Malinow, R., Madison, D. V., & Tsien, R. W. (1988). Persistent protein kinase activity underlying long-term potentiation. *Nature, 335,* 820–824.

Manabe, T., & Nicoll, R. A. (1994). Long-term potentiation: Evidence against an increase in transmitter release probability in the CA1 region of the hippocampus. *Science, 265,* 1888–1893.

Mathews, C. K., & van Holden, K. E. (1990). *Biochemistry.* Redwood City, CA: Benjamin/Cummings.

Mayer, M. L., MacDermott, A. B., Westbrook, G. L., Smith, S. J., & Barker, J. L. (1987). Agonist- and voltage-gated calcium entry in cultured mouse spinal cord neurons under voltage clamp measured using arsenazo III. *J. Neurosci., 7,* 3230–3244.

Miller, K. D. (1994). A model for the development of simple cell receptive fields and the ordered arrangement of orientation columns through activity-dependent competition between ON- and OFF-center inputs. *J. Neurosci., 14,* 409–441.

Morris, R. G., Anderson, E., Lynch, G. S., & Baudry, M. (1986). Selective impairment of learning and blockade of long-term potentiation by an *N*-methyl-*D*-aspartate receptor antagonist, AP5. *Nature, 319,* 774–776.

Mulkey, R. M., Herron, C. E., & Malenka, R. C. (1993). An essential role for protein phosphatases in hippocampal long-term depression. *Science, 261,* 1051–1055.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Math. Biol., 15,* 267–273.

Oliet, S. H. R., Malenka, R. C., & Nicoll, R. A. (1996). Bidirectional control of quantal size by synaptic activity in the hippocampus. *Science, 271,* 1294–1297.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C.* Cambridge: Cambridge University Press.

Sejnowski, T. J. (1977). Storing covariance with nonlinearly interacting neurons. *J. Math. Biol., 4,* 303–321.

Stanton, P. K., & Sejnowski, T. J. (1989). Associative long-term depression in the hippocampus induced by Hebbian covariance. *Nature, 339,* 531–533.

Stent, G. S. (1973). A physiological mechanism for Hebb's postulate of learning. *Proc. Natl. Acad. Sci. USA, 70,* 997–1001.

Stevens, C. F., & Wang, Y. (1994). Changes in reliability of synaptic function as a mechanism for plasticity. *Nature, 371,* 704–707.

Stryker, M. P., & Harris, W. A. (1986). Binocular impulse blockade prevents the formation of ocular dominance columns in cat visual cortex. *J. Neurosci., 6,* 2117–2133.

Takeuchi, A. (1958). The long-lasting depression in neuromuscular transmission of frog. *Jpn. J. Physiol., 8,* 102–113.

Thies, R. E. (1965). Neuromuscular depression and the apparent depletion of transmitter in mammalian muscle. *J. Neurophysiol., 28,* 427–442.

von der Malsburg, C. (1973). Self organization of orientation selective cells in the striate cortex. *Kybernetik, 14,* 85–100.

Walsh, C. T. (1977). *Enzymatic reaction mechanisms.* San Francisco: W. H. Freeman.

Walton, M. K., Schaffner, A. E., & Barker, J. L. (1993). Sodium channels, $GABA_A$ receptors, and glutamate receptors develop sequentially on embryonic rat spinal cord cells. *J. Neurosci., 13,* 2068–2084.

Wexler, E. M., & Stanton, P. K. (1993). Priming of homosynaptic long-term depression in hippocampus by previous synaptic activity. *NeuroReport, 4,* 590–594.

Wiesel, T. N., & Hubel, D. H. (1963). Single-cell responses in striate cortex of kittens deprived of vision in one eye. *J. Neurophysiol., 26,* 1003–1017.

Xie, X., Berger, T. W., & Barrionuevo, G. (1992). Isolated NMDA receptor-mediated synaptic responses express both LTP and LTD. *J. Neurophysiol., 67,* 1009–1013.

Yang, X.-D., & Farber, D. S. (1991). Initial synaptic efficacy influences induction and expression of long-term changes in transmission. *Proc. Natl. Acad. Sci. USA, 88,* 4299–4303.

Received October 22, 1996; accepted June 21, 1997.

