Statistical Coding and Short-Term Synaptic Plasticity:
A Scheme for Knowledge Representation in the Brain

Christoph von der Malsburg
Abteilung für Neurobiologie, Max-Planck-Institut für Biophysikalische Chemie
D-3400 Göttingen W.Germany
and
Elie Bienenstock
Laboratoire de Neurobiologie du Développement
Bâtiment 440, Université de Paris-Sud
F-91405 Orsay Cedex France

Abstract. This work is a theoretical investigation of some consequences of the hypothesis that transmission efficacies of synapses in the Central Nervous System (CNS) undergo modification on a short time-scale. Short-term synaptic plasticity appears to be an almost necessary condition for the existence of activity states in the CNS which are stable for about 1 sec., the time-scale of psychological processes. It gives rise to joint “activity-and-connectivity” dynamics. This dynamics selects and stabilizes particular high-order statistical relationships in the timing of neuronal firing; at the same time, it selects and stabilizes particular connectivity patterns. In analogy to statistical mechanics, these stable states, the attractors of the dynamics, can be viewed as the minima of a hamiltonian, or cost function. It is found that these low-cost states, termed synaptic patterns, are topologically organized. Two important properties of synaptic patterns are demonstrated: (i) synaptic patterns can be “memorized” and later “retrieved”, and (ii) synaptic patterns have a tendency to assemble into compound patterns according to simple topological rules. A model of position-invariant and size-invariant pattern recognition based on these two properties is briefly described. It is suggested that the scheme of a synaptic pattern may be more adapted than the classical cell-assembly notion for explaining cognitive abilities such as generalization and categorization, which pertain to the notion of invariance.

We are still in almost complete ignorance of how the brain works: there are some very good cues that the Central Nervous System (CNS) is a probabilistic distributed highly non-linear dynamical system—or asynchronous network of “automata”—but we have practically no conceptual tools for studying such knowledge representation devices. The remarkable performances of our cognitive apparatus have received little convincing explanation on the basis of neuronal functioning. One of the most intriguing aspects of cognition is perhaps best described by the notion of an “invariant”. Many of the outstanding abilities of the brain, such as categorization and generalization, are directly related to it and could probably be better understood if we had a solid theory of how the brain generates and manipulates invariants. This would require as a first step clarifying the notion of invariance itself. The studies of perception by the Gestalt psychologists at the turn of the century have popularized the idea that

NATO ASI Series, Vol. F20
Disordered Systems and Biological Organization
Edited by E. Bienenstock et al.
© Springer-Verlag Berlin Heidelberg 1986
invariances result from construing objects as structured sets of relationships between their parts, rather than as mere sums or juxtapositions. Modern work in artificial intelligence indeed concentrates on the elaboration of flexible schemes for representing and handling relational information of various types. It should be expected that an understanding of how the brain represents such information using neurons and synapses would constitute a major breakthrough on both the theoretical and the experimental levels.

1. The Assembly as Dynamical Object of the Brain

According to the currently dominating view the relevant signal for the function of the brain is contained in the first-order statistics of the electrical activity of neurons: neural signals are independent Poisson processes the rates of which encode signals which can be extracted as a short-term mean. The time interval over which the mean is to be taken may depend on the location and function of the neuron. This hypothesis of "rate-coding" is strongly rooted in the field of neurophysiology, where it is an everyday experience that the response elicited in a single cortical neuron by peripheral sensory stimulation can hardly if ever be exactly reproduced. There remains an "irreducible" variability in the discharge of the neuron, attributed mainly to fluctuations in the amount of transmitter released in the synaptic cleft upon arrival of a spike in the axon terminal. This variability is a nuisance in most neurophysiological experiments; one eliminates it by averaging over many stimulation trials (technique of "post- or peri-stimulus-time-histogram"). It is also regarded as a nuisance to brain function by most authors. Since the brain cannot afford the time necessary to average over many trials, the hypothesis of rate coding is usually complemented by the notion of using redundancy, i.e., averaging over many neurons carrying out the same function simultaneously. Ensemble averaging, instead of temporal averaging, is thus thought to be the means by which the brain performs "reliable computation in the presence of noise".

The monograph in which Hebb introduced in 1949 the notion of the "cell-assembly" and emphasized its associative-memory properties has inspired much of later brain theory. Roughly, a cell-assembly is a subset of neurons with many mutual excitatory inter-connections. For most authors, including Hebb himself, the functioning of cell-assemblies is to be understood within a strict rate-coding framework. Once a state is initiated in which the mean rate of firing of most or all cells belonging to a given assembly is higher than average, it tends to persist for some time because of the strong excitatory connections between cells within the assembly. Thus, many models use the notion of a preferred or persistent activity state, which is thought to be more stable than a random activity configuration. The brain would spend about half a second or less second—the "observed duration of a single content in perception"—in one of these states before switching to another persistent state. One may refer to this characteristic range of intervals as to the "psychological time-scale". The internal structure of the space of attractors, or locally stable states, as well as their relationships to external objects, are determined by the connectivity of the system, i.e., the strengths, or efficacies, of the various synapses between the neurons. Connectivity is first globally laid down during early development, and later tuned during various stages of learning.

The description in terms of attractors emphasizes the aspects of brain dynamics linked to discrete events such as single "percepts". It provides a useful model of associative, or content-addressable, also called "error-correcting", memory. The attractors are the memories of the system. Suppose the system is started at a point in the activity space which is near one memory; the system will then with high likelihood settle to this particular memory. One may then say that a content-addressed memory has been retrieved, that appropriate associations have been made, or that errors have been corrected. Alternatively, the information to be processed can be fed into the network by "clamping" part of the activity variables, letting all the others undergo the usual dynamics; the system may be used in this way to solve constraint-satisfaction tasks.

Against the notion of the assembly the argument may be raised that the brain seems not at all to be designed to support persistent configurations of firing rates. We know that neuronal integration is highly non-linear: several spikes in excitatory afferents to a neuron which impinge in close neighborhood both in time and in space (location of the nerve endings over the dendritic arborization) are much more effective in triggering an action potential in the postsynaptic cell than the same number of spikes distributed over a somewhat larger spatio-temporal domain. The neuron may consequently be regarded as a "coincidence-detector" in the domain of the millisecond. This timescale is set by membrane time-constants, by axonal conduction times and by synaptic delays, which are all in the millisecond range. On the other hand, mean firing rates in cerebral cortex are low, typically between 5 and 20 spikes per second, and the excitation level of cells is bound to fluctuate wildly. There is, therefore, a high probability that fluctuations in the actual spike rate would extinguish an assembly which should remain on or would trigger an assembly which should be off. This instability of the assembly is a consequence of the apparent gap between the psychological and the physiological time-scales, half a second and a few milliseconds.

Arguments criticizing the assembly as the fundamental "data format" of the brain have been raised elsewhere. The main point is a lack of flexible representation of relational knowledge by the assembly. Together with the stability arguments above these considerations have lead us to an altogether different approach to brain dynamics. This new approach is based on higher-order statistics as being the essential signal in the brain. As was just mentioned, neurons are extremely sensitive to highly coincident signals, they are strongly activated by such "favorable events". But how can favorable events be reliably produced? An arbitrary network with as dense interconnections as the cortical one is likely to produce flat high-order statistics. Networks of special type are needed to produce favorable events.

It has been suggested that coincidences could be propagated in pathways of particular architecture, called "synfire chains". Neurons in each link of the chain converge and also diverge—onto neurons in the next link. Whenever sufficiently many cells at one link become simultaneously active, this synchronized activity elicits synchronized activity in the next link, and so on. In such a network coincidences don't get lost since neuronal activity propagates like a coherent traveling wave. A stable state could be obtained in a synfire chain which would be closed on itself, somewhat reminiscent of
the old "reverberating circuita" of Lorente de No. Any cell could be part of several synfire chains, a particular one being selected by the precise set of other neurons with which the cell happens to fire in synchrony. Difficulties with the synfire chain concept arise because (i) long regular chains of this kind are unlikely to develop in the CNS; (ii) the firing of cells in such a chain would be periodic, yet neurons exhibit only little periodicity other than that linked to respiration or blood-flow; and most importantly because (iii) such a model supports only those stable activity modes which are rigidly determined by the structure of the network.

