NEW ENCYCLOPEDIA OF NEUROSCIENCE

Theoretical and Computational Neuroscience:
Attractor network models (MS number: 1397)

Xiao-Jing Wang
Department of Neurobiology, Kavli Institute for Neuroscience,
Yale University School of Medicine,
333 Cedar Street, New Haven, CT 06510
Email: xjwang@yale.edu

Synopsis. The term ‘attractor’, when applied to neural circuits, refers to dynamical states of neural populations that are self-sustained and stable against perturbations. It is part of the vocabulary for describing neurons or neural networks as dynamical systems. This concept helps to quantitatively describe self-organized spatiotemporal neuronal firing patterns in a circuit, during spontaneous activity or underlying brain functions. Moreover, the theory of dynamical systems provides tools to examine the stability and robustness of a neural circuit’s behavior, proposes a theory of learning and memory in terms of the formation of multiple attractor states or continuous attractors, and sheds insights into how variations in cellular/synaptic properties give rise to a diversity of computational capabilities.

Key words: Neuronal spatiotemporal firing patterns, spontaneous Up and Down states, multistability, molecular switch, associative memory, dynamical system, persistent activity, working memory, continuous attractor, time integration, decision making, prefrontal cortex.
Introduction

The term ‘attractor’ is being increasingly used by neurophysiologists to characterize stable, stereotyped spatio-temporal neural circuit dynamics. Examples include readily identifiable rhythmic activity in a central pattern generator; well-organized propagation patterns of neuronal spike firing in cortical circuits \textit{in vivo} or \textit{in vitro}; self-sustained persistent activity during working memory; neuronal ensemble representation of associative long-term memory. In these examples, the rich and complex neural activity patterns are generated largely through regenerative mechanism(s), and emerge as collective phenomena in recurrent networks. In part, interests in attractor networks arise from our growing appreciation that neural circuits are typically endowed with an abundance of feedback loops, and that the attractor theory may provide a conceptual framework and technical tools for understanding such strongly recurrent networks.

The characterization of being ‘stable and stereotyped’ is sometimes taken to imply that an attractor network is not sensitive to external stimuli and difficult to be reconfigured. On the contrary, as was shown by recent studies, attractor networks are not only responsive to inputs, but may in fact be instrumental to the slow time integration of sensory information in the brain. Moreover, attractors can be created or destroyed by inputs, hence a same network can subserve different functions (such as working memory and decision making) depending on the inputs and cognitive control signals. The ‘attractor landscape’ of a neural circuit is readily modifiable by changes in cellular and synaptic properties, which forms the basis of the attractor model for associative learning.

In this article, I will first introduce the basic concepts of dynamical systems, attractors and bistability, using simple single neuron models. Then, I will discuss attractor network models for associative long-term memory, working memory, decision making. Modeling work and experimental evidence will be reviewed, open questions will be outlined. It will be shown that attractor networks are capable of time integration and memory storage over timescales much longer than the biophysical time constants of fast electrical signals in neurons and synapses. Therefore, strongly recurrent attractor networks are especially relevant to memory and higher cognitive functions.

The neuron is a dynamical system

A passive nerve membrane is a dynamical system described by a simple RC circuit equation:

\[
C_m \frac{dV_m}{dt} = -g_L(V_m - E_L) + I_{app}
\]

where \(V_m\) is the transmembrane voltage, \(C_m\) the capacitance, \(g_L\) the leak conductance (the inverse
of the input resistance), $E_L$ the leak reversal potential, and $I_{app}$ the injected current. In the absence of an input, the membrane is at the resting state, $V_{ss} = E_L$, say $-70$ mV. This steady state is stable: if $V_m$ is transiently depolarized or hyperpolarized by a current pulse, after the input offset it will evolve back to $V_{ss}$ exponentially with a time constant $\tau_m = C_m/g_L$ (typically 10-20ms). Thus, $V_{ss}$ is the simplest example of an attractor. More generally, for any sustained input drive $I_{app}$, the membrane always has a steady state $V_{ss} = V_L + I_{app}/g_L$, given by $dV_m/dt = 0$ (i.e. $V_{ss}$ does not change in time). The behavior of this passive membrane changes quantitatively with input current, but remains qualitatively the same: the dynamics is always an exponential time course (determined by $\tau_m$) toward the steady state $V_{ss}$, regardless of how high is the current intensity $I_{app}$, or how large is the capacitance $C_m$ or the leak conductance $g_L$. Moreover, the response to a combination of two stimuli $I_1$ and $I_2$ is predicted by a linear sum of the individual responses to $I_1$ or $I_2$ presented alone. These characteristics are generally true for a linear dynamical system, such as a differential equation with only linear dependence on $V_m$.

