17
Memory and attractor dynamics

Humans remember important events in their lives. You might be able to recall every detail of your first exam at college, or of your first public speech, or of your first day in kindergarten, or of the first time you went to a new school after your family moved to a new city. Human memory works with associations. If you hear the voice of an old friend on the phone, you may spontaneously recall stories that you had not thought of for years. If you are hungry and see a picture of a banana, you might vividly recall the taste and smell of a banana ... and thereby realize that you are indeed hungry.

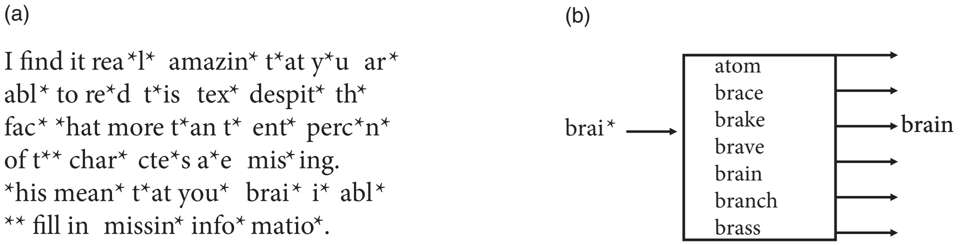
In this chapter, we present models of neural networks that describe the recall of previously stored items from memory. In Section 17.1 we start with a few examples of associative recall to prepare the stage for the modeling work later on. In Section 17.2 we introduce an abstract network model of memory recall, known as the Hopfield model. We take this network as a starting point and add, in subsequent sections, some biological realism to the model.

17.1 Associations and memory

A well-known demonstration of the strong associations which are deeply embedded in the human brain is given by the following task. The aim is to respond as quickly as possible to three questions. Think of the first answer that comes to mind! Are you ready? Here are the questions: (i) Can you give me an example of a color? (ii) Can you give me an example of a tool? (iii) Can you give me an example of a fruit? For each of these, what was the very first example that came to your mind? Chances are high that your examples are “red” for color and “hammer” for tool. In fact, most humans have particularly strong associations from tool to hammer and from color to red. Regarding fruit, the cultural background plays a more important role (apple, orange, banana), but since the text at the beginning of this chapter mentioned bananas, you probably had a slightly stronger bias toward banana at the moment when you answered the above questions than what you would have had under normal circumstances. This bias through an earlier word or context is a highly significant effect, called “priming” in psychophysics.

Not surprisingly, the word “red” is associated with seeing the color red and vice versa. If you read a list of words that contains names of colors, you are normally fast in doing

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press



I find it rea*l* amazin* t*at y*u ar*
abl* to re*d t*is tex* despit* th*
fac* *hat more t*an t* ent* perc*n*
of t** char* cte*s a*e mis*ing.
his mean t*at you* brai* i* abl*
** fill in missin* info* matio*.

Fig. 17.1 Memory recall cued by partial information. (a) Read it! (b) Schematic view of the recall process. Your brain has memorized a list of words. Based on partial information and the context, your brain is able to complete the missing characters.

so and do not experience any particular difficulty. Similarly, you can easily name the color of objects. However, people find it difficult to name the ink color in lists of words that contain entries such as *red, green, blue*, but are written in colors that are inconsistent with the word (e.g., the word *red* is written in green, whereas the word *green* is written in blue). In this case responses in the color-naming task are slower compared to naming the color of geometric objects. The measurable difference in reaction time in naming the color of (inconsistent) words compared to the color of objects is called the Stroop effect (Stroop, 1935; MacLeod, 1991). The association of the color "red" with the word *red* makes it difficult to name the ink color (e.g., green) in which the word *red* is written.

In this chapter we mainly focus on association in the sense of completing partial information. Take a look at Fig. 17.1a. Nearly all words are incomplete, but your brain is able to cope with this situation, just as you are able to follow a phone conversation over a noisy line, recognize a noisy image of a handwritten character or associate the picture of an orange with its taste to retrieve your concept of an orange as a tasty fruit.

17.1.1 Recall, recognition, and partial information

If half of an orange is hidden behind a coffee mug, you can still recognize it as an orange based on the partial information you have. Recognition works because you have seen oranges before and have memorized the concept "orange," including a prototypical image of this fruit. More generally, when you see a noisy image of a known object (e.g., the letter "T") your brain is able to retrieve from memory the prototype version of the object (e.g., an idealized "T"). Thus recognizing an object in a noisy environment involves the process of "memory recall."

A highly simplified schematic view of memory recall based on partial information is shown in Fig. 17.1b. The input (e.g., an incomplete word) is compared to a list of all possible words. The most likely entry (i.e., the one which is most similar to the input) in the list is given as the output of memory recall.

Similarly, noisy images of objects are recognized if the brain finds, among the memorized items, one which is highly similar (Fig. 17.2). Let us call the "pure" noise-free memory item a prototype p^{μ}, where the index $1 \leq \mu \leq M$ labels all different memory

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

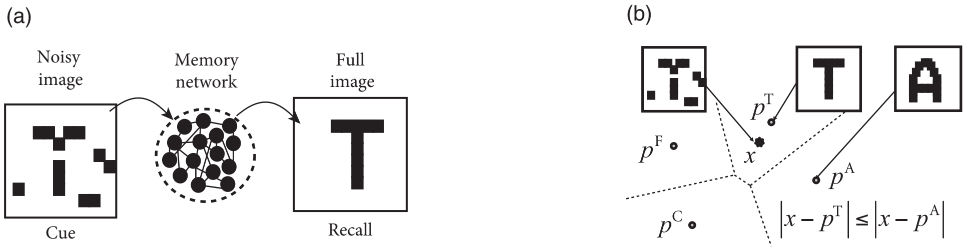


Fig. 17.2 Recall and recognition as search for nearest prototype. (a) A letter “T” in a noisy image (left) serves as a cue in order to recall a noise-free prototype letter “T” from the memory embedded in a neural network. (b) Recognition of the input x (black star, representing a noisy “T”) can be interpreted as an algorithm that searches for the nearest prototype p^α such that $|x - p^\alpha| \leq |x - p^\mu|$ for all μ, and p^μ denotes all possible prototypes (gray circles). The dashed lines are the sets of points with equal distance to two different prototypes.

items. The prototype can be visualized as a point in some high-dimensional space. A noisy input cue x corresponds to another point in the same space. Suppose that we have a similarity measure which enables us to calculate the distance $|x - p^\mu|$ between the input cue and each of the prototypes. A simple method of memory recall is a search algorithm that goes through the list of all available prototypes to find the nearest one. More formally, the output of the recall process is the prototype p^α with

$$|x - p^\alpha| \leq |x - p^\mu| \quad \text{for all } \mu\,, \tag{17.1}$$

which gives rise to a simple geometric picture (Fig. 17.2b).

The aim of this chapter is to replace the explicit algorithmic search for the nearest prototype by the dynamics of interacting neurons. Instead of an *explicit* algorithm working through a list of stored prototypes, the mere cross-talk of neurons embedded in a large network will find the prototype that corresponds best to the noisy cue – in a highly distributed and automatic fashion, reminiscent of what we believe is happening in the brain (Fig. 17.2a). Brain-style computation implements an *implicit* algorithm, as we shall see in Sections 17.2 and 17.3.

17.1.2 Neuronal assemblies

Neural assemblies play a central role in the implicit algorithm for memory retrieval that we will discuss in Section 17.2. Neuronal assemblies (Hebb, 1949) are subnetworks of strongly connected neurons that, together, represent an abstract concept. For example, your mental concept of a “banana” containing the mental image of its form, color, taste, and texture could be represented by one assembly of strongly connected neurons, while another

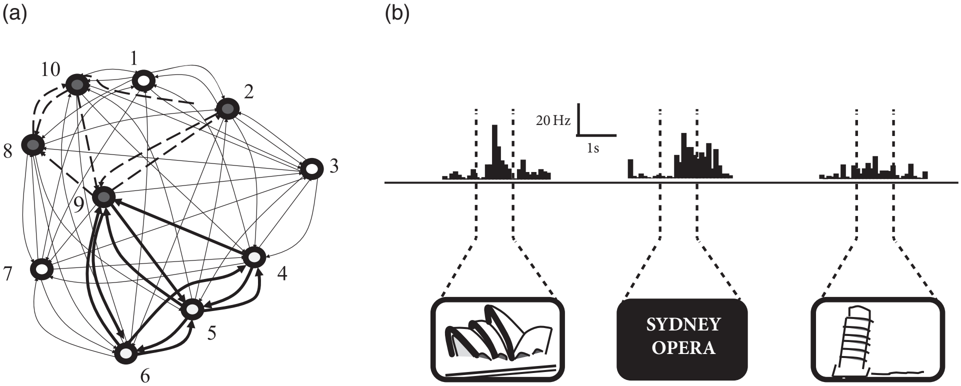


Fig. 17.3 Assemblies and responses to abstract concepts. (a) Schematic diagram of a network of 10 neurons containing two assemblies, defined as strongly connected subgroups (thick solid and dashed lines, respectively). Note that neuron 9 participates in both assemblies. Assemblies could represent abstract mental concepts. (b) Response of a single unit in the human hippocampus (Quiroga *et al*., 2005). The same neuron responds strongly to an image of the Sydney opera house and the words "Sydney opera," but much more weakly to images of other landmarks such as the Pisa tower. Vertical lines indicate the one-second period of stimulation with images of Sydney opera house, words, or Pisa tower, respectively. The photographic images used in the real experiment are replaced here by sketches. Adapted from Quiroga *et al*. (2005).

might represent your concept of Paris with mental pictures of the Eiffel Tower and the Louvre, and yet another your concept of Sydney with its famous opera house.

