned} \tag{2.33}$$

The $i = j$ terms are of no consequence because $S_i^2 = 1$; we chose $W_{ii} = 0$ and in any case they just contribute a constant to E. The energy function depends on 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.

2.11.1 The energy never increases

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, which we associate with memorized patterns or so-called retrieval states, are at local minima of the energy surface (Figure 25A,B).

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

$$E = -\sum_{(ij)}^{N} W_{ij} S_i S_j + \text{constant} \tag{2.34}$$

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:

$$S_i' = sgn\left(\sum_{j\neq 1}^{N} W_{ij} S_j\right) \quad . \tag{2.35}$$

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

$$\begin{aligned} E' - E \quad &= -\sum_{j\neq i}^{N} W_{ij} S_i' S_j + \sum_{j\neq i}^{N} W_{ij} S_i S_j \\ &= 2S_i \sum_{j\neq i}^{N} W_{ij} S_j. \end{aligned} \tag{2.36}$$

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

A final point is that the addition of asymmetric connections will lead to flow that does not converge to a minimum. Weak asymmetry preserves converge to minima form small distances, but leads to drift for large distances from minima (Figure 25C)

2.11.2 Hebb minimizes the energy

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 ξ_i is largest. So we choose

$$E = -\frac{1}{2N}\sum_{k=1}^{P}\left(\sum_{i=1}^{N} S_i\xi_i^k\right)^2 \quad . \tag{2.37}$$

Multiplying this out gives

$$\begin{aligned} E \quad &= -\frac{1}{2N}\sum_{k=1}^{P}\left(\sum_{i=1}^{N} S_i\xi_i^k\right)\left(\sum_{j=1}^{N} S_j\xi_j^k\right) \\ &= -\frac{1}{2}\sum_{i\neq j}^{N}\left(\frac{1}{N}\sum_{k=1}^{P}\xi_i^k\xi_j^k\right)S_iS_j \end{aligned} \tag{2.38}$$

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_iS_j.

2.11.3 The issue of spurious attractors

We have shown that the Hebb prescription gives us, for small enough P, a dynamical system that has attractors, i.e., local minima of the energy function, i.e., for the desired states $\vec{\xi}^k$. But we have not shown that these are the only attractors. And indeed there are others, as discovered by Amit, Gottfried and Sompolinsky (1985).

- The reversed states $-\vec{\xi}^k$ 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.

- There are stable **mixture states** $\vec{\xi}^{mix}$, 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:

$$\xi_i^{mix} = sgn(\pm\xi_i^1 \pm \xi_i^2 \pm \xi_i^3) \quad . \tag{2.39}$$

  All $2^3 = 8$ sign combinations are possible, but we consider for definiteness the case where all the signs are chosen as +'s, i.e., $\xi_i^{mix} = sgn(\xi_i^1 + \xi_i^2 + \xi_i^3)$. The other cases

are similar. Observe that on average ξ_i^{mix} has the same sign at ξ_i^1 three times out of four; only if ξ_i^2 and ξ_i^3 both have the opposite sign is the overall sign be reversed. So ξ_i^{mix} is Hamming distance $N/4$ from ξ_i^1, and of course from ξ_i^2 and ξ_i^3 too; the mixture states lie at points equidistant from their components. This also implies that $\sum_i \xi_i^1 \xi_i^{mix} = 3N/4$ - $N/4 = N/2$ on average, as opposed to $\sum_i \xi_i^1 \xi_i^1 = N$, so the depth of the energy minimum is reduced by a factor of 4.