More interesting behaviors become possible, when nonlinearity is introduced by the inclusion of voltage-gated ionic currents. For instance, if we add in the RC circuit a noninactivating sodium current $I_{NaP} = g_{NaP}m_{NaP}(V_m)(V_m - E_{Na})$, where the conductance exhibits a nonlinear (sigmoid) dependence on $V_m$, the membrane dynamics becomes

$$C_m \frac{dV_m}{dt} = -g_L(V_m - E_L) - g_{NaP}m_{NaP}(V_m)(V_m - E_{Na}) + I_{app} \tag{2}$$

This system is endowed with a self-excitatory mechanism: a higher $V_m$ leads to more $I_{NaP}$ which in turn produces a larger depolarization. If $g_{NaP}$ is small, the weak positive feedback affects the membrane dynamics only slightly ((Fig. 1B, red). With a sufficiently large $g_{NaP}$, the steady state at $V_{Down} \simeq -70$ mV is still stable, because at this voltage $I_{NaP}$ is not activated. however, the strong positive feedback gives rise to a second, depolarized plateau potential (at $V_{Up} \simeq -20$ mV) (Fig. 1B, blue). Therefore, the membrane is bistable: a brief input can switch the system from one to another attractor state (Fig. 1A). As a result, a transient stimulus can now be remembered for a long time, in spite of the fact that the system has only a short biophysical time constant (20ms). Unlike linear systems, in a nonlinear dynamical system gradual changes in a parameter ($g_{NaP}$) can give rise to an entirely new behavior (bistability).

Attractor states are stable under small perturbations, and switching between the two can be induced only with sufficiently strong inputs. How strong is strong enough? The answer can be found by plotting the total ion current $I_{tot} = I_L + I_{NaP} - I_{app}$ against $V_m$, called the I-V curve (Fig. 1B, upper panel with $I_{app} = 0$). Obviously, a $V_m$ is a steady state if $I_{tot}(V_m) = 0$ (thus $dV_m/dt = 0$). As seen in Fig. 1B (blue), the two attractors (filled circles) are separated by a third steady state (open circle). The latter is unstable: if $V_m$ deviates slightly from it, the system will
not returns back but converge to either of the two attractors. Indeed, if $V_m$ is slightly smaller, $I_{tot}$ is positive (hyperpolarizing), so $V_m$ decreases towards $V_{Down}$. Conversely, if $V_m$ is slightly larger, $I_{tot}$ is negative (derpolarizing), so $V_m$ increases towards $V_{Up}$. Therefore, an external input must be strong enough to bring the membrane potential beyond the unstable steady state, in order to switch the system from one attractor state to the other.

It is worth noting that the ‘attractor landscape’ not only depends on the strength of the feedback mechanism (the value of $g_{NaP}$), but also sustained inputs. As shown in Fig. 1C, with a fixed $g_{NaP}$, a constant applied current $I_{app}$ shifts the I-V curve up or down. Either a hyperpolarization or depolarization can destroy the bistability phenomenon. This simple example demonstrates that neuronal bistable dynamics can be readily reconfigured by external inputs. As we shall see below, this is generally true for neural networks as well, and has important computational implications.

Attractors need not be steady states. In neurons, a plateau potential is typically not stable as a steady state. Instead, upon depolarization the Hodgkin-Huxley-type sodium and potassium currents produce repetitive action potentials, which represent another (oscillatory) type of attractor behavior. The attractor nature of periodic spiking is shown in Fig. 2 using the classical Hodgkin-Huxley model: regardless of the initial states, the membrane system always converges to the same periodic attractor state. If the system is perturbed by a transient stimulus, it would resume the same firing pattern after the stimulus offset, except for a shift of spiking time, or the phase of the periodic attractor state. Thus, a periodic attractor is robust (the amplitude and periodicity are not sensitive to transient perturbations); at the same time it it sensitive to phase shift, hence the clock can be readily reset. This general property of nonlinear oscillators is a key for understanding synchronization among neurons and coherent brain rhythms. It is also a cornerstone of the neurobiology of biological clocks, such as the circadian rhythm, sleep cycle, or central pattern generators for locomotion like walking, swimming, and breathing. In each of these systems, the rhythm is amenable to be reset by a transient input, which leads to a phase shift in time but otherwise does not alter the stereotypical activity pattern in the network.

The oscillatory attractor illustrates that the behavior of a dynamical system (the Hodgkin-Huxley model) is determined by direct observables (the membrane voltage) as well as internal dynamical variables (ion channel gating variables). The space of all dynamical variables form a ‘phase space’ of the system, which is typically multi-dimensional. A neural circuit in the mammalian brain consists of many thousands of cells, its phase space is enormous. Dynamics of such systems can be very complex, exhibiting a wide gamut of spatiotemporal activity patterns. It is generally not possible to define an energy function for such systems. Nevertheless, as we shall see below, the concept of computational energy landscape is still helpful for developing intuitions about attractor networks.
Synaptic plasticity and associative memory