The assembly as a subgroup of strongly connected neurons has been an influential theoretical notion, introduced by Hebb (1949). Do such assemblies exist? The short answer is: We don't know. Neurons belonging to an assembly do not have to be neighbors but can be widely distributed across one, or even several, brain areas. Experimentally, it is therefore difficult to check for the presence of an assembly as a *group* of neurons. However, Quiroga *et al*. (2005) found individual neurons in human patients that code for abstract mental concepts such as the Sydney opera house. These patients suffer from severe treatment-resistant epilepsy which makes a surgical intervention necessary. In order to precisely locate the focus of the epilepsy in relation to important brain areas (such as those for speech or motor control), electrophysiological recordings are made while the patient performs various tasks. In contrast to neurons in the visual cortex which respond in the presence of an appropriate visual stimulus, single neurons in the medial temporal lobe of the human cortex (in particular, in the hippocampus) do not respond to a specific stimulus, but to a much broader set of stimuli that are linked to the same mental concept. For example, the written word "Sydney" and a picture of the opera house in Sydney both cause a response in the same neuron (Fig. 17.3), which we can therefore interpret as one of the neurons belonging to the assembly of neurons encoding the mental concept "Sydney."

Three aspects are worth emphasizing. First, it is unlikely that the neuron responding to the Sydney opera house is the only one to do so. Therefore, we should not think of a *single* neuron as representing a concept or memory item, but rather a group of neurons.

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

The idea that a single neuron represents one concept is sometimes called the "grandmother cell" code: if the cell coding for grandmother were to die in our brain, our memory of grandmother would disappear as well. At the current stage of research, neural codes based on groups of cells are a more likely code than a grandmother cell code.

Second, the same neuron participates in several assemblies. In the recording sessions of Quiroga *et al.* where a large collection of pictures of famous individuals and landmarks were used, each unit showed strong responses to about 3% of the stimuli (Quiroga *et al.*, 2005).

Third, some, but not all, of the neurons showed prolonged responses that persisted after the end of stimulus presentation. This could potentially indicate that a memory item is retrieved and kept in the brain even after the stimulus has disappeared. All three aspects play a role in the memory model discussed in Section 17.2.

17.1.3 Working memory and delayed matching-to-sample tasks

In contrast to long-term memory, items in working memory do not have to be kept for a lifetime. For example, humans use their working memory when they write down a phone number that they just received or search in the supermarket for items on their shopping list. Neural activity during working memory tasks has been recorded in monkeys, in the particular in the prefrontal and inferotemporal cortex (Miyashita, 1988a; Fuster and Jervey, 1982; Miller and Cohen, 2001). In a delayed matching-to-sample task, a monkey has to indicate whether a second stimulus is, or is not, identical to a first stimulus received one or several seconds earlier.

To correctly perform the task, the monkey has to remember the sample stimulus during the delay period where no stimulation is given. Some neurons in the prefrontal cortex show sustained activity during the delay period (Fig. 17.4a). This has been interpreted as a neural signature of working memory. During the delay period, the time course of neural activity varies widely between different objects for one neuron (Fig. 17.4b). and across a population of neurons (Rainer and Miller, 2002), which indicates that simple models such as the ones discussed in this chapter do not explain all aspects of working memory.

17.2 Hopfield model

The Hopfield model (Hopfield, 1982), consists of a network of N neurons, labeled by a lower index i, with $1 \leq i \leq N$. Similar to some earlier models (McCulloch and Pitts, 1943; Little, 1974; Willshaw *et al.*, 1969), neurons in the Hopfield model have only two states. A neuron i is "ON" if its state variable takes the value $S_i = +1$ and "OFF" (silent) if $S_i = -1$. The dynamics evolve in discrete time with time steps Δt. There is no refractoriness and the duration of a time step is typically not specified. If we take $\Delta t = 1$ ms, we can interpret $S_i(t) = +1$ as an action potential of neuron i at time t. If we take $\Delta t = 500$ ms, $S_i(t) = +1$ should rather be interpreted as an episode of high firing rate.

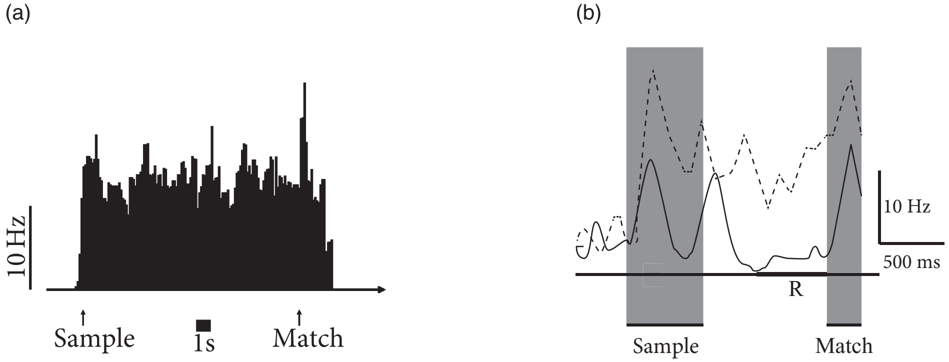


Fig. 17.4 Delayed matching-to-sample task. (a) PSTH of a neuron in the anterior ventral temporal cortex in a visual working memory task. The monkey has to indicate whether a first stimulus (sample, presented for 0.2 s at time marked by arrow) is identical to a second one which can be either a matching (arrow) or an unfamiliar stimulus; adapted from Miyashita (1988a). (b) PSTH of a single neuron in the prefrontal cortex in response to two different images, one object (dashed line) and one different noise pattern (solid line). Sample stimuli were presented for 650 ms. After a delay period of 1 s, a matching stimulus was presented. "R" marks a period when responses tend to recover after a transient dip. Vertical axis: firing rate measured with respect to baseline activity. Adapted from Rainer and Miller (2002).

Neurons interact with each other with weights w_{ij}. The input potential of neuron i, influenced by the activity of other neurons is

$$h_i(t) = \sum_j w_{ij} S_j(t) \,. \tag{17.2}$$

The input potential at time t influences the probabilistic update of the state variable S_i in the next time step:

$$\text{Prob}\{S_i(t+\Delta t) = +1|h_i(t)\} = g(h_i(t)) = g\left(\sum_j w_{ij} S_j(t)\right) , \tag{17.3}$$

where g is a monotonically increasing gain function with values between zero and 1. A common choice is $g(h) = 0.5[1+\tanh(\beta h)]$ with a parameter β. For $\beta \to \infty$, we have $g(h) = 1$ for $h > 0$ and zero otherwise. The dynamics are therefore deterministic and summarized by the update rule

$$S_i(t+\Delta t) = \text{sgn}[h(t)]. \tag{17.4}$$

For finite β the dynamics are stochastic. In the following we assume that in each time step all neurons are updated synchronously (parallel dynamics), but an update scheme where only one neuron is updated per time step is also possible.

The aim of this section is to show that, with a suitable choice of the coupling matrix w_{ij}, memory items can be retrieved by the collective dynamics defined in Eq. (17.3), applied to all N neurons of the network. In order to illustrate how collective dynamics can lead

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

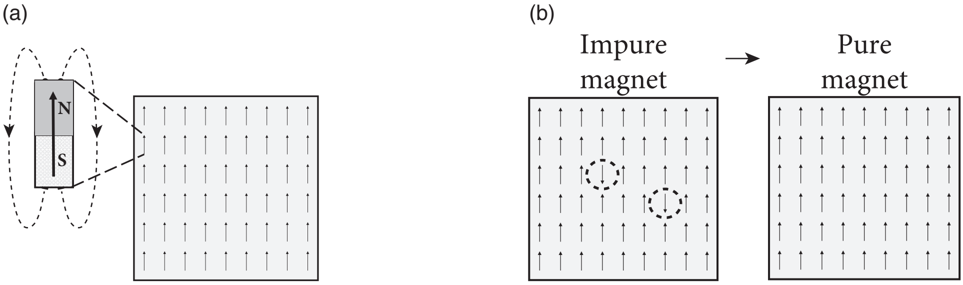


Fig. 17.5 Physics of ferromagnets. (a) Magnetic materials consist of atoms, each with a small magnetic moment, here visualized as an arrow, a symbol for a magnetic needle. At low temperature, all magnetic needles are aligned. Inset: Field lines around one of the magnetic needles. (b) At high temperature, some of the needles are misaligned (dashed circles). Cooling the magnet leads to a spontaneous alignment and reforms a pure magnet. Schematic figure.

to meaningful results, we start, in Section 17.2.1, with a detour through the physics of magnetic systems. In Section 17.2.2, the insights from magnetic systems are applied to the case at hand, i.e., memory recall.

