2

Dynamic Field Theory: Foundations

GREGOR SCHÖNER AND ANNE R. SCHUTTE

In Chapter 1 we introduced the notions of activation variables, $u(t)$, and their neural dynamics, $\dot{u} = -u + h + \text{inputs} + \text{interaction}$. Activation variables characterize the inner state of the central nervous system (CNS). They may be coupled to other activation variables through interaction. They may also receive inputs directly from the sensory surfaces. And they may provide input to other activation variables and, ultimately, have an impact on motor systems (in ways we will study in depth in Chapter 4). In Chapter 1 we advanced the notion that activation variables "stand for" something outside the CNS that is ultimately specified by the links of their dynamics to the sensory or motor surfaces, be they direct or through other activation variables. In this chapter we need to make this intuition explicit and address directly how activation variables may come to represent states of affairs outside the CNS.

This raises the question, of course, of the kind of states outside the CNS that need to be represented inside the CNS. We will argue that those states form continua that span the many different possible percepts, the many possible motor actions, and, ultimately, the many possible thoughts. Seemingly discrete states such as object categories or different categories of motor acts are often embedded in continua. Recognizing a letter as a category, for instance, we also perceive its continuous variations such as size, orientation, contrast, or any of the other manifold visual dimensions. In fact, this is true even in what is sometimes called *categorical perception*. In categorical perception, two stimuli are only discriminated if they fall into different categories. Different versions of a stimulus that both fall into the same category are not discriminated. The question is whether there is ever truly categorical perception (Pisoni, 1973). Today, most researchers soften the concept of categorical perception by requiring only that discrimination between stimuli be enhanced when they fall into different categories, not if they fall into the same category (Goldstone & Hendrickson, 2009). It is typically found that discrimination of stimuli that fall into the same category is never fully abolished.

In summary, dynamic field theory (DFT) is founded on the hypothesis that the continuous states of the world are primary. How the CNS breaks continua into categories then requires an account that must go beyond merely postulating that discrete activation variables stand for discrete categories. The critical question, therefore, is how activation variables represent continua. In this chapter, we will introduce the idea of continuous sets of activation variables that form activation fields. These activation fields are linked through continuous mappings to sensory and motor surfaces. We will apply the neural dynamics of activation variables to activation fields and will re-encounter the instabilities analyzed in Chapter 1, the detection and the selection instabilities. Generalizing neural dynamics to fields will enable us to differentiate between different paths through the detection instability, depending on whether localized or global input is the driving force. We will also be able to more clearly establish in what sense sustained activation is a mechanism for working memory of metric information.

A major theoretical advance that the move from activation variables to activation fields enables is a better understanding of how learning may shape neural representations. We will look at the simplest learning mechanism within DFT, the laying down of a memory trace that facilitates activation of field locations previously activated. Through the memory trace, the history of activation preshapes fields, so that all field locations are no longer equal. We will discuss how this might build a bridge from the hypothesized fundamental continuity of neural representations toward the neural representation of categorical states.

So this chapter is quite ambitious. It presents the core ideas of DFT that permeate the entire book. It reviews the associated conceptual commitments while also trying to be pedagogical and clear. If the going gets rough, go to the end of the chapter. There we will make the ideas concrete and practical in a set of worked-through examples. The dynamic field model we will review invokes all the instabilities introduced earlier as well as the memory trace to account for sensory-motor decision-making and perseverative reaching in infancy and early childhood.

SPACES

It is quite intuitive that there would be infinitely many different things we could potentially see. Think about an object, say, a bottle standing on the table in front of you. The bottle might vary in size, shape, color, and surface texture. It might be positioned at different locations on the table. If someone held up the bottle, its orientation relative to you, the observer, might vary. All these variations are, a priori, continuous in nature: location, orientation, color, shape, texture—all may vary in a graded way. Visual morphing software makes such continuous variation directly accessible to computer graphics.

How might we formalize these continua of possible percepts? Let's use a minimal setting that would be typical of a psychophysics experiment: a single spot of brightness moving on a computer screen. The observer perceives the moving spot while fixating on a location marked by a cross. A continuum of instantaneous motion percepts is possible: The spot can move through different locations in different directions. This continuum can be described using a mathematical space that is spanned by coordinate axes. A possible set of coordinates includes the two-dimensional location of the spot on the retina and the direction of motion on the retina relative to a fixed axis, say, the horizontal axis (Figure 2.1). This yields a three-dimensional space of possible motion percepts of a single spot of light. Each location in that space represents one possible motion percept. Visual object motion may vary along additional dimensions such as speed, rigid body rotation, motion in depth, and so on. There is probably no single best way for how to describe the set of possible motion percepts. The dimensions we need to include may be dictated by the questions we ask an observer in an experiment. We might ask an observer to discriminate between motions that differ in movement direction, or ask the observer to point a joystick in the direction of motion perceived. In this case, motion direction is a critical dimension that needs to be accounted for. In a more complex setting, we might ask an observer to intercept a moving object. This probes multiple dimensions of motion perception, including direction but also speed and timing.

How many dimensions are needed to describe a real-world percept? An extreme view, taken in mathematical models of computer vision, is to sample the image by "pixel" (picture elements) and describe each pixel by a few coordinate axes that can capture, for instance, the intensity in the three color channels red, green, and blue. An image resolution that human observers find convincing may

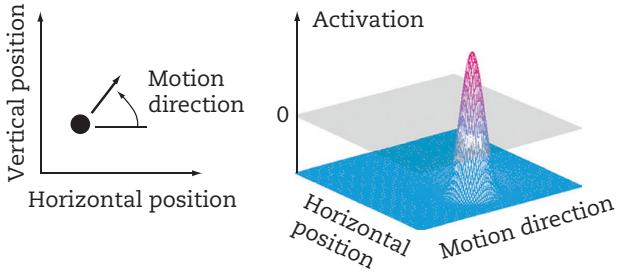

FIGURE 2.1: *Left:* Possible perceptual manifestations of a single moving spot of brightness, marked by a filled circle, moving in the direction marked by an arrow, can be described by a small number of continuous dimensions, including the location of the motion in the visual array (horizontal and vertical in a retinal reference frame) and the direction of motion. *Right:* For two of these dimensions, the representation of a single motion in an activation field is illustrated. The motion induces a single peak of positive activation located at the appropriate location in the space of possible motions, while all other locations in the field have negative levels of activation. Note that this activation pattern represents the location and the direction of motion of the spot of brightness at one moment in time. If we were to follow the spot of brightness as it moves on the retina, the peak would track that movement, shifting to a new retinal location at every moment in time.

be as high as 1000 × 1000 pixels, which would imply that the image as a possible percept has about 3 million dimensions. Now that is a questionable count. First of all, most variations of an individual pixel lead to visual noise, not to new visual percepts. The range of possible images created by looking at the world is constrained by properties of the world. For instance, surfaces tend to be continuous and their orientation in space tends to vary continuously. This creates reflectance maps in which brightness varies continuously. In fact, it is possible to estimate shape from shading based on such constraints (Koenderink & van Doorn, 2003). Moreover, visual perception is constrained by attention. Only a small portion of the image is in the attentional foreground at any given moment in time. In fact, human observers may be blind to changes in nonattended parts of the visual array if the transients used to induce change are masked (Simons, 2000).

So counting the dimensions of an image might not be a good estimate of the dimension of the space of possible percepts. Although the example we used in Figure 2.1 is a simplified laboratory setting, possible percepts may be best captured by visual feature dimensions that characterize individual objects in the perceptual foreground. The neurophysiology of the visual system suggests that there is a limited number of cortical maps representing such visual features, perhaps not more than 40 to 60 (Swindale, 2000). DFT is based on the hypothesis that neural representations in the brain can be captured by continua spanning a limited number of dimensions. We typically use coordinate systems that are consistent with the known cortical feature maps. This link to neurophysiology will be expanded on in Chapter 3.

That the set of possible voluntary limb movements is similarly of modest dimensionality is, perhaps, more directly intuitive. Consider, for instance, the set of possible voluntary movements of the hand that are oriented to an object (Figure 2.2). Such movements may vary in direction and extent, perhaps also in the amount and direction of mechanical resistance, or in the peak velocity of the movement. Neurons in motor and pre-motor cortex are tuned to such movement parameters, which span the space of possible movements (Georgopoulos, 1986). Each location in that space corresponds to one particular hand movement.

The visual array is a two-dimensional space that is an important component of the descriptions of both possible percepts and possible actions. This is obvious when one thinks of eye movements in which gaze is shifted toward different locations in the visual array. A visual scene is captured by its spatial layout, typically along the two spatial dimensions that describe a surface such as a tabletop or the floor on which we stand. In addition to their spatial location we may remember the colors of objects, their shape, or their orientation. If we lump these feature dimensions together, we can think of objects as being represented by a location in an appropriate space that combines visual space with feature dimensions. Sets of objects are sets of such locations. Later we will see how this embedding of percepts and actions in the two-dimensional visual array can play a role in organizing higher-dimensional representations through binding (see Chapters 5 and 8). We can use the same style of thinking for more abstract properties of the world. For instance, an "ordinal" dimension can be used to characterize the spatial or temporal order of events (this idea will be elaborated on in Chapter 14).

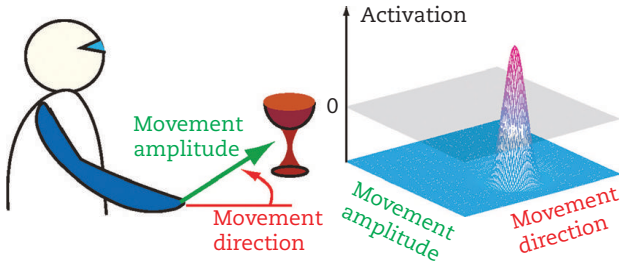

FIGURE 2.2: *Left:* Illustration of the movement parameters' direction and amplitude: By varying the direction of end-effector motion in space, together with the movement amplitude, a set of possible targeted hand movements can be described. *Right:* Activation defined over these two dimensions represents through a single peak the presence of a movement plan. The location of the peak indicates which movement amplitude and direction is planned. Activation in the peak is positive while elsewhere it is negative, so that only activation variables inside the peak may impact downstream neuronal networks that may be driving the motor action.

ACTIVATION FIELDS

What might a neural representation of a continuous space look like? Go back to Figure 2.1, which illustrates the three-dimensional space of the possible visual motions of a single spot of brightness. This space can be represented by a continuum of activation variables, one for each location in the three-dimensional space. These activation variables are labeled with an index that has continuous values. Mathematically, this makes them a field, a field of activation. This mathematical concept of a field is precisely analogous to how fields are used in physics, such as in the gravitational field, the electrical field, or the flow field inside a fluid or gas. The gravitational field, for instance, assigns to every location in three-dimensional Euclidian space a gravitational potential that can be assessed by observing the force exerted on a test mass. At any location, that force points in the direction in space in which the gravitational potential decreases most strongly, computed as the gradient of the gravitational field. The link between activation fields and measurement or observation is similarly based on the spatial pattern generated in the activation field. This is illustrated on the right half of Figure 2.1 for the activation field defined over the horizontal position and the direction of a visual motion (the vertical position is omitted to make the graphical representation practical). The field has an activation pattern with a single peak of action. Its center specifies the location and direction of the single perceived visual motion.

Not only the location of maximal activation but also the width of the peak is meaningful and can be assessed in an experiment. Psychophysical experiments on visual motion, for instance, can probe the range of activation around a particular location in the location/direction space by inducing an initial activation pattern through a first motion stimulus—say, a horizontal motion (an activation pattern centered on 0°). This may then be followed by a second stimulus that probes neighboring locations of the location/direction space, for example, by specifying motion at an angle of 67.5° (= 90° – 22.5°) from horizontal and another at an angle of 112.5° (= 90° + 22.5°) from horizontal. Motion perception will be typically selective, so that only one of the two motions is seen. If the 67.5° motion is preferred over the 112.5° motion, then we infer that the prior pattern of activation centered at 0° overlaps more with input at 67.5° than with input at 112.5°, biasing motion perception toward the closer angle. This was confirmed in experiments characterized by the label "motion inertia" (Anstis & Ramachandran, 1987) and were referred to in Chapter 1. The experiments show that the activation peak representing horizontal motion at 0° must reach out to at least 67.5°. Paradigms of perceptual hysteresis provide similar signatures of the metric range over which previous perceptual experience, represented by patterns of activation, impacts new perceptual experience (Hock, Kelso, & Schöner, 1993; Hock & Schöner, 2010).