Bistable switches, or more generally multiple attractors, can be realized on all scales, from the molecular machinery of individual synapses, electrical activity of single neurons, to large neural circuits. It is well known that synapses that form connections between neurons are highly plastic, and experience-dependent synaptic modifications are believed to be a physiological substrate of learning and memory. A single synapse comprises hundreds of proteins that interact with each other in a highly connected signal transduction network, therefore a synapse is a dynamical system. The biochemical time constants in such a network, and the typical protein lifetimes, range from seconds to hours. So how can synapses store memories that may be retained for many years? One possibility is that the expression of memory maintenance involves changes in the molecular composition of the synapse that are mechanically stable over a long time. Alternatively, memories could be stored in molecular switches, with two or more states that are stable over durations beyond the intrinsic molecular time constants. Recent physiological experiments have yielded evidence for switch-like behavior during synaptic modifications. Molecular studies and modeling have revealed several candidate protein kinases that may exhibit switch like behavior at single synapses, such as calcium/calmodulin-dependent protein kinase II (CaMKII), protein kinase C (PKC) or mitogen-activated protein kinase (MAPK). Memory switches do not necessarily have an infinite lifetime, since an active state may be turned off by subsequent synaptic changes during ongoing neural activity, by protein turnover, or by molecular fluctuations due to the small synaptic volume (about 0.1 femtoliter). Nevertheless, the key point here is that, if individual synapses exhibit stable self-sustained active states (attractors), the lifetime of a memory trace is not directly limited by the biochemical time constants of synaptic signaling pathways.

Whereas synaptic plasticity provides a structural basis for memory formation, the stored information is encoded in a distributed manner by the synaptic connection patterns in a neural circuit. Theoretically, it has been proposed that associative memories are learned through the creation of stable neural activity patterns (attractors). In this view, a memory network has many attractor states, each representing a particular memory item and has its own ‘basin of attraction’ within which other states evolve dynamically into the attractor state (Fig. 3). The stability of attractor states ensures that memory storage is robust against small perturbations. These memory states are imprinted in the network by long-lasting changes of synaptic connections through Hebbian learning, and much theoretical work has been devoted to the analysis of ‘storage capacity’, the number of memory items that can be stored and retrieved reliably as a function of the network size. Memories thus established are ‘associative’, since a partial cue would bring the network into the basin of attraction of an attractor with the information content close to that of the sensory cue.
Thus, memory retrieval can be done ‘by association’ between a cue and the corresponding memory item, and the recall process is error-correcting (incomplete information still leads to the correct memory retrieval). This capability for pattern completion is a hallmark of associative memory. At the same time, an attractor model is also capable of pattern separation, in the sense that two slightly different input patterns, near the boundary of two basins of attraction, may drive the network to two distinct attractor patterns, leading to the retrieval of two separate memories.

To experimentally test the attractor model, many neurons must be simultaneously monitored so that distributed activity patterns in memory circuits can be assessed. One of such circuits is the hippocampus, which is known to be critical to the formation of episodic memory and spatial memory. In rodents, during exploration in a familiar environment, pyramidal ‘place cells’ in the hippocampus are activated when the animal passes through a specific location, called a ‘place field’. This spatial selectivity is characterized by a bell-shaped tuning curve (neuronal firing rate as a function of the animal’s location). Place fields of hippocampal cells cover the entire surface of the environment with about equal density, so that the neural ensemble firing can be decoded to readout the animal’s location in space. The hippocampus thus implements a neuronal representation of a spatial map. Moreover, when the animal is exposed to another environment, place fields of cells undergo great changes (‘remapping’), until a representation of the new environment is established in the hippocampus. Therefore, the hippocampus stores distinct maps, with each map being activated as the rat enters a different environment. It has been proposed that these spatial representations reflect distinct attractor states, based on the observation that activity of place cells is preserved in the dark (without external visual cues). The attractor model has recently been directly tested by combining a clever navigation task design with simultaneous recording from many single cells in the hippocampus. In one experiment, rats were trained to forage in two distinct environments, a square and a circle that differed in color, texture and shape. This led to remapping of place fields in the majority of cells recorded in the hippocampal area CA1. Then, in probe trials, the rat commuted between a series of environments of the same color and texture, but the shape morphed gradually between the square and the circle. These environments were chosen randomly from trial to trial. Remarkably, it was found that the place fields of the recorded cells abruptly and coherently changed from square-like to circle-like (Fig. 4). This observation provides strong support for the attractor model, which predicts that the hippocampal coding of space should exhibit both pattern completion (minor changes from a familiar environment—a square or a circle—do not alter place fields), and pattern separation (similar inputs intermediate between square-like and circle-like result in drastic different place fields, due to a switch between the two learnt maps).

Persistent neural firing observed in a brain area may not necessarily be generated locally, but is a mere reflection of mnemonic activity elsewhere. The area CA1 is commonly viewed as an readout
circuit, whereas memories are believed to be stored upstream. It has long been hypothesized that autoassociative memories are formed in the area CA3, which projects to CA1 and is endowed with strong recurrent excitatory connections among pyramidal cells, a prerequisite for the generation of attractor states. Furthermore, recent discoveries have drawn attention to the entorhinal cortex, the major input area for the hippocampus. Cells in the dorsolateral entorhinal cortex fire whenever the rat is on any vertex of a triangular lattice spanning the whole surface of the environment. This 'grid cell activity' is preserved in the dark, when the visual cues are absent. These findings have led to the proposal that the entorhinal cortex embeds an attractor network with a grid-like representation of the environment, and the hippocampus transforms this periodic firing pattern in the input into a non-periodic firing pattern that encodes the animal’s current spatial position. If so, it remains to be seen whether the hippocampus still needs to operate as an attractor network, and whether spatial navigation is subserved by interconnected attractor circuits.