17.2.1 Detour: magnetic analogy

Magnetic material contains atoms which carry a so-called spin. The spin generates a magnetic moment at the microscopic level visualized graphically as an arrow (Fig. 17.5a). At high temperature, the magnetic moments of individual atoms point in all possible directions. Below a critical temperature, however, the magnetic moment of all atoms spontaneously align with each other. As a result, the microscopic effects of all atomic magnetic moments add up and the material exhibits the macroscopic properties of a ferromagnet.

In order to understand how a spontaneous alignment can arise, let us study Eqs. (17.2) and (17.3) in the analogy of magnetic materials. We assume that $w_{ij} = w_0 > 0$ between all pairs of neurons $i \neq j$, and that self-interaction vanishes, $w_{ii} = 0$.

Each atom is characterized by a spin variable $S_i = \pm 1$ where $S_i = +1$ indicates that the magnetic moment of atom i points “upward.” Suppose that, at time $t = 0$, all spins take a positive value ($S_I = +1$), except that of atom i which has a value $S_i(0) = -1$ (Fig. 17.5a). We calculate the probability that, at time step $t = \Delta t$, the spin of neuron i will switch to $S_i = +1$. This probability is according to Eq. (17.3)

$$\text{Prob}\{S_i(t+\Delta t) = +1 | h_i(t)\} = g(h_i(t)) = g\left(\sum_{j=1}^{N} w_{ij}\, S_j(t)\right) = g(w_0\,(N-1)), \qquad (17.5)$$

where we have used our assumptions. With $g(h) = 0.5[1 + \tanh(\beta h)]$ and $w_0 = \beta = 1$, we find that, for any network of more than three atoms, the probability that the magnetic moments of all atoms would align is extremely high. In physical systems, β plays the role

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

of an inverse temperature. If β becomes small (high temperature), the magnetic moments no longer align and the material loses its spontaneous magnetization.

According to Eq. (17.5) the probability of alignment increases with the network size. This is an artifact of our model with all-to-all interaction between all atoms. Physical interactions, however, rapidly decrease with distance, so that the sum over j in Eq. (17.5) should be restricted to the nearest neighbors of neuron i, e.g., about 4 to 20 atoms depending on the configuration of the atomic arrangement and the range of the interaction. Interestingly, neurons, in contrast to atoms, are capable of making long-range interactions because of their far-reaching axonal cables and dendritic trees. Therefore, the number of topological neighbors of a given neuron is in the range of thousands.

An arrangement of perfectly aligned magnetic elements looks rather boring, but physics offers more interesting examples as well. In some materials, typically consisting of two different types of atoms, say A and B, an anti-ferromagnetic ordering is possible (Fig. 17.6). While one layer of magnetic moments points upward, the next one points downward, so that the macroscopic magnetization is zero. Nevertheless, a highly ordered structure is present. Examples of anti-ferromagnets are some metallic oxides and alloys.

To model an anti-ferromagnet, we choose interactions $w_{ij} = +1$ if i and j belong to the same class (e.g., both are in a layer of type A or both in a layer of type B), and $w_{ij} = -1$ if one of the two atoms belongs to type A and the other to type B. A simple repetition of the calculation in Eq. (17.5) shows that an anti-ferromagnetic organization of the spins emerges spontaneously at low temperature.

The same idea of positive and negative interactions w_{ij} can be used to embed an arbitrary pattern into a network of neurons. Let us draw a pattern of black and white pixels corresponding to active ($p_i = +1$) and inactive ($p_i = -1$) neurons, respectively. The rule extracted from the anti-ferromagnet implies that pixels of opposite color are connected by negative weights, while pixels of the same color have connections with positive weight. This rule can be formalized as

$$w_{ij} = p_i\, p_j\,. \tag{17.6}$$

This rule forms the basis of the Hopfield model.

17.2.2 Patterns in the Hopfield model

The Hopfield model consists of a network of N binary neurons. A neuron i is characterized by its state $S_i = \pm 1$. The state variable is updated according to the dynamics defined in Eq. (17.3).

The task of the network is to store and recall M different patterns. Patterns are labeled by the index μ with $1 \le \mu \le M$. Each pattern μ is defined as a desired configuration $\left\{p_i^\mu = \pm 1; 1 \le i \le N\right\}$. The network of N neurons is said to correctly represent pattern μ, if the state of all neurons $1 \le i \le N$ is $S_i(t) = S_i(t+\Delta t) = p_i^\mu$. In other words, patterns must be fixed points of the dynamics (17.3).

For us as human observers, a meaningful pattern could, for example, be a configuration

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

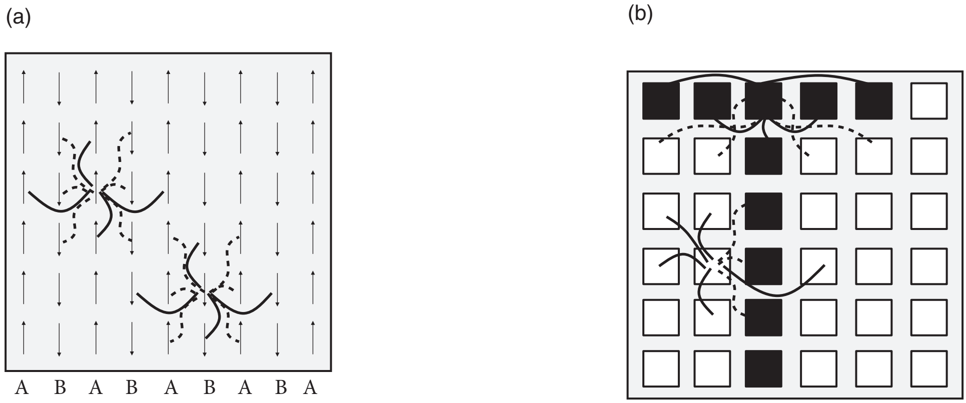


Fig. 17.6 Storing patterns. (a) Physical anti-ferromagnets consist of layers of atoms A and B. All magnetic moments are aligned within a layer of identical neurons, but exhibit different orientations between layers. A model where interactions within atoms of the same type are positive (solid lines) and interactions between atoms of different type are negative (dashed lines) can explain the spontaneous order in the arrangement of magnetic moments. The interaction scheme for two atoms with their ten nearest neighbors is indicated. (b) If we replace magnetic moments by black and white pixels (squares), represented by active and inactive neurons, respectively, the neuronal network can store a pattern, such as T. Interactions are positive (solid lines) between pixels of the same color (black-to-black or white-to-white) and negative otherwise. Only a few representative interactions are shown. Schematic figure.

in form of a “T,” such as depicted in Fig. 17.6b. However, visually attractive patterns have large correlations between each other. Moreover, areas in the brain related to memory recall are situated far from the retinal input stage. Since the configuration of neurons in memory-related brain areas is probably very different from those at the retina, patterns in the Hopfield model are chosen as fixed random patterns; see Fig. 17.7.

During the set-up phase of the Hopfield network, a random number generator generates, for each pattern μ, a string of N independent binary numbers $\{p_i^\mu = \pm 1; 1 \leq i \leq N\}$ with expectation value $\langle p_i^\mu \rangle = 0$. Strings of different patterns are independent. The weights are chosen as

$$w_{ij} = c \sum_{\mu=1}^{M} p_i^\mu \, p_j^\mu \,, \tag{17.7}$$

with a positive constant $c > 0$. The network has full connectivity. Note that for a single pattern and $c = 1$, Eq. (17.7) is identical to the connections of the anti-ferromagnet, Eq. (17.6). For reasons that become clear later on, the standard choice of the constant c is $c = 1/N$.

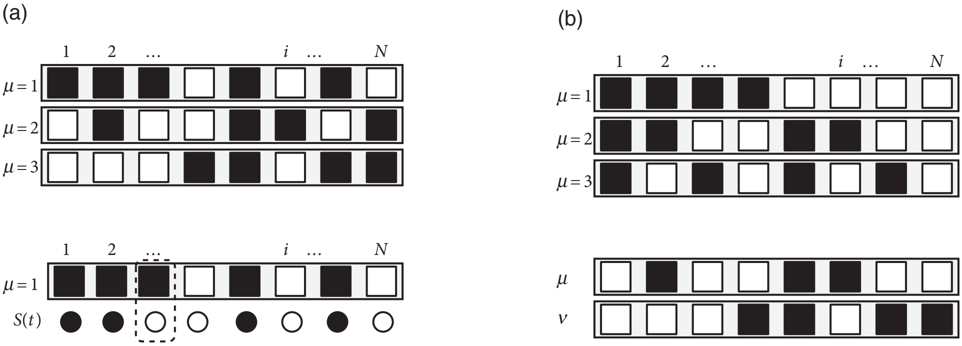


Fig. 17.7 Hopfield model. (a) Top: Three random patterns $\mu = 1,2,3$ in a network of $N = 8$ neurons. Black squares ($p_i^\mu = +1$) and white squares ($p_i^\mu = -1$) are arranged in random order. Bottom: The overlap $m^1 = (1/N)\sum_i p_i^1 S_i(t)$ measures the similarity between the current state $S(t) = \{S_i(t); 1 \le i \le N\}$ and the first pattern. Here only a single neuron exhibits a mismatch (dotted line). The desired value in the pattern is shown as black and white squares, while the current state is indicated as black and white circles. Schematic figure. (b) Orthogonal patterns have a mutual overlap of zero so that correlations are $C^{\mu\nu} = (1/N)\sum_i p_i^\mu p_i^\nu = \delta^{\mu\nu}$ (top) whereas random patterns exhibit a small residual overlap for $\mu \neq \nu$ (bottom).