NOTE Communicated by Richard Cone

Axon Guidance: Stretching Gradients to the Limit

Geoffrey J. Goodhill
Georgetown Institute for Cognitive and Computational Sciences, Georgetown University Medical Center, Washington, DC 20007, U.S.A.

Herwig Baier
Department of Biology, University of California, San Diego, La Jolla, CA 92093-0366, U.S.A.

Neuronal growth cones, the sensory-motile structures at the tips of developing axons, navigate to their targets over distances that can be many times greater than their diameter. They may accomplish this impressive task by following spatial gradients of axon guidance molecules in their environment (Bonhoeffer & Gierer, 1984; Tessier-Lavigne & Placzek, 1991; Baier & Bonhoeffer, 1994). We calculate the optimal shape of a gradient and the distance over which it can be detected by a growth cone for two competing mechanistic models of axon guidance. The results are surprisingly simple: Regardless of the mechanism, the maximum distance is about 1 cm. Since gradients and growth cones have coevolved, we suggest that the shape of the gradient in situ will predict the mechanism of gradient detection. In addition, we show that the experimentally determined dissociation constants for receptor-ligand complexes implicated in axon guidance are about optimal with respect to maximizing guidance distance. The relevance of these results to the retinotectal system is discussed.

1 Introduction

The mechanisms that guide axons to appropriate targets in the developing brain are largely unknown. A popular notion, first suggested by Cajal, is that spatial gradients of axon guidance molecules are detected by the growth cone and provide directional information. Experimental evidence for the existence of such mechanisms is gradually mounting. However, so far there has been little consideration of the theoretical limits on axon guidance by gradients imposed by physical limits on the detection of a concentration difference across a small sensing device. Here, using a few pieces of experimental data and some simple approximations, we address these limits.

For a growth cone to be guided by a gradient, it must be able to sense a sufficiently large difference in ligand concentration over its length. The ligand may be attractive or repellent, and may be substrate bound, freely

diffusing, or a combination of both. Two possible mechanisms for gradient detection by a growth cone are (1) internal amplification of a small percentage change in external concentration across the width w of the growth cone (Bonhoeffer & Gierer, 1984; Gierer, 1987), and (2) a shifting internal baseline that reduces the effective concentration at one edge of the growth cone to zero (Walter, Allsop, & Bonhoeffer, 1990). Gradient detection by the former mechanism requires a sufficiently high percentage change p in concentration over distance w, while the latter requires a sufficiently high absolute concentration difference ΔC over w. Three additional constraints limit gradient detection. First, the local external concentration must be less than a critical value C_{high}, at which most receptors are saturated. Second, it must be greater than a critical value C_{low}, at which an insufficient number of receptors are bound to overcome noise. C_{low} and C_{high} vary relative to the dissociation constant k_d for the receptor-ligand complex. Third, the local concentration must also be greater than a physical limit C_{noise}, which is k_d independent. At this concentration, the number of ligand molecules in the vicinity of the growth cone is so small that over the time scales of relevance to the growth cone, thermally induced fluctuations wash out the gradient signal (Tranquillo & Lauffenburger, 1987).

2 Maximum Guidance Distance

What is the maximum range r_{max} for which guidance is possible for the two mechanisms above? The optimal gradient for case 1 has a constant fractional change across the width of the growth cone w for all positions: an exponential gradient. Consider $C(r) = C_0 e^{-ar}$ where C is concentration, r is distance, and C_0 and a are constants. Requiring a percentage change of p ($= \Delta C/C$) across distance w yields $a = p/w$. The maximum distance for which $C \geq C_{low}$ is achieved when $C_0 = C_{high}$. This gives

$$r_{max} = \frac{w}{p} \log_e \frac{C_{high}}{C_{low}}. \tag{2.1}$$

The optimal gradient for case 2 has a constant absolute concentration change across the width of the growth cone: a linear gradient. Consider $C(r) = C_0 - ar$. Requiring a concentration change of ΔC over distance w yields $a = \Delta C/w$. Again the optimal value of C_0 is C_{high}. For the analogous case of leukocyte chemotaxis, it is known that sensitivity to gradients is optimized when the external concentration is equal to the dissociation constant k_d of the relevant receptor (Devreotes & Zigmond, 1988), which yields $\Delta C = pk_d$. This gives