In the motor domain, behavioral signatures of the width of activation peaks may be observed through the variance of movements from trial to trial. In the timed movement initiation paradigm, participants are trained to initiate movements at a fixed time, paced by a metronome (Ghez et al., 1997). Which movement out of a range of possible movements must be performed is cued only a short moment before the metronome signal. This stimulus–response time is experimentally varied. When the possible movements are metrically close, say, closer than 60° for movement direction, then the distributions of movement directions across trials observed for short stimulus response times is monomodal and centered on the mean movement direction. When the different possible movements are metrically far from each other, farther than 60° for movement direction, then the distributions are multimodal, each maximum centered on one of the possible movement directions (Favilla, 1997). With increasing stimulus–response interval, the monomodal distributions sharpen and become centered on the correct, cued movement direction. In the multimodal distributions, one peak centered on the correct movement direction sharpens and grows, whereas the other peaks decay. The transition from monomodal to multimodal initial distributions of movement parameters gives an indication for the width of the underlying activation peaks in the space of movement directions (Erlhagen & Schöner, 2002). In fact, it is possible to directly observe such distributions from the neural activity of populations of neurons tuned to movement direction (Georgopoulos, Schwartz, & Kettner, 1986). The width of distributions of population activation is consistent with the estimate from the behavioral data (Erlhagen, Bastian, Jancke, Riehle, & Schöner, 1999). This link between activation fields and population activity in the brain will be reviewed in detail in Chapter 3.

Peaks of activation are the fundamental units of representation in DFT. Peaks signify two things.

First, because the level of activation within a peak exceeds the threshold of the sigmoid function, the peak reflects the fact that an instance has been created within the activation field that is now capable of impacting any other neural networks that the field projects onto. This may include the motor system, so that peaks ultimately drive behavior in DFT (exactly how they do this is the topic of Chapter 4). In a sense, peaks are thus "go" signals for whatever process is driven by the field. Secondly, the location of a peak represents metric information along the dimensions that span the activation field. Through its location, a peak thus signifies an estimate of a perceptual state, of a movement parameter, or of other metric feature dimensions.

If perceptual information along the dimension of an activation field is multivalued, peaks of activation may represent different kinds of perceptual decisions. Figure 2.3 gives an example from the perception of apparent motion (Giese, 1999). When a point of light is first shown and then replaced by two points of light at different locations, one of three things may happen: Visual motion may be perceived from the first point of light in the direction that averages the directions to the two target lights (fusion). A splitting visual motion may be perceived, starting at the first light and ending at the two new locations (transparency). Or a single visual motion may be seen from the first to only one of the two new locations (selection). (See Kim and Wilson, 1993, for psychophysics of this kind.) An activation field representing movement direction may represent all three states of affairs. It may generate a single peak centered over the two targets (fusion). It may generate two peaks, each centered over the direction to one target (transparency). Or it may generate a single peak centered over one of the two targets (selection).

In Chapter 6 we will see that the number of peaks that can be simultaneously activated is limited by inhibitory interaction, a constraint that provides a neural account for capacity limits. So, the typical picture in DFT is that only a small number of activation peaks are present at any time.

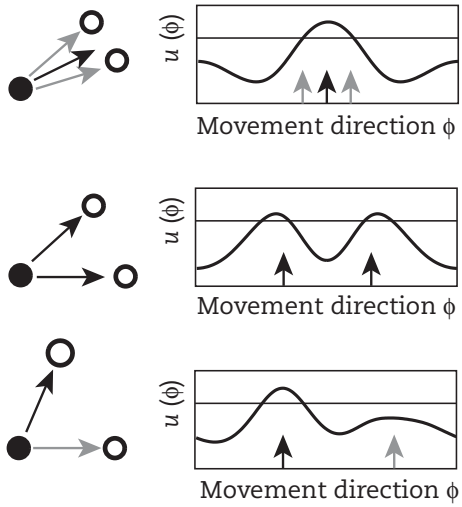

FIGURE 2.3: The *left* column illustrates three stimuli of apparent motion in which a spot of brightness (filled circle) is extinguished and two spots of brightness (open circles) appear elsewhere. Such displays may generate a percept of apparent visual motion as indicated by the arrows. Depending on the angular distance between stimulated motions, the perceived visual motion (black arrows) is either a single fused motion (*top*) in the direction of the average of the two stimulated motions (gray arrows), or consists of two transparent motions in the stimulated direction (*middle*), or is a single motion at one of the two stimulated locations (*bottom*). The *right* column shows the activation field defined over movement direction that represents these perceptual outcomes. *Top*: The fused motion (black arrow) is represented by a peak positioned near the average direction of the two inputs, whose locations are marked by gray arrows. *Middle*: Two motions perceived at the same time (transparency) are represented by two peaks located each over a stimulated movement direction. *Bottom*: One motion is represented by a single peak located at the site corresponding to its movement direction, while activation at the other stimulated site is suppressed. Adapted from Giese, 1999.

FIELD DYNAMICS

In DFT, activation fields are postulated to form dynamical systems. This means that an activation field, $u(x, t)$, defined over dimension, x, evolves in time, t, as described by a differential equation. This equation has a form analogous to that used for individual activation variables in Chapter 1. It links the rate of change of activation, $\dot{u}(x,t)$, at any location, x, through a $-u(x,t)$ term to the current level of activation, $u(x, t)$. This is the stabilization mechanism that limits growth of activation at positive levels and decay of activation at negative levels. The resting level, $h < 0$, is assumed to be the same for all field locations, while localized input, $s(x, t)$, may vary along the field dimension and in time. Thus, the first three terms in

$$\tau\dot{u}(x,t) = -u(x,t) + h + s(x,t) + \int k(x-x')\,g\big(u(x',t)\big)\,dx' \qquad (2.1)$$

are identical to the dynamics of individual activation variables, except that the discrete index that numbers the different activation variables has been replaced by the continuous variable, x, that spans the field dimension. As before, the parameter, τ, determines the overall timescale of the temporal evolution of $u(x, t)$.

What is different for activation fields compared to activation variables is the mathematical format of neural interaction. The integral is a continuous version of the sum over all field sites, x'. Each site, x', contributes only to the extent to which activation at that site exceeds a threshold as mediated by a sigmoidal function, $g(u(x',t))$. The threshold for coupling is, by convention, at $u = 0$, although the sigmoid function may be soft enough to allow activations slightly below zero to also contribute. The strength with which supra-threshold activation at site x' contributes to the rate of change of activation, $\dot{u}(x,t)$, at site x is a function, $k(x-x')$, of the distance between the two sites. Interaction is excitatory $(k(x-x')>0)$ for close distances, and inhibitory $(k(x-x')<0)$ for larger distances. This dependence of coupling strength on the distance between field sites makes the dynamics a homogeneous integrodifferential equation: The dynamics looks the same everywhere along the dimension of the field (see Box 2.1). With a solution, $u(x, t)$, any shifted version of this solution is also a solution. Only localized inputs, $s(x, t)$, that differ at different field locations break the homogeneity.

BOX 2.1 CONVOLUTIONS

COAUTHORED WITH SEBASTIAN SCHNEEGANS

Activation fields are continuous in space, but when we numerically solve the integrodifferential equations of DFT, we approximate continuous space in discrete steps, just as we did for continuous time (Box 1.4). This box explains how the convolution of the field with the interaction kernel is computed, which gives us the opportunity to help create a better understanding of the meaning of the convolution. We are referring to this contribution to the neural dynamics, Equation 2.1:

$$\left[k * g(u)\right](x) = \int k(x - x')g\left(u(x',t)\right)dx', \tag{B2.1}$$

where k is the interaction kernel listed in Equation A2.3 and g is the sigmoidal threshold function of Equation A2.2. The interaction kernel is analogous in DFT to synaptic weights in neural networks. These would be the weights with which "neurons" at locations x' project onto the "neuron" at location x. The integral has a particular form. It is a function of one argument, x, and integrates over the product of two functions. One function depends only on the integration variable, x', the other depends on the difference between the outer variable, x, and the integration variable, x'. Integrals with this form are called *convolutions*. The asterisk in the new notation, $\left[k * g(u)\right](x)$, stands for "convolve," here, convolve the kernel, k, with the function, $g(u)$.

The range over which the integral extends is not marked, implying that it extends over the entire space spanned by the variable x'. In some cases, such as for spatial memory, this may be a linear space, for example, the spatial positions along a line that may, a priori, extend to infinity in both directions. In other cases, this may be a circular space, for example, the space of heading directions, in which case it extends over the complete circle. In either case, we would like the boundary of the space over which the activation field is defined to play no particular role, as, in most cases we model, nothing is known about boundary effects. Your visual field, for instance, is limited, but the boundaries play no particular role. Vision just diminishes near the boundary.

When we compute the integral of Equation B2.1 concretely, we need to commit to a particular range of integration and address the boundary issue. This is true, in particular, when the integral is computed numerically. The best way to make the boundaries "neutral" is to impose periodic boundary conditions on the activation field: Activation at the left boundary

of the field is equal to activation at the right boundary of the field. This is natural for circular space, in which there is no boundary, so the cut we make when we compute the integral should not matter. It is useful also for spaces at the boundaries of which activation diminishes. The periodic boundary condition is the most neutral one, in a sense. And if activation values are low near the boundary, the precise boundary condition doesn't matter.

How do we work with periodic boundary conditions? Figure 2.4 illustrates the key idea. At the top of the figure is a field over a finite range, here from 0° to 180°. What is plotted is already the supra-threshold activation field, $g(u(x'))$, as a function of x'. The interaction kernel, plotted in the third row, has the same size, ranging from –90° to +90°. Now, let's say we try to compute the convolution integral for a particular value, x, of the outer variable, say, $x = 50°$, as suggested in the figure. In the graphical depiction of this computation, we have to align the center of the interaction kernel with this point in the field. The following problem arises: The kernel extends on the left into portions of the field that lie outside the boundaries. And the field extends on the right beyond the reach of the kernel. We can solve this problem by expanding the space over which the supra-threshold field is defined. This is illustrated in the top two rows. We simply copy the left half of the field and attach that half on the right, and copy the right half of the field and attach it on the left. This imposes periodic boundary conditions on the center part, which is the true field we are trying to model. And it now makes values available to those parts of the kernel or of the field that reach beyond the boundaries. At the bottom of the figure are the matching parts of kernel and supra-threshold field plotted on top of each other. Computing the convolution now simply consists of multiplying these two curves with each other at each field location and then integrating across the shown range.

This becomes even clearer when we replace the mysterious concept of "integrating" with "summing" by going to a discrete numerical approximation. On the computer, we sample

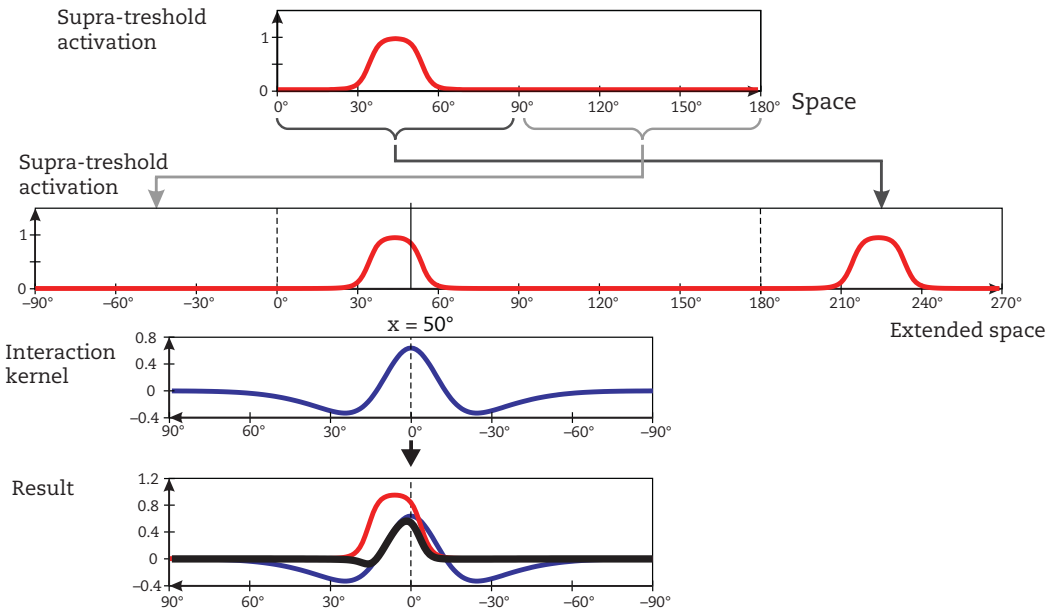

FIGURE 2.4: *Top:* Supra-threshold activation, $g(u(x'))$, of a field is shown over a finite range (from 0 to 180°). *Second from top:* The field is expanded to twice that range by attaching the left half of the field on the right and the right half on the left, imposing periodic boundary conditions. *Third from top:* The kernel has the same size as the original field and is plotted here centered on one particular field location, $x = 50°$. *Bottom:* The matching portions of supra-threshold field (red line) and kernel (blue line) are plotted on top of each other. Multiplying the values of these two functions at every location returns them to the black line. The integral over the finite range of the function shown in black is the value of the convolution at the location $x = 50$.

the continuous field dimension, x', by discrete steps in space, $x_i = i\Delta x$, where $i = 0,1,2,\ldots,n$ and $n = L / \Delta x$ (where we choose Δx such that n is an odd integer number). Here we have assumed that the range of x' is $[0, L]$ ($L = 180$ in the figure). The convolution is then approximated as

$$\left[k * g(u)\right](x_m) = \sum_{i=m-l}^{i=m+l} k(x_m - x_i)\, g\left(u(x_i)\right) \qquad \text{(B2.2)}$$

where $l = (n - 1)/2$ is the half-width of the kernel. The sum extends to indices outside the original range of the field (e.g., for $m = 0$ at $i = -l$). But that doesn't cause problems because we extended the range of the field, as shown in Figure 2.18.

