Physics 178/278 - David Kleinfeld - Winter 2022

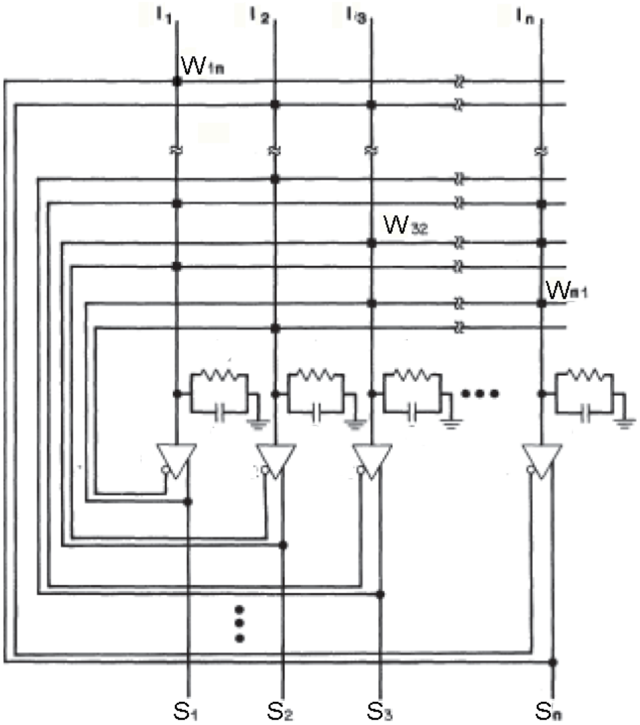
Lecture 3
Revised 13 January 2022 01:20

3 Recurrent neuronal networks: Associative memory II

3.1

In the last session, we found that an associate memory based on a recurrent Hopfield network (figure 1)can store a number of memories that scales more weakly than the number of neurons, i.e., $P \propto N/log(N)$, if one can tolerate at most a single error upon recall. This scaling was for of "worst case" analysis, the kind typically associated with computer science. Today we will review the dynamics to learn some of the quirks of this system, mainly through the perspective of an Energy function, then consider the form of the scaling for a "typical case", where a fraction of errors are tolerated; this is the kind of analysis associated with statistical physics.

Figure 1: The Hopfield recurrent network. From Hertz, Krogh and Palmer 1991 following Hopfield 1982



3.2 Energy description and convergence

This section was abstracted from Ch. 2 of "Introduction to the Theory of Neural Computation" (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

$$E = -\frac{1}{2} \sum_{ij;i\neq j}^{N} W_{ij} S_i S_j \quad . \tag{3.1}$$

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 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.

3.2.1 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 2A,B). 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, although experimentally symmetric synapses occur more than expected by chance (Figure 3). Nonetheless, 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.

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.

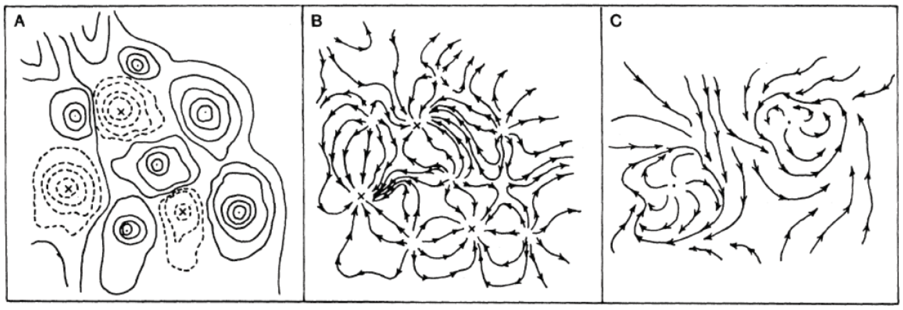
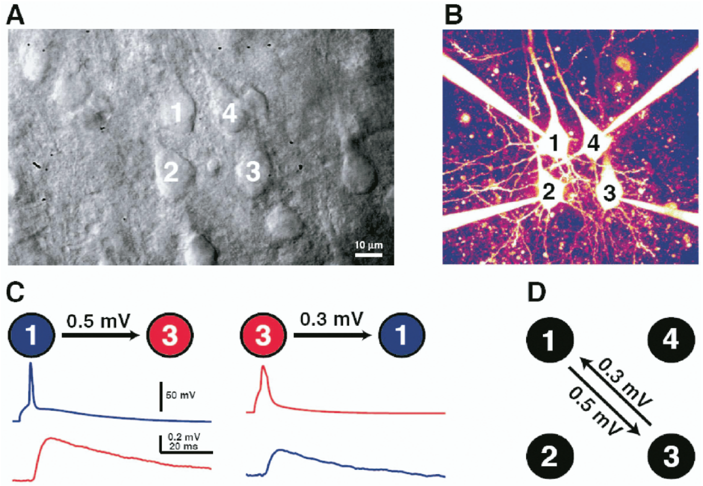