The approach taken here is of a slightly more radical kind: the particular network structure needed to support a particular signal structure is created by the signals themselves on the psychological time scale in a bootstrapping fashion. This is possible with the help of a synaptic mechanism for remembering, at any given synapse, events that occurred a short time ago at this synapse. Such a mechanism would stabilize good timing relationships and thus favor the re-occurrence of events which were successful from the viewpoint of transmission of excitation through the synapse. The ensuing network structures and signals constitute a data format which is very appropriate for typical operations performed in the brain, as will be argued in later sections of this paper.

2. Short-Term Plasticity

As was originally proposed elsewhere by one of the authors, we assume that within the CNS the following rules apply.

Rule A: successful synaptic events enhance the transmission efficacy of the synapse.

Rule B: transmission failures such as presynaptic firing without postsynaptic firing, and possibly also "failures" of the inverse type, i.e., postsynaptic without presynaptic firing, depress synaptic efficacy.

Both types of plastic change become effective within a few milliseconds.

These rules can be modified in an obvious way to include the case of inhibitory synapses. For instance, firing of an inhibitory afferent followed by postsynaptic firing is a transmission failure; such an event depresses the efficacy of the concerned synapse.

Short-term plasticity is restricted to a small range of strengths: the absolute value of the strength saturates after a small number of similar events. In case of successful transmission, it reaches a maximal level. In case of failures, it settles at a minimum.

Short-term plasticity lets connectivity and activity evolve on the same time-scale, leading to the notion of a joint activity-and-connectivity state. It is proposed that the events which underlie brain function are best described by such compound states, rather than by mere activity states. The dynamics of this new type is characterized by positive feedback: a successful event occurring at a given synapse increases the efficacy of this synapse, which in turn increases the likelihood of re-occurrence of the same or a similar event again. Positive feedback and saturation of synaptic weights ensure stability of the joint activity-connectivity state. We shall also see that positive feedback operates as a selection mechanism which favors and stabilizes a particular type of activity-connectivity configurations.

Short-term plasticity is related to hebbian synaptic modification: "When an axon of cell A is near enough to excite a cell B and repeatedly and persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells so that A's efficiency as one of the cells firing B, is increased." However, Hebb's rule requiring repeated and persistent coincidences, hebbian modification is generally understood as depending on the correlation between pre- and post-synaptic activities, estimated over a long period of time. It is a slow and long-lasting modification of the network's connectivity, a developmental or learning process which is invoked to explain the "growth of the assembly." In contrast, short-term synaptic modification is in force after a single synaptic event, or a small number of them. It is also reversible, for an unsuccessful event may occur shortly after a successful one and totally undo what the first event did. Long-term plasticity can now be more precisely formulated, as time-integrated short-term modification.

A randomly connected cortex-like network can be described as an entanglement of pathways of different length and delay, so that any fine temporal structure becomes washed out by the dynamics and flat high-order statistics is to be expected. Short-term plasticity as summarized in Rules A and B selects out activity-connectivity states which depart in a systematic way from both independence of activity fluctuations in different neurons and randomness of connectivity. An excitatory synapse can only be successful if there are other active synapses converging on the same target cell and receiving synchronous impulses, thus producing a strong postsynaptic effect; also, few or no activated inhibitory afferents should have fired in the last milliseconds. All synapses participating in a successful event are strengthened, thus making reoccurrence of the event more likely. Clearly then, successful events signify a departure from random firing. They also signify departure from random connectivity: there must be something in the current connectivity state of the network which makes it likely that all fibres in a particular group fire in synchrony, and makes it unlikely that fibres belonging to another group fire at the same time. Stable activity-connectivity states differ from random states in that there is a good match between activity and connectivity, that is, between successful events and activated synapses, and between failures and deactivated synapses. The positive feedback dynamics achieves such optimal matching under the constraints set by the physical connections and by first-order statistics.

How can stable connectivity states be characterized in distinction to random ones? The following idea may be relevant. In an isolated network, coincidences can be generated by "amplifying" single spikes if there are many parallel alternative pathways of equal length between a pair of neurons. (The length of a multi-synaptic pathway between two neurons in the CNS is obtained by adding up all the axonal conduction times, synaptic delays and somatic-dendritic integration times involved.) Whenever the common "source" i emits a spike, this signal diverges along the currently active pathways out of neuron i, some of which converge again on the common "target" j, with accurate
increases the likelihood of coincident events in comparison to a random network. We contend that connectivity states with many such micro-configuration.s are the most favorable to the occurrence of coincidences. Notice that there should exist an optimal length for the alternative parallel pathways in the favorable micro-configurations. Very short paths don't contribute much because there simply cannot be many of them. Long paths on the other hand are not reliable and introduce scatter in the timing relationships. We leave open the issue of a quantitative estimation of the optimal length.

The number of favorable micro-configurations could be trivially maximized by activating all available excitatory links. This would lead to global synchronization in the network. Inhibition acts to block this route, by limiting the number of cells which are simultaneously active. The number of synapses required to connect any two neurons in cortex—the graph theoretic distance between them—is very small, possibly 5 on the average\(^1\). If all excitatory synapses were activated, single spikes would be amplified to enormous numbers within a few synaptic delays. Inhibition would counteract this amplification, thereby leading to transmission failure in many synapses and reducing the number of activated synapses. It may therefore be assumed that in a valid network configuration only a small number of synapses to or from each cell is activated. In consequence, stable connectivity states are to be characterized as being a compromise between maximal number of favorable micro-configurations and sparsity. We will later give a more global characterization in terms of topological structure.

From now on, we will use the term synaptic pattern for the stable states, or attractors, of the activity-and-connectivity dynamics. The word "pattern" is meant to suggest the existence of internal structure. This stands in contrast with the "assembly" where no internal structure exists apart from stronger-than-average excitatory connectivity. This distinction will become clearer in the next sections.

Synfire chains are a particular instance of synaptic pattern: they clearly contain many of the required micro-configurations, of length between 2 and the length of the chain. They form only a small subclass of the space of synaptic patterns, because their architecture is very regular. The generic synaptic pattern has no global order. It is an irregular high-dimensional mosaic of small multiple-parallel-alternative-paths. The image of a mosaic is, however, slightly misleading, for these preferred micro-configurations are not just laid down on the side of each other like stones in a mosaic. Rather, they are entangled in a complicated, essentially random way: each source-cell or target-cell in one such subgraph is an intermediate cell from the standpoint of others. There are two equivalent ways to look at the stable activity-and-connectivity states: using either connectivity or activity. One could try to characterize the organization in these states by high-order statistical moments, which would tell how the activity propagates in the network. In the case of a synfire chain, there is a well-defined topology which allows the activity to propagate like a traveling wave. In the generic synaptic pattern no such global order exists, yet we shall see that local topological structure is still there. This allows for a type of "propagation" which is much less coherent than in a synfire chain, yet with more coincidences than if all firing processes were statistically independent. In all other respects, the activity in a synaptic pattern looks totally random; it is perfectly aperiodic, has low first-order moments, i.e., average firing rates, and cannot be distinguished from random activity by looking at only one neuron at a time.