Note again that to determine the interaction effects for the whole field, this computation has to be repeated for each point x_m. In COSIVINA all of these problems have been solved for you, so you don't need to worry about figuring out the indices in Equations like B2.2 ever again!

Activation peaks are inherently attractors of this neural dynamics. As illustrated in Figure 2.5, local excitatory interaction among locations within a peak of activation stabilizes the peak from decaying. If this were the only form of interaction, however, activation at the boundaries of a peak would keep rising, leading to unbounded expansion of the peak. Inhibitory interaction over longer distances in the field stabilizes peaks against this expansion. Thus, excitatory and inhibitory interaction together stabilize the shape of activation peaks. Amari (1977) showed this mathematically. His and subsequent analyses help us solve the "inverse" dynamics problem. In the typical "forward" dynamics problem we are taught in math courses, we find the solutions of a given equation. Modeling entails inverse dynamics, finding an equation that has the desired solutions. In DFT, we seek equations that have peaks of activation as attractor solutions. The mathematical analysis shows that the Amari neural dynamics is a possible equation that has peaks as attractors, and we adopt that equation as a possible mathematical formalization of DFT on that basis.

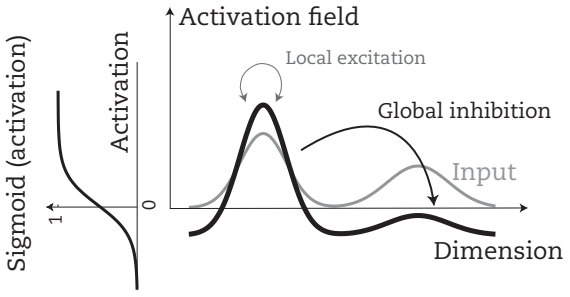

FIGURE 2.5: *Left:* A sigmoid function, $g(u)$, approaches zero for sufficiently negative values, and a positive constant for sufficiently positive values of activation, u. *Right:* As mediated by the sigmoid function, activated regions in the field interact by exciting nearby locations (light gray arrow), stabilizing peaks from decay, and inhibiting locations farther removed (dark gray arrow), stabilizing peaks against diffusion.

Through their positive levels of activation, peaks signal the decision in which an instance is created along the underlying dimension. This decision is stabilized by neural interaction. Neural interaction does not stabilize peaks against shifts along the field dimension. In the absence of localized input, the field dynamics is homogeneous so that any shifted version of an activation peak is also a possible solution. We shall see later in this chapter that drift along the field dimension is psychophysically real. Localized input may limit or stop such drift.

The two contributions to neural interaction, excitatory and inhibitory, are related to the two forms of interaction discussed for discrete activation variables in Chapter 1. Local excitatory interaction is a generalization of the self-excitation studied there, while global inhibition is a generalization of the mutual inhibitory coupling studied for two activation variables. Figure 2.6 illustrates these analogies by showing the relationship between the activation fields and discrete activation variables. One may think of the discrete activation variables as representing the total activation within a region in the field that approximately covers an activation peak. In this picture we only keep track of locations that receive input at some point in a task setting. In Chapter 1, only two locations were ever stimulated, and that is why two

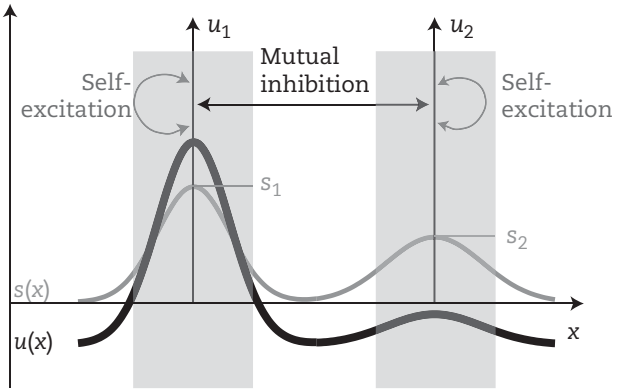

FIGURE 2.6: An activation field, $u(x)$ (solid dark line), is stimulated by input, $s(x)$ (solid gray line), with two local maxima. The field dynamics can be captured qualitatively by keeping track of activation only within the two regions (highlighted by gray shading) that receive input. Total activation in each region is described by an activation variable, u_1 and u_2, respectively; total input into each region by input strengths, s_1 and s_2, respectively. In this approximation, local excitatory interaction within each region becomes self-excitation of the activation variables, while global inhibitory interaction becomes mutual inhibition between the two activation variables.

activation variables were sufficient. Local excitatory interaction summed within a region shows up in the neural dynamics of the activation variable as self-excitation. Inhibitory interaction only gathers contributions from locations at which activation may become positive. For two activation variables, these are the two regions captured by the two variables, so that mutual inhibitory coupling of the two activation variables captures global inhibition. This analogy underscores, once more, that local populations rather than individual neurons are the substrate for representation. The question of how a particular activation variable with a discrete index may come to stand for a particular perceptual or motoric state is answered by embedding the activation variables in activation fields. The discrete variables are merely samples of an underlying continuous metric dimension.

ATTRACTORS AND THEIR INSTABILITIES

In Chapter 1, we discussed attractors and instabilities in some detail for the neural dynamics of one or two activation variables. The mathematical concept of stability and the mechanisms of bifurcation are really the same for activation fields, but they are less intuitive and more difficult to visualize. We shall look now at the two classes of attractor solutions of the dynamics of activation fields, the subthreshold and the self-stabilized activation patterns, and examine the instabilities that separate them. Lifting the dynamics from discrete activation variables to activation fields will provide new insight into the meaning of the instabilities and the situations in which they may arise. The exercises at the end of this chapter invite you to reproduce all instabilities discussed here, making use of an interactive simulator of dynamic fields.

Detection

The simplest stable state of the equation arises when activation is below zero and only weak inputs are present. In that limit case, no portion of the field is activated enough to return positive values from the sigmoid. Interaction is therefore not engaged and the field dynamics is now independent at each location, x, of the field

$$\tau \dot{u}(x,t) = -u(x,t) + h + s(x,t) \qquad (2.2)$$

Figure 2.7 illustrates this dynamics at one location. At its zero-crossing, $\dot{u}(x,t) = 0$, lies the stationary solution,

$$u_0(x,t) = h + s(x,t) \qquad (2.3)$$

that represents *the subthreshold attractor state,* essentially just the input, $s(x, t)$, shifted downward by $h < 0$. As in Chapter 1, we can read the stability of this solution off the negative slope of the rate of change at the zero-crossing. Activation grows if it lies below, decays if it lies above this stationary state. If input varies over time, activation will thus track the subthreshold solution with a delay that reflects the timescale, τ, of the field dynamics. (Strictly speaking, the subthreshold solution is not stationary then.)

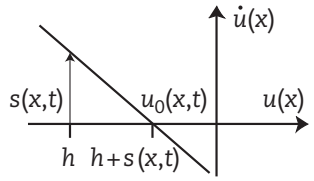

FIGURE 2.7: The dynamics of activation, $u(x)$, at a single field location, x, is illustrated. This dynamics is independent of activation at other locations as long as interaction is not engaged. That is the case around the subthreshold attractor, $u_0(x,t) = h + s(x,t) < 0$, that emerges as the zero-crossing of the rate of change, $\dot{u}(x)$. The subthreshold attractor becomes unstable and disappears if input $s(x, t)$ becomes sufficiently strong so that it pushes the subthreshold attractor toward zero from below and engages interaction.

Interaction is engaged as soon as activation approaches zero from below anywhere along the field dimension. Let's look at a location where input drives activation toward the threshold. We approximate the input pattern, $s(x, t)$, as a Gaussian centered on that location. Figure 2.8 traces the attractors of the neural dynamics when the strength of that localized input pattern increases. We start out with weak input, at which the only stable stationary state is the subthreshold attractor, a copy of the input pattern shifted down by the resting level, as discussed earlier. For a single Gaussian input function, this attractor is a subthreshold "hill" of activation. As input strength increases, activation in that attractor reaches threshold from below, engaging excitatory interaction, which pulls up the activation within the hill. In a recurrent cycle, increasing activation levels within the hill engage local excitatory interaction more strongly, which in turn increases activation levels. Through this growth cycle the subthreshold hill of activation becomes unstable in what we call the *detection instability*.

What solution does the activation field converge to once the subthreshold state has become unstable? Inhibitory interaction eventually limits the growth of the activated region, leading to a new balance of excitatory and inhibitory interaction. This is the self-stabilized peak attractor that is fundamental to DFT. Within the peak, the balance of excitation and inhibition leads to a positive level of activation, so that this attractor is an instance of the dimension represented by the field in the sense discussed earlier. Outside the peak, the inhibitory influence from the peak is unopposed by excitatory interaction, leading to a negative level of activation below the resting level.

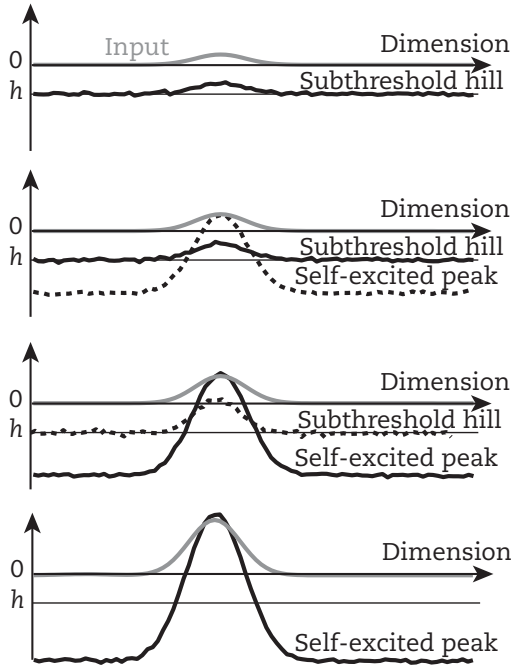

FIGURE 2.8: For a localized input pattern (gray solid line) that increases in strength (from top to bottom), the attractor states of a dynamic activation field are shown. *Top:* At low input strength, the only attractor is the subthreshold "hill" of activation (black solid line) that mirrors input shifted down by the negative resting level of the field. *Second from top:* At a larger input level, the subthreshold hill of activation continues to be stable but coexists with a self-excited peak of activation (black dashed line). This self-excited peak is close to the reverse detection instability: If input were weakened a little, the peak would decay and the system would return from this bistable regime to the monostable regime illustrated above it. *Second from bottom:* For stronger input, the subthreshold hill of activation (black dashed line) becomes unstable at detection instability, the upper limit of the bistable regime. *Bottom:* At even stronger input, the self-excited peak of activation is the only remaining attractor. The system is again monostable.

The possibility of a self-excited peak does not appear just as the subthreshold hill becomes unstable. This attractor has been around at levels of localized input below the detection instability. There is a range of input levels within which both the subthreshold hill and the self-stabilized peak of activation are stable. For input levels within this range, the neural dynamics is bistable. Only one of the two stable states can be realized at any one time. Which state the system is in depends on the history of activation. In the previous narrative, the neural dynamics starts in the subthreshold hill state and input strength is then increased.