Figure 3: Experimental evidence for symmetric synapses based on pairwise recordings form L5 pyramidal neurons in mouse brain slice. Only one pair of neurons are connected in these measurements. From Song, Sjostrom, Reigl, Nelson and Chklovskii, 2005



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

$$E = -\sum_{(ij)}^{N} W_{ij} S_i S_j + \text{constant} \tag{3.2}$$

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:

$$S_i' = sgn\left(\sum_{j\neq 1}^{N} W_{ij} S_j\right) \quad . \tag{3.3}$$

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

$$\begin{aligned} E' - E \quad &= -\sum_{j\neq i}^{N} W_{ij} S_i' S_j + \sum_{j\neq i}^{N} W_{ij} S_i S_j \\ &= 2S_i \sum_{j\neq i}^{N} W_{ij} S_j . \end{aligned} \tag{3.4}$$

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 2C)

3.2.2 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

$$E = -\frac{1}{2N} \sum_{k=1}^{P} \left(\sum_{i=1}^{N} S_i \xi_i^k \right)^2 \quad . \tag{3.5}$$

Multiplying this out gives

$$\begin{aligned} E &= -\frac{1}{2N} \sum_{k=1}^{P} \left(\sum_{i=1}^{N} S_i \xi_i^k \right) \left(\sum_{j=1}^{N} S_j \xi_j^k \right) \qquad (3.6) \\ &= -\frac{1}{2} \sum_{i \neq j}^{N} \left(\frac{1}{N} \sum_{k=1}^{P} \xi_i^k \xi_j^k \right) S_i S_j \end{aligned}$$

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 $S_i S_j$.

3.2.3 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:

$$\xi_i^{mix} = sgn(\pm\xi_i^1 \pm \xi_i^2 \pm \xi_i^3) \quad . \tag{3.7}$$

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:

$$\begin{aligned} \mu_i^{mix} &= \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1} \xi_i^k \xi_j^k \xi_j^{mix} \\ &= \frac{1}{2}\xi_i^1 + \frac{1}{2}\xi_i^2 + \frac{1}{2}\xi_i^3 + \text{crossterms} \quad . \end{aligned} \tag{3.8}$$

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 $\vec{\xi^k}$.

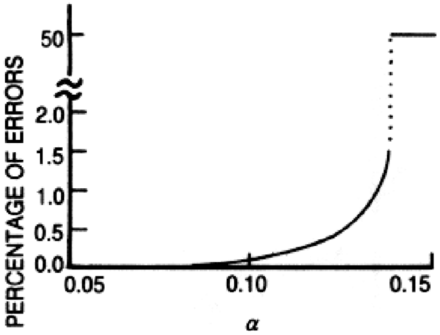
3.3 The phase diagram of the Hopfield model

This section was abstracted from Ch. 2 of "Introduction to the Theory of Neural Computation" (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

$$\alpha_C \equiv \frac{P}{N}|_{\text{critical}} \approx 0.138 \quad . \tag{3.9}$$

The jump in the number of memory states is considerable: from near-perfect recall to zero (Figure 4). This tells us that with no internal fast, or thermal, noise the system jups 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 σ^2 in our notation, but the

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



statistical mechanics literature uses T for temperature) (Figure 5). There is a roughly triangular region where the network functions as a memory device (Figure 5A,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

$$m^{\mu} \equiv \langle\langle \frac{1}{N} \sum_{i=1}^{N} \xi_i^{\mu} \langle S_i \rangle \rangle\rangle \tag{3.10}$$

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.