$$r_{max} = \frac{w}{p} \frac{C_{high} - C_{low}}{k_d}. \tag{2.2}$$

What are plausible parameter values? We assume a growth cone diameter w (including filopodia) of 20 μm. Direct evidence (Baier & Bonhoeffer, 1992), analogous data for leukocyte chemotaxis (Devreotes & Zigmond, 1988), and theoretical considerations (Tranquillo & Lauffenburger, 1987) suggest that p is about 2 percent. Data for leukocyte chemotaxis suggest that $C_{low} \approx k_d/100$ and $C_{high} \approx 10k_d$ (the asymmetry is due to down-regulation of receptors at high external concentrations) (Zigmond, 1981). Assuming $C_{low} > C_{noise}$ yields $r_{max} \approx 0.7$ cm for the exponential case (see equation 2.1) and $r_{max} \approx 1$ cm for the linear case (see equation 2.2). Note that these values scale linearly with the size of the growth cone and do not depend on k_d. The calculation assumes that the growth cone can detect $p = 2$ percent for $C_{high} \geq C \geq C_{low}$, whereas in fact it is likely that p needs to be much larger away from $C = k_d$. Correcting for this would reduce r_{max} in both cases. Similarly if growth cones employ a combination of the two mechanisms, r_{max} would again be reduced: 1 cm is thus an upper bound.

Three obvious scenarios for how axons could be guided over larger distances are as follows. First, there could exist a series of spaced gradients of different ligands, each binding to the same or different receptors and guiding the growth cone over only a portion of the full distance. Second, there could exist overlaid gradients of different ligands, each competing for occupancy of the same receptor. Appropriate differences in affinity would allow guidance in multiple regions. Third, there could exist multiple receptors on the growth cone for the same ligand, with different affinities. Each would guide the growth cone over the segment of the gradient lying within its appropriate concentration range. Note that these considerations apply to attractant as well as repellent guidance molecules, or to combinations of both.

3 Noise Limits to Receptor-Ligand Affinity

To maximize guidance distance, it is clearly necessary to choose $C_{low} > C_{noise}$. An accurate calculation of C_{noise} requires knowledge of parameters such as the length of time over which an axon integrates signals from its receptors before assessing a gradient value, which has not been measured. Here instead a conservative order of magnitude estimate for C_{noise} is made. We assume, as an extremely rough estimate, that 100 molecules in the vicinity of the growth cone are sufficient for a 2 percent gradient to be detected. This means that the growth cone can distinguish 50 molecules on one side from 51 on the other. Imagine that the growth cone plus filopodia occupies a cube of side length 20 μm; this has a volume of approximately 10^{-11} liters. One hundred ligand molecules in this volume correspond to a ligand concentration $C_L \approx 0.01$ nM (note that the proportion of the cube occupied by the body of the growth cone, and thus unavailable to the ligand molecules, is small). Equating this with the lower limit due to the dynamics of receptor binding, $C_{low} = k_d/100$, yields $k_d \approx 1$ nM. We suggest that a k_d of very

roughly 1 nM represents a lower limit for axon guidance receptor-ligand complexes.[1] A receptor-ligand affinity significantly higher than this (i.e., $k_d \ll 1$ nM) would not improve the accuracy of gradient reading. A significantly lower affinity would require comparatively large amounts of factor to be produced. An alternative reverse-engineering argument based on the same principle is that the k_d of the receptor-ligand complex could predict the actual signal-to-noise requirements of gradient reading.

4 Applications to the Retinotectal System

Two recently identified repellent axon guidance molecules are believed to be involved in the formation of the retinotectal projection: ephrin-A5 (Drescher et al., 1995) and ephrin-A2 (Cheng, Nakamoto, Bergemann, & Flanagan, 1995; Nakamoto et al., 1996). Both are expressed as gradients in the chick optic tectum, and both bind to one family of receptors, some members of which are expressed on retinal growth cones (for review, see Friedman & O'Leary, 1996a). The ephrin-A2 gradient spans the entire tectum, while the ephrin-A5 gradient is shifted posteriorly in the tectum, being absent from the anterior tectum (where retinal axons enter) (Cheng et al., 1995; Drescher et al., 1995; Nakamoto et al., 1996). k_d values have recently been measured in vitro for ephrin-A5 and ephrin-A2 for three growth cone receptors: EphA3, EphA5, and EphA4. These values are as follows, for ephrin-A5 and ephrin-A2, respectively: EphA3: 0.144 nM/0.86 nM; EphA5: 0.616 nM/8.62 nM; EphA4: 0.622 nM/12.7 nM (Monschau et al., 1997). In each case, the value for ephrin-A2 is roughly an order of magnitude higher than that for ephrin-A5.