In order for an animal to navigate without reference to external cues, the neural instantiation of a spatial map must be constantly updated by the integration of linear and angular self-motion. Whether an attractor model is capable of carrying out such computations robustly is not well understood. Finally, little is known about precisely how the attractor landscape in the space of neuronal firing patterns is shaped by the details of the animal training process. Another study found that, using a somewhat different learning and probe procedure, but also with morphed environmental shapes, place fields of CA3 and CA1 cells switched more gradually and less coherently. This finding does not necessarily mean that the attractor paradigm is incorrect, as we expect from computational work that different learning history gives rise to different sets of attractor states. Elucidation of the general principles and cellular mechanisms of this learning process, in future experiments and modeling, will help determine whether the attractor model represents a sound theoretical framework for the neurobiology of learning and memory.

**Persistent activity and working memory**

Attractor models have also been applied to working memory, our brain’s ability to actively hold information ‘online’ for a brief period of time (seconds). Neurons that maintain ‘working memory’ must be manifestly active in a sustained manner, how can such persistent activity be generated in the absence of direct external input? R. Lorente de Nó in the 1930’s and D.O. Hebb in the late 1940’s proposed that the answer lies in the feedback loop connections. Thus, in a working memory network, every cell receives excitatory drive from both afferent inputs and intrinsic synaptic connections. Inputs activate neurons in a selective cell assembly, the triggered spike activity reverberates through excitatory synaptic circuit which is enough to sustain an elevated firing when the inputs
are withdrawn. This general idea has been made rigorous in attractor models, according to which a working memory circuit exhibits multiple attractor states (each coding a particular memory item) that coexist with a background (resting) state. All the attractor states are self-maintained and relatively stable in the face of small perturbations or noise. Yet, memory states can be turned on or switched off by brief external stimuli.

Stimulus-selective neural persistent activity has been observed in awake animals performing delayed response tasks that depend on working memory. For example, in a delayed match-to-sample task, two visual objects are shown consecutively, separated by a time gap of a few seconds, and the subject is asked to judge whether the two stimuli are the same. Or, after the presentation of the first object and the delay, an array of visual objects are shown, and the subject must indicate by a motor response which of them is identical to the first visual cue. In both cases, the subject’s performance relies on the working memory about the first object across the delay. The stored information involves a collection of discrete items. Other delayed response tasks engage working memory of an analog quantity, such as spatial location or stimulus amplitude. While a monkey is performing such a task, neurons in the prefrontal, posterior parietal, inferotemporal, and premotor cortices were found to exhibit elevated persistent activity that is selective to stimuli. Three types of mnemonic coding has been observed: (a) object working memory cells are tuned to one or a few of discrete memory items; (b) spatial working memory cells typically exhibit a bell-shaped (Gaussian) tuning function of the spatial location or directional angle; (c) cells that store ‘parametric working memory’ of magnitudes (like vibration stimulus frequency) are characterized by a monotonic tuning function of the encoded feature.

These experimental observations lend support to the attractor model, inasmuch as stimulus-selective persistent firing patterns are sustained internally in the absence of direct sensory input, dynamically stable, and approximately tonic in time (e.g. across a delay). However, experiments show that delay neural activity is often not tonic but exhibits time variations such as ramping up or ramping down. How to account for the heterogeneity and time courses of mnemonic persistent activity represents a challenge to the attractor network model. Moreover, it remains an open question as to what cellular or circuit mechanisms are responsible for the generation of persistent activity. This question has begun to be addressed using biologically-constrained models of persistent activity. Fig. 5 shows such a recurrent network model for spatial working memory, in which spiking neurons and synapses are calibrated by the known cortical electrophysiology. The key feature is an abundance of recurrent connections (‘loops’) between neurons, according to a Mexican-hat type architecture: localized recurrent excitation between pyramidal cells with similar preference to spatial cues, and broader inhibition mediated by interneurons (Fig. 5A). In a simulation of a delayed oculomotor task (Fig. 5B), the network is initially in a resting state in which
all cells fire spontaneously at low rates. A transient input drives a subpopulation of cells to fire at high rates. As a result they send recruited excitation to each other via horizontal connections. This internal excitation is large enough to sustain elevated activity, so that the firing pattern persists after the stimulus is withdrawn. Synaptic inhibition ensures that the activity does not spread to the rest of the network, and persistent activity has a bell shape (‘bump attractor’). At the end of a mnemonic delay period the cue information can be retrieved by reading out the peak location of the persistent activity pattern; and the network is reset back to the resting state. In different trials, a cue can be presented at different locations. Each cue triggers a persistent firing pattern of the same bell-shape but peaked at a different location (Fig. 5C). A spatial working memory network thus displays a continuous family of ‘bump attractors’.