17.2.3 Pattern retrieval

In many memory retrieval experiments, a cue with partial information is given at the beginning of a recall trial. The retrieval of a memory item is verified by the completion of the missing information.

To mimic memory retrieval in the Hopfield model, an input is given by initializing the network in a state $S(t_0) = \{S_i(t_0); 1 \le i \le N\}$. After initialization, the network evolves freely under the dynamics (17.3). Ideally the dynamics should converge to a fixed point corresponding to the pattern μ which is most similar to the initial state.

To measure the similarity between the current state $S(t) = \{S_i(t); 1 \le i \le N\}$ and a pattern μ, we introduce the overlap (Fig. 17.7a)

$$m^\mu(t) = \frac{1}{N}\sum_i p_i^\mu S_i(t)\,. \tag{17.8}$$

The overlap takes a maximum value of 1 if $S_i(t) = p_i^\mu$, i.e., if the pattern is retrieved. It is close to zero if the current state has no correlation with pattern μ. The minimum value $m^\mu(t) = -1$ is achieved if each neuron takes the opposite value to that desired in pattern μ.

The overlap plays an important role in the analysis of the network dynamics. In fact, using Eq. (17.2) the input potential h_i of a neuron i is

$$h_i(t) = \sum_j w_{ij} S_j(t) = c\sum_{j=1}^{N}\sum_{\mu=1}^{M} p_i^\mu\, p_j^\mu\, S_j(t) = c N \sum_{\mu=1}^{M} p_i^\mu\, m^\mu(t), \tag{17.9}$$

where we have used Eqs. (17.7) and (17.8). To make the results of the calculation

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

independent of the size of the network, it is standard to choose the factor $c = 1/N$, as mentioned above. In what follows we always take $c = 1/N$ unless indicated otherwise. For an in-depth discussion, see the scaling arguments in Chapter 12.

To close the argument, we now use the input potential in the dynamics equation (17.3) and find

$$\text{Prob}\{S_i(t+\Delta t) = +1|h_i(t)\} = g\left[\sum_{\mu=1}^{M} p_i^\mu\, m^\mu(t)\right]. \tag{17.10}$$

Equation (17.10) highlights that the M macroscopic similarity values m^μ with $1 \le \mu \le M$ completely determine the dynamics of the network.

Example: Memory retrieval

Let us suppose that the initial state has a significant similarity with pattern $\mu = 3$, for example an overlap of $m^\mu(t_0) = 0.4$ and no overlap with the other patterns $m^\nu = 0$ for $\nu \neq 3$.

In the noiseless case Eq. (17.10) simplifies to

$$S_i(t_0+\Delta t) = \text{sgn}\left[\sum_{\mu=1}^{M} p_i^\mu\, m^\mu\right] = \text{sgn}\left[p_i^3\, m^3(t_0)\right] = p_i^3 \quad \text{for all } i. \tag{17.11}$$

Hence, each neuron takes, after a single time step, the desired state corresponding to the pattern. In other words, the pattern with the strongest similarity to the input is retrieved, as it should be.

For stochastic neurons we find

$$\text{Prob}\{S_i(t_0+\Delta t) = +1|h_i(t)\} = g[p_i^3\, m^3(t_0)]. \tag{17.12}$$

We note that, given the overlap $m^3(t_0)$, the right-hand side of Eq. (17.12) can take only two different values, corresponding to $p_i^3 = +1$ and $p_i^3 = -1$. Thus, all neurons that *should* be active in pattern 3 share the same probabilistic update rule:

$$\text{Prob}\{S_i(t_0+\Delta t) = +1|h_i(t)\} = g[m^3(t_0)] \quad \text{for all } i \text{ with } p_i^3 = +1. \tag{17.13}$$

Similarly all those that *should* be inactive share another rule:

$$\text{Prob}\{S_i(t_0+\Delta t) = +1|h_i(t)\} = g[-m^3(t_0)] \quad \text{for all } i \text{ with } p_i^3 = -1. \tag{17.14}$$

Thus, despite the fact that there are N neurons and M different patterns, during recall the network breaks up into two macroscopic populations: those that should be active and those that should be inactive. This is the reason why we can expect to arrive at macroscopic population equations, similar to those encountered in Part III of the book.

Let us use this insight for the calculation of the overlap at time $t_0 + \Delta t$. We denote the size of the two populations (active, inactive) by N_+^3 and N_-^3, respectively, and find

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

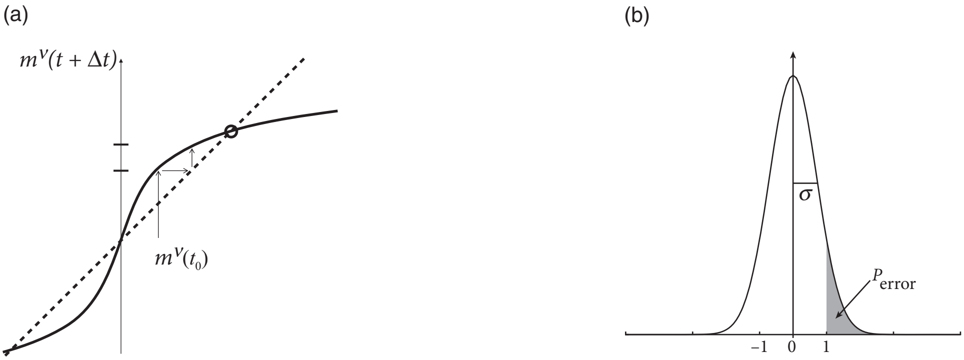


Fig. 17.8 Memory retrieval in the Hopfield model. (a) The overlap $m^{\nu}(t+\Delta t)$ with a specific pattern ν is given as a function of the overlap with the same pattern $m^{\nu}(t)$ in the previous time step (solid line); see Eq. (17.16). The overlap with the $M-1$ other patterns is supposed to vanish. The iterative update can be visualized as a path (arrow) between the overlap curve and the diagonal (dashed line). The dynamics approach a fixed point (circle) with high overlap corresponding to the retrieval of the pattern. (b) The probability P_{error} that during retrieval an erroneous state flip occurs corresponds to the shaded area under the curve; see Eq. (17.20). The width σ of the curve is proportional to the pattern load M/N. Schematic figure.

$$
\begin{aligned}
m^3(t_0+\Delta t) &= \frac{1}{N}\sum_i p_i^3 S_i(t_0+\Delta t) \qquad (17.15)\\
&= \frac{N_+^3}{N}\left[\frac{1}{N_+^3}\sum_{i \text{ with } p_i^3=+1} S_i(t_0+\Delta t)\right] - \frac{N_-^3}{N}\left[\frac{1}{N_-^3}\sum_{i \text{ with } p_i^3=+1} S_i(t_0+\Delta t)\right].
\end{aligned}
$$

We can interpret the two terms enclosed by the square brackets as the average activity of those neurons that should, or should not, be active, respectively. In the limit of a large network ($N \to \infty$) both groups are very large and of equal size $N_+^3 = N_-^3 = N/2$. Therefore, the averages inside the square brackets approach their expectation values. The technical term, used in the physics literature, is that the network dynamics are "self-averaging." Hence, we can evaluate the square brackets with probabilities introduced in Eqs. (17.13) and (17.14). With $\text{Prob}\{S_i(t_0+\Delta t) = -1|h_i(t)\} = 1 - \text{Prob}\{S_i(t_0+\Delta t) = +1|h_i(t)\}$, we find

$$m^3(t_0+\Delta t) = \frac{1}{2}\{2\,g[m^3(t_0)]-1\} - \frac{1}{2}\{2\,g[-m^3(t_0)]-1\}. \qquad (17.16)$$

In the special case that $g(h) = 0.5[1+\tanh(\beta h)]$ Eq. (17.16) simplifies to an update law

$$m^3(t+\Delta t) = \tanh[\beta\, m^3(t)], \qquad (17.17)$$

where we have replaced t_0 by t, in order to highlight that updates should be iterated over several time steps.

We close with three remarks. First, the dynamics of N neurons has been replaced, in a mathematically precise limit, by the iterative update of one single macroscopic variable, i.e., the overlap with one of the patterns. The result is reminiscent of the analysis of the

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

macroscopic population dynamics performed in Part III of the book. Indeed, the basic mathematical principles used for the equations of the population activity $A(t)$ are the same as the ones used here for the update of the overlap variable $m^\mu(t)$.

Second, if $\beta > 1$, the dynamics converge from an initially small overlap to a fixed point with a large overlap, close to 1. The graphical solution of the update of pattern $\nu = 3$ (for which a nonzero overlap existed in the initial state) is shown in Fig. 17.8. Because the network dynamics is "attracted" toward a stable fixed point characterized by a large overlap with one of the memorized patterns (Fig. 17.9a), the Hopfield model and variants of it are also called "attractor" networks or "attractor memories" (Amit, 1989; Barbieri and Brunel, 2008).