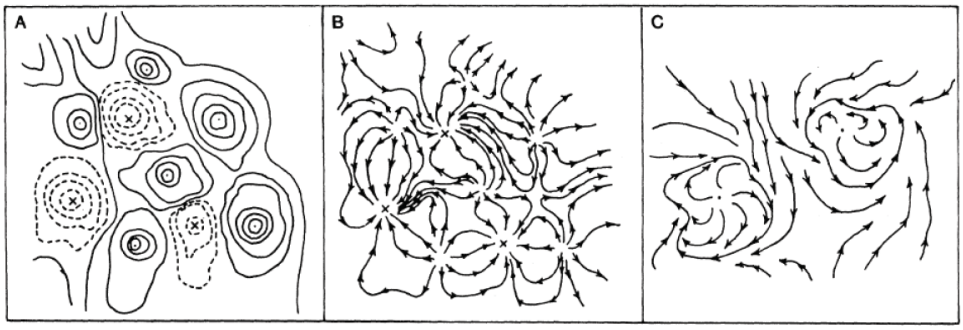
To check the stability, pick out the three special states with $k = 1$, 2, and 3, still with all + signs, to find:

$$\begin{aligned} \mu_i^{mix} &= \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1} \xi_i^k \xi_j^k \xi_j^{mix} \\ &= \frac{1}{2}\xi_i^1 + \frac{1}{2}\xi_i^2 + \frac{1}{2}\xi_i^3 + \text{cross terms} \quad . \end{aligned} \tag{2.40}$$

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 for some neurons, whereas the neurons must have nonzero inputs to have defined outputs of ± 1.

- For large P there are spurious local minima that are not correlated with any finite number of the original patters $\vec{\xi}^k$.

Figure 25: 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.



Box 3. Capacity of a linear network

How many states can be stored in a recurrent network with linear interactions? We make use of a parallel, clocked updating scheme in which we explicitly note the time steps, *i.e.*,

$$r_i(t) = \sum_{j=1}^{N} W_{ij} r_j(t-1). \tag{2.41}$$

In vector notation, this is

$$\vec{\mathbf{r}}(t) = \mathbf{W}\ \vec{\mathbf{r}}(t-1). \tag{2.42}$$

We now iterate, the synchronous equivalent of recurrence, starting from time $t = 0$:

$$\begin{aligned} \vec{\mathbf{r}}(1) &= \mathbf{W}\ \vec{\mathbf{r}}(0) \\ \vec{\mathbf{r}}(2) &= \mathbf{W}\ \vec{\mathbf{r}}(1) \\ \vec{\mathbf{r}}(3) &= \mathbf{W}\ \vec{\mathbf{r}}(2) \\ &\cdot \\ &\cdot \\ \vec{\mathbf{r}}(n) &= \mathbf{W}\ \vec{\mathbf{r}}(n-1). \end{aligned} \tag{2.43}$$

This becomes

$$\vec{\mathbf{r}}(n) = \mathbf{W}^n\ \vec{\mathbf{r}}(0). \tag{2.44}$$

Noting the unitary transform (Box 4)

$$\mathbf{W} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathbf{T}} \tag{2.45}$$

the iterative expression for $\vec{\mathbf{r}}(n)$ becomes,

$$\begin{aligned} \vec{\mathbf{r}}(n) \quad &= \left(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathbf{T}}\right)^n\ \vec{\mathbf{r}}(0) \\ &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathbf{T}}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathbf{T}} \cdots \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathbf{T}}\ \vec{\mathbf{r}}(0) \\ &= \mathbf{U}\mathbf{\Lambda}^{\mathbf{n}}\mathbf{U}^{\mathbf{T}}\ \vec{\mathbf{r}}(0). \end{aligned} \tag{2.46}$$

But the diagonal matrix $\mathbf{\Lambda^n}$, when rank ordered so that λ_1 is the dominant eigenvalue, becomes

$$\mathbf{\Lambda}^n = \begin{pmatrix} \lambda_1^n & 0 & 0 & \cdots \\ 0 & \lambda_2^n & 0 & \\ 0 & 0 & \lambda_3^n & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \end{pmatrix} = \lambda_1^n \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & \left(\frac{\lambda_2}{\lambda_1}\right)^n & 0 & \\ 0 & 0 & \left(\frac{\lambda_3}{\lambda_1}\right)^n & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \end{pmatrix} \overrightarrow{n \to \infty}\ \lambda_1^n \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \end{pmatrix}.$$