No experimental evidence exists at the time in support of short-term synaptic modification. It is in general quite hard to unambiguously demonstrate a change in the permeability of a synapse in the CNS of higher animals. There are several reasons for this, the chief one being the highly complex connectivity which makes it difficult to tell what it is that caused an observed change in the "response properties" of a cell. On the other hand, the hypothesis of short-term plasticity is consistent with current knowledge on molecular processes at the synapse. It has, for instance, been suggested that the dendritic spine undergoes fast geometrical changes\(^12\), which could have significant effects on the strength of the transmitted impulse\(^13,14,15\). The receptor molecule in the post-synaptic membrane could also be subject to fast reversible conformation changes which would affect its sensitivity to the neurotransmitter; in some cases, such "allosteric" effects are well documented\(^16,17\). It remains to be proven that such changes may take place on a short time-scale, and obey Rule A and Rule B for enhancement and depression of synaptic efficacy. At any rate, it is a crucial requirement that modification depend on both pre- and post-synaptic activities. The latter is meant in a broad sense: it could be a local signal originating from neighboring synapses on the dendrite, which would not necessarily involve propagated activity, i.e., spiking, in the postsynaptic neuron. Also, it is not required that all synapses in the system be modifiable: short-term plasticity could be restricted to excitatory synapses or to a certain sub-category of these. Whether plasticity of inhibitory synapses would suffice deserves further investigation. The only absolute prerequisite for the theory proposed here is the existence of molecular mechanisms which result in local short-term positive feedback between the occurrence of "successful events" at a synapse and the transmission efficacy of that synapse.

3. A Simplified Formulation

It will not be possible for some time to give a precise and realistic formulation of brain dynamics. Most of the detail of interest is not accessible to experiment at present. Even if we had exact knowledge of all relevant neural mechanisms it would not form a convenient basis for mathematical treatment. Those mechanisms are the result of a long history of evolutionary optimization and are bound to be complicated. We shall base all further discussion on a simplified formulation of neural dynamics. This is done in the spirit of statistical physics, which abounds in radical simplifications. The Ising model of magnetism, for instance, is a caricature of real-world physical processes, yet it has proven to be very successful in the study of phase transitions. The elucidation of the fundamental notion of a "universality class" would have been impossible without simplifying steps which capture the essence of the interactions involved.

As a first simplification we restrict explicit treatment to excitatory synaptic weights and will consider neural signals and inhibitory synapses as "hidden variables", the effect of which on the excitatory weights being taken into account implicitly. We denote the instantaneous efficacy of the synapse between cell \(j\) and cell \(i\) by \(w_{ij}\). We assume symmetry, \(w_{ij} = w_{ji}\), and exclude self-coupling: \(w_{ii} = 0\). We furthermore restrict synaptic weights to the values 1 or 0. Dendritic geometry, which probably restricts direct communication between synapses to local pools, is neglected, all synapses converging onto one cell being treated as equivalent. As a consequence of these simplifications, the
current state of the system is entirely described by its connectivity which is simply a non-directed graph, with no kinks and no parallel edges. This graph has the fixed set of vertices \(V = \{1, \ldots , n\}\), which corresponds to the set of neurons in the network. The set of links, \(\{w_{ij} | i, j = 1, \ldots , n\}\), is the dynamic object.

For a connection to be modified within short time it must exist physically. Let us denote by \(s_{ij} = 1\) the fact that cells \(j\) and \(i\) are connected by an axon and a synapse (neglecting, again, the unidirectionality of synapses). \(s_{ij} = 0\) then stands for a non-existing physical connection. We correspondingly have the constraint \(w_{ij} \leq s_{ij}\). In cortex, a typical neuron has between \(10^3\) and \(10^8\) synaptic contacts.

The effects of the "hidden variables" are described as rules of cooperation and competition between synaptic weights. The effect of inhibition is to produce competition between the links at each node and keep their number below a certain parameter \(p\), where \(p\) is much smaller than the number of synaptic contacts of neurons. Neural signals induce cooperation between sets of links which for constellations favorable for the production of synchronous events. As indicated above, we try to capture this effect by introducing cooperativity between alternative paths of equal length between two cells \(i\) and \(j\). Since we also ignore the orientation of edges, two such paths simply form a cycle (a closed circuit) of even length. We further assume that only cycles of a given optimal length \(q\), or \(q\)-cycles, contribute (\(q\) would be twice the length of the alternative pathways between a source-neuron and a target-neuron in an oriented-graph version).

To formalize the above dynamic rules we shall adopt the well-known Gibbs-Boltzmann formalism. There, a central role is played by a function of the state \(H(w)\) called the energy, or Hamiltonian function. The function \(H(w)\) is a sum of local contributions corresponding to all the interactions between the variables in the system. Intuitively, \(H(w)\) is a measure of the extent to which these interactions are "satisfied" in the present global state \(w\). It may therefore also be considered a "cost function" of an optimization problem. In our case, \(H(w)\) is the negative sum of the number of \(q\)-cycles:

\[
H(w) = - \sum_{(i_1, \ldots , i_q)} w_{i_1, i_2} \cdots w_{i_{q-1}, i_q} w_{i_q, i_1}
\]

with the constraints

\[
\sum_j w_{ij} \leq p \quad i = 1, \ldots , n
\]

\[
\sum_i w_{ij} \leq s_{ij} \quad i, j = 1, \ldots , n
\]

Constraint (2) limits the number of links at each node, whereas (3) makes sure the link between cells \(i\) and \(j\) cannot be activated if there is no pre-existing physical connection \(s_{ij}\) between \(i\) and \(j\). The dynamics of the system is now formulated simply by requiring that the probability of finding the system in state \(w\) obeys a Gibbs-Boltzmann distribution \(P(w) = \exp(-H(w)/T)/Z\) —where \(Z\) is a normalization constant and \(T\) is the "temperature" of the system—with the proviso that only states valid under the constraints (2) and (3) are admitted. (The constraints could also be implemented as additive terms in \(H(w)\). In our simulations we preferred to impose them kinematically, as formulated here.) According to this formula, low-energy states are more likely than high-energy ones. The shape of the resulting probability landscape also depends in a global way on \(T\): lowering \(T\) sharpens the peaks and troughs, whereas increasing it flattens the landscape out; at high \(T\), all states are nearly equally likely, whereas at low \(T\), high-energy states have essentially zero probability. The effect in the present case is to give states with high numbers of \(q\)-cycles high probability, subject to the constraints.

An algorithm which is often used to study computer simulations the equilibrium behaviour of such systems at a given finite \(T\) is the Metropolis algorithm.\(^{16}\) It consists of a sequence of tentative moves of the state \(w\). The criterion for accepting or rejecting a move is probabilistic and depends on the change in \(H(w)/T\): the larger the increase in \(H(w)\), the less likely one is to accept that move. In simulated annealing, one starts the process with high \(T\), thus avoiding trapping in local minima, and slowly reduces \(T\) to find one of the absolute energy minima, or states whose energy-level is very close to the absolute minimum.\(^{19}\)

The result of the dynamics described are synaptic patterns in the form of sparse graphs with many \(q\)-cycles. The formulation given contains non-local interactions between links. This non-locality is the result of suppressing the signal variables. It should be stressed that \(H\) and \(T\) are by no means a physical energy and a physical temperature. The analogy with physical systems is simply a convenient tool to simulate the dynamics on the computer. In physical systems, \(T\) measures the amount of thermal noise, relative to the strength of the local interactions that make up the function \(H\). The role of noise in cortex can be played by a non-specific diffuse afferent system to the cortex which injects uncorrelated activity interfering with the mechanism of coincidence detection. The amount of injected noise could even be varied in an appropriate way to achieve an effect analogous to annealing in the brain.

The reader will already have noticed the analogy of the system presented here to some systems studied in statistical physics. There, the term frustration\(^{20}\) is used to designate large systems of interacting "particles" with the following two properties: (i) there is a certain amount of conflict between local interactions at any point of the system, and (ii) there is quenched disorder, i.e., these local interactions are of random type. The prototypical frustrated system in solid-state physics is the spin glass, but many complex optimization problems such as the traveling salesman problem are also frustrated systems and can be studied within a similar statistical mechanics framework. Frustrated systems exhibit several interesting properties; one of them is their high degeneracy: there exist many almost-minimum energy states which, despite their being very different—they are quite distant from each other in configuration space—have nearly identical energy levels. As solutions to an optimization problem these many different states are virtually as good as the optimal solution(s). The problem of minimizing (1) under constraints (2) can be regarded as a frustrated problem.
4. Graph Structures

The low-energy states of (1-2) can be studied under various constraints \( s \) in (3). Numerical studies use simulated annealing; an attempted move consists either in a single link activation if it is allowed by constraints (2) and (3), or in the deactivation of a single link. We first consider the "generic" patterns, i.e., the minima of \( H \) under constraints (2), the constraints (3) being void: \( s_{ij} = 1 \) for all \( i \neq j \); that is, the graph of permanent links is the complete graph on \( V \). We shall see that several simple topological properties of graphs are related to being a low-\( H \) state. The argument is intuitive, strongly supported by computer simulations, but by no means a strict mathematical proof. It nevertheless involves some mathematical notions about random graphs.