Finally, the assumption that, apart from pattern 3, all other patterns have an initial overlap exactly equal to zero is artificial. For random patterns, we expect a small overlap between arbitrary pairs of patterns. Thus, if the network is exactly in pattern 3 so that $m^3 = 1$, the other patterns have a small but finite overlap $|m^\mu| \neq 0$, because of spurious correlations $C^{\mu\nu} = (1/N)\sum_i p_i^\mu p_i^\nu$ between any two random patterns μ and ν; Fig. 17.7b. If the number of patterns is large, the spurious correlations between the patterns can generate problems during memory retrieval, as we shall see now.

17.2.4 Memory capacity

How many random patterns can be stored in a network of N neurons? Memory retrieval implies pattern completion, starting from a partial cue. An absolutely minimal condition for pattern completion is that at least the dynamics should not move *away* from the pattern, if the initial cue is *identical* to the complete pattern (Hertz *et al.*, 1991). In other words, we require that a network with initial state $S_i(t_0) = p_i^\nu$ for $1 \leq i \leq N$ stays in pattern ν. Therefore pattern ν must be a fixed point under the dynamics.

We study a Hopfield network at zero temperature ($\beta = \infty$). We start the calculation as in Eq. (17.9) and insert $S_j(t_0) = p_j^\nu$. This yields

$$\begin{aligned} S_i(t_0 + \Delta t) &= \operatorname{sgn}\left[\frac{1}{N}\sum_{j=1}^{N}\sum_{\mu=1}^{M} p_i^\mu\, p_j^\mu\, p_j^\nu\right] \\ &= \operatorname{sgn}\left[p_i^\nu\left(\frac{1}{N}\sum_{j=1}^{N} p_j^\nu\, p_j^\nu\right) + \frac{1}{N}\sum_{\mu\neq\nu}\sum_{j} p_i^\mu\, p_j^\mu\, p_j^\nu\right], \end{aligned} \tag{17.18}$$

where we have separated the pattern ν from the other patterns. The factor in parentheses on the right-hand side adds up to 1 and can therefore be dropped. We now multiply the second term on the right-hand side by a factor $1 = p_i^\nu\, p_i^\nu$. Finally, because $p_i^\nu = \pm 1$, a factor p_i^ν can be pulled out of the argument of the sign-function:

$$S_i(t_0 + \Delta t) = p_i^\nu \operatorname{sgn}\left[1 + \frac{1}{N}\sum_j \sum_{\mu\neq\nu} p_i^\mu\, p_i^\nu\, p_j^\mu\, p_j^\nu\right] = p_i^\nu \operatorname{sgn}[1 - a_{i\nu}]\,. \tag{17.19}$$

The desired fixed point exists only if $1 > a_{i\nu} = \frac{1}{N}\sum_j \sum_{\mu \neq \nu} p_i^\mu \, p_i^\nu \, p_j^\mu \, p_j^\nu$ for all neurons i. In other words, even if the network is initialized in perfect agreement with one of the patterns, it can happen that one or a few neurons flip their sign. The probability of moving away from the pattern is equal to the probability of finding a value $a_{i\nu} > 1$ for one of the neurons i.

Because patterns are generated from independent random numbers $p_i^\mu = \pm 1$ with zero mean, the product $p_i^\mu \, p_i^\nu \, p_j^\mu \, p_j^\nu = \pm 1$ is also a binary random number with zero mean. Since the values p_i^μ are chosen independently for each neuron i and each pattern μ, the term $a_{i\nu}$ can be visualized as a random walk of $N\,(M-1)$ steps and step size $1/N$. For a large number of steps, the positive or negative walking distance can be approximated by a Gaussian distribution with zero mean and standard deviation $\sigma = \sqrt{(M-1)/N} \approx \sqrt{M/N}$ for $M \gg 1$. The probability that the activity state of neuron i erroneously flips is therefore proportional to (Fig. 17.86)

$$P_{\text{error}} = \frac{1}{\sqrt{2\pi}\sigma} \int_1^\infty e^{\frac{-x^2}{2\sigma^2}} \, dx \approx \frac{1}{2}\left[1 - \text{erf}\left(\sqrt{\frac{N}{2M}}\right)\right], \tag{17.20}$$

where we have introduced the error function

$$\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_0^x e^{-y^2} \, dy. \tag{17.21}$$

The most important insight is that the probability of an erroneous state flip increases with the ratio M/N. Formally, we can define the storage capacity C_{store} of a network as the maximal number $M^{\max}$ of patterns that a network of N neurons can retrieve

$$C_{\text{store}} = \frac{M^{\max}}{N} = \frac{M^{\max}\,N}{N^2}\,. \tag{17.22}$$

For the second equality sign we have multiplied both numerator and denominator by a common factor N which gives rise to the following interpretation. Since each pattern consists of N neurons (i.e., N binary numbers), the total number of bits that need to be stored at maximum capacity is $M^{\max}\,N$. In the Hopfield model, patterns are stored by an appropriate choice of the synaptic connections. The number of available synapses in a fully connected network is N^2. Therefore, the storage capacity measures the number of bits stored per synapse.

Example: Erroneous bits

We can evaluate Eq. (17.20) for various choices of P_{error}. For example, if we accept an error probability of $P_{\text{error}} = 0.001$, we find a storage capacity of $C_{\text{store}} = 0.105$.

Hence, a network of 10 000 neurons is capable of storing about 1000 patterns with $P_{\text{error}} = 0.001$. Thus in each of the patterns, we expect that about 10 neurons exhibit erroneous activity. We emphasize that the above calculation focuses on the *first* iteration step only. If we start in the pattern, then about 10 neurons will flip their state in the first iteration. But these flips could in principle cause further neurons to flip in the second iteration and eventually initiate an avalanche of many other changes.

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

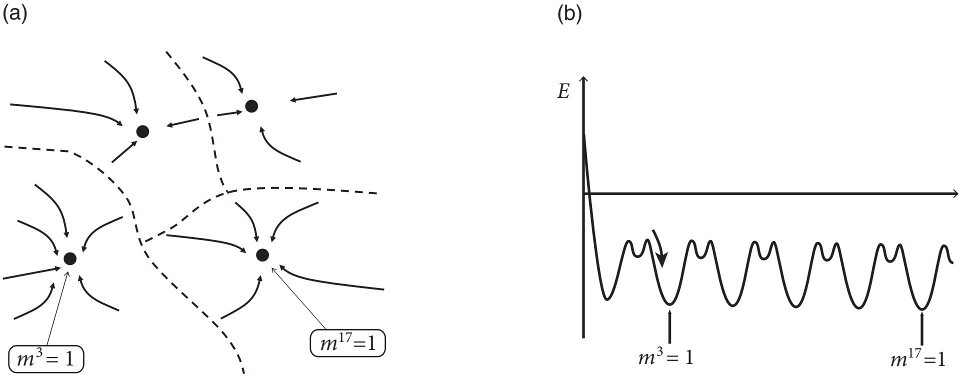


Fig. 17.9 Attractor picture and energy landscape. (a) The dynamics are attracted toward fixed points corresponding to memory states (overlap $m^\nu = 1$). Four attractor states are indicated. The dashed lines show the boundaries of the basin of attraction of each memory. (b) The Hopfield model has multiple equivalent energy minima, each one corresponding to the retrieval (overlap $m^\nu = 1$) of one pattern. Between the main minima, additional local minima (corresponding to mixtures of several patterns) may also exist.

A more precise calculation shows that such an avalanche does not occur if the number of stored patterns stays below a limit such that $C_{\text{store}} = 0.138$ (Amit *et al.*, 1985, 1987b).

17.2.5 The energy picture

The Hopfield model has symmetric interactions $w_{ij} = w_{ji} = c\sum_{\mu=1}^{M} p_i^\mu\, p_j^\mu$. We now show that, in any network with symmetric interactions and asynchronous deterministic dynamics

$$S_i(t+\Delta t) = \operatorname{sgn}[h(t)] = \operatorname{sgn}\left[\sum_j w_{ij}\, S_j(t)\right], \tag{17.23}$$

the energy

$$E = -\sum_i \sum_j w_{ij}\, S_i\, S_j \tag{17.24}$$

decreases with each state flip of a single neuron (Hopfield, 1982).

In each time step only one neuron is updated (asynchronous dynamics). Let us assume that after application of Eq. (17.23) neuron k has changed its value from S_k at time t to $S'_k = -S_k$ while all other neurons keep their value $S'_j = S_j$ for $j \neq k$. The prime indicates values evaluated at time $t+\Delta t$. The change in energy caused by the state flip of neuron k is

$$E' - E = -\sum_i w_{ik}\, S_i\, (S'_k - S_k) - \sum_j w_{kj}\, S_j\, (S'_k - S_k)\,. \tag{17.25}$$

First, because of the update of neuron k, we have $S'_k - S_k = 2S'_k$. Second, because of the symmetry $w_{ij} = w_{ji}$, the two terms on the right-hand side are identical, and $\sum_i w_{ik} S_i = \sum_i w_{ki} S_i = h_k$. Third, because of Eq. (17.23), the sign of h_k determines the new value S'_k of

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

neuron k. Therefore the change in energy is $E' - E = -4h_k \operatorname{sgn} h_k < 0$. In other words, the energy E is a Liapunov function of the deterministic Hopfield network.