Biophysically-realistic models have specific predictions about the circuit properties required for the generation of stimulus-selective persistent activity. In particular, it was found that a network with strong recurrent loops is prone to instability if excitation (positive feedback) is fast compared to negative feedback, as is expected for a nonlinear dynamical system in general. This is the case when excitation is mediated by the AMPA receptors, which are about 2-3 times faster than inhibition mediated by GABA\textsubscript{A} receptors (time constant 5-10 ms). The interplay between AMPA and GABA\textsubscript{A} receptors in a excitatory-inhibitory loop naturally produces fast network oscillations. In a working memory model, the large amount of recurrent connections, needed for the generation persistent activity, often lead to excessive oscillations that are detrimental to network stability. Working memory function can be rendered stable, if excitatory reverberation is slow, i.e. contributed by the NMDA receptors (time constant 50-100 ms) at recurrent synapses. Thus, the model predicts a critical contribution of NMDA receptors to working memory. Other processes with time constants of hundreds of milliseconds, such as short-term synaptic facilitation or intrinsic ion channels in single cells, may also contribute to reverberatory dynamics underlying working memory.

On the other hand, the feedback mechanism cannot be too slow. An alternative to the attractor model is the scenario in which persistent activity actually is not stable but represents slowly decaying ‘afterdischarges’ mediated by some intracellular mechanisms like second messenger pathways or kinetics of protein kinases. However, this scenario predicts that triggering inputs must be long lasting (for seconds), which is incompatible with physiological experiments in which working memory states have been shown to be switchable quickly by brief external inputs (a few hundreds of milliseconds). The recurrent (attractor) network mechanism achieves stability and long persistence time of memory storage, as well as rapid flexible memory encoding and erasure that are behaviorally desirable.
Time integration and categorical decision making

Cortical areas that are engaged in working memory – like the prefrontal cortex – are also involved in other cognitive functions such as decision making, selective attention, behavioral control. This suggests that microcircuit organization in these areas is equipped with the necessary properties to subserve both internal representation of information and dynamical computations of cognitive types. As it turns out, models originally developed for working memory can account for decision making processes as well. An example is shown in Fig. 6 from model simulations of a visual motion discrimination experiment. In this two-alternative forced choice task, monkeys are trained to make a judgment about the direction of motion (say, left or right) in a stochastic random dot display, and to report the perceived direction with a saccadic eye movement. A percentage of dots (called motion strength) move coherently in the same direction, so the task can be made easy or difficult by varying the motion strength (close to 100% or 0%) from trial to trial. While a monkey is performing the task, single-unit recordings revealed that neurons in the posterior parietal cortex and prefrontal cortex exhibit firing activity correlated with the animal’s perceptual choice. For example, in a trial where the motion strength is low (say 6.4%), if the stimulus direction is left whereas the monkey’s choice is right, the response is incorrect. In that case, cells selective for right display a higher activity than those selective for left, hence the neural activity signals the animal’s perceptual decision rather than the actual sensory stimulus. This experiment can be simulated using the same model designed for working memory. The only difference between a working memory simulation and a decision simulation is that, whereas for a delayed response task only one stimulus is presented, for a perceptual discrimination task conflicting sensory inputs are fed into competing neural subpopulations in the circuit. This is schematically depicted in Fig. 6A, where the relative difference in the inputs $c' = (I_A - I_B)/(I_A + I_B)$ mimicks the motion strength in the visual motion discrimination experiment. Fig. 6B shows a simulation with $c' = 6.4%$. At the stimulus onset, the firing rates of the two competing neural populations $r_A$ and $r_B$ initially ramp up together for hundreds of milliseconds, before diverging from each other when one increases while the other declines. The perceptual choice is decided based on which of the two neural populations wins the competition. Therefore, consistent with the physiological observations from the monkey experiment, decision process proceeds in two steps. Sensory data are first integrated over time in a graded fashion, which in the model is instantiated by the NMDA receptor dependent slow reverberation. This is followed by winner-take-all competition produced by synaptic inhibition, leading to a categorical (binary) choice.

Fig. 6C shows attractor dynamics in the ‘decision space’ where $r_A$ is plotted against $r_B$. In the absence of stimulation (left panel), three attractors coexist (filled circles): a spontaneous state
(when both $r_A$ and $r_B$ are low), and two persistent activity states (with high $r_A$ and low $r_B$, or vice versa). Upon the presentation of a stimulus (right panel), the attractor landscape is altered, and the spontaneous steady state disappears, so that the system is forced to evolve toward one of the two active states which represent perceptual decisions (A or B). In this graph, the sensory evidence is in favor of the choice A (with $c' = 6.4\%$), so the attractor A has a larger basin of attraction (orange) than that of the attractor B (brown). The system is initially in the spontaneous state which now falls in the basin of attraction A, and evolves toward the decision state A in a correct trial (blue). However, at low $c'$ the bias is not strong, and noise can induce the system’s trajectory to travel across the boundary of the two attraction basins, in which case the system eventually evolves to the decision state B in an error trial (red). The crossing of a boundary between attraction basins is slow, which explains why the reaction times are longer in error trials than in correct trials, as was observed in the monkey experiment. After the offset of the stimulus, the system’s configuration reverts back to that on the left panel. Because a persistently active state is self-sustained, the perceptual choice (A or B) can be stored in working memory for later use, to guide behavior. In this way, the attractor dynamics model offers an unified account for working memory and decision making computations.