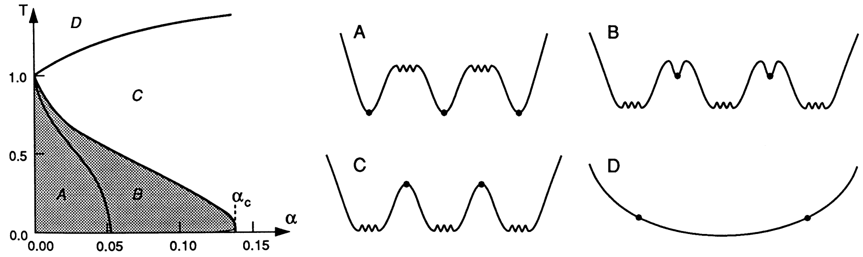
For small enough α and *variance* there are also mixture states that are correlated with an odd number of the patterns as discussed earlier. These always have higher free energy than the desired states. Each type of mixture state is stable in the triangular region defined by "A" and "B", but with smaller 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 that distinguished between a fixed or "frozen" system and one that perpetually drifts may not be zero, i.e.,

$$q \equiv \langle\langle \frac{1}{N} \sum_{i=1}^{N} \langle S_i \rangle^2 \rangle\rangle \tag{3.11}$$

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

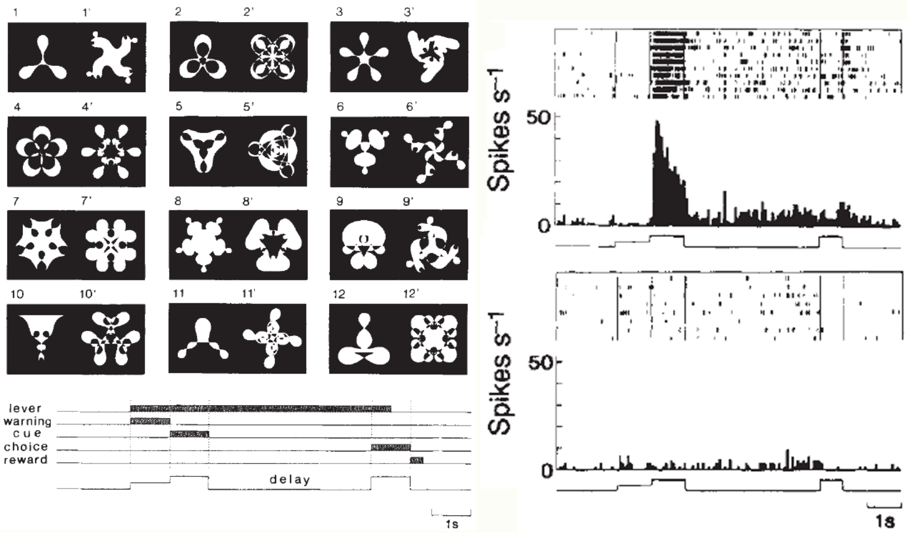
Figure 5: The phase diagram of the Hopfield model. From Hertz, Krogh and Palmer 1991, following Amit, Gutfreund and Sompolinsky 1985.



3.4 Can we relate stable states to a task?

The previous Neuropixels data by Carandini and Harris implies states, 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 6). The monkey signals if the two are part of a matched set.

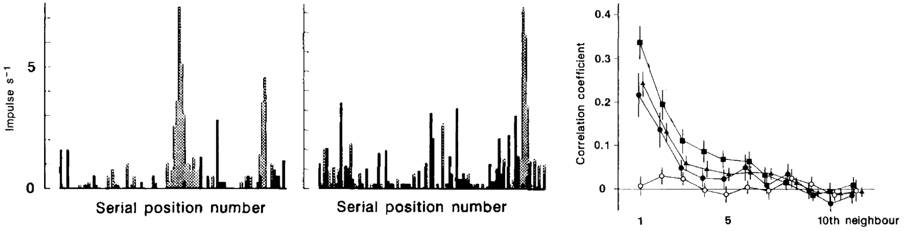
Figure 6: Delayed match after sample task in monkey recording from IT cortex. From Sakai and Miyashita 1991



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 through-out 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 7), 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.,

Figure 7: Overlap of firing of two neurons for the fixed sequence for patterns used for training. The Correlation is shown for 5 different cells. From Miyashita 1988



$$W_{ij} = \frac{1}{N}\sum_{k=1}^{P}\xi_i^k\xi_j^k + \frac{a}{N}\sum_{k=1}^{P}\xi_i^k\xi_j^{k+1} \quad . \tag{3.12}$$

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.