Since the dynamics leads to a decrease of the energy, we may wonder whether we can say something about the global or local minimum of the energy. If we insert the definition of the connection weights into the energy function (17.24), we find

$$E = -\sum_i \sum_j \left(c \sum_\mu p_i^\mu p_j^\mu \right) S_i S_j = -c N^2 \sum_\mu (m^\mu)^2 \,, \tag{17.26}$$

where we have used the definition of the overlap; see Eq. (17.8).

The maximum value of the overlap with a fixed pattern ν is $m^\nu = 1$. Moreover, for random patterns, the correlations between patterns are small. Therefore, if $m^\nu = 1$ (i.e., recall of pattern ν) the overlap with other patterns $\mu \neq \nu$ is $m^\mu \approx 0$. Therefore, the energy landscape can be visualized with multiple minima of the same depth, each minimum corresponding to retrieval of one of the patterns (Fig. 17.9b).

17.2.6 Retrieval of low-activity patterns

There are numerous aspects in which the Hopfield model is rather far from biology. One of these is that, in each memory pattern, 50% of the neurons are active.

To characterize patterns with a lower level of activity, let us introduce random variables $\xi_i^\mu \in \{0,1\}$ for $1 \leq i \leq N$ and $1 \leq \mu \leq M$ with mean $\langle \xi_i^\mu \rangle = a$. For $a = 0.5$ and $p_i^\mu = 2\xi_i^\mu - 1$ we are back to the patterns in the Hopfield model. In the following we are, however, interested in patterns with an activity $a < 0.5$. To simplify some of the arguments below, we suppose that patterns are generated under the constraint $\sum_i \xi_i^\mu = N a$ for each μ, so that all patterns have *exactly* the same target activity a.

The weights in the Hopfield model of Eq. (17.7) are replaced by

$$w_{ij} = c' \sum_{\mu=1}^{M} (\xi_i^\mu - b)\,(\xi_j^\mu - a)\,, \tag{17.27}$$

where a is the mean activity of the stored patterns, $0 \leq b \leq 1$ a constant, and $c' = [2a(1-a)N]^{-1}$. Note that Eq. (17.7) is a special case of Eq. (17.27) with $a = b = 0.5$ and $c' = 2c$.

As before, we work with binary neurons $S_i = \pm 1$ defined by the stochastic update rule in Eqs. (17.2) and (17.3). To analyze pattern retrieval we proceed analogously to Eq. (17.10). Introducing the overlap of low-activity patterns

$$m^\mu = \frac{1}{2a(1-a)N} \sum_j (\xi_j^\mu - a)\, S_j, \tag{17.28}$$

we find

$$\text{Prob}\{S_i(t+\Delta t) = +1 | h_i(t)\} = g\left[\sum_{\mu=1}^{M} (\xi_i^\mu - b)\, m^\mu(t) \right]. \tag{17.29}$$

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

Example: Memory retrieval and attractor dynamics

Suppose that at time t the overlap with one of the patterns, say pattern 3, is significantly above zero while the overlap with all other patterns vanishes $m^\mu \approx m\,\delta^{\mu 3}$, where δ^{nm} denotes the Kronecker-δ. The initial overlap is $0.1 < m \leq 1$. Then the dynamics of the low-activity networks split up into two groups of neurons, i.e., those that should be "ON" in pattern 3 ($\xi_i^3 = 1$) and those that should be "OFF" ($\xi_i^3 = 0$).

The size of both groups scales with N: there are $a \cdot N$ "ON" neurons and $(1-a) \cdot N$ "OFF" neurons. For $N \to \infty$, the population activity A^{ON} of the "ON" group (i.e., the fraction of neurons with state $S_i = +1$ in the "ON" group) is therefore well described by its expectation value

$$A^{\text{ON}}(t+\Delta t) = g[(1-b)\,m^3(t)]\,. \tag{17.30}$$

Similarly, the "OFF" group has activity

$$A^{\text{OFF}}(t+\Delta t) = g[(-b)\,m^3(t)]\,. \tag{17.31}$$

To close the argument we determine the overlap at time $t+\Delta t$. Exploiting the split into two groups of size $a \cdot N$ and $(1-a) \cdot N$, respectively, we have

$$\begin{aligned} m^3(t+\Delta t) &= \frac{1}{2a(1-a)N}\left[\sum_{j \text{ with } \xi_j^3=1} (1-a)\,S_j(t+\Delta t) + \sum_{j \text{ with } \xi_j^3=0} (-a)\,S_j(t+\Delta t)\right] \\ &= \frac{1}{2}\left[A^{\text{ON}}(t+\Delta t) - A^{\text{OFF}}(t+\Delta t)\right]. \end{aligned} \tag{17.32}$$

Thus, the overlap with pattern 3 has changed from the initial value $m^3(t) = m$ to a new value $m^3(t+\Delta t)$. Retrieval of memories works if iteration of Eqs. (17.30)–(17.32) makes m^3 converge to a value close to unity while, at the same time, the other overlaps m^ν (for $\nu \neq 3$) stay close to zero.

We emphasize that the analysis of the network dynamics presented here does not require symmetric weights but is possible for arbitrary values of the parameter b. However, a standard choice is $b = a$, which leads to symmetric weights and to a high memory capacity (Tsodyks and Feigelman, 1986).

17.3 Memory networks with spiking neurons

The Hopfield model is an abstract conceptual model and rather far from biological reality. In this section we aim at pushing the abstract model in the direction of increased biological plausibility. We focus on two aspects. In Section 17.3.1 we replace the binary neurons of the Hopfield model with spiking neuron models of the class of Generalized Linear Models or Spike Response Models; see Chapter 9. Then, in Section 17.3.2 we ask whether it is possible to store multiple patterns in a network where excitatory and inhibitory neurons are functionally separated from each other.

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

17.3.1 Activity of spiking networks

Neuron models such as the Spike Response Model with escape noise, formulated in the framework of Generalized Linear Models, can predict spike times of real neurons to a high degree of accuracy; see Chapters 9 and 11. We therefore choose the Spike Response Model (SRM) as our candidate for a biologically plausible neuron model. Here we use these neuron models to analyze the macroscopic dynamics in attractor memory networks of spiking neurons.

As discussed in Chapter 9, the membrane potential u_i of a neuron i embedded in a large network can be described as

$$u_i(t) = \sum_f \eta(t - t_i^f) + h_i(t) + u_{\text{rest}}, \tag{17.33}$$

where $\eta(t - t_i^f)$ summarizes the refractoriness caused by the spike-afterpotential and $h_i(t)$ is the (deterministic part of the) input potential

$$h_i(t) = \sum_j w_{ij}\varepsilon(t - t_j^f) = \sum_j w_{ij} \int_0^\infty \varepsilon(s)\, S_j(t-s)\mathrm{d}s\,. \tag{17.34}$$

Here i denotes the postsynaptic neuron, w_{ij} is the coupling strength from a presynaptic neuron j to i, and $S_j(t) = \sum_f \delta(t - t_j^f)$ is the spike train of neuron j.

Statistical fluctuations in the input as well as intrinsic noise sources are both incorporated into an escape rate (or stochastic intensity) $\rho_i(t)$ of neuron i,

$$\rho_i(t) = f(u_i(t) - \vartheta)\,, \tag{17.35}$$

which depends on the momentary distance between the (noiseless) membrane potential and the threshold ϑ.

In order to embed memories in the network of SRM neurons we use Eq. (17.27) and proceed as in Section 17.2.6. There are three differences compared to the previous section: First, while previously S_j denoted a binary variable ±1 in *discrete* time, we now work with spikes $\delta(t - t_j^f)$ in *continuous* time. Second, in the Hopfield model a neuron can be active in every time step while here spikes must have a minimal distance because of refractoriness. Third, the input potential h is only one of the contributions to the total membrane potential.

Despite these differences the formalism of Section 17.2.6 can be directly applied to the case at hand. Let us define the *instantaneous* overlap of the spike pattern in the network with pattern μ as

$$m^\mu(t) = \frac{1}{2a(1-a)N} \sum_j (\xi_j^\mu - a)\, S_j(t), \tag{17.36}$$

where $S_j(t) = \sum_f \delta(t - t_j^f)$ is the spike train of neuron j. Note that, because of the Dirac δ-function, we need to integrate over m^μ in order to arrive at an observable quantity. Such

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

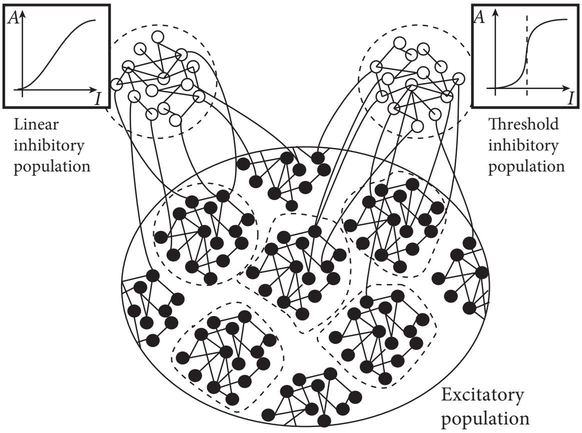


Fig. 17.10 A population of excitatory neurons interacts with two populations of inhibitory neurons. Memory patterns are embedded as Hebbian assemblies in the excitatory population. All neurons are integrate-and-fire neurons. Theory predicts that the first inhibitory population should be activated to levels where the gain function (left inset) is approximately linear. The second inhibitory population is activated if the total input is above some threshold value (right inset).

an integration is automatically performed by each neuron. Indeed, the input potential Eq. (17.34) can be written as

$$\begin{aligned} h_i(t) &= \sum_j \left(\frac{1}{2a(1-a)N} \sum_{\mu=1}^{M} (\xi_i^\mu - b)\,(\xi_j^\mu - a) \right) \int_0^\infty \varepsilon(s)\, S_j(t-s) \mathrm{d}s \\ &= \sum_{\mu=1}^{M} (\xi_i^\mu - b) \int_0^\infty \varepsilon(s) m^\mu(t-s)\, \mathrm{d}s, \end{aligned} \tag{17.37}$$

where we have used Eqs. (17.27) and (17.36).