The topological characterization of patterns involves the comparison with random graphs in the thermodynamic limit, i.e., when \( n \), the size of the graph, goes to infinity. Large random graphs contain very few cycles of any given order \( k \), when \( k \) is small with respect to \( n \). Stated accurately, the distribution of the number of cycles of a given fixed order \( k \) in a random graph obeys, in the thermodynamic limit, a Poisson law, the mean of which depends on \( k \) and on the average number of edges per node but not on \( n \). In contrast, in a regular lattice in a euclidean space, the number of \( k \)-cycles for a given \( k \) grows linearly with \( n \)—provided the lattice contains \( k \)-cycles, e.g., \( k \) is an even integer not less that 4 in a rectangular lattice with nearest-neighbor connections. Such a lattice, which contains many short cycles, is an instance of a low-\( H \) state, for appropriate values of \( q \). As was already mentioned, the graph-theoretic distance between two nodes in any graph is the length of the shortest path between them. In a random graph, the size of the set of nodes within distance \( k \) from a given node grows rapidly: it is an exponential function of \( k \)—as long as \( k \) is small with respect to the size of the graph. In contrast, in a regular lattice in a euclidean space, the size of a neighborhood of a given node increases only as a power function of the radius of this neighborhood. For any graph, the exponent in this function is called the fractal dimensionality of the graph. Fractal dimensionality of euclidean lattices is finite—it is identical to the dimension of the space—while it is infinite for random graphs.

It is argued that patterns are in many respects closer to euclidean lattices than to random graphs. Computer simulations show that the links in a pattern obey the following empirical rule: when several short indirect paths connect node \( i \) to node \( j \), it is likely that the direct connection between \( i \) and \( j \) exists as well. This is due to the cooperation of edges in \( q \)-cycles, which tends to make the connectivity "consistent with itself." The neighborhood relationships defined by such a connectivity are quite different from those found in a random graph; in contrast, there is a good match between these relationships and low-dimensional euclidean topology. Parts of a pattern may indeed be embedded in \( \mathbb{R}^2 \)—for instance—in such a way that the graph-theoretic distance and the euclidean distance agree with each other (see ref. 22 for details on this embedding problem). The simulations support the conjecture that the fractal dimensionality of patterns is finite. The chief difference between a pattern and a euclidean lattice is that the former is quite irregular; as a result of this, the embedding in a low-dimensional space cannot be achieved for the pattern as a whole.

The local minima of \( H \). Given the set of tentative moves used in the annealing, a local minimum of \( H \) is a graph such that there are exactly \( p \) edges out of each node and each edge belongs to at least one \( q \)-cycle. In general, this does not ensure low energy; one may achieve lower energy under constraints (2), each edge being part of several \( q \)-cycles as is the case in a rectangular lattice. Slow annealing is precisely required to avoid the local minima of high energy. Let it be mentioned for the sake of completeness that the absolute minima of \( H \) are regular graphs of a very particular type. They are disjoint unions of graphs of small size, namely of the order of \( p \) or \( q \), which in general exhibit strong symmetries. For some values of \( p \) and \( q \), each of these small graphs is a clique, i.e., a completely connected subgraph. In practice, as soon as \( n \) is much larger than \( p \) and \( q \), these absolute minima are too hard to find, even with very slow annealing. A typical pattern is a connected graph with the topological properties outlined above.

Given \( n, p \), and \( q \), we now define the set of \( (n, p, q) \)-patterns as the set of sub-optimal, i.e., almost best, solutions to the problem of minimizing (1) under constraint (2)—again not a strict mathematical definition. There are two important facts we need to know about \( (n, p, q) \)-patterns: (i) the edges in an \( (n, p, q) \)-pattern define a topological structure which is locally that of a low-dimensional euclidean space; and (ii) low-dimensional regular lattices with connections to the \( p \)-nearest-neighbors are instances of \( (n, p, q) \)-patterns, for appropriate \( p \) and \( q \). The notion of a pattern is deliberately ill-defined, yet each occurrence of the word "pattern" may from now on be replaced by the more restrictive "\( p \)-nearest-neighbor lattice" with no appreciable loss.

Remarkable collective properties of our dynamical system arise when constraints \( s \) are imposed which already contain patterns. We shall study two such types of constraints. In the first one, the graph \( s \) of permanent links is a random mixture of several patterns. We shall focus on the case of a mixture of two isomorphic patterns (see Fig. 1), which may be easily generalized. Let \( s^1 \) be an \((n, p, q)\)-pattern. Let \( P \) be a random permutation of the set of vertices \( V = \{1, \ldots, n\} \), and let \( s^2 \) be defined as follows: for any \( i \) and \( j \) in \( V \), \( s^2_{ij} = 1 \) if and only if \( \tilde{s}^1_{P(i)P(j)} = 1 \). Let now \( s = s^1 \cup s^2 \), i.e., for any \( i \) and \( j \) in \( V \), \( s_{ij} = 1 \) if and only if \( s^1_{ij} = 1 \) or \( s^2_{ij} = 1 \). Since the permutation \( P \) is random and \( p \ll n \), the overlap between the graphs \( s^1 \) and \( s^2 \) is small, and there are exactly \( 2p \) links at almost each \( i \) in the graphs \( s \). Constraints (2) and (3) therefore imply that at each node \( i \), \( p \) links have to be selected out of the \( 2p \) available.

Proposition 1. Under the above conditions on \( s \), the problem of minimizing \( H(w) \) under constraints (2) and (3) has two optimal solutions, namely \( w^1 = s^1 \), and \( w^2 = s^2 \).

We do not attempt here to make a mathematical theorem of this proposition. The physical intuition behind Prop. 1 is the following. Since the permutation \( P \) is random and \( p \) and \( q \) are small with respect to \( n \), the frequency of "mixed" \( q \)-cycles using edges belonging to both \( s^1 \) and \( s^2 \) is vanishingly small. Hence, at each node \( i \), \( s^1 \)-edges compete with \( s^2 \)-edges for activation. On the other hand, \( s^1 \)-edges belonging to a common \( q \)-cycle cooperate with each other, and so do \( s^2 \)-edges. Cooperation and competition propagate in the system in the following sense. Once a "local decision" is taken at node \( i \) to activate \( s^1 \)-edges rather than \( s^2 \)-edges, it is more rewarding to stick to this decision and activate \( s^1 \)-edges at nodes which are neighbors of \( i \) in the topology.
of $s^1$, since this allows to utilize each edge in a number of cycles as large as can be possibly achieved.