3.5 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 8). First, the circuitry in areas CA3 has heavy recurrence (Figure 9). 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 10), suggestive of the formation of an attractor. In fact, attractors models of the hippocampus are always in fashion.

Figure 8: Schematic of brain-wide circuitry centered on hippocampus.

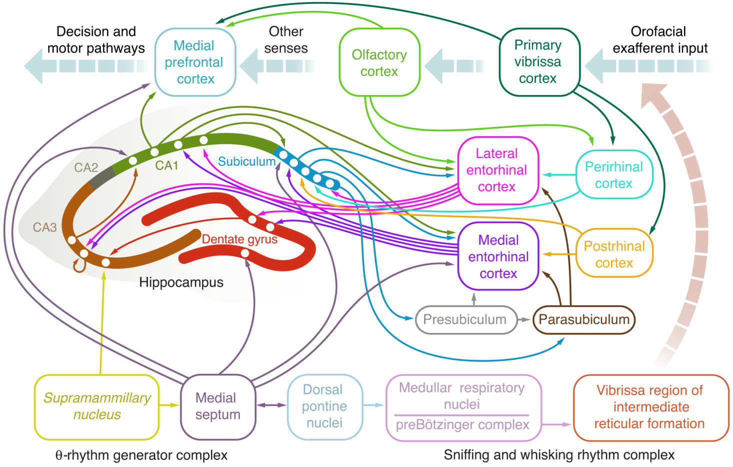
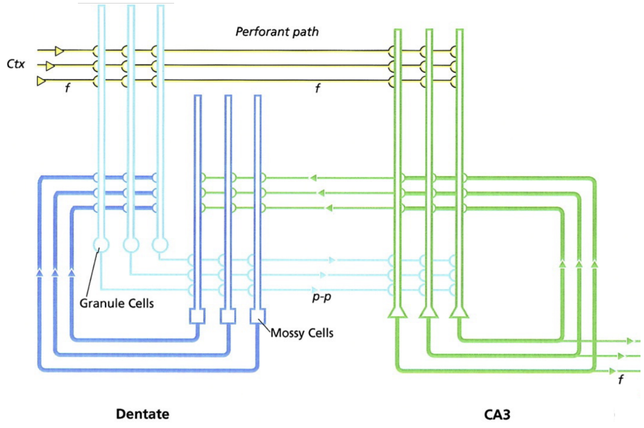
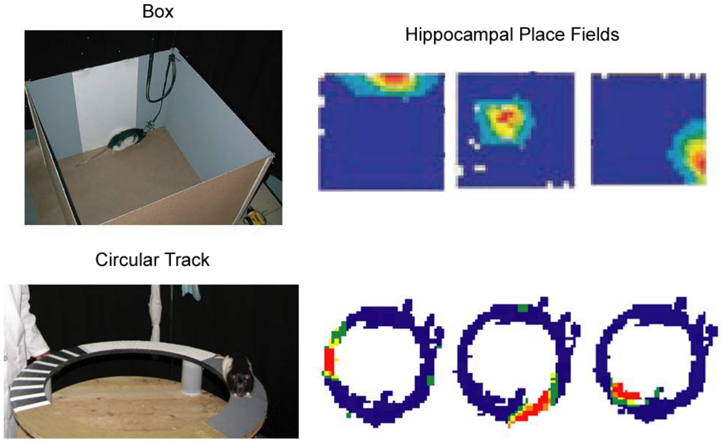


Figure 9: Schematic of brain circuitry centered on hippocampal area CA3.



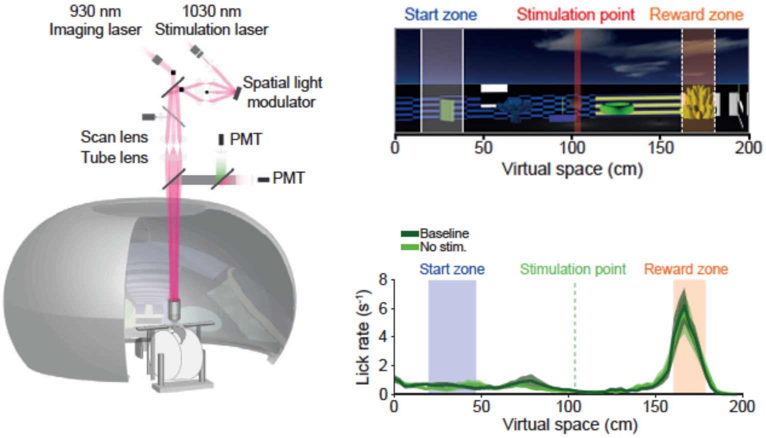
Hausser recorded from neurons in hippocampus that responded to locations all along a virtual linear track (Figure 11). They selected on one location to focus their interest and stacked the deck