Thus, in a network of N neurons (e.g., $N = 100\,000$) which has stored M patterns (e.g., $M = 2000$) the input potential is completely characterized by the M overlap variables, which reflects an enormous reduction in the complexity of the mathematical problem. Nevertheless, each neuron keeps its identity for two reasons:

(i) Each neuron i is characterized by its "private" set of past firing times t_i^f. Therefore each neuron is in a different state of refractoriness and adaptation which manifests itself by the term $\sum_f \eta(t - t_i^f)$ in the total membrane potential.

(ii) Each neuron has a different functional role during memory retrieval. This role is defined by the sequence $\xi_i^1, \xi_i^2, \ldots, \xi_i^M$. For example, if neuron i is part of the active assembly in patterns $\mu = 3, \mu = 17, \mu = 222, \mu = 1999$ and should be inactive in the other 1996 patterns, then its functional role is defined by the set of numbers $\xi_i^3 = \xi_i^{17} = \xi_i^{222} = \xi^{1999} = 1$ and $\xi_i^\mu = 0$ otherwise. In a network that stores M different patterns there are 2^M different functional roles so it is extremely unlikely that two neurons play the same role. Therefore each of the N neurons in the network is different!

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

However, during retrieval we can reduce the complexity of the dynamics drastically. Suppose that during the interval $t_0 < t < t_0 + T$ all overlaps are negligible, except the overlap with one of the patterns, say pattern ν. Then the input potential in Eq. (17.37) reduces for $t > t_0 + T$

$$h_i(t) = (\xi_i^\nu - b) \int_0^\infty \varepsilon(s) m^\nu(t-s)\,\mathrm{d}s, \tag{17.38}$$

where we have assumed that $\varepsilon(s) = 0$ for $s > T$. Therefore, the network with its N different neurons splits up into two homogeneous populations: the first one comprises all neurons with $\xi_i^\nu = +1$, i.e., those that should be "ON" during retrieval of pattern ν; and the second comprises all neurons with $\xi_i^\nu = 0$, i.e., those that should be "OFF" during retrieval of pattern ν.

In other words, we can apply the mathematical tools of population dynamics that were presented in Part III of this book to analyze memory retrieval in a network of N different neurons.

Example: Spiking neurons without adaptation

In the absence of adaptation, the membrane potential depends only on the input potential and the time since the last spike. Thus, Eq. (17.33) reduces to

$$u_i(t) = \eta(t - \hat{t}_i) + h_i(t) + u_{\text{rest}}, \tag{17.39}$$

where $\hat{t}_i$ denotes the last firing time of neuron i and $\eta(t - \hat{t}_i)$ summarizes the effect of refractoriness. Under the assumption of an initial overlap with pattern ν and no overlap with other patterns, the input potential is given by Eq. (17.38). Thus, the network of N splits into an "ON" population with input potential

$$h^{\text{ON}}(t) = (1-b) \int_0^\infty \varepsilon(s) m^\nu(t-s)\,\mathrm{d}s \tag{17.40}$$

and an "OFF" population with input potential

$$h^{\text{OFF}}(t) = (-b) \int_0^\infty \varepsilon(s) m^\nu(t-s)\,\mathrm{d}s\,. \tag{17.41}$$

For each of the populations, we can write down the integral equation of the population dynamics that we saw in Chapter 14. For example, the "ON"-population evolves according to

$$A^{\text{ON}}(t) = \int_{-\infty}^{t} P_I(t|\hat{t}) A(\hat{t}) \mathrm{d}\hat{t} \tag{17.42}$$

with

$$P_I(t|\hat{t}) = \rho(t) \exp\left[- \int_{\hat{t}}^{t} \rho(t')\,\mathrm{d}t' \right], \tag{17.43}$$

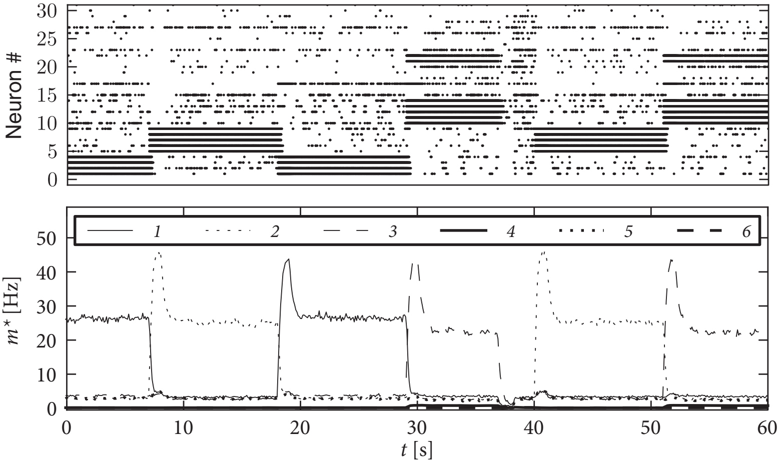


Fig. 17.11 Attractor network with spiking neurons. Memory retrieval in a network of 8000 excitatory neurons which stores 90 different patterns. Top: The spike raster shows 30 neurons selected and relabeled so that the first five neurons respond to pattern 1, the second group of five neurons to pattern 2, etc. Bottom: Overlap defined here as $m^{\mu *} = A^{\text{ON}}(t)$ with the first six patterns $1 \leq \mu \leq 6$. After a partial cue (t = 8,18.5, 19.5, 40, 51 s), one of the patterns is retrieved and remains stable without further input during a delay period of 10 seconds. Occasionally a global input to the inhibitory neurons is given leading to a reset of the network ($t = 38$ s). After the reset, the network remains in the spontaneously activity state.

where $\rho(t) = f(\eta(t-\hat{t}) + h^{\text{ON}}(t) + u_{\text{rest}} - \vartheta)$. An analogous equation holds for the "OFF"-population.

Finally, we use Eq. (17.36) to close the system of equations. The sum over all neurons can be split into one sum over the "ON"-population and another over the "OFF"-population, of size $a \cdot N$ and $(1-a) \cdot N$, respectively. If the number N of neurons is large, the overlap is therefore

$$m^{\nu}(t) = \frac{1}{2}[A^{\text{ON}}(t) - A^{\text{OFF}}(t)]\,. \tag{17.44}$$

Thus, the retrieval of pattern ν is controlled by a small number of macroscopic equations.

In an analogous sequence of calculations one needs to check that the overlap with the other patterns μ (with $\mu \neq \nu$) does not increase during retrieval of pattern ν.

17.3.2 Excitatory and inhibitory neurons

Synaptic weights in the Hopfield model can take both positive and negative values. However, in the cortex, all connections originating from the same presynaptic neuron have the same sign, either excitatory or inhibitory. This experimental observation, called Dale's law, gives rise to a primary classification of neurons as excitatory or inhibitory.

In Chapter 16 we started with models containing separate populations of excitatory and inhibitory neurons, but could show that the model dynamics are, under certain conditions,

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

equivalent to an effective network where the excitatory populations excite themselves but inhibit each other. Thus explicit inhibition was replaced by an effective inhibition. Here we take the inverse approach and transform the effective mutual inhibition of neurons in the Hopfield network into an *explicit* inhibition via populations of inhibitory neurons.

To keep the arguments transparent, let us stick to discrete time and work with random patterns $\xi_i^\mu \in \{0,1\}$ with mean activity $(\sum_i \xi_i^\mu)/N = a$. We take weights $w_{ij} = c' \sum_\mu (\xi_i^\mu - b)(\xi_j^\mu - a)$ and introduce a discrete-time spike variable $\sigma_i = 0.5(S_i + 1)$ so that $\sigma_i = 1$ can be interpreted as a spike and $\sigma_i = 0$ as the quiescent state. Under the assumption that each pattern μ has exactly $a \cdot N$ entries with $\xi_i^\mu = 1$, we find that the input potential $h_i = \sum_j w_{ij} S_j$ can be rewritten with the spike variable σ

$$h_i(t) = 2c' \sum_j \sum_\mu (\xi_i^\mu - b)\, \xi_j^\mu\, \sigma_j - 2c' \sum_j \sum_\mu (\xi_i^\mu - b)\, a\, \sigma_j\,. \tag{17.45}$$

In what follows we choose $b = 0$ and $c' = 1/4N$. Then the first sum on the right-hand side of Eq. (17.45) describes excitatory and the second one inhibitory interactions.