Prop. 1 is stated in the general case where the graph $s^1$ used to define the constraints $s$ is any $(n, p, q)$-pattern. One can make from Prop. 1 a mathematical statement by restricting the definition of $s^1$ to regular lattices. (The statement would have to be probabilistic, for a random permutation enters the definition of $s$, and would be made in the thermodynamic limit.) Prop. 1 can also be stated as follows: with an adequate choice of parameters, the dynamics of local interactions results in global competition between randomly superimposed patterns. It is crucial that the stored graphs $s$ be patterns. If the $s$ were random graphs, they could by no means be separated after they had been mixed: in that case, the optimization problem would be highly degenerate, all the "attractors" being random mixtures of the stored random graphs. Prop. 1 generalizes to the superposition of more than two patterns, and to the superposition of patterns which are not isomorphic to each other. It should be mentioned that as soon as more than one pattern is stored in a network, "spurious attractors", i.e., local minima appear. These "solutions" are essentially made of pieces of stored patterns glued together. Simulated annealing is necessary if one wants to avoid these local minima. If more and more patterns are stored in a network of fixed size $n$, "mixed" $q$-cycles become more and more numerous. In a first step, this forces one to use slower and slower annealing schedules. Eventually, when too many patterns are stored, the ability to retrieve them even with very slow annealing schedules breaks down altogether: some patterns which are mixtures of the stored patterns contain more $q$-cycles, hence are of lower energy, than the "pure" ones. Computer simulations have been performed using one-dimensional lattices: the stored patterns $s$ are isomorphic, randomly permuted replicas of the first-$p$-neighbors lattice on the circle. Typical values of the parameters are $n = 150$, $p = 6$, and $q = 4$. Such simulations clearly show the existence of the absolute minima of Prop. 1, and of many local minima.

We now turn to another type of constraint $s$, the last to be studied here. For the sake of clarity, we first state things informally: the system selects out graphs which form continuous maps between stored patterns whenever it is given a chance to do so. Formally, the constraints are the following (see Fig. 2). Let the set of vertices $V$ be divided into $p$ disjoint subsets: $V = V^1 \cup V^2$, with $V^1 \cap V^2 = \emptyset$. $V^1$ and $V^2$ are termed "layers". Let the graph $s$ restricted to $V^1 \times V^1$ be a pattern and let the graph $s$ restricted to $V^1 \times V^2$ be an isomorphic replica of this pattern. Let the graph $s$ restricted to $V^1 \times V^2$ be complete, i.e., $s_{ij} = 1$ for all $(i, j)$ in $V^1 \times V^2$. Finally, let the dynamic graph $w$ restricted to $V^1 \times V^1$ be "clamped" and equal to $s$, and similarly for $V^2$. The dynamics concerns the inter-layer part of $w$, i.e., those links connecting $V^1$ to $V^2$. All existing $q$-cycles contribute to $H$.

Proposition 2. Under the above conditions on $s$, the activated links in a minimum $w$ of $H$ under constraints (2) and (3) define a bi-continuous map (both the direct and the inverse maps are continuous) between $V^1$ and $V^2$, where the topology on each of the two layers is induced from the graph $s$.

The intuition in Prop. 2 is the following. Suppose first that only one inter-layer edge is allowed at each $i$ in $V^1$, and similarly at each $j$ in $V^2$ (This is the case illustrated in
Fig. 2. Any graph \( w \) with precisely one such edge per node defines a bijection (a map which is one-to-one and onto) between \( V^1 \) and \( V^2 \). If the decision is taken to activate link \( w_{ij} \), with \( i \in V^1 \) and \( j \in V^2 \), this decision propagates in both layers: the inter-layer link from a nearest-neighbor of \( i \) in \( V^1 \) should connect it to a nearest-neighbor of \( j \) in \( V^2 \), in order to activate as many “mixed” cycles as possible. Ultimately, the bijection between \( V^1 \) and \( V^2 \) in the optimal solution is an isomorphism of graphs. It is therefore an isometry, hence a bi-continuous map. In general, more than one inter-layer edge is allowed at each node: the resulting map is not bijective, yet the same argument shows that it connects neighbors to neighbors, i.e., it is bi-continuous. The “slope” of the projection depends on the parameter \( p \), which may be given different values in the two layers. If that is the case, the mapping does not preserve the graph-theoretic distance, i.e., it is not an isometry. This fact is useful when applying Prop. 2 to size-invariant pattern recognition, as is done in the next section. (For some values of the parameters, the optimal map is made of many separate projections between small disconnected patches in the two layers, but we need not be concerned with this case.)

Prop. 2 can be generalized in several ways. In particular, there is no need for two disjoint layers \( V^1 \) and \( V^2 \); what is really required is that part of the edges be clamped and contain a pattern. Note that the resulting total graph on \( V = V^1 \cup V^2 \), which includes both the dynamical inter-layer edges and the clamped edges in each layer, is itself a pattern. Thus, Prop. 2 may be regarded as a generic way to obtain compound patterns out of simple ones. There is also an analogy between Prop. 2 and the development of retinoptic maps. The retina is a 2-dimensional nervous tissue which projects onto central structures in the brain. These structures too are essentially 2-dimensional and it is found that the projection in the mature system preserves the topology: neighboring retinal ganglion cells project to neighboring cells in the Lateral Geniculate Nucleus of the thalamus, and these in turn project to neighboring cells in primary visual cortex. Preserving neighborhood relationships of external objects in higher centers is certainly important for vision, but such bi-continuous maps are also found in many other parts of the brain: they are in general termed topographic projections. A central issue in neurobiology is the development of topographic maps during embryonic and early postnatal life. A large literature exists on this topic, both experimental and theoretical. The reader familiar with it will immediately recognize the similarities with Prop. 2. In both cases, the problem is that of constructing a continuous map between two pieces of nervous tissue. In the case of retinotopy the topology is that of the 2-dimensional euclidean space, and the time-scale is that of ontogenetic development; the map consists of synaptic contacts which are to stay there throughout the life of the organism. In contrast, Prop. 2 deals with “patterns”, where the topology is not always as clearly defined. In particular, no strict dimensionality exists, yet if one were to introduce one, it would by no means be restricted to 2. The time-scale is very different, for such maps should be constructed in about 1sec. The analogy is still important because the underlying mechanisms are closely related to each other. Short-term modification is something like a very fast version of hebbian modification, which is almost certainly involved in the development of retinotopy. This makes short-term plasticity plausible from the point of view of evolution: it was certainly easier to modify an already existing mechanism, than to “invent” short-term plasticity from scratch.

Figure 2. A schematic illustration of Prop. 2 in the 1-dimensional case. Solid lines are clamped edges: they define, in each of the two “layers”, a circular graph as in Fig. 1. Only the inter-layer edges are dynamic, with the constraint that at most one such edge is allowed at each node. The dashed lines represent a stable state, that is, a minimum of \( H \) under this constraint; this state evidently realizes an isomorphic projection between the two patterns; it achieves the highest possible number of 4-cycles.

Figure 3. A model for size- and position-invariant pattern recognition. For explanation see text. Two types of local feature, or label, are represented by o and x. The figure illustrates that inter-layer connections are label-preserving.
One may summarize the constraint-free behavior of the system together with Propositions 1 and 2, by saying that the dynamics always generates patterns, and that in doing so it takes advantage of already existing patterns whenever this is possible. It is an interesting question to what degree the simplifying assumptions implicit in the definition of $H$ are essential for establishing the collective properties captured by Propositions 1 and 2. Behavior as stated in Proposition 1 has been demonstrated in a system with full joint dynamics of excitatory and inhibitory signals, transmitted by directed links, and fast-modifying synapses. A very special cooperative configuration has, however, been used in that study. Earlier work has validated Proposition 2 in systems with signal-and-synaptic-modification dynamics, and with multiple parallel pathways.

5. A model of position-and-size-invariant pattern recognition

Propositions 1 and 2 are general rules for manipulating patterns. We now give an example of how these two rules can be used to solve a particular type of problem, namely invariant pattern recognition. For this purpose, a model of quite specific architecture is proposed. It is one possible solution among many others within the framework given here, and it is meant to illustrate the use of synaptic patterns as a knowledge representation scheme.