**Neural integrators**

Qualitatively speaking, working memory requires neurons to convert a transient input pulse into a sustained persistent activity, like a time integral of the stimulus. Similarly, in perceptual decisions, approximate linear ramping activity, at a rate proportional to input strength, can also be conceptualized as time integration. It is worth noting, however, a genuine integrator implies that, after a transient input is turned off, the activity is persistent at a firing rate proportional to the input strength, spanning a continuous range. This is not the case in Fig. 6, where after the stimulus offset the neural activity is binary (representing one of the two categorical choices), independent of the input motion strength. This is what has been observed in posterior parietal neurons, and is the kind of neural signals needed to accomplish categorical choices.

There are two known coding schemes by neural integrators. In a ‘location code’ neurons exhibit a bell-shaped Gaussian tuning curve of an encoded feature, therefore distinct neural groups are engaged in storing different values of the analog feature. By contrast, according to a ‘rate code’, persistent firing rate of each neuron varies linearly with the encoded feature. As a result, if rates of different neurons are plotted against each other, they fall on a straight line in the ‘firing-rate space’. This observation led to the theoretical concept of ‘line attractors’. In other words, the stimulus feature is specified by ‘which neurons are active’ in a location code, and by ‘how high are the neural
activities’ in a rate code. Mathematically, a perfect integrator is described by \( dX/dt = I(t) \), where \( I(t) \) is the input, \( X \) is either the peak locus of network activity (in a location code) or the neural firing activity level (in a rate code).

Fig. 7 shows an neural integrator model in which the encoded feature is specified by the peak location of the network activity profile. This model was proposed for head direction (HD) cells. HD neurons are part of the spatial navigation system and signal the animal’s directional heading. When an animal turns its head, the angular velocity signal carried by vestibular inputs is integrated over time by HD cells into a positional signal, and the latter is sustained internally when the animal keeps the head direction fixed. HD cells are selective for angular head direction according to a Gaussian (bell-shaped) tuning curve. Interestingly, available evidence indicates that HD cells are generated in a neural circuit characterized by a paucity of local excitatory connections. Consistent with this observation, the model of Fig. 7 proposes a cross-inhibition mechanism, without recurrent excitation, that generates direction-selective ‘hill of persistent activity’. Moreover, the model surmises a shift mechanism by ‘rotation cells’ consisting of two inhibitory cell populations (Fig. 7A). When the head direction is fixed, the inputs from the two inhibitory neural populations are balanced with each other, and the hill of activity of the excitatory neural population is maintained fixed (Fig. 7B, time epochs between input pulses). As the head turns, the angular velocity signal increases the firing of one inhibitory population, while decreases the firing of the other. The resulting asymmetric inhibitory inputs induce the activity pattern in the excitatory neural population to move, at a constant speed proportional to the input amplitude (angular velocity) (Fig. 7B, time epochs during input pulses). Computation performed by the model is quite close to an integral operation in the sense of Calculus. When the input intensity is doubled (second versus first input pulse), the hill of activity moves twice as fast. Moreover, the network can integrate both positive (first and second pulses) and negative (third pulse) inputs. In fact, with the third input of an amplitude half as that of the second input, but lasting twice as long, the hill of activity moves back to the position before the second pulse, as expected by perfect integration.

Like the bump attractor network model for spatial working memory, this HD cell model assumes that the network is symmetrical and homogeneous. How can this be realized in the brain remains unclear. Also, models of perfect line attractors require fine-tuning of network parameters. To see why this is the case, consider the following simple firing-rate equation

\[
\frac{dr}{dt} = \frac{(-r + w_{rec}r)}{\tau} + I(t)
\]

where \( r \) is a firing rate, \( \tau \) is a typical biophysical (membrane or synaptic) time constant, and \( w_{rec} \) is the strength of recurrent connections. The effective time constant of the system is given by \( \tau_{eff} = \tau_{syn}/(1 - w_{syn}) \), which is longer than \( \tau \) in the presence of \( w_{rec} \). For instance, if \( \tau = 100 \)
ms and $1 - w_{rec} = 0.05$, then $\tau_{eff} = 2$ sec. When $w_{rec} = 1$ (which implies fine-tuning of the parameter $w_{rec}$), $\tau_{eff} = \infty$, and the system becomes a perfect integrator. How neural integrators can be realized robustly by plausible biophysical mechanisms remain a topic of active current research. Further progress depends on new experiments to put more constraints, and establish a firmer mechanistic basis, for computational modeling in this area.

**Concluding remarks**

In summary, the language of attractors is natural for describing the electrical activity of neurons and neural circuits. It provides a plausible theoretical framework for both short-term working memory and long-term associative memory. Importantly, nonlinearity due to feedback loops makes it possible that graded changes of a cellular or synaptic parameter lead to the emergence of qualitatively different behaviors (e.g. with or without persistent activity). The functional implications of this insight are potentially far reaching, as it suggests that cortical areas dedicated to distinct functions (e.g. sensory processing versus working memory) may share a same canonical cortical circuit layout, but with subtle differences in the cellular/molecular makeup, connectivity properties, neuromodulatory influence.

Whether and how attractor networks are realized in the brain has still not been proven experimentally. An open question concerning working memory is whether persistent activity is primarily generated by local circuit dynamics, single cell property, or large-scale network composed of several brain regions. For both working memory and long-term memory, a major challenge is to elucidate biological substrates that underly the robustness of continuous attractors and integrators. Furthermore, there is a dichotomy between rapidly switchable attractors on one hand, and intracellular signaling network with multiple time constants on the other hand. The interplay between cellular processes and collective network dynamics will turn out to be an exciting topic in future research.