To interpret the second term as arising from inhibitory neurons, we make the following assumptions. First, inhibitory neurons have a linear gain function and fire stochastically with probability

$$\mathrm{Prob}\{\sigma_k = +1 | h_k^{\mathrm{inh}}\} = g(h_k^{\mathrm{inh}}(t))\Delta t = \gamma h_k^{\mathrm{inh}}(t)\,, \tag{17.46}$$

where the constant γ takes care of the units and k is the index of the inhibitory neuron with $1 \le k \le N^{\mathrm{inh}}$. Second, each inhibitory neuron k receives input from C excitatory neurons. Connections are random and of equal weight $w^{E \to I} = 1/C$. Thus, the input potential of neuron k is $h_k^{\mathrm{inh}} = (1/C) \sum_{j \in \Gamma_k} \sigma_j$ where Γ_k is the set of presynaptic neurons. Third, the connection from an inhibitory neuron k back to an excitatory neuron i has weight

$$w_{ik}^{I \to E} = \frac{a}{\gamma N^{\mathrm{inh}}} \sum_\mu \xi_i^\mu\,. \tag{17.47}$$

Thus, inhibitory weights onto a neuron i which participates in many patterns are stronger than onto one which participates in only a few patterns. Fourth, the number N^{inh} of inhibitory neurons is large. Taken together, the four assumptions give rise to an average inhibitory feedback to each excitatory neuron proportional to $\sum_j \sum_\mu \xi_i^\mu\, a\, \sigma_j$. In other words, the inhibition caused by the inhibitory population is equivalent to the second term in Eq. (17.45).

Because of our choice $b = 0$, patterns are only in weak competition with each other and several patterns can become active at the same time. In order to also limit the total activity of the network, it is useful to add a second pool of inhibitory neurons which turn on whenever the total number of spikes in the network exceeds $a \cdot N$. Note that biological cortical tissue contains many different types of inhibitory interneurons, which are thought to play different functional roles; Fig. 17.10.

Figure 17.11 shows that the above argument carries over to the case of integrate-and-fire neurons in continuous time. We emphasize that the network of 8000 excitatory and

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

two groups of inhibitory neurons (2000 neurons each) has stored 90 patterns of activity $a \approx 0.1$. Therefore each neuron participates in many patterns (Curti *et al.*, 2004).

In practice, working memory models with spiking neurons require some parameter tuning. Adding to working models a mechanism of synaptic short-term facilitation (see Chapter 3) improves stability of memory retrieval during the delay period (Mongillo *et al.*, 2008).

17.4 Summary

The Hopfield model is an abstract model of memory retrieval. After a cue with a partial overlap with one of the stored memory patterns is presented, the memory item is retrieved. Because the Hopfield model has symmetric synaptic connections, memory retrieval can be visualized as downhill movement in an energy landscape. An alternative view is that of memories forming attractors of the collective network dynamics. While the energy picture does not carry over to networks with asymmetric interactions, the attractor picture remains applicable even for biologically more plausible network models with spiking neurons.

Attractor networks where each neuron participates in several memory patterns can be seen as a realization of Hebb's idea of neuronal assemblies. At the current state of research, it remains unclear whether increased spiking activity observed in the cortex during delayed matching-to-sample tasks has a relation to attractor dynamics. However, the ideas of Hebb and Hopfield have definitely influenced the thinking of many researchers.

Literature

Precursors of the Hopfield model are the networks models of Willshaw *et al.* (1969), Kohonen (1972), Anderson (1972), and Little (1974). The model of associative memory of Willshaw *et al.* (1969) was designed for associations between a binary input pattern $\xi^{\mu,A}$ and a binary output vector $\xi^{\mu,B}$, where $\xi_i^{\mu,A/B} \in \{0,1\}$ and interactions weights w_{ij} are taken to vary as $\xi_i^{\mu,B}\xi_j^{\mu,A}$ where j is a component of the input and i a component of the desired output pattern. A recurrent network can be constructed if the dimensionality of inputs and outputs match and the output from step n is used as input to step $n+1$.

Intrinsically recurrent models of memory were studied with linear neurons by Kohonen (1972) and Anderson (1972) and with stochastic binary units by Little (1974). The latter showed that, under some assumptions, persistent states that can be identified with potential memory states can exist in such a network.

Hopfield's (1982) paper has influenced a whole generation of physicists and is probably the most widely cited paper in computational neuroscience. It initiated a wave of studies of storage capacity and retrieval properties in variants of the Hopfield model using the tools of statistical physics (Amit *et al.*, 1985, 1987b) including extensions to low-activity patterns (Amit *et al.*, 1987a; Tsodyks and Feigelman, 1986), sparsely connected networks (Derrida *et al.*, 1987) and temporal sequences (Sompolinsky and Kanter, 1986; Herz *et al.*, 1989). The energy function in Hopfield (1982) requires symmetric interactions, but the dynamics

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

can also be analyzed directly on the level of overlaps. The book by Hertz *et al.* (1991) presents an authoritative overview of these and related topics in the field of associative memory networks.

The transition from abstract memory networks to spiking network models began after 1990 (Amit and Tsodyks, 1991; Gerstner, 1991; Gerstner and van Hemmen, 1992; Treves, 1993; Amit and Brunel, 1997a,b) and continued after 2000 (Curti *et al.*, 2004; Mongillo *et al.*, 2008), but a convincing memory model of spiking excitatory and inhibitory neurons where each neuron participates in several memory patterns is still missing. The relation of attractor network models to persistent activity in electrophysiological recordings in monkey prefrontal cortex during memory tasks is discussed in the accessible papers of Barbieri and Brunel (2008) and Balaguer-Ballester *et al.* (2011).

Exercises

1. **Storing one or several patterns**.

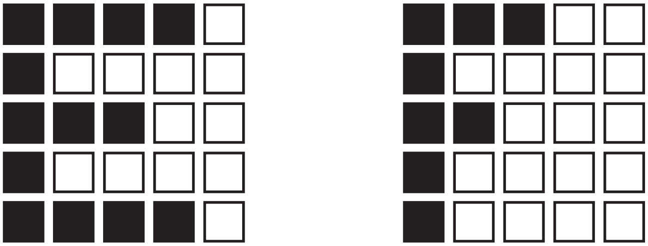


Fig. 17.12 Patterns for "E" and "F" in a Hopfield network of 25 neurons.

(a) In a Hopfield network of 25 binary neurons (Fig. 17.12), how would you encode the letter E? Write down couplings w_{ij} from arbitrary neurons onto neuron i if i is either the black pixel in the lower left corner of the image or the white pixel in the lower right corner.
(b) Suppose the initial state is close to the stored image, except for m pixels which are flipped. How many time steps does the Hopfield dynamics take to correct the wrong pixels? What is the maximum number of pixels that can be corrected? What happens if 20 pixels are flipped?
(c) Store as a second pattern the character F using the Hopfield weights $w_{ij} = \sum_\mu p_i^\mu p_j^\mu$. Write down the dynamics in terms of overlaps. Suppose that the initial state is exactly F. What is the overlap with the first pattern?

2. **Mixture states.** *Use the rule of the Hopfield network to store the six orthogonal patterns such as those shown in Fig. 17.7b but of size* 8×8.
(a) Suppose the initial state is identical to pattern $\nu = 4$. What is the overlap with the other patterns $\mu \neq \nu$?
Hint: Why are these patterns orthogonal?
(b) How many pixels can be wrong in the initial state, so that pattern $\nu = 4$ is retrieved with deterministic dynamics?
(c) Start with an initial state which has overlap with two patterns: $m^1 = (1-\alpha)m$ and $m^2 = \alpha m$ and $m^\mu = 0$ for $\mu \geq 3$. Analyze the evolution of the overlap over several time steps, using deterministic dynamics.
(d) Repeat the calculation in (c) but for a mixed cue $m^1(t) = m^2(t) = m^3(t) = m < 1$ and $m^\nu(t) = 0$ for $\nu > 3$. Is the mixture of three patterns a stable attractor of the dynamics?

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press

(e) Repeat the calculation in (d), for stochastic dynamics.

3. **Binary codes and spikes.** *In the Hopfield model, neurons are characterized by a binary variable $S_i = \pm 1$. For an interpretation in terms of spikes it is, however, more appealing to work with a binary variable $\sigma_i \in \{0,1\}$.*
   (a) Write $S_i = 2\sigma_i - 1$ and rewrite the Hopfield model in terms of the variable σ_i. What are the conditions so that the the input potential in the rewritten model is simply $h_i = \sum_j w_{ij}\sigma_j$?
   (b) Repeat the same calculation for low-activity patterns and weights $w_{ij} = c' \sum_\mu (\xi_i^\mu - b)(\xi_j^\mu - a)$ with some constants a, b, c' and $\xi_i^\mu \in \{0,1\}$. What are conditions such that $h_i = \sum_j w_{ij}\sigma_j$?

https://doi.org/10.1017/CBO9781107447615.021 Published online by Cambridge University Press