The activation pattern tracks the change of input strength within the subthreshold solution as indicated by Equation 2.3. Only when the subthreshold hill becomes unstable at the detection instability does the activation pattern switch to the alternate attractor, the self-stabilized peak of activation. Conversely, if the system starts out in an activation pattern near the self-stabilized peak, it converges to that attractor and stays in that attractor as input changes. This may happen, for instance, if the system has been pushed through the detection instability by a strong input which is then reduced in strength. Once the system has switched to the self-stabilized peak, it persists in this state even as input strength is reduced back below the critical level of the detection instability.

As long as there is enough positive activation within the peak to keep the peak afloat through local excitatory interaction within the peak, the stabilization mechanism of the peak attractor remains viable. When the level of localized input falls below a critical level, this mechanism begins to fail. The reverse detection instability occurs, delimiting the range of bistability on the side of low levels of input (Figure 2.8).

In summary, when the strength of localized input varies, the dynamics of activation fields goes through three regimes: monostable with the subthreshold hill of activation as sole attractor at low levels of input; bistable with both subthreshold hill and self-stabilized peak of activation as attractors at intermediate levels of input strength; and monostable with the self-stabilized peak of activation as sole attractor at high levels of input strength. Within the bistable region, which attractor is observed depends on the history of activation and, thus, on the history of input strength. Increasing input strength leads to persistence of the subthreshold hill of activation up to the detection instability. Decreasing input strength leads to the persistence of the self-stabilized peak of activation down to the reverse detection instability. This is the same hysteresis discussed in Chapter 1, in the approximation where we described the dynamics around the stimulated location of the field by a single activation variable with self-excitatory interaction (see Figure 1.17).

The name we chose, *detection instability*, suggests that the switch from the subthreshold hill to a self-excited peak of activation could be viewed as a detection decision. The peak indicates that an instance of whatever the field represents has been created and is now capable of affecting downstream parts of the neural dynamics because the activation levels are sufficient to drive sigmoidal coupling functions above zero. The bistability of the dynamics just below the detection instability implies that the detection decision remains stable even if the input that induced it fluctuates in strength. This is a significant feature of decision-making in neural dynamics that may be contrasted with the notion of threshold piercing common in neural network models. According to this notion, a detection is registered whenever an activation variable exceeds a particular detection threshold (Schall, 2004). When this threshold is first crossed, fluctuations in the input signal may often lead to activation falling below the threshold, again in close temporal vicinity to the first detection. Crossing of the threshold is thus not a stable mechanism for making detection decisions when these are linked to fluctuating sensory signals. The detection instability, in contrast, makes it possible to make stable detection decisions in the face of time-varying and fluctuating sensory input.

Another conceptual implication of the detection instability has to do with continuous versus discrete time. As an organism moves through an environment, sensory inputs typically vary continuously over time. Out of such time-continuous sensory data, the detection instability creates an event at a discrete time, the moment when the rapid transition from a subthreshold hill to a self-stabilized peak signifies a decision. Embedded in a complete sensory-motor system, this event may ultimately trigger motor actions. The discrete moments in time at which such actions are initiated thus emerge autonomously from the time-continuous neural dynamics.

After the discrete decision event, the self-stabilized peak remains coupled to continuously varying sensory input, however. One way this can be seen comes from the fact that the peak is centered on the localized input, as analyzed mathematically by Amari (1977). The position of the peak may be viewed as an estimate of the location at which localized input is maximal. When the input pattern moves, the peak tracks the moving input. The peak will typically lag behind the moving input, just like any low-pass filter does, and for input that moves too fast it may fail to track (the peak then decays at the old location and a new peak is induced at the new location). But within these constraints, the peak

stays connected to time-varying input that is sufficiently strong.

Working Memory

The reverse detection instability does not always occur; there are conditions under which even at zero strength of localized input the self-excited peak attractor persists. This may happen, for instance, for sufficiently large resting levels, $h < 0$, which alone can be sufficient to keep activation in the self-excitatory loop that sustains the peak. At a given resting level, this may happen when the strength of local excitatory interaction is sufficiently large. Under these conditions, whenever a peak has somehow been induced, the peak persists, sustained entirely by interaction, in the absence of any localized external input into the field.

To see the functional significance of self-sustained peaks consider a scenario in which a peak is first induced by a detection instability at a location, x_0, at which localized input was maximal. When the localized input is removed, the peak persists and thus effectively is a memory of the previous detection decision (Figure 2.9). Its positive level of activation represents a memory of the fact that significant input to this field has existed at some point. Its location represents a memory of the location of that previous input. Sustained peaks of activation of this nature are the commonly accepted image of how working memory comes about in neural populations, consistent with neurophysiological evidence for sustained firing of neurons in working memory tasks (Fuster, 2005; Fuster & Alexander, 1971). This will be discussed at length in Chapter 6, where we will also address capacity limits and how information is brought into and out of working memory.

Sustained peaks of activation are really the same attractors as self-stabilized peaks of activation. We speak of sustained peaks after the localized input has been removed. Whether or not a peak is sustained in the absence of input depends on dynamic parameters. Figure 2.9 illustrates one form of the *memory instability*, a transition in dynamic regime in the absence of localized input. For a sufficiently negative resting level, h (left column in the figure), the neural dynamics is monostable with the subthreshold attractor in the absence of localized input. At higher (but still negative) resting level, h (right column in the figure), the neural dynamics is bistable in the absence of localized input. Both the subthreshold state and sustained peak are attractors of the field dynamics. The sustained peak will be observed when the dynamics starts out with a self-excited peak state as shown in the figure. In this bistable regime, the sustained peak is actually a family of infinitely many possible attractors, which are marginally stable because they can be shifted along the field dimension. Drift along the marginally stable direction is possible in the presence of noise. Any small inhomogeneity breaks the marginal stability and leads to the emergence of a single attractor that is localized over any local maximum of input. The drift and breaking of marginal stability are psychophysically real and can be observed in human working memory for metric information as discussed later in this chapter. (Strictly speaking, marginally

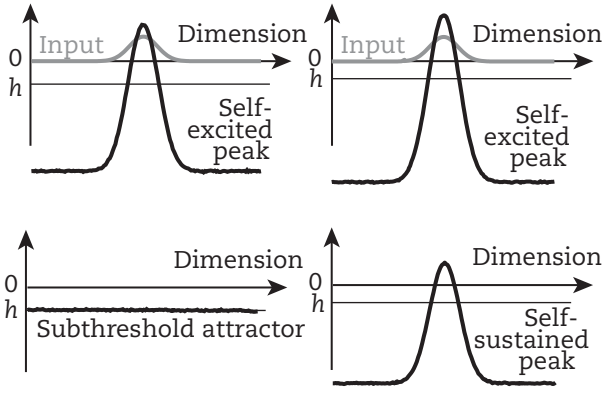

FIGURE 2.9: The memory instability is illustrated by contrasting a condition in which peaks of activation are not sustained when localized input is removed (*left*) with a situation in which peaks are sustained (*right*). In each case, a localized input (gray solid line) induces a self-stabilized peak (*top*) and is then removed (*bottom*). When peaks are not sustained, the system switches to subthreshold attractor upon removal of localized input (*bottom left*). When peaks are sustained, the self-excited peak becomes a self-sustained peak (*bottom right*). The resting level, $h < 0$, is more negative on the left than on the right. Increasing resting level may push the system through the memory instability into the regime of sustained peaks.

stable sustained peaks are not attractors, but it is common practice to still refer to them this way, as they resist all perturbations except lateral shift.).

Selection

Now let's look at slightly more complex input patterns, minimally an input with two local maxima (Figure 2.10). Generically, a self-stabilized peak arises at only one of the two locations. Activation at the other location is suppressed by inhibitory interaction that comes from the activated peak. The location with suppressed activation cannot conversely inhibit the activated peak because its activation is insufficient to return positive values of the sigmoid. The timing of activation controls which location "wins" this selective competition. A location at which activation rises earlier reaches supra-threshold levels of activation first and begins to inhibit activation at other locations. Locations at which activation arises later are inhibited before they can reach supra-threshold levels. The temporal advantage of a location may arise because inputs arrive asynchronously. This is the case, for instance, if one location was previously stimulated and prior activation from that previous stimulation biases the selection when a new stimulus arrives. The competitive advantage of a location may also arise because inputs of different strengths impinge on different locations. The input function, $s(x, t)$, may favor one location over another as suggested in Figure 2.10. As a result, activation at the location that receives stronger input rises faster and reaches threshold earlier, engaging interaction and suppressing the further increase of activation at competing locations. In the models discussed so far, we have not specified exactly how input profiles arrive. In neural networks, the pattern of synaptic connectivity from a sensory surface to the network determines how sensitively a neuron responds to a particular input. Input patterns that best match the pattern of synaptic connectivity provide the strongest input to a given neuron (Haykin, 2008). This core mechanism of neural networks is lumped into the input function, $s(x, t)$, in DFT. "Good match" of an input pattern is thus captured by large levels of input for a particular location, leading to early rise of activation at that location and a competitive advantage of that location. The selection mechanism of DFT is thus a possible process implementation of the connectionist conception in which the neuron is selected that responds maximally because its connectivity best matches an input pattern.

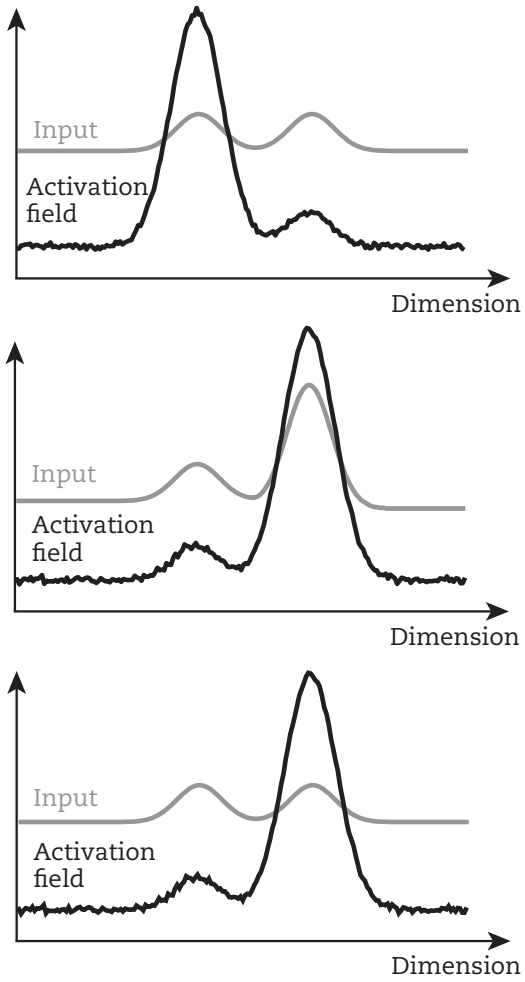

FIGURE 2.10: Input functions (solid gray lines) and stable activation patterns (solid black lines) are shown as functions of the field dimension in three situations. *Top:* Input is bimodal, with identical maximal level of input at two locations. An activation peak centered on the left mode is a stable state that may have emerged because activation was initially higher on the left from the leftmost mode being presented first, or by chance from fluctuations in input. *Middle:* When input to the rightmost location is much stronger than to the leftmost location, the peak centered on the left location is no longer stable and the system switches in selection instability to a peak centered on the rightmost location. *Bottom:* If input is then returned to symmetric levels for both modes, the peak centered on the right mode remains stable, an instance of the stabilization of selection decisions.

The determination of selection by temporal order implies that selection choices are stabilized when input varies. Once a self-excited peak has been erected over a particular local maximum of input, inhibitory interaction from this peak to all other locations prevents other peaks from arising over other stimulated locations even if input to those locations becomes stronger than

input to the selected location. This can be seen in Figure 2.10: Activation is suppressed at the alternate field location even though input to either location is of the same strength. The stabilization of selection decisions makes it possible to continuously link an activation field to sensory input while at the same time preventing the selection decisions from fluctuating each time the location of maximal input varies. Contrast this to an algorithm, which would select at every moment in time the location of maximal activation. That location could vary from moment to moment across multiple stimulated locations. In a sense, stable selection is a form of robust estimation, in that components of input that are metrically close to the location of the selected peak contribute to the estimate that peak represents, while components that are metrically far from the selected peak are suppressed.