The chick optic tectum extends over 6–9 mm during formation of the retinotectal map. The distance that the farthest projecting retinal growth cones have to travel across its (bent) surface is well over 1 cm. Our calculations predict that if retinal axons are guided within the tectum solely by gradient mechanisms, then some method for extending guidance must be operating.[2] We suggest that retinal growth cones could use the same receptor(s) for both ephrin-A2 and ephrin-A5, with the low-affinity ephrin-A2 gradient providing guidance in the anterior tectum, the high-affinity ephrin-A5 gradient providing guidance in the posterior tectum, and a combination of both gradients providing guidance in the middle. In addition,

[1] This calculation applies to both substrate-bound and freely diffusing ligands and also analogously to the sensing of a gradient on a two-dimensional surface.

[2] The situation is apparently more involved: only the nasal-most retinal axons traverse the entire tectum. The more temporal the axons' site of origin in the retina, the farther anteriorly they terminate in the tectum. This graded response to tectal cues, such as ephrin-A5, is possibly reflected by a gradient of receptor level, such as EphA3 (Drescher et al., 1995), in the retina. However, temporal axons are able to navigate to their appropriate tectal target if misrouted or surgically displaced, suggesting that they can utilize gradient information in tectal regions that they normally do not encounter.

the affinity values for ephrin-A5 and ephrin-A2 given above are all within an order of magnitude of our theoretical lower limit of 1 nM, which is reasonable agreement given the crudeness of our calculation. (However, these are in vitro measurements, which may differ from values in vivo.)

5 Regulation of Gradient Shape

An unresolved issue of both biological and theoretical interest is how gradient shape can be regulated in an embryonic field (Crick, 1970). Some axon guidance molecules, like netrin-1 (Kennedy et al., 1994; Serafini et al., 1994), are diffusible factors that are secreted by target cells (Tessier-Lavigne & Placzek, 1991; Kennedy et al., 1994). Simple diffusion yields gradients that are inefficient when growth cones have to traverse distances greater than 1 mm (Tessier-Lavigne & Placzek, 1991; Goodhill, 1997). Binding of the factor to the substrate (e.g., the extracellular matrix) could modify the shape of the gradient to maximize the distance and optimize the accuracy of guidance. The positional information conferred by the gradients of ephrin-A2 and ephrin-A5 in the tectum is initially set up by gradients of morphogens (Crick, 1970) and by transcription factors such as en-1 or en-2 (Itasaki & Nakamura, 1996; Logan et al., 1996; Friedman & O'Leary, 1996b). The local concentrations of these have to be translated into local concentrations of guidance molecules. The translation mode is unknown, but we expect, given the size constraints discussed here, that nature has made some effort to optimize it.

6 Conclusions

For the two possible mechanisms of gradient detection across the width of the growth cone (measuring a fractional change versus a difference from an adjustable baseline), the maximum guidance distance is surprisingly similar (0.7–1.0 cm). However, the shape of the optimal gradient is different in the two cases (exponential versus linear). Therefore, it should be possible to predict the actual gradient-reading mechanism by accurately measuring the shape of gradients of axon guidance protein in situ. Our result also has important implications for the scalability of axon guidance mechanisms to animals substantially larger than the rats and chickens that are most commonly studied.

Acknowledgments

We thank Uwe Drescher, Friedrich Bonhoeffer, and colleagues for sharing results prior to publication. G. J. G. thanks Dennis D. M. O'Leary for very helpful discussions. H. B. thanks Bill Harris and Christine Holt for helpful discussions and support. H. B. was funded by the Alexander von Humboldt Foundation.

References

Baier, H., & Bonhoeffer, F. (1992). Axon guidance by gradients of a target-derived component. *Science, 255,* 472–475.