**Acknowledgments.** I thank K.-F. Wong for making Fig. 6C, and P. Song for making Fig. 7. This work is supported by the NIH grant MH62349.
Further reading


Figure Captions

Figure 1. Positive feedback and attractor dynamics in a simple neural membrane model. The membrane voltage is described by an RC circuit with the addition of a fast noninactivating sodium current \( I_{NaP} = g_{NaP} m_{NaP}(V_m - E_{Na}) \), where \( m_{NaP}(V_m) = 1/(1 + \exp(-(V_m + 45)/5)) \) is a sigmoid function of \( V_m \). The interplay between \( I_{NaP} \) and membrane depolarization produces an excitatory regenerative process. (A) Bistability with \( g_{NaP} = 0.015 \mu S \). The system is initially at rest \( (V_{Down} \simeq -70 \text{ mV}) \). A depolarizing current pulse switches the membrane to a plateau potential \( (V_{Up} \simeq -20 \text{ mV}) \) which persists after the input offset. A second, hyperpolarizing current pulse switches the membrane back to the resting state. (B) Upper panel: I-V curve (the total current \( I_{tot} = I_L + I_{NaP} - I_{app} \) as a function of \( V_m \), with \( I_{app} = 0 \)); a steady state is given by an intersection with \( I_{tot} = 0 \). Lower panel: computational energy function (Lyapounov function) \( U(V_m) \), in which a steady state corresponds to a maximum (unstable) or a minimum (stable). For \( g_{NaP} = 0 \) (black, passive membrane) or \( g_{NaP} = 0.08 \) (red), there is only one steady state \( (\simeq -70 \text{ mV}) \). For \( g_{NaP} = 0.015 \) (blue), there are three steady states, two are stable (filled circles) and the third is unstable (open circle). (C) Bistability is modulated by inputs, with \( g_{NaP} = 0.015 \) fixed. The injected current intensity is varied (blue, violet and green for \( I_{app} = 0, -0.5 \) and 0.5 nA respectively). Same format as (B). Other parameter values are \( C_m = 0.5 \text{ nF}, g_L = 0.025 \mu S, V_L = -70 \text{ mV}, V_{Na} = 55 \text{ mV} \). The energy function \( U(V_m) \) is defined by rewriting the RC circuit Eq. (2) as \( dV_m/dt = F(V_m) = -dU/dV_m \), hence the energy function \( U(V_m) \) is the integral of \( -F(V_m) \). For instance, with \( g_{NaP} = 0 \), Eq. (2) is reduced to Eq. (1) and can be rewritten as \( dV_m/dt = (V_{ss} - V_m)/\tau_m \equiv F(V_m) \). By integrating \( -F(V_m) \), we have \( U(V_m) = (V_{ss} - V_m)^2/(2\tau_m) + U_0 \), with an arbitrary constant \( U_0 \). Therefore, \( U(V_m) \) is a parabola with \( V_{ss} \) at the bottom of the valley of the energy function (Fig. 1B, black). For \( g_{NaP} = 0.015 \) (Fig. 1B, blue), the energy function \( U(V_m) \) displays two valleys at \( V_{Down} \) and \( V_{Up} \), separated by a peak at the unstable steady state.

Figure 2. Oscillatory attractor in the original Hodgkin-Huxley model of action potentials. The model has a leak current \( I_L \), a fast sodium current \( I_{Na} \) and a nonactivating potassium current \( I_K \). Its dynamics is described by four coupled differential equations (the membrane voltage \( V_m \), the activation and inactivation gating variables \( m \) and \( h \) for \( I_{Na} \), and the activation variable \( n \) for \( I_K \)). (A) Repetitive firing of action potentials with a current pulse \( I_{app} = 9 \mu A/cm^2 \). (B) The potassium activation variable \( n \) is plotted against \( V_m \), showing that the oscillation forms a closed trajectory in this projected ‘phase space’. The trajectory roughly consists of three portions: the upstroke of
an action potential (when both $V_m$ and $n$ increase), the downstroke (when $V_m$ decreases while $n$ keeps increasing then starts to decrease), and the refractory period (when $n$ continues to decrease while $V_m$ starts to increase). Different colors correspond to five different initial conditions (with different $V_m$ and $n$ values at $t = 0$), in all cases the system dynamically evolves into the oscillatory attractor state. The Hodgkin-Huxley model exhibits bistability between a steady resting state and an oscillatory state, which is not shown for the sake of clarity.

**Figure 3.** Illustration of an attractor model for associative memory. Computational energy function is depicted as a landscape of hills and valleys, plotted against the neural activity states (on the XY plane). The synaptic connections and other properties of the circuit, as well as external inputs, determine its contours. The circuit computes by following a path that decreases the computational energy until the path reaches the bottom of a valley. In an associative memory circuit, the valleys correspond to memories that are stored as associated sets of information (the neural activities). If the circuit is cued to start out with approximate or incomplete information, it follows a path downhill to the nearest valley (red), which contains the complete information. From Tank DW and Hopfield JJ (1987), Collective computation in neuronlike circuits. Sci Amer 257: 104-114.