The stabilization of selection decisions has limits. When input strengths are sufficiently different, an initially established selection decision may be reversed. In the top panel of Figure 2.10, the leftmost peak has been selected in some way. When the rightmost input becomes much larger than the input to the leftmost peak (in the middle panel), this selection decision can be overturned. A peak at the rightmost location emerges and suppresses by inhibition the peak at the leftmost location. This switch involves an instability, which we call the *selection instability*. Just as for detection, this instability occurs at the boundary of a bistable region in which two attractors coexist: A peak centered on either input is stable. Beyond the selection instability, the system is monostable; only the peak centered over the more strongly stimulated location remains stable.

This capacity to select a location from a multimodal input pattern generalizes beyond just two locations. Whether or not selection leads to a single self-excited peak or whether multiple peaks can coexist depends on the interaction kernel—in particular, its inhibitory portion. When inhibition levels off at larger distances, then peaks that are sufficiently far apart from each other can coexist. Generally, as more peaks are induced, the total amount of inhibition projected onto other locations increases. This limits the number of peaks that can be stabilized, providing an account for capacity limits of working memory, as discussed in Chapter 6.

There are additional instabilities hidden here. Transitions may occur from a dynamic regime in which multiple peaks can be stable to a regime in which a single peak is selected. Transitions may occur between dynamic regimes in which the number of peaks that can coexist changes. In each case, these instabilities can be brought about by changes in the strength and range of contributions to interaction within fields, but may also depend on the metric and strength of inputs and on the resting level. In principle, the number of such instabilities is unlimited. Another kind of transition occurs within the selective regime. For instance, when the neural dynamics is bistable, with a peak positioned over either of two local maxima of input, a transition may occur to a monostable regime when the two locations move close to each other. This results in a single peak positioned over an averaged location (Kopecz & Schöner, 1995).

One final instability needs to be addressed here, a variant of the detection instability linked also to selection. This instability has broad implications for DFT in particular, for its link to learning, which will be discussed next. Consider again a situation with a few localized inputs that are now quite weak. We might think of these inputs as inhomogeneities of the field that may arise through sensory input from the layout of the scene or from learning processes that give some field locations higher resting levels than others (see later discussion in this chapter). As illustrated in Figure 2.11, these small inhomogeneities preshape the field in the subthreshold state. The detection instability may now amplify this preshape into a full, self-stabilized peak. The input that induces the detection instability may be homogeneous, that is, contain no specific information about the location at which a peak is to be generated. What happens is that such a homogeneous boost to the activation level of the field first drives the field through the threshold at one of the locations that are a little more activated than the rest of the field. Interaction engages and brings about a detection instability around that location. If inhibition is global, the emergent peak will drive selection so that other, slightly less preactivated locations cannot generate peaks. Even if the boost is present for a brief moment only, the bistability of subthreshold and self-stabilized peaks below the detection instability helps stabilize the full peak once it has been activated. So the boost-driven detection instability amplifies small inhomogeneities in the field into complete self-excited peaks that represent decisions and impact downstream neural dynamics. Conversely, the boost-driven detection

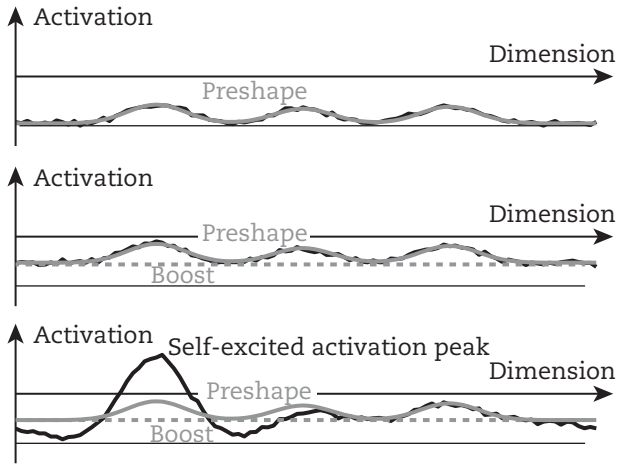

FIGURE 2.11: *Top:* An activation field is preshaped at three locations, so the subthreshold attractor has small hills of activation there. *Middle:* An input that is constant across the field boosts the activation pattern, pushing activation toward zero from below, here very close to the detection instability. *Bottom:* The field has gone through the detection instability, in which the subthreshold attractor has vanished, and has activated a self-stabilized peak localized over one of the three preactivated regions.

instability alleviates the demands on sensory input and on learning processes: These processes need to deliver only small, graded inhomogeneities that can then be amplified into full decisions without further specific information. This may help bootstrap fields from the sensory-motor domain in which inputs tend to be strong and stable to the cognitive domain in which inputs are internally generated and may be transient and weak. Using "boosts" to activate items is a topic addressed throughout the book, culminating in Chapter 14, where we will leave the sensory-motor domain farthest behind.

MEMORY TRACE

The neural dynamics discussed so far take place on a timescale at which inputs vary and decisions are made. Sustained peaks of activation, however, transform events on that fast timescale to longer timescales at which working memory resides. As working memory, sustained peaks are susceptible to capacity limits and interference which limit the persistence of these activation states when inputs vary in time. Interference arises through the selection instability when new sensory information competes with the existing sustained peaks.

A more general neural dynamics at the longer timescale of memory is a dynamics of learning. The simplest form of such learning is, perhaps, habit formation, as postulated by William James (1899). Habits are formed when particular behaviors are experienced often enough. They make it easier to reproduce the same behaviors. While the modern understanding of habit formation is both more complex and more specific (Yin & Knowlton, 2006), the Jamesian metaphor can be translated into DFT as an elementary and generic form of learning: Any instance of neural representation, a self-excited peak of activation, leaves a memory trace that facilitates the re-emergence of the same activation peak in the future (Erlhagen & Schöner, 2002). Figure 2.12 illustrates the mechanism: For a given activation field, the memory trace is a second layer of dynamics that evolves on the slower timescale of learning. Any supra-threshold activation in the field provides excitatory input into the memory trace. Locations at which activation is above threshold thus grow a memory trace. As the memory trace at an activated location grows, it decays at all other locations where there is currently no supra-threshold activation. In the absence of any supra-threshold activation, however, the memory trace remains unchanged, neither growing nor decaying. This form of a dynamic memory trace generates a representation of the history of supra-threshold activation in the field. The memory trace, in turn, provides weak excitatory input into the activation fields. This is how the memory trace facilitates peak formation at the locations where peaks have previously been generated.

A mathematical formalization of the memory trace invokes a second layer of dynamics for a field of memory trace levels, $u_{\text{mem}}(x,t)$:

$$\tau_{\text{mem}}\dot{u}_{\text{mem}}(x,t) = -u_{\text{mem}}(x,t) + g\big(u(x,t)\big) \qquad (2.4)$$

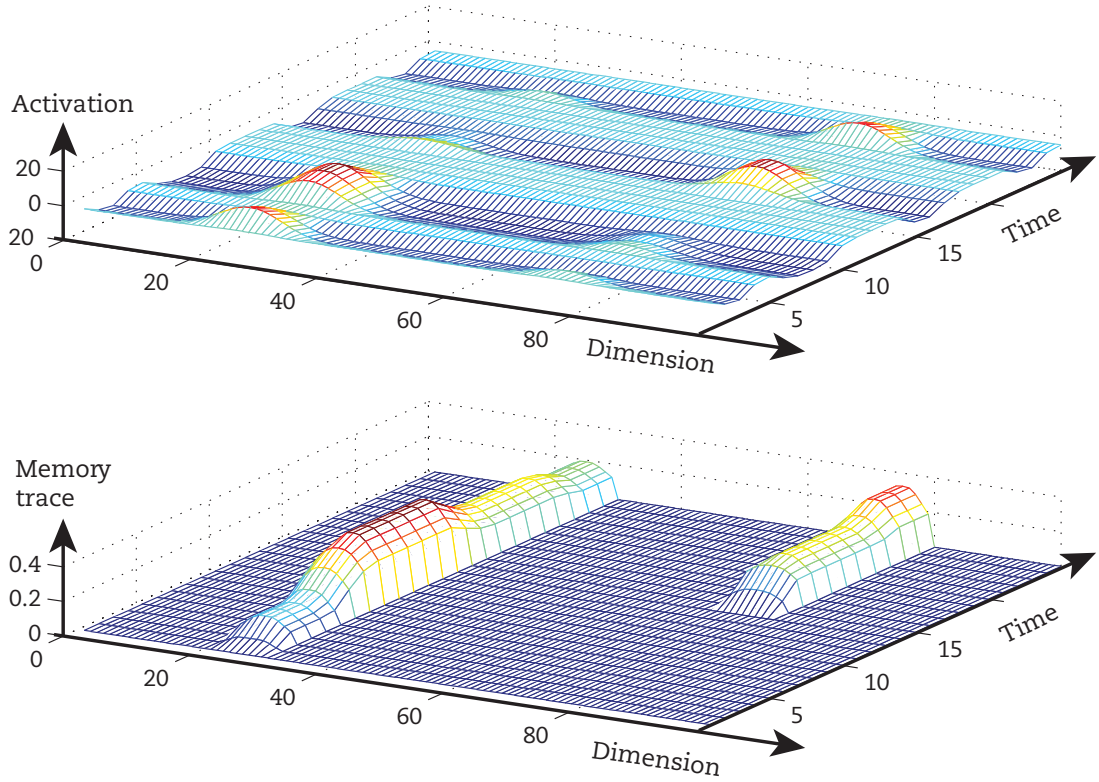

FIGURE 2.12: Evolution over time of an activation field (*top*) and its memory trace (*bottom*). The field receives time-varying input at two locations that induces a self-stabilized peak at these locations at different moments in time, interspersed with time intervals during which activation is below threshold everywhere along the field dimension. Supra-threshold activation drives the memory trace up at the matching location, for example, on the left for the first 10 seconds. At competing locations, the memory trace decays, for example, on the left around 15 seconds, as the trace grows on the right. In the absence of supra-threshold activation, the memory trace remains unchanged, for example, between 8 and 12 seconds and again between 18 and 20 seconds.

that evolves on the slower timescale, $\tau_{\text{mem}} \gg \tau$. The memory trace couples to the field dynamics according to

$$\begin{aligned}\tau \dot{u}(x,t) = &-u(x,t) \\ &+ h + s(x,t) + c_{\text{mem}} u_{\text{mem}}(x,t) \\ &+ \int k(x-x') g\big(u(x',t)\big) dx' \end{aligned} \qquad (2.5)$$

with strength, c_{mem}. The memory trace does not evolve (right-hand side of Equation 2.4 set to zero) when no location in the activation has supra-threshold levels of activation. More complex learning dynamics may have a faster timescale for the building of a memory trace than for its decay.

Erlhagen and Schöner (2002) showed how the dynamics of the memory trace generates a representation of the probability of events. Consider a two-choice motor task in which the frequency with which each choice occurs varies across different conditions. Response times covary with the probability of each choice according to the Hyman law (Hyman, 1953): Response times are shorter for the more frequent choice. In their dynamic field model of the task, Erlhagen and Schöner represented the movement choices as values of a movement parameter encoded in an activation field. The imperative stimulus specifies which choice to select and also serves as the "go" signal, authorizing the participant to respond. That stimulus was modeled as localized input to that field. This input drives the field through the detection instability, inducing a peak at the location that encodes the cued movement parameter value. Over time, peaks arise at the two locations, as illustrated in Figure 2.12. The probability of each choice determines the frequency with which the peaks occur. The memory trace at the two locations representing the two movements converges across trials to levels that reflect the frequency of each choice, a higher level being for the more frequent movement. These levels feed into the activation field, preactivating the field at the two locations. On any given trial, the imperative stimulus encounters, therefore, different

initial activation levels. The more probable choice starts from a higher initial level of activation and thus reaches threshold earlier, leading to shorter response times. A detailed mathematical analysis predicts the Hyman law, in which response times increase with the logarithm of choice probability. (The logarithm comes from the exponential time course of activation as it relaxes to the attractor. Inverting the exponential to compute the time at which threshold is reached leads to a logarithmic dependence on initial activation levels. See the appendix in Erlhagen and Schöner, 2002, for a derivation). The memory trace could thus be viewed as a process of how neural representations build probabilistic priors from their history of activation, as postulated by adherents to Bayesian thinking in cognition.