Baier, H., & Bonhoeffer, F. (1994). Attractive axon guidance molecules. *Science, 265,* 1541–1542.

Bonhoeffer, F., & Gierer, A. (1984). How do retinal axons find their target on the tectum? *Trends Neurosci., 7,* 378–381.

Cheng, H. J., Nakamoto, M., Bergemann, A. D., & Flanagan, J. G. (1995). Complementary gradients in expression and binding of Elf-1 and Mek4 in development of the topographic retinotectal projection map. *Cell, 82,* 371–381.

Crick, F. H. C. (1970). Diffusion in embryogenesis. *Nature, 255,* 420–422.

Devreotes, P. N., & Zigmond, S. H. (1988). Chemotaxis in eukaryotic cells: A focus on leukocytes and *Dictyostelium. Ann. Rev. Cell. Biol., 4,* 649–686.

Drescher, U., Kremoser, C., Handwerker, C., Loschinger, J., Noda, M., & Bonhoeffer, F. (1995). In-vitro guidance of retinal ganglion-cell axons by RAGS, a 25 KDa tectal protein related to ligands for Eph receptor tyrosine kinases. *Cell, 82,* 359–370.

Friedman, G. C., & O'Leary, D. D. M. (1996a). Eph receptor tyrosine kinases and their ligands in neural development. *Curr. Opin. Neurobiol., 6,* 127–133.

Friedman, G. C., & O'Leary, D. D. M. (1996b). Retroviral misexpression of engrailed genes in the chick optic tectum perturbs the topographic targeting of retinal axons. *J. Neurosci., 16,* 5498–5509.

Gierer, A. (1987). Directional cues for growing axons forming the retinotectal projection. *Development, 101,* 479–489.

Goodhill, G. J. (1997). Diffusion in axon guidance. *European Journal of Neuroscience, 9,* 1414–1421.

Itasaki, N., & Nakamura, H. (1996). A role for gradient en expression in positional specification on the optic tectum. *Neuron, 1,* 55–62.

Kennedy, T. E., Serafini, T., de al Torre, J. R., & Tessier-Lavigne, M. (1994). Netrins are diffusible chemotropic factors for commissural axons in the embryonic spinal cord. *Cell, 78,* 425–435.

Logan, C., Wizenmann, A., Drescher, U., Monschau, B., Bonhoeffer, F., & Lumsden, A. (1996). Rostral optic tectum acquires caudal characteristics following ectopic engrailed expression. *Current Biology, 6,* 1006–1014.

Monschau, B., Kremoser, C., Ohta, K., Tanaka, H., Kaneko, T., Yamada, T., Handwerker, C., Hornbereger, M. R., Löschinger, J., Pasquale, E. B., Siever, D. A., Verderame, M. F., Müller, B., Bonhoeffer, F., & Drescher, U. (1997). Shared and distinct functions of RAGS and ELF-1 in guiding retinal axons. *EMBO Journal, 16,* 1258–1267.

Nakamoto, M., Cheng, H. J., Friedman, G. C., Mclaughlin, T., Hansen, M. J., Yoon, C. H., O'Leary, D. D. M., & Flanagan, J. G. (1996). Topographically specific effects of ELF-1 on retinal axon guidance in-vitro and retinal axon mapping in-vivo. *Cell, 86,* 755–766.

Serafini, T., Kennedy, T. E., Galko, M. J., Mirzayan, C., Jessell, T. M., & Tessier-Lavigne, M. (1994). The netrins define a family of axon outgrowth-promoting

proteins homologous to C. Elegans UNC-6. *Cell, 78,* 409–424.
Tessier-Lavigne, M., & Placzek, M. (1991). Target attraction—are developing axons guided by chemotropism? *Trends Neurosci., 14,* 303–310.
Tranquillo, R. T., & Lauffenburger, D. A. (1987). Stochastic-model of leukocyte chemosensory movement. *J. Math. Biol., 25,* 229–262.
Walter, J., Allsop, T. E., & Bonhoeffer, F. (1990). A common denominator of growth cone guidance and collapse? *Trends Neurosci., 11,* 447–452.
Zigmond, S. H. (1981). Consequences of chemotactic peptide receptor modulation for leukocyte orientation. *J. Cell. Biol., 88,* 644–647.