As in all models of perception, an essential element is the feature-detector cell. Each such cell is responsible for a specific "elementary" feature—or "primitive"—at a particular position of the sensory surface. Cells of this kind exist in living organisms, and a good deal of sensory physiology is devoted to studying their properties. For instance, some cortical neurons respond selectively to a visual stimulus in the form of an elongated bar or an intensity gradient in a well-defined location of the retina, with a well-defined orientation. We shall not specify here the type of feature-detectors to be employed: an adequate repertoire has to be chosen in each particular application, depending on the stimulus one wishes to process and on the desired invariances. Let $r$ be the number of different types of local features. The model contains 3 levels of cells connected in a hierarchical way (see Fig. 3). The lowest level is the retina $R$ where patterns are actually displayed. The intermediate layer $L^1$ contains feature-detector cells arranged in a retinotopic way: each such cell is active if and only if the visual pattern currently displayed on $R$ contains the appropriate local feature at the appropriate location. The connectivity between $R$ and $L^1$ is responsible for these two properties: (i) being a "retinotopic" projection from $R$ to $L^1$, it endows $L^1$ with the 2-dimensional topology of $R$, and we shall assume that $L^1$ is a regular $m \times m$ array; (ii) by extracting local features from the patterns displayed on $R$, it endows each cell in $L^1$ with feature-detector properties. This connectivity is hard-wired and fixed, and not the original part of the model. There are $r$ feature-detectors at each of the $m^2$ positions in $L^1$, and we shall assume that exactly one of them is active at any given time. We also assume that the patterns we have to learn and recognize all have the same first-order statistics in $L^1$: the number of active features of each type is the same for all patterns.

The term pattern is used in the present section in its conventional meaning of a visual stimulus. There is no danger of confusion: patterns on $L^1$ are also patterns according to the graph-theoretic definition of the previous section. We assume that whenever two neighboring feature-detectors in $L^1$ are active, their activities are correlated on a short time-scale, whereas distant units are uncorrelated. Thus, correlation of activity encodes proximity in $L^1$. (This is in good agreement with experimental data on correlations between cells in the retina. Proximity-dependent correlations also play a crucial role in the ontogenetic development of retinotopic projections, see above.) Again, the fine temporal scale on which correlations are calculated (a few milliseconds) is not explicit in the model; we imply correlations by assuming that each cell in $L^1$ is connected to all $L^1$-cells in the $p$ first neighboring sites. This results in $r \cdot p$ links per cell since there are $r$ cells at each site in $L^1$. This connectivity is a fixed one: $w_{ij} = a_{ij}$ within $L^1$. In summary, upon presentation of a stimulus, or visual pattern, on the retina, a subset of the cells in $L^1$ is activated, one of the $r$ cells in each location. The links between the activated cells form a 2-dimensional synaptic pattern. Since each of the $m^2$ vertices in $L^1$ is labeled by one of the $r$ features one may call the active object in $L^1$ a labeled pattern.

Layer $L^2$ is the third hierarchical level in the model, a non-retinotopic set of $m^2$ feature-detectors. Before any learning has taken place, there is no topological order in $L^2$, and there are no connections at all within $L^2$. Inter-layer connections, i.e., connections between $L^1$ and $L^2$, are only allowed between cells of the same feature-type (see Fig. 3). Save for this constraint, inter-layer permanent connectivity—the set of links denoted by $s$—is complete: one feature-detector of a given type in $L^1$ is connected to all feature-cells of the same type in $L^2$, and vice versa. Thus, any $L^1$-cell is potentially activated by an appropriate local feature at any position on $R$. We shall assume that the partition of the $m^2$ $L^2$-cells into the $r$ feature-types is such that there are as many cells of a given type as there are active cells of this type in $L^1$ when a pattern is displayed on the retina; this number was assumed above to be the same for all patterns. These two assumptions will eventually allow us to establish a one-to-one correspondence between cells in $L^2$ and local features in $L^1$. Being by no means essential, they are very convenient for the presentation of the model. In summary, $L^2$-cells are as feature-specific as $L^1$-cells, yet they are totally unselective regarding retinotopic location. In real brains, the transition from a topographic representation to a non-topographic one is done in progressive steps: again the model is a caricature of the CNS.

The dynamics of the model consists in learning and in recognition. Learning affects the "permanent" synaptic weights $s$ within $L^2$. $L^2$ is thus the memory-layer of the system, where patterns are stored. When a pattern is presented on $R$, a labeled pattern results in $L^1$. Storing this labeled pattern in the memory-layer $L^2$ requires the choice of a one-to-one map between $L^1$ and $L^2$. Let $f$ be an arbitrary bijection from the set of active feature-cells in $L^1$ onto layer $L^2$. The existence of such bijections is guaranteed by our assumptions. Storage is then done as follows: for any two active cells $i$ and $j$ which are neighbors in $L^1$—i.e., the site of $j$ is one of the $p$-first-neighbors of the site of $i$—the link $s_{f(i)f(j)}$ is set equal to 1. This is done for all pairs $i$ and $j$, including cells of different types. In summary, the s-graph that is stored in $L^2$ is a 2-D $m \times m$ regular lattice with $p$-nearest-neighbor connections; in addition it is labeled pattern-isomorphic to the current labeled pattern on $L^1$, that is, there exists a bijective map between the two which is an isomorphism with respect to the graph-structure and which preserves the labels.
Further learning is done by superimposing patterns on top of each other in the memory-layer $L^2$, that is, by taking the union of the $s$-graphs, as in Prop. 1 of the previous section. We assume for simplicity that the learning sequence consists of $M$ patterns which are totally unrelated to each other; it is then a reasonable policy to take a new arbitrary map $f$ for each new pattern stored. When the $M$ patterns have been stored in $L^2$, permanent connectivity within $L^2$ is a random mixture—in the sense of Prop. 1—of $M$ isomorphic replicas of one $2$-$D$ pattern, the array of sites of $L^1$. However, these $M$ graphs are not labeled pattern-isomorphic to each other: no mapping exists between two of them which would preserve both the graph-structure and the labels, i.e., the feature types.

The translation of our learning algorithm back to the language of signals and synapses—the activity-and-connectivity variables—is straightforward. We first have to find a map $f$. This is done by using inhibition to reduce the set of active inter-layer links at each node to the required size; in practice, this size may be much larger than 1. Storage in $L^2$ is then achieved through modification of the permanent synaptic weights as follows. Since cells $i$ and $j$ are neighbors in $L^1$, they have correlated activities. As a result of activating inter-layer links which form a sparse projection $f$ between $L^1$ and $L^2$, these correlations are carried over to $L^2$: cells $f(i)$ and $f(j)$ now have correlated activities. We then increase the permanent weight $s$ of excitatory synapses between cells with correlated activity. This synaptic modification rule is quite "classical", yet not strictly hebbian: rather than collecting statistics for a long time, the model performs "one-shot learning". The required modification of the $s$-weights need not be much slower than short-term modification of $u$-weights. Yet, whereas modification of $u$ is transitory, immediately reversible and part of current brain function, modification of $s$ is long-lasting. The latter is a process by which a trace of the current (activity-and-) connectivity state $u$ of the brain is stored in its permanent structure $s$. This requires a gating signal or "Now Print" command, for it would be disastrous to alter the brain's permanent structure at all times without discriminating whether the current state is a "favored" one from the standpoint of the fast joint activity-and-connectivity dynamics, and whether it is of biological importance. Chemical or electrical signals which control synaptic modification in a global way are studied in various parts of cortex.