Thus the system converges to a numerical factor times the dominant eigenvector of $\mathbf{W}$, i.e.,

$$\vec{\mathbf{r}}(n)\ \overrightarrow{n \to \infty}\ \lambda_1^n \begin{pmatrix} \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \vec{\mu}_1 & \vec{\mu}_2 & \cdots & \vec{\mu}_N \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \end{pmatrix} \begin{pmatrix} \cdots & \vec{\mu}_1 & \cdots \\ \cdots & \vec{\mu}_2 & \cdots \\ \cdots & \vec{\mu}_3 & \cdots \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ \cdots & \vec{\mu}_N & \cdots \end{pmatrix} \vec{\mathbf{r}}(0)$$

which becomes

$$\vec{\mathbf{r}}(n)\ \overrightarrow{n \to \infty}\ \lambda_1^n\ [\vec{\mu}_1 \cdot \vec{\mathbf{r}}(0)]\ \vec{\mu}_1 \tag{2.47}$$

and thus only a single state is supported in an iterative network comprised of linear neurons.

Box 4. Review of Unitary Transforms

Recall that a matrix $\mathbf{W}$ satisfies an eigenvalue equation

$$\mathbf{W}\ \vec{\mu}_k = \lambda_k \vec{\mu}_k \tag{2.48}$$

where k labels labels the eigenvalue with $k = 1, ..., N$ and includes the case of potential degenerate eigenvectors. The eigenvalues are real numbers when $\mathbf{W}$ is a symmetric matrix whose elements are real. The spectral theorem states that a symmetric matrix whose elements are real can be diagonalized by a matrix transformation by a unity transformation that rotates $\mathbf{W}$ and preserves the eigenvalues, *i.e.*,

$$\mathbf{W} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathbf{T}} \tag{2.49}$$

where $\mathbf{U}$ is a unitary matrix defined through $\mathbf{U}\mathbf{U}^{\mathbf{T}} = \mathbf{I}$ and $\det(\mathbf{U}) = 1$. Each column in $\mathbf{U}$ is one of the eigenvectors $\vec{\mu}_k$, i.e.,

$$\mathbf{U} = \begin{pmatrix} \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \vec{\mu}_1 & \vec{\mu}_2 & \cdots & \vec{\mu}_N \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \end{pmatrix} \quad \text{and} \quad \mathbf{U}^{\mathbf{T}} = \begin{pmatrix} \cdots & \vec{\mu}_1 & \cdots \\ \cdots & \vec{\mu}_2 & \cdots \\ \cdots & \vec{\mu}_3 & \cdots \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ \cdots & \vec{\mu}_N & \cdots \end{pmatrix}$$

and the rotated eigenvectors, $\mathbf{U}^{\mathbf{T}}\vec{\mu}$, are of the form

$$\mathbf{U}^{\mathbf{T}}\vec{\mu}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \end{pmatrix} \qquad \mathbf{U}^{\mathbf{T}}\vec{\mu}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \end{pmatrix} \cdot ..$$

since $\mathbf{W}\ \vec{\mu}_k = \lambda_k \vec{\mu}_k$ implies $\mathbf{\Lambda}\ \mathbf{U}^{\mathbf{T}}\vec{\mu}_k = \lambda_k \mathbf{U}^{\mathbf{T}}\vec{\mu}_k$, the $\mathbf{U}^{\mathbf{T}}\vec{\mu}_k$ are the eigenvectors of the diagonalized (rotated) system. The diagonal matrix $\mathbf{\Lambda}$ contains the eigenvalues along the diagonal, such that

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \\ 0 & 0 & \lambda_3 & \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \end{pmatrix}$$