The history of activation may, more dramatically, lead to the emergence of categories. In Figure 2.12 we suggested that activation peaks occur repeatedly in different, non-overlapping locations. The memory trace thus consists of distinct patches that preshape the activation field in distinct locations. We have already argued that the boost-driven detection instability may amplify such preshaping into full-blown, self-stabilized peaks. Figure 2.13 illustrates that this may lead to categorical responding, so that the memory trace becomes a mechanism for category formation. In the figure, the field is preshaped by a memory trace with subthreshold hills at two locations. The imperative stimulus contains both a boost (a homogeneous input to the entire field) and a small, localized input that overlaps with one of the two preactivated locations. The localized input is sufficient to bias the field toward selecting the location with which this input overlaps instead of the alternative location, but is not sufficient to drive peak formation and is weaker than input from the memory trace. As a result, the field generates a self-stabilized peak positioned over the location preactivated by the memory trace, rather than the location specified by the localized input. Were we to vary the precise location of the localized input, the location of the self-stabilized peak would remain largely invariant, dictated by the pattern of preshaping. Only when the cue shifts enough to now bias the field toward selection of the alternate choice does the self-stabilized peak shift. In this sense, the field responds categorically to the imperative stimulus, the categories being the distinct locations at which the memory trace has been built up, preshaping the activation field.

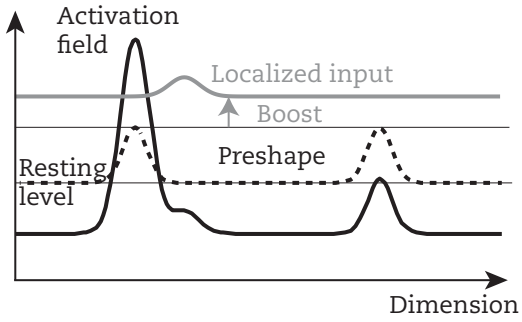

FIGURE 2.13: Categorical responding based on the memory trace: A field is preshaped (dashed line) by a memory trace at two locations at which peaks of activation have been frequently encountered. Other regions of the field are at resting level. When a weak localized input is applied jointly with a boost to the field (gray solid line), a self-stabilized peak (black solid line) is generated at the preactivated location that best overlaps with the small, localized input. Elsewhere, the field is suppressed below resting level, including at the precise location of the small, localized input.

The memory trace is an unsupervised form of learning, analogous to the Hebbian principle, in which the activation patterns experienced in a neural network change the network's functionality. Unlike the Hebbian rule, the memory trace is not based on correlation but only on activation itself. It could be viewed as a first-order form of facilitation that drives "bias" units of activation variables, while the Hebbian rule is a second-order form of facilitation that drives connections between inputs and activation variables. Continuous-time versions of Hebbian learning rules analogous to the memory trace used here have been proposed from the earliest days of neural network modeling (Grossberg, 1970). In Chapter 14 we will unify Hebbian and memory-trace learning through a formally analogous dynamics. Learning is covered extensively in Part 3 of this book.

ILLUSTRATION: DYNAMIC FIELD MODEL OF PERSEVERATIVE REACHING

To illustrate how dynamic fields and the associated memory trace can be used to understand elementary forms of embodied cognition, we take you now through an exemplary model, the DFT account for perseverative reaching in the A-not-B task. This example is particularly attractive, because it happens to involve all four basic

instabilities—detection, selection, memory, and boost-driven detection—as well as the dynamics of the memory trace.

The A-not-B task was first developed by Piaget as a measure of infants' understanding of object permanence (Piaget, 1954). In the canonical task, infants watch as an experimenter hides a toy in one of two wells in the top of a box. After a delay, the experimenter pushes the box forward and allows the infant to search for the toy. In the first couple of "A" trials, the toy is hidden in one well, the "A" location, and most infants successfully reach for it. Then the experimenter switches to a "B" trial, hiding the toy in the other well at the "B" location. Young infants who make the A-not-B error reach to the A location on the B trials, despite having just seen the toy hidden at B. This only happens when a delay of a few seconds is imposed between hiding the toy and enabling the infant to reach for it. Around 1 year of age infants stop making the error and search correctly at B on the B trials.

Smith, Thelen, Titzer, and McLin (1999) developed a variant of the A not B task in which, instead of hiding a toy, they simply waved a lid, put it down, and allowed the infant to reach. Infants typically reach for one of the lids, lift it up, and sometimes put it into their mouths. In this version of the task there is no hidden toy. This toyless version of the task is thus simply about how infants decide where to reach when there are two possible targets that afford reaching and grasping.

Thelen, Schöner, Scheier, and Smith (2001) proposed a dynamic field model of the A-not-B task. The motor planning field represents the possible reaching directions and is governed by Equation 2.1, with four sources of input illustrated in Figure 2.14. The evolution of the motor planning field over the course of an A trial is illustrated in Figure 2.15, together with the time courses of three of the sources of input. Task input has two modes, each stimulating movement directions oriented toward the two locations of the two lids or objects. The specific input is centered on the movement direction toward the cued location and is only transiently presented while the cuing occurs. The memory trace reflects the history of activation of the field and preactivates the movement direction of earlier reaches. These inputs are integrated over time in the motor planning field. At the start of the trial, before the cue is provided, only task input and input from the memory trace are present, together not strong enough to generate a self-stabilizing peak, so that the field remains in the subthreshold state. When specific input arrives, it pushes the field through a detection instability. The field generates a peak at the cued location in the motor planning

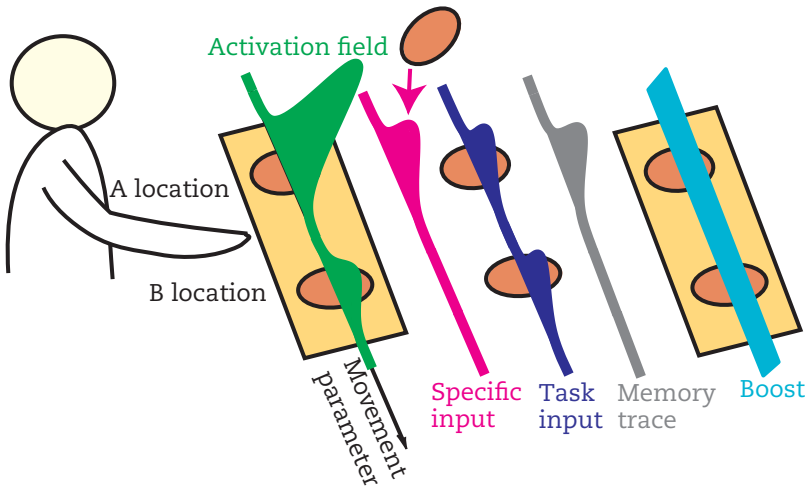

FIGURE 2.14: The A-not-B task entails a baby reaching for one of two objects (here, brown lids) presented on a movable box. The motor plan is represented by an activation field (green) defined over movement direction. A self-stabilized peak, here shown at the A location, drives reaching. Four sources of input to the field are sketched. Specific input arises (red) when attention is drawn to one location, for instance, by waving the object before setting it down on the box (here, at the A location). Task input (violet) reflects the visual layout of the scene, in which the two objects provide input at their respective locations. The memory trace (gray) preactivates field locations at which peaks have previously been induced (here, the A location). The boost (blue) broadly excites all field sites as soon as the box is pushed into the reaching range of the baby.

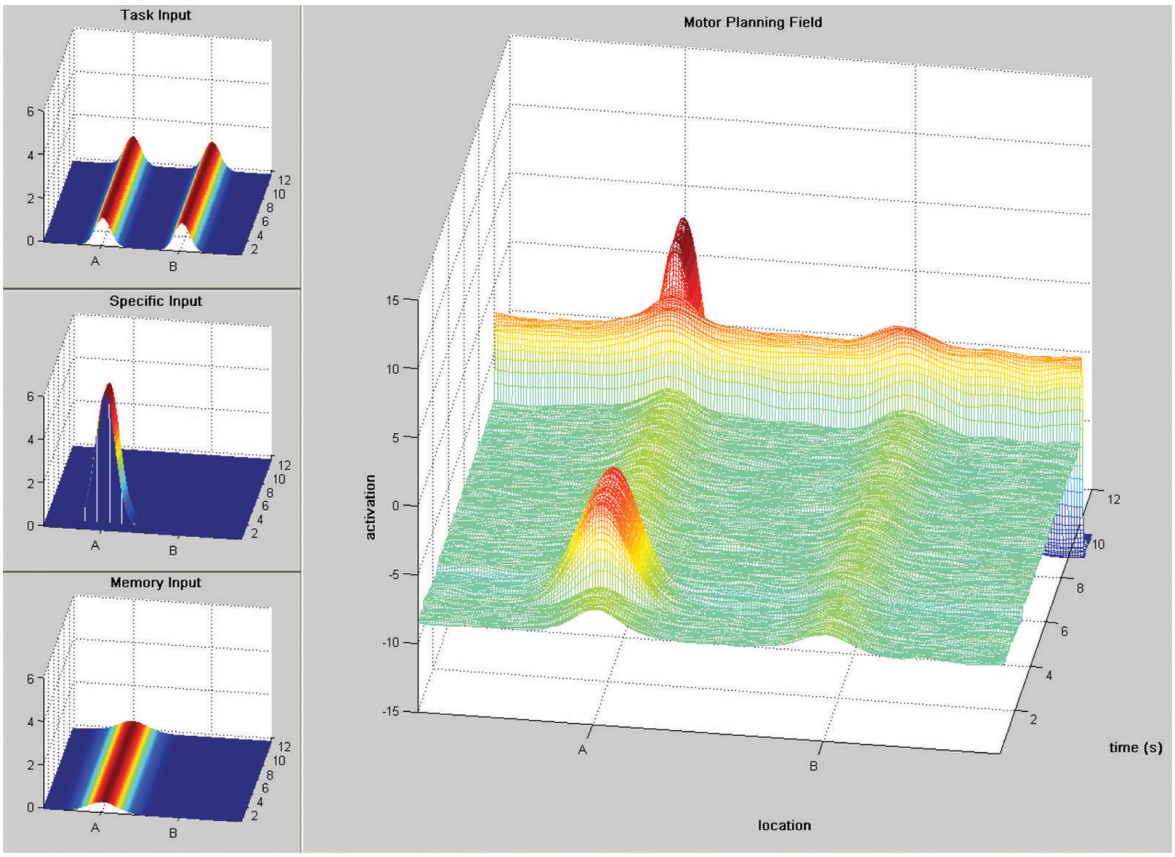

FIGURE 2.15: Time courses of inputs and activation field of the model of perseverative reaching. This is a simulation for an A trial that models the behavior of young infants. The large frame shows the activation field defined over movement direction (horizontal axis) evolving over time (from front to back). Task input (small panel on top left) and input from the memory trace (small panel bottom left) preshape the field at A (left) and B (right) locations. Transient-specific input (small panel middle left) induces a peak early in the trial (peak on the left in front), which decays again after specific input has been removed. The homogeneous boost supplied late in the trial pushes activation up broadly. This induces detection instability and a peak at the A location re-emerges.

field. In the model of the young infants who make perseverative errors, we postulate that interactions in the field are not strong enough to sustain the peak after the specific input ceases at the end of the cueing action. Thus during the delay, the field goes through a reverse detection instability, the peak decays, and the field returns to the subthreshold solution. At the end of the delay, the box is pushed into the reaching space of the infant. We model this by supplying an additive, homogeneous boost to the entire field (Schöner & Dineva, 2007). This moves the field through a boost-driven detection instability, and a peak is generated at the location with the most preactivation, the A location. In other words, the field makes the decision to reach to A.

The first B trial for the model of young infants' behavior is shown in Figure 2.16. At the start of the trial, the memory trace and the task input preshape the field such that there are two subthreshold hills of activation, one centered over each hiding location. The peak at the A location, however, is stronger due to the input from the memory trace that has built up over the A trials. When the specific input stimulates the B location, a self-excited peak is built there, which again decays once specific input ends. When the boost is provided at the end of the delay, the field again generates a peak at the A location, at which preactivation is highest. The model thus makes the A-not-B error.