The dynamics of recognition concerns $L^1 - L^2$ $u$-links as well as $u$-links within $L^2$—the $u$'s within $L^1$ are always clamped. One of the $M$ learned patterns is presented on $R$, which elicits a labeled pattern in $L^1$. This labeled pattern is generally quite different from the one that was experienced during learning, because the retinotopic position of the stimulus is different. Yet it is isomorphic to it, as a labeled pattern. (For this to be strictly true, "periodic" boundary conditions in $L^1$ should be used: $L^1$ is then a torus.) A minimum of $H(u)$ is then sought under the usual constraints, where only active cells in $L^1$ are taken into consideration. Constraints (2) are as follows: one inter-layer link is allowed at each active $L^1$-cell and at each $L^2$-cell, and $p L^2$-links are allowed at each $L^2$-cell. Constraints (3) are determined in $L^1$ by the fixed $L^1$-connectivity and by the pattern currently presented on $R$—since this pattern determines the set of active cells on $L^1$—and in $L^2$ by the patterns previously stored there. Proposition 1 gives solutions to the problem of minimizing $H$ within the sole layer $L^2$, disregarding $L^1$ constraints: since the $s$-graph in $L^2$ is a random mixture of isomorphic $(m^2, p, q)$-patterns, there are $M$ minima, which are precisely the $M$ stored patterns. Labels are irrelevant to this fact. On the other hand we see from Proposition 2 that a minimum of $H(u)$ in the complete network includes an isomorphism between $L^1$ and $L^2$, if this is made possible by the patterns activated in $L^1$ and in $L^2$. This isomorphism of graphs is necessarily label-preserving too (because the $L^1 - L^2$-connections are). Such a map indeed exists, but only if the activated pattern in $L^2$ is the correct "memory", i.e., the graph which was stored when the current pattern was experienced for the first time. No label-preserving isomorphism exists with the other stored patterns. In summary, the global minimum of $H(u)$ consists of the right pattern in $L^2$—the stored memory—together with the projection between $L^1$ and $L^2$ which achieves the isomorphism. Since this is true irrespective of the position of the stimulus on $R$, position-invariant pattern recognition has been achieved.

Note that neither the invariant representation nor the map can be found unless the other one is found at the same time. Like some other models, the present one has the nice feature that the invariant representation and the map between the two representations are found simultaneously: information flows both top-down and bottom-up. The chief advantage of our solution is that it does not require any "map-units", which specifically code for pieces of maps between the two layers. Such units have a type of specificity which makes them somewhat unlikely to exist in the CNS. In contrast, short-term synaptic modification is a simple local mechanism which ensures that isomorphisms of graphs—actually homeomorphisms—are "automatically" found whenever they exist.

Computer simulations of this model have been performed using the one-dimensional topology of the circle. Feature-extraction being a secondary issue, the model consisted only of the two layers $L^1$ and $L^2$. Features are then mere labels numbered from 1 to $r$, and their only significance is to restrict $L^1 - L^2$-links to cells with identical labels. The following set-up was used for instance. $L^1$ is a circular array of $m = 80$ sites, with connections to the 6 nearest neighbors; there are $r = 8$ label-cells per site; each pattern on $L^1$ consists in activating 16 blocks of 5 adjacent cells with the same label. There are two blocks of each label in a pattern, and a pattern is therefore equivalent to a permutation of these 16 blocks. These permutations were chosen at random to obtain different patterns. The block-structure of patterns captures the continuity of objects in the real world: when appropriately discretized, objects can be described as clusters of local features of the same type. Also, the block-structure allows one to contract and dilate the patterns and thereby to address the problem of size-invariant recognition (see below). Contributions to $H$ come from rectangles, i.e., $q = 4$. In this set-up, it was found that one could store and reliably retrieve up to 4 patterns. This storage capacity is small but it must be borne in mind that a pattern stored upon presentation in an arbitrary position can be recognized when presented in any of the 80 positions on the circle. The storage capacity of $L^2$ can be expected to grow with its size as rapidly as the number of statistically independent permutations of its cells.

Size-invariant recognition follows from Proposition 1 and 2 in the same way as position-invariance. In principle, one may infer form Prop. 2 that recognition in our model is invariant under the broader class of all homeomorphic—bijective and bicontinuous—transformations. In practice, there are several rather trivial constraints which restrict the invariance to a class of reasonable deformations. Computer simulations using
1-dimensional block-structured patterns with parameters as above or slightly different, showed that a dilatation or a contraction by a factor of 2 of the pattern used for learning does not affect its recognition. A thorough description of the model will appear elsewhere, including a more detailed account of size-invariance. An application of the model to the recognition of real-world 2-D images after feature-extraction is currently in progress.

The successful recognition of a visual stimulus has to be communicated to other parts of the brain. This is no difficulty if all objects in the brain have the generic form of patterns. Interactions between them may take place according to rules similar to Propositions 1 and 2.

6. Discussion

With the available experimental techniques it seems to be difficult or impossible to gain access to the working principle of the brain. Experiment therefore must be complemented by the constructive approach of reverse engineering, i.e., we should try to work backwards from the performance of the brain to its working principles. In contrast to the situation in physics, where simple paradigms can be constructed—and simple laws abstracted from them—the brain confronts us with extremely complicated structures, interposed between basic principles and observable functional properties. Any discussion of working principles must be complemented by hypothetical constructions, e.g., network structures, to establish a connection to the functional level. Such arguments can always be called in question, because 'the brain perhaps does it quite differently'.

The problem is made worse by the difficulties of functional verification of hypothetical structures. By necessity, all interesting and relevant structures are numerically complicated, non-linear and unstable. It is therefore far from trivial to derive the global behavior of a model the local structure of which is hypothetically given. It is very likely that progress with the brain problem is dictated by progress with the functional verification problem. In the past, this step has mostly been based on analogy or even on pure imagination, and much of the discussion has been amateurish in style. New methodological developments are of the utmost importance. The availability of the computer for simulation opened the way for a big step forward, although even computer simulations have their own specific limitations. For one thing the sequential computer still sets very narrow quantitative limits, into which models of the nervous system have to be squeezed with art. Another difficulty is to discriminate numerical artifacts from `real', i.e., robust, phenomena. The observed model behavior may be of measure zero in structure space.

The Boltzmann approach and simulated annealing must in this context be considered a major advance. The systematic introduction of "thermal" noise acts as a filter which passes only essential, generic, behavior of a given model. More importantly, a global probabilistic description of the system can be given directly in terms of local structure—formulated as additive terms of a Hamiltonian—thus providing a firm basis for intuitive reasoning. The method also opens access to the great wealth of experience physics has gathered with self-organizing thermodynamic systems. The method also seems to stress important properties of the nervous system which are not rendered correctly by deterministic models, although, however, it is still possible that fundamental discrepancies will be discovered between the nervous system and systems with Gibbs-Boltzmann statistics.

Computer simulations (and attempted technological applications) have repeatedly hit upon enormous difficulties with seemingly trivial problems with, for instance, low-level vision, figure-ground separation and invariant pattern representation. Many models would work in impoverished, noiseless and unambiguous environments but fail completely under realistic conditions. One of the reasons for failure is local ambiguity, which can only be reduced by imposing systems of mutual constraint, such that solutions have to be found iteratively. Another reason for model failure lies in the fact that realistic environments transcend any pre-established system of classification. Both types of problem are fundamental in nature and stress the necessity of a radical shift in the conceptual framework within which models and theories are constructed. The problems of brain organization and of the underlying data structure are to be evaluated anew.

Both of these problems are addressed in this paper. The data structure of the brain must form a natural basis for its important operations. One such operation has already been mentioned, the storage and iterative imposition of systems of constraints (this aspect is nicely presented and discussed in ref. 6). Another important operation is the flexible composition of complex objects from pre-existing parts. This is important, for instance, when composing a visually perceived scene from familiar patterns. If the fundamental data structure is the assembly—a set of simultaneously activated cells—then a composite object should be represented by a superassembly, the set-theoretic union of the constituent assemblies. This scheme—although it is the basis for the overwhelming majority of brain theories—has the fundamental flaw that the set-theoretic union cannot be undone: it is impossible to recover the superposed sets from the union. Suppose, for instance, you are imagining a scene involving your grandmother, who happens to frown, and your grandfather, who happens to smile. Suppose, furthermore, your grandmother, your grandfather, the frown and the smile are represented by four assemblies. Then the superassembly representing the whole scene—the union of all four assemblies—is ambiguous and could equally well stand for a situation in which your grandfather had the frown on his face! The cause for this difficulty is the complete lack of internal structure in the assembly (for a thorough discussion see ref. 8).