Received February 14, 1997; accepted July 7, 1997.

NOTE Communicated by Read Montague

Equivalence of a Sprouting-and-Retraction Model and Correlation-Based Plasticity Models of Neural Development

Kenneth D. Miller
Departments of Physiology and Otolaryngology and Neuroscience Graduate Program, W. M. Keck Center for Integrative Neuroscience, Sloan Center for Theoretical Neurobiology at University of California, San Francisco, CA 94143-0444, U.S.A.

A simple model of correlation-based synaptic plasticity via axonal sprouting and retraction (Elliott, Howarth, & Shadbolt, 1996a) is shown to be equivalent to the class of correlation-based models (Miller, Keller, & Stryker, 1989), although these were formulated in terms of weight modification of anatomically fixed synapses. Both models maximize the same measure of synaptic correlation, subject to certain constraints on connectivity. Thus, the analyses of the correlation-based models suffice to characterize the behavior of the sprouting-and-retraction model. More detailed models are needed for theoretical distinctions to be drawn between plasticity via sprouting and retraction, weight modification, or a combination.

The model of Elliott et al. involves stochastic search through allowed weight patterns for those that improve correlations. That of Miller et al. instead follows dynamical equations that determine continuous changes of the weights that improve correlations. The identity of these two approaches is shown to depend on the use of subtractive constraint enforcement in the models of Miller et al. More generally, to model the idea that neural development acts to maximize some measure of correlation subject to a constraint on the summed synaptic weight, the constraint must be enforced subtractively in a dynamical model.

1 Introduction

Models of activity-dependent, correlation-based mechanisms of neural development (Miller 1990a, 1996a) have, for simplicity, typically used weight modifications of anatomically fixed synapses (but see von der Malsburg, 1979; Fraser & Perkel, 1990; Montague, Gally, & Edelman, 1991; Colbert, Fall, & Levy, 1994; Elliott, Howarth, & Shadbolt, 1996a, b). However, anatomical changes in connectivity—for example, via synaptic sprouting and retraction guided or stabilized by correlation-based rules—also may play important roles in activity-dependent development and learning.

For example, retraction of axons and dendrites plays a role in many forms of activity-dependent development. In ocular dominance plasticity, thala-

mocortical afferents withdraw from regions of cortex that come to be dominated by afferents serving the opposite eye (LeVay, Stryker, & Shatz, 1978). Dendrites of postsynaptic neurons also have been shown to avoid inappropriate ocular dominance columns (Katz & Constantine-Paton, 1988; Katz, Gilbert, & Wiesel, 1989; Kossel, Löwel, & Bolz, 1995). At the neuromuscular junction, activity-dependent competition normally leads to anatomical retraction of the axons of all but one input to each muscle fiber (Purves & Lichtman, 1985). However, retraction of an axon at the neuromuscular junction is preceded by significant loss of its physiological synaptic strength (Colman, Nabekura, & Lichtman, 1997). Thus, these regressive anatomical changes might simply follow physiological change and not play a leading or guiding role in activity-dependent development.

Stronger evidence for a central role for anatomical changes is found in optic tectum, where there is a continuous retinotopic reorganization of the visual map throughout life. This implies that both retinotopic and, when they occur, ocular dominance maps in this system must be continually maintained amid constant anatomical rearrangement of inputs (Constantine-Paton, Cline, & Debski, 1990; Debski, Cline, & Constantine-Paton, 1990). Striking anatomical changes also occur during visual cortical development. During the normal development of ocular dominance columns in cat visual cortex, there is a huge increase in the overall number of synapses in visual cortex (Cragg, 1975). At the same time, the terminal arbors of thalamic afferents to visual cortex increase in density and branching complexity, though not in overall extent (Antonini & Stryker, 1993a). However, in both optic tectum and visual cortex, control of synapse number during development appears to be activity independent (Hayes & Meyer, 1989; Bourgeois, Jastreboff, & Rakic, 1989), suggesting that activity-dependent processes help determine which connections survive but do not influence the overall number of survivors.