Figure 10: Intracellular recording from CA1 in the behaving, free-ranging mouse. From Lee, Manns, Sakmann and Brecht, 2006



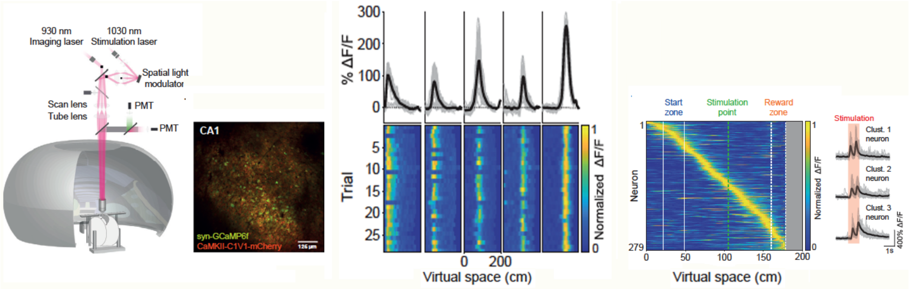
by asking the mouse to lick at this location on the track, designated the reward location. Thus a readily observable behavior was linked to a place.

Figure 11: Set up of virtual record and stimulation task. From Robinson, Descamps,Russell, Buchholz, Bicknell, Antonov, Lau, Nutbrown, Schmidt-Hieber and Hausser, 2020.



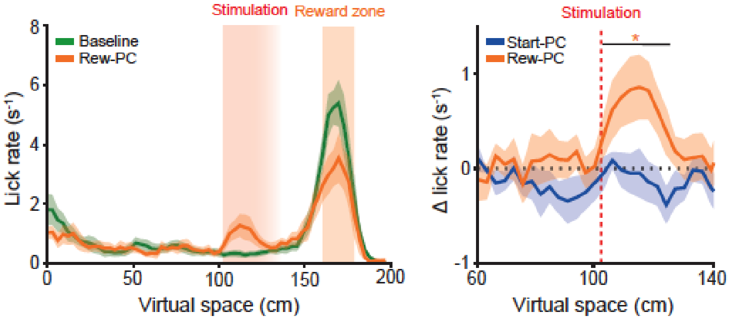
Hausser demonstrated that cells were excited at all phases along the virtual track (Figure 12). And that he could target cells for stimulation. Thus he could potentially initiate convergence to a state, more than less. During a test trial, Hausser stimulated a fraction (≈ 10) of the cells that responded 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 13. It was as though the animal thought it was at the reward location, although it was elsewhere.

Figure 12: Imaing shows neurns that respond to all locations along the track. From Robinson, Descamps,Russell, Buchholz, Bicknell, Antonov, Lau, Nutbrown, Schmidt-Hieber and Hausser, 2020.



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 13: 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.

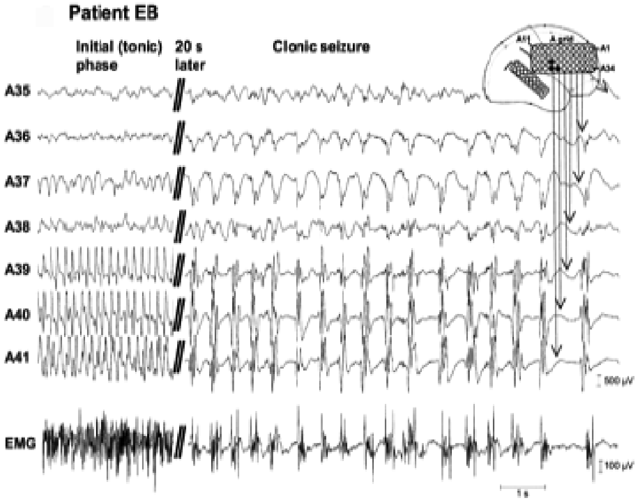


3.6 Noise and spontaneous excitatory states as a model for epilepsy

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 (Figure 14). Epilepsy typically followed a loss or reduction in inhibition, so that a particularly simple model is a network with only 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