The absence of internal structure in the assembly makes it difficult or impossible to solve the invariance problem on this basis. Invariance requires that patterns with similar internal structure but on different sets of cells evoke the same response. Now, the optimum of structure in the connections between assemblies can be attained if cells are labeled by feature types, and connections are restricted to pairs of cells with the same feature type. Two assemblies then interact maximally if they are described by identical lists of features. This type of dynamical relationship is too unspecific and generalizes to too large a class of patterns: If pattern a activates assembly A (in a structure corresponding to our L1) which in turn activates an assembly Z (in a structure corresponding to our L2) on the basis of identical composition of feature types, then another pattern b which activates an assembly B with a feature-composition
identical to that of $A$ will also activate $Z$. Pattern $b$ could be a transformed (shifted) version with the same internal structure as $a$, but the problem is that $b$ could also be a totally different pattern which happens to be composed of the same types of local features in the same relative proportions. This over-generalization is due to a lack of internal structure in the assembly. In order to show that this problem is not a man of straw, let us cite an example. It has often been proposed to solve the invariance problem by representing visual patterns by the power-spectrum, the absolute value of the spatial Fourier-transform (for references see 31). Fourier components play the role of features, taking the absolute value instills position invariance by destroying information on relative phases. The original image cannot be recovered because very different images may have the same power-spectrum, which illustrates the point made here.

The assembly, hypothesized as universal data format of the brain, with its lack of internal structure, is to be contrasted with virtually all data formats explicitly employed by man. Written and spoken language use the system of neighborhoods inherent in sequential order; mathematical symbols (and in fact all visual images!) use the topology of visual space as their system of relations; and data structures employed by Artificial Intelligence use pointers and adjacency in address-space to establish relations among primitives. It is not to be expected that exactly brain itself should do without a system of relations.

Why should relations in the brain be encoded by temporal coincidences and by dynamic synapses? Aren’t there other possibilities, especially ones that can be accommodated within the rate-coding framework? The usual approach represents sets of cells which together form an object by object-specific cells, often called cardinal cells or pontifical cells. This ‘solution’ leads to inflexibility and to enormous administrative problems, and has been discussed elsewhere. A second possibility would be to let activity in neurons encode low-level (binary, ternary...) relations in a way analogous to the representation by statistical moments discussed in this paper. This must be considered a serious possibility. However, relatively special wiring patterns would be required to implement the scheme. For instance, the interactions of two cells $a$ and $b$ with each other and with third parties have to be controlled by cell $c$ if the activity of $c$ encodes the presence of a binary relation between $a$ and $b$. This machinery is of a kind which cannot be expected to be ubiquitous to the nervous system. It rather would have to be created where needed. The organization of these special circuits is a problem in itself, see below.

Representation of relations within the brain by temporal coincidences of signals and by short-term modifying synapses has the great advantage that the necessary machinery is ubiquitous and that no specialized circuits are required. If two cells can interact at all with each other, or with third parties, the necessary synapses are there to be activated or inactivated, and the corresponding coincidence rates can be enhanced or suppressed. Coincidence rates in turn control nervous interactions in the desired way because neurons are coincidence detectors. Without requirement of any special structures the system forms a natural basis for organization and almost automatically leads to the emergence of connectivity-and-activity patterns which are of a very useful kind, as illustrated by the solution to the invariance problem described in section 5.

Connectivity dynamics can, with the help of slow plasticity, develop circuit patterns which form a basis for rate-coding of relations as discussed two paragraphs ago. This is best illustrated with the structure of section 5. Each cell in $L^1$ has connections to a number of cells with the same feature specificity in $L^2$. During a particular stimulation, this multiple projection must be reduced, for each active cell in $L^1$, to a single $L^1-L^2$ connection and a single cell in $L^2$, thus allowing the system a certain freedom of choice. Suppose a new pattern is activated in $L^1$, part of which is homeomorphic to part of a pattern which has already been stored in $L^2$. The system will use its freedom of choice for making the newly activated and stored pattern in $L^2$ part-identical with the old one. This can also be expressed by saying that the system recognizes familiar sub-patterns in new stimuli. There is thus a tendency for cells in $L^2$ to specialize to particular local connectivity structures, and activity in a cell of $L^2$ may be interpreted as encoding relationships between certain sets of other cells. One may, however, argue that even if the mature brain uses a system of rate-coding of relations, high-order statistical coding and short-term plasticity are essential for ontogenesis of the system and for the flexibility needed to handle new situations.

Regarding the organizational aspect mentioned above, the analogy with theories of evolution and of immunology, which has already been proposed in neurobiology,33,34 is very relevant. Selection is indeed a central notion in the present theory: the conflicting excitatory and inhibitory interactions result in the selection of sparse connectivity graphs $\tilde{\omega}$, which are subgraphs of the permanent graph $\omega$. We saw that the states which win in this selection process are those which yield the best possible match between successful synaptic events and activated synapses. In the simplified description which uses only connectivity variables, the favored states are characterized by being as consistent as possible with themselves; this self-consistency can be expressed mathematically by topological properties of graphs. $H(\omega)$ is an extremely simplified—intrinsic measure of the self-consistency of the connectivity state $\omega$, which can be thought of as playing the role of the fitness coefficient in a darwinian evolution. Additional constraints specify this general criterion to a form which is adapted to the problem at hand. For instance, in the situation discussed in section 5, the $M$ patterns stored in $L^2$ stand in mutual competition, and the presentation of a stimulus on $L^1$ results in the selection of one of these $M$ patterns, namely the one which is homeomorphic, as a labeled pattern, to the pattern currently presented on $L^1$. This may be viewed as a "resonance" mechanism, which acts to select one out of many competing patterns. Note that the frustration in the problem of minimizing $H(\omega)$ in the absence of any constraint on the graph $\omega$, hence the degeneracy of the solution, allows the dynamics to act, on a slow time-scale, as the "generator of diversity" required in selective theories: a "repertoire" of patterns may be created prior to the interaction with the environment. Yet, the main thrust of the present theory is in the faster selection processes. One could say that in the present theory much of what is classically thought to be achieved by ontogenetic, maybe even phylogenetic, evolution processes is passed on to fast connectivity dynamics. The flexibility of such a scheme could perhaps be one of the keys to the remarkable performances of our brain.

Direct experimental verification could concentrate on existence and properties either of high-order statistics in signals, or of short-term modification in synapses. One possible
experiment would be the following. In a waking animal, the signals of several sensory cells are recorded which can be activated a) by a single stimulus (which is considered by the trained animal as a unit) or b) by several stimuli, which are present simultaneously but which are manifestly different and independent of each other. In case a the signals should have a higher rate of coincidences in their fine temporal structure than in case b. The experiment is a difficult one (especially if the relevant statistics is of rather high order) but would open the window to an entirely new universe of signals in the brain. The experiment may be more feasible on the basis of mass-potentials, created by thousands of cells, if a situation can be found in which the topology of networks and the topology of ordinary space are sufficiently close to each other. Another kind of experiment would concentrate on the detection of short-term synaptic modulation. This may be possible in tissue culture. One would have to create a situation consisting of two cells with a direct synaptic connection. By proper experimental control of the cells' signals, and determination (e.g., by patch-clamp techniques) of the size of the EPSP created in one cell by activity in the other, one could hope to show that coincident activity in the two cells increases the size of the EPSP within a fraction of a second, and asynchronous activity decreases its size. The controlling signal on the post-synaptic side may be the membrane potential, but it may also be the concentration of a second messenger which is influenced by high-order convergence of pre-synaptic signals.

An altogether different "proof" of the feasibility of the system discussed here would consist in the implementation and test in an electronic machine. The machine would have to be massively parallel in order to be able to run networks of realistic size in finite time. Such a machine would differ from the conventional computer in being adapted to exactly one "program", namely the one incorporating the principles of organization implicit in the dynamic laws for individual cells and synapses\textsuperscript{35}. It may perhaps be expected that non-Von-Neumann-architecture machines of this or similar kinds will play a significant role in future technology.

Acknowledgements: This work has benefited from many discussions, in particular with S. Geman. Numerical simulations were done on an FPS 164 array processor, a computing facility of GREQ 70 ("Expérimentation Numérique").

References