Thelen and colleagues (2001) modeled development by postulating that older infants had higher resting levels of the motor planning field. A higher resting level (h in Equation 2.1) means that activation can more easily reach the threshold level of the

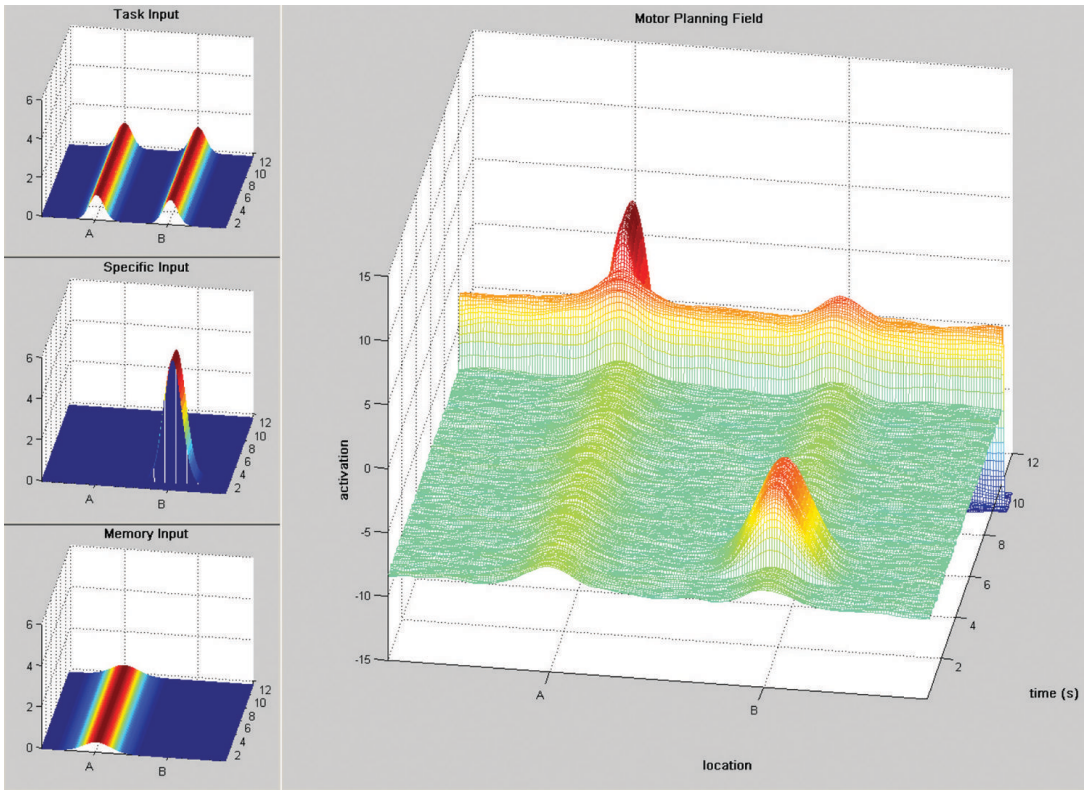

FIGURE 2.16: Time courses of the inputs and activation field of model of perseverative reaching as in Figure 2.15, but now for a B trial of the "young" model.

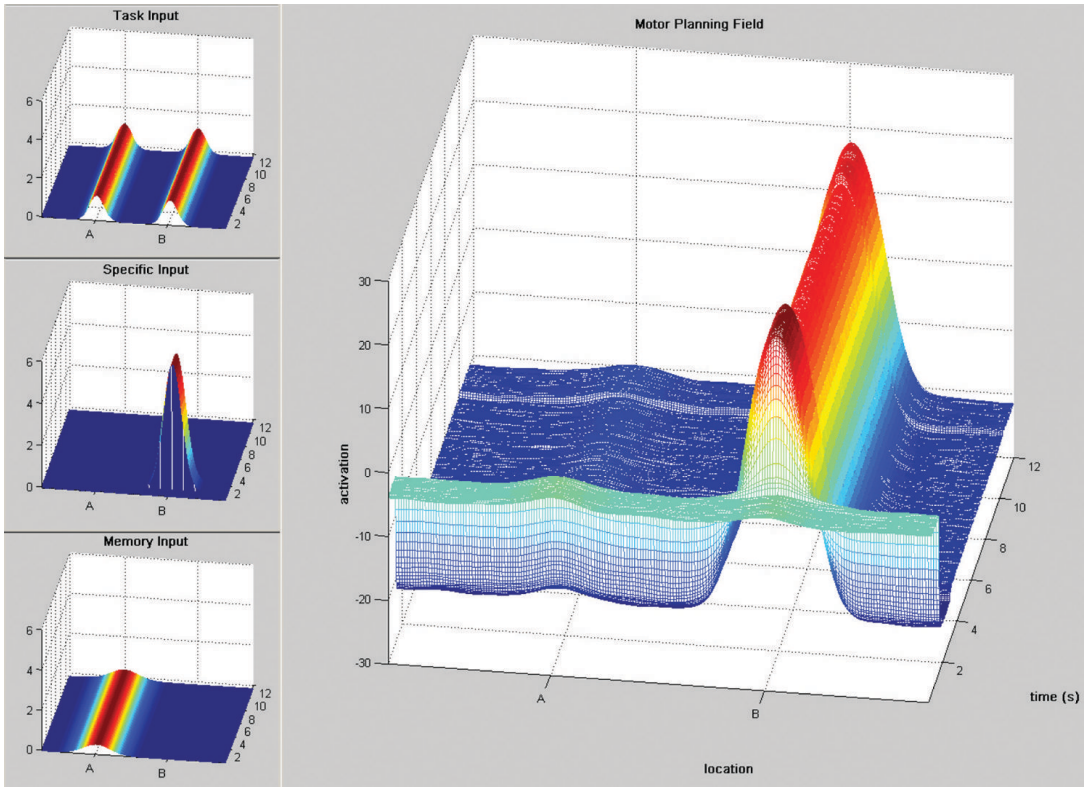

FIGURE 2.17: Time courses of the inputs and activation field of model of perseverative reaching as in Figure 2.15, but now for a B trial of the "old" model.

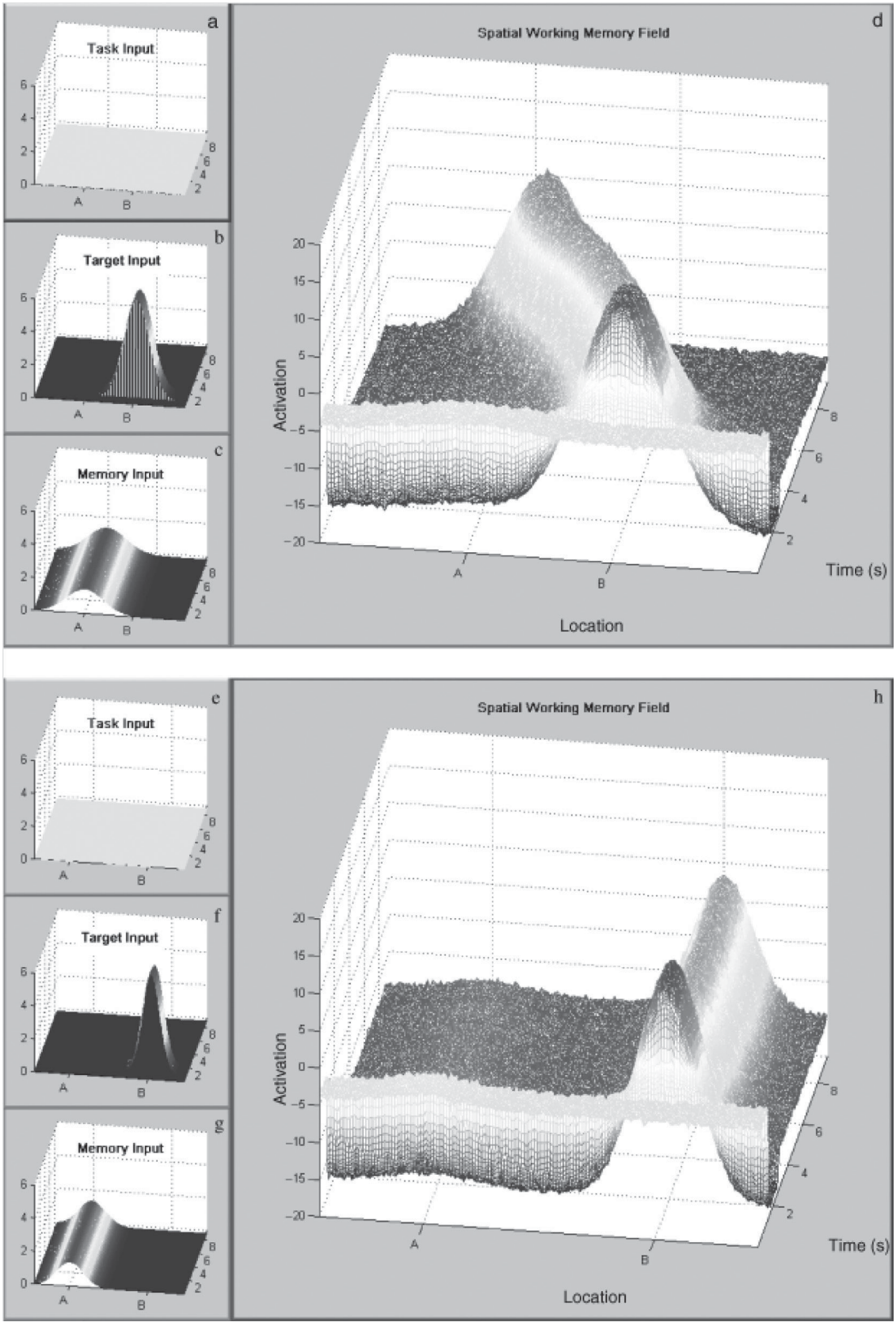

FIGURE 2.18: Time courses of the inputs and activation field of model of the sandbox version of the A-not-B task, using the same conventions as in Figure 2.15. Through the absence of task input in the sandbox (small panel top left in both parts of the figure), the peak is not locked in place. *Top:* A and B locations relatively close to each other. *Bottom:* A and B locations farther removed form each other. Note that the memory trace is a little broader in the top portion of the figure: the drifting peak leaves a broader memory trace.

sigmoid and interaction can be engaged more easily. The shift to higher resting level is thus a shift to stronger interaction and may push the system through the memory instability, beyond which sustained peaks of activation in the absence of localized input become possible. Figure 2.17 shows the first B trial for such an "older" model. At the start of the trial, task input and memory trace preshape the field as before. Specific input at B induces a peak at B through the detection instability. When specific input ends, however, a sustained peak remains at the B location, as the system is now in the regime that enables working memory. When the boost is supplied at the end of the delay, the peak at B is further strengthened and a correct reach to B is implied.

This model has been used to make several predictions that have been tested empirically. One prediction is that spontaneous errors, in which infants reach to B on an A trial, will influence whether or not the infant makes the A-not-B error (Schöner & Dineva, 2007). This prediction probes a core property of DFT. The dynamic field model provides a process account for making the decision to reach to either A or B. A macroscopic neural state is formed when that decision occurs, a peak positioned over either location. This macroscopic neural event leaves a trace—literally, the memory trace—which then in turn may impact future decisions. Thus, in the model, noise may induce a peak to form at the B location rather than the A location on an A trial, inducing a spontaneous error (Dineva, 2005). That peak lays down a memory trace at the B location. This makes it more likely that the spontaneous error will be repeated on later A trials, and it reduces the probability that the infant will make the A-not-B error. On the first B trial, both A and B locations have some preactivation from the respective memory traces there, so that the boost does not necessarily induce a peak at A.

This is in contrast to many connectionist models in which the selection of one out of multiple possible choices is often assumed to occur in a "read-out" process. For instance, an alternative connectionist model of the A-not-B error (Munakata, McClelland, Johnson, & Siegler, 1997) features two neurons that represent the two choices: one neuron standing for reaches to A, the other for reaches to B. The activation levels of the two neurons at the end of the delay are then interpreted as the probabilities with which either reach is realized. A spontaneous error occurs when the less activated neuron is selected, on read out, to determine the outcome of the trial. Clearly, such a decision taken outside the model does not leave a memory trace and thus does not impact future outcomes.

Schutte, Spencer, and Schöner (2003) extended the dynamic field model of perseverative reaching to capture the behavior of older children in an A-not-B sandbox task. In the task, children watch as a toy is buried in a long, narrow sandbox. There is a short delay and then the child searches for the toy. In the first six trials the toy is buried at one location, the A location. In the last three trials it is buried at a second location, the B location. Even the youngest children tested in this task, 18-month-olds, would not make the A-not-B error in the canonical A-not-B task. In the sandbox version, they dig for the toy on a B trial at a location that is strongly shifted toward the A location. Four-year-olds show this metric attraction to A and, under some conditions, even children as old as 6 years show the bias.