**Figure 4.** Abrupt and coherent expression of spatial representation of hippocampal neurons. A rat explores a series of environments with morphed shapes (top icons), between a square (the three leftmost columns) and a circle (the three rightmost columns). Twenty single cells were simultaneously recorded, shown in rows from top to bottom. Warm colors indicate locations (place field) of the rat when the cell’s spiking activity is high. Each field is scaled to peak firing rate shown as red. The 17 of 20 simultaneously recorded place cells with different (remapped) firing patterns in the square and the circle almost all switch from the square-like to circle-like pattern between the h and f octagons. Eight cells had fields in the circle but not the square (cells 1 to 8); four in the square but not the circle (9 to 12); five fired in both but in different places (13 to 17); and three did not reach the criterion for remapping (18 to 20). From Wills TJ, Lever C, Cacucci F, Burgess N, O’Keefe J (2005), Attractor dynamics in the hippocampal representation of the local environment. Science 308: 873-876.

**Figure 5:** Bump attractor model for spatial working memory. (A) Mexican-hat type connectivity of the model circuit. Pyramidal neurons are arranged according to their preferred cues (0 to 360 degrees). Recurrent excitatory connections are strong between cells with similar cue preference, and decrease with the difference in their preferred cues. Local excitation is counteracted by broad synaptic inhibition mediated by GABAergic interneurons. (B) Spatiotemporal activity pattern of the pyramidal cell population in a simulation of delayed oculomotor response task. Each dot is
an action potential. Abscissa: time, ordinate: neurons (labeled by their preferred cues). C: cue period, D: delay period, R: response period. The elevated and localized neural activity is triggered by the cue stimulus at 180°, and persists during the delay period. On the right is shown the spatial pattern, where the average firing rate during the delay period is plotted versus the neuron’s preferred cue. The memory of the spatial cue is stored by the peak location of this bell-shaped persistent activity pattern (‘bump attractor’). From Compte A, Brunel N, Goldman-Rakic PS and Wang X-J (2000) Synaptic mechanisms and network dynamics underlying visuospatial working memory in a cortical network model. Cerebral Cortex 10: 910-923. (C) Temporal evolution of the peak location of mnemonic persistent activity pattern in 20 trials with transient stimuli at different locations. The memory of the initial cue is well preserved during the 6 second delay period. From Renart A, Song P and Wang X-J (2003) Robust spatial working memory in a heterogeneous network model with homeostatic synaptic scaling. Neuron 38: 473-485.

Figure 6: A spiking neuron circuit model for two-alternative forced-choice tasks. (A) Model scheme. There are two pyramidal cell groups, each of which is selective to one of the two directions (A=left, B=right) of random moving dots in a visual motion discrimination experiment. Within each pyramidal neural group there are strong recurrent excitatory connections which can sustain persistent activity triggered by a transient preferred stimulus. The two neural groups compete through feedback inhibition from interneurons. (B) A network simulation with \( c' = \frac{I_A - I_B}{I_A + I_B} = 6.4\% \). Population firing rates \( r_A \) and \( r_B \) exhibit an initial slow ramping (time integration) followed by eventual divergence (categorical choice). In this sample trial, although the input A is larger, the decision is B (an erroneous choice). (C) Decision dynamics shown in the two-dimensional plane where firing rates \( r_A \) and \( r_B \) are plotted against each other. Left panel: in the absence of stimulation, there are three attractor states (filled circles) and two unstable steady states (open circles). Colored regions are the basins of attractions (maroon for the resting state, orange and brown for the persistent activity states). Right panel, with a stimulus of \( c' = 6.4\% \) in favor of choice A, the decision trajectory is shown for two trials (correct trial in blue, error trial in red). See text for further description. (A-B) from Wang (2002). (C): Simulation by K. F. Wong using the model published in Wong KF and Wang X-J (2006) A recurrent network mechanism for time integration in perceptual decisions. J Neurosci 26, 1314-1328.

Figure 7: Time integration by a bump attractor network. (A) Model scheme. An excitatory neural network, encoding a directional angle (0 to 360 degrees), receives inputs from two inhibitory neural populations that are balanced with each other when there is no input. A velocity signal increases input to one of the inhibitory population (+I) and decreases input to the other (−I), leading to a bias for the excitatory network. (B) Network firing pattern (Top) in response to a
series of input signal steps (Bottom). Neurons in the network are aligned along the y-axis, labeled by their preferred directional angles. X-axis is time. Each dot is a spike. The network activity pattern has the form of a bell-shaped profile (Upper panel, right), its peak location encodes the directional angle (white line in the rastergram). In the absence of an external input, the directional information is maintained by persistent firing pattern. An input induces the hill of activity to move at a speed proportional to the stimulus amplitude, hence the network performs a time integral of the input. See text for more detailed discussion. Simulation by P. Song using the model published in Song P and Wang X-J (2005) Angular path integration by moving ‘hill of activity’: a spiking neuron model without recurrent excitation of the head-direction system. J Neurosci 25: 1002-1014.
Figure 1
Figure 2
Figure 3
Figure 4
Figure 6
Figure 7