Figure 14: The onset of epilepsy recorded in the human brain with indwelling surface electrodes. From Hamer, LuEders, Knake, Fritsch, Oertel and Rosenow 2003



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 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

$$\mu_i \;=\; W_0 \sum_{j=1}^{N} S_j \;+\; I^{ext}. \tag{3.13}$$

The energy per neuron, denoted ϵ_i, is then

$$\begin{aligned}\epsilon_i \;&=\; -S_i\,\mu_i \qquad (3.14)\\ &=\; -S_i\,W_0 \sum_{j=1}^{N} S_j \;-\; S_i\,I^{ext}\end{aligned}$$

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

$$< S > = \frac{1}{N} \sum_{j=1}^{N} S_j. \tag{3.15}$$

so that

$$\epsilon_i \;=\; -S_i\,(W_0 N < S > \;+ I^{ext}). \tag{3.16}$$

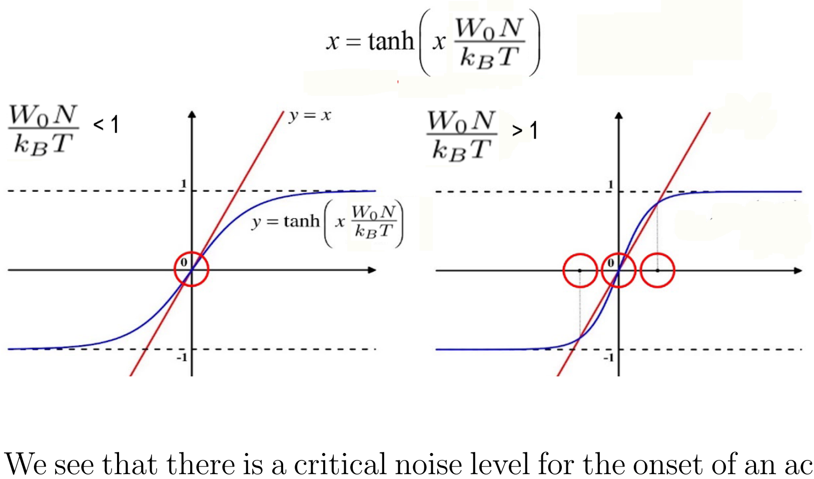
We can now use the expression for the value of the energy in term of the average spike rate, $< S >$, to solve self consistently for $< S >$. We know that the average rate is given by a Boltzman factor over all of the S_i. Thus

$$\begin{aligned} < S > \quad &= \frac{\sum_{S_i=\pm 1} \; S_i \; e^{-\epsilon_i/k_BT}}{\sum_{S_i=\pm 1} \; e^{-\epsilon_i/k_BT}} \qquad (3.17) \\ &= \frac{\sum_{S_i=\pm 1} \; S_i \; e^{S_i(W_0N<S>+I^{ext})/k_BT}}{\sum_{S_i=\pm 1} \; e^{S_i(W_0N<S>+I^{ext})/k_BT}} \\ &= \frac{e^{-(W_0N<S>+I^{ext})/k_BT} \; - \; e^{(W_0N<S>+I^{ext})/k_BT}}{e^{-(W_0N<S>+I^{ext})/k_BT} \; + \; e^{(W_0N<S>+I^{ext})/k_BT}} \\ &= \tanh\left(\frac{W_0N < S > +I^{ext}}{k_BT}\right). \end{aligned}$$

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 $W_0N/(k_BT)$, which pits the connection strength W_0 against the noise level T/N (Figure 15). We also see how the input-output function $tanh\{x\}$ naturally arises.

- For $W_0N/(k_BT) < 1$, the high noise limit, there is only the solution $< S >= 0$ in the absence of an external input h_0.
- For $W_0N/(k_BT) > 1$, 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 $< S >= sgn(I^{ext}/k_BT)$ independent of the interactions.

Figure 15: The graphical solution to the activity, denoted x rather than $< S >$ in the figure.



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).