An important difference between this task and the canonical A-not-B task is, of course, that no lids mark the hiding locations. Therefore, the location at which children search for the toy is a graded measure of their representation of the planned motor act. At the developmental stage of these children, it is plausible that they are already able to create a working memory of a planned action. The model should, therefore, be in the regime in which it may sustain peaks without localized input. Figure 2.18 shows simulations of the model on the first B trial. There is no task input. Specific input at the B location is transient early in the trial, and input from the memory trace around the A location reflects previous searches. Specific input induces a self-stabilized peak at the B location that is sustained after specific input ends. When the A and B locations are sufficiently close to each other (top of Figure 2.18), the sustained peak at B is affected by input from the memory trace at the A location. That input drives activation up on the side of the peak that overlaps with the A location. This increases activation at the peak so that inhibitory interaction compensates, suppressing the side turned away from the A location more than that turned toward the A location due to the asymmetry of input. The peak is slowly attracted to the A location. This drift induces the metric bias toward the A location, which is a signature characteristic of the A-not-B error. Note that the cause of this form of the A-not-B error is

different from that for the canonical task. Rather than "forgetting" about the cue at the B location, working memory for the motor intention drifts over the delay toward the A location because there is no input at the B location to keep the peak anchored there.

When the A and B locations are placed farther apart (bottom of Figure 2.18), the sustained peak at B does not overlap the memory trace input at A. Preactivation around the A location is suppressed by the inhibition from the peak at B, and that peak remains stationary at the B location. The model does not make an error.

Both signatures are seen in experiments. Young children show strong metric bias, and the bias increases as the delay increases. When the A and B locations are farther apart, metric bias toward A is reduced.

CONCLUSION

This chapter has introduced the core concepts of dynamic field theory: (1) the continuous spaces of possible percepts, possible actions, and possible representations; (2) the time-space continuous activation fields and their neural dynamics; (3) self-stabilized activation peaks as units of representation and the instabilities through which peaks emerge and bring about detection and selection decisions, working memory, and categorization; and (4) the dynamics of the memory trace as the simplest form of learning. In the next chapter we will show how DFT is firmly grounded in neurophysiology—essentially, by capturing the dynamics of population activity in the higher nervous system.

That the units of representation in DFT are stable states is of central importance to DFT. In Chapter 4, the last chapter in this first part of the book, about the foundations of DFT, we will see how the stability of activation peaks enables the linking of representations to sensory and motor processes and thus supports the embodiment of cognition. Stability is linked to robustness: When the neural dynamics of an activation field changes, for instance, through coupling to other parts of a larger neural architecture, stable peak solutions resist change. This makes it possible for dynamic fields to retain their dynamic regime, enabling detection, selection, and working memory, even as they are coupled to neural architectures. This will be a theme in Part 2 of the book. Stability is also critical for learning. In this chapter we showed how instabilities of the subthreshold states of dynamic fields can amplify small inputs or in homogeneities in the field into full, self-stabilized peaks. This changes what learning processes need to achieve. They need to nudge neural processes to self-stabilize new representations, rather than learn such representations completely. This theme will be important in Part 3 of the book.

REFERENCES

Amari, D. S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics, 27,* 77–87.

Anstis, S. M., & Ramachandran, V. S. (1987). Visual inertia in apparent motion. *Vision Research, 27,* 755–764.

Dineva, E. (2005). *Dynamical Field Theory of Infants Reaching and its Dependence on Behavioral History and Context*. Doctoral Dissertation, International Graduate School in Neuroscience, Ruhr-Universität Bochum, Germany.

Erlhagen, W., Bastian, A., Jancke, D., Riehle, A., & Schöner, G. (1999). The distribution of neuronal population activation (DPA) as a tool to study interaction and integration in cortical representations. *Journal of Neuroscience Methods, 94,* 53–66.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological review, 109*(3), 545–572.

Favilla, M. (1997). Reaching movements: Concurrency of continuous and discrete programming. *Neuroreport, 8,* 3973–3977.

Fuster, J. M. (2005). *Cortex and Mind—Unifiying Cognition*. Oxford University Press.

Fuster, J. M., & Alexander, G. E. (1971). Neuron Activity Related to Short-Term Memory. *Science, 173,* 652–654.

Georgopoulos, A. P. (1986). On reaching. *Annual Reviews of Neuroscience, 9,* 147–170.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neural population coding of movement direction. *Science, 233,* 1416–1419.

Ghez, C., Favilla, M., Ghilardi, M. F., Gordon, J., Bermejo, R., & Pullman, S. (1997). Discrete and continuous planning of hand movements and isometric force trajectories. *Experimental Brain Research, 115,* 217–233.

Giese, M. A. (1999). *Dynamic neural field theory of motion perception*. Dordrecht: Kluwer Academic Publishers.

Goldstone, R. L., & Hendrickson, A. T. (2009). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science,* 1(1), 69–78.

Grossberg, S. (1970). Some networks than can learn, remember, and reproduce any number of

complicated space-time patterns, II. *Studies in Applied Mathematics, XLIX*(2), 135–166.

Haykin, S. O. (2008). *Neural networks and learning machines* (3rd ed.). Upper Saddle Brook, NJ: Prentice Hall.

Hock, H. S., Kelso, J. A. S., & Schöner, G. (1993). Perceptual stability in the perceptual organization of apparent motion patterns. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 63–80.

Hock, H. S., & Schöner, G. (2010). A neural basis for perceptual dynamics. In V. Jirsa & R. Huys (Eds.), Nonlinear Dynamics in Human Behavior (151–177). Berlin: Springer-Verlag.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology, 45*, 188–196.

James, W. (1899). *Principles of psychology* (Vol. I). New York: Henry Holt.

Kim, J., & Wilson, H. R. (1993). Dependence of plaid motion coherence on component grating directions. *Vision Research, 33*, 2479–2489.

Koenderink, J., & van Doorn, A. (2003). Shape and shading. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 1090–1105). Cambridge, MA: MIT Press.

Kopecz, K., & Schöner, G. (1995). Saccadic motor planning by integrating visual information and pre-information on neural, dynamic fields. *Biological Cybernetics, 73*, 49–60.

Munakata, Y., McClelland, J. L., Johnson, M. H., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review, 104*, 686–719.

Piaget, J. (1954). *The construction of reality in the child*. New York: Basic Books.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics, 13*, 253–260.

Schall, J. D. (2004). On building a bridge between brain and behavior. *Annual Reviews of Psychology, 55*, 23–50.

Schöner, G., & Dineva, E. (2007). Dynamic instabilities as mechanisms for emergence. *Developmental Science, 10*(1), 69–74.

Schutte, A. R., Spencer, J. P., & Schöner, G. (2003). Testing the dynamic field theory: Working memory for locations becomes more spatially precise over development. *Child Development, 74*(5), 1393–1417.

Simons, D. J. (2000). Current approaches to change blindness. *Visual Cognition, 7*(1-3), 1–15.

Smith, L. B., Thelen, E., Titzer, R., & McLin, D. (1999). Knowing in the context of acting: The task dynamics of the A-not-B error. *Psychological Review, 106*(2), 235–260.

Swindale, N. V. (2000). How many maps are there in visual cortex? *Cerebral Cortex, 10*(7), 633–643.

Thelen, E., Schöner, G., Scheier, C., & Smith, L. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Brain and Behavioral Sciences, 24*, 1–33.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews. Neuroscience, 7*(6), 464–476.

EXERCISES FOR CHAPTER 2

The interactive simulator `launcherOneLayerField_preset` solves numerically the dynamic field Equation 2.1 with added random noise, repeated here in full detail:

$$\begin{aligned}\tau\dot{u}(x,t) &= -u(x,t)+h+s(x,t)\\ &\quad +\int k(x-x')g\big(u(x',t)\big)dx'\\ &\quad +q\xi(x,t)\end{aligned} \tag{A2.1}$$

where the sigmoidal function is given by

$$g(u)=\frac{1}{1+\exp(-\beta u)}. \tag{A2.2}$$

The interaction kernel is given by

$$\begin{aligned}k(x-x') &= \frac{c_{\text{exc}}}{\sqrt{2\pi}\sigma_{\text{exc}}}\exp\left[-\frac{(x-x')^2}{2\sigma_{\text{exc}}^2}\right]\\ &\quad -\frac{c_{\text{inh}}}{\sqrt{2\pi}\sigma_{\text{inh}}}\exp\left[-\frac{(x-x')^2}{2\sigma_{\text{inh}}^2}\right]\\ &\quad -c_{\text{glob}}.\end{aligned} \tag{A2.3}$$

Note that in this formulation of the kernel, the amplitudes of the two Gaussian components are normalized, such that a change in the interaction widths σ does not change the total strength of the interaction. Localized input is supplied in the form

$$s(x,t)=\sum_i a_i\exp\left[-\frac{(x-p_i)^2}{2w_i^2}\right]. \tag{A2.4}$$

Sliders at the bottom of the graphical user interface (GUI) provided by the program enable one to control the widths, w_{si}, locations, p_{si}, and amplitudes, a_{si}, of three such inputs ($i=1,2,3$). Sliders are also available to vary the parameters h, q, c_{exc}, c_{inh}, and c_{glob}. Additional parameters can be accessed via the Parameters button. Predefined sets of parameter values can be loaded by clicking on the pop-up

menu on the bottom right of the GUI, highlighting the appropriate choice, and then clicking the Select button.

The state of the field is shown in the top set of axes in the GUI. The blue line shows the current distribution of activation, $u(x, t)$. The green line is the input shifted by the resting level, $h + s(x, t)$, and the red line shows the field output (sigmoidal function of the field activation) at each position, $g(u(x, t))$, scaled up by a factor of 10 for better visibility. In the bottom set of axes, the shape of the interaction kernel is displayed. Note that the kernel is plotted over distances in the feature dimension, with zero at the center of the plot. This interaction pattern is then applied homogenously for all positions in the field.

The goal of this exercise is to explore and reproduce the instabilities discussed in the chapter.

Exercise 1: Detection Instability

This exercise works best with the predefined parameter set "stabilized." Start out with the field in the resting state (the default) and introduce a localized input by increasing one of the stimulus amplitudes. For small input strengths, observe how the field (blue line) tracks the changing input (green line); this is the subthreshold solution. When activation first reaches zero from below, the field output at that location rises (red line). Observe how at this point very small changes in input strength lead to a new solution, the self-stabilized peak, which has more activation at its peak than input (blue line exceeds green line).

a) Show that, up to the detection instability, the system is bistable, by lowering input again to a level at which you previously saw the subthreshold solution. You can reset the field to the initial condition by pressing the `Reset` button. You will find that from the resting level the field converges to the subthreshold solution again.
b) While a self-stabilized peak stands in the field, move the inducing input laterally with the slider that changes the location of the input function. If you do this slowly enough, the peak will track input. If you do this too fast, the peak disappears at the old location in a reverse detection instability and reappears at the new location in a detection instability.
c) After having induced a peak again by increasing localized input, observe the reverse detection instability by lowering the input strength gradually. Close to where activation reaches zero from above you may observe the collapse of the self-stabilized peak and a quick relaxation to the subthreshold solution.

Exercise 2: Memory Instability

Vary the resting level, h, increasing it step-wise. At each level, induce a peak as in the first exercise and then try to destabilize it through the reverse detection instability by returning localized input strength to zero. At a critical value of the resting level, you will find that the peak decays slowly, then not at all after you have returned the localized input strength to zero. This is the memory instability, leading to a regime in which peaks can be sustained without localized input.

a) You can load a convenient parameter set within the memory regime by selecting the predefined parameter set "memory." Induce a peak, remove localized input, then reintroduce this input in a location close to the sustained peak. In which way is the peak updated?
b) Do the same, but now reintroduce input at a location far from the sustained peak. What happens?

Exercise 3: Selection

Choose the predefined parameter set "selection." Provide two localized inputs by increasing two stimulus amplitudes to intermediate values (between 6 and 8). Observe how only the location first receiving input develops a peak.

a) Increase input strength at the second location until you observe the selection instability.
b) Return that input strength to the original values. Show that the system is bistable.
c) Do the symmetric exercise, increasing input strength at the first location.
d) Adjust two input strengths to be exactly the same, making sure that there is some random noise in the field ($q > 0$). Use the `Reset` button to restart the field from the

resting level. Observe how one of the two locations with input is selected. Repeat several times and convince yourself that selection is stochastic.

Exercise 4: Boost-Induced Detection

Supply small subthreshold input at three locations that is not sufficient to induce peaks. Then slowly increase the resting level until a detection instability is triggered somewhere in the field. Observe how a peak is generated at one of the three locations that have small input. Try to see how small you can make that localized input and still observe the peak at one of the three locations. You can do this with or without noise.