Recent results suggest that activity-dependent processes can directly influence axonal sprouting in visual cortex, at least under abnormal conditions causing denervation of a cortical region. In kitten visual cortex, closure of one eye (monocular deprivation) for short times leads to dramatic loss of arborizations of the thalamocortical afferents corresponding to the closed eye (Antonini & Stryker, 1993b). Reverse deprivation (opening the closed eye and closing the open one) following similar periods of deprivation leads to a strong anatomical shift in favor of the originally closed eye (Movshon & Van Sluyters, 1981). Thus, thalamocortical afferents from the originally closed eye appear to show significant axonal sprouting after reverse deprivation, although this sprouting might occur only into regions that have already been denervated by the newly closed eye (Mioche & Singer, 1989). Similarly, following retinal lesions in adult cats, intracortical axonal sprouting appears to occur into the visual cortical region that formerly responded to the lesioned area (Darian-Smith & Gilbert, 1994).

A rapidly accumulating set of evidence suggests a possible role for neu-

rotrophins in activity-dependent synaptic plasticity in visual cortex and hippocampus (e.g., Cabelli, Wohn, & Shatz, 1995; Fiorentini, Berardi, & Maffei, 1995; Kang & Schuman, 1995; Korte et al., 1995; McAllister, Lo, & Katz, 1995; all reviewed in Thoenen, 1995, and Bonhoeffer, 1996). This also suggests a role for anatomical synaptic rearrangements, because in many systems neurotrophins play a strong role in influencing sprouting and retraction of axonal and dendritic branches (e.g., Purves & Lichtman, 1985, McAllister et al., 1995). However, it remains unclear whether neurotrophins play a specific instructional role in activity-dependent plasticity, as opposed to, say, a nonspecific role in regulating overall levels of sprouting. Furthermore, neurotrophins are also involved in weight modification of anatomically fixed synapses (Figurov, Pozzo-Miller, Olafsson, Wang, & Lu, 1996; Kang & Schuman, 1995; Korte et al., 1995).

Finally, many other experiential modifications can lead to anatomical changes in synaptic connectivity in a variety of other neural structures (Weiler, Hawrylak, & Greenough, 1995).

Mechanisms of activity-dependent synaptic sprouting and retraction have not been explicitly included in previous correlation-based models for at least two reasons. First, formulation of the dynamics of such mechanisms has seemed forbidding in the absence of improved experimental characterization. Second, the study of anatomically fixed synapses seems potentially adequate to understand the behavior of more general correlation-based models. For example, if the range of retinotopically allowed axonal exploration in a correlation-based sprouting-and-retraction model is equivalent to the range of initial axonal connections in a similar model using fixed synapses, then both models would explore the same space of possible connections, and both should converge to the same "most-correlated" set of connections within that space.

This article first shows that there is a precise such equivalence between one very simple sprouting-and-retraction model—that of Elliott et al., (1996a)—and previously formulated models using fixed synapses (Miller, 1990a). Then it shows that the use of resource limitations or competitive constraints (Miller & MacKay, 1994) that are linear in the synaptic weights, such as weight normalization, in the sprouting model naturally corresponds to subtractive implementation of such constraints in the fixed-synapse models. More generally, subtractive implementation of linear constraints emerges as a natural result of an energy-minimization viewpoint, whereas it appeared quite arbitrary from the viewpoint of dynamical models.

2 Equivalence of a Sprouting-and-Retraction Model and a Fixed-Synapse Model

Elliott et al. (1996a) consider a sprouting-and-retraction model of ocular dominance development (see Figure 1A). A two-dimensional grid of cortical cells receives synapses from two two-dimensional input layers, one layer

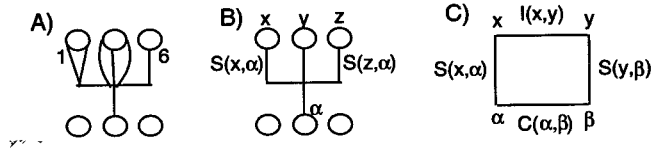


Figure 1: Equivalence of the models of Elliott et al. and Miller et al. For simplicity, the models are illustrated for the case of only a single input type (e.g., a single eye).

(A) Model of Elliott et al. Each synapse is given an individual label. Here, six synapses are shown from the central presynaptic cell to the three postsynaptic cells; these synapses are labeled from 1 to 6 (only the labels 1 and 6 are shown). The activity at each synapse is $\sigma_i \in \{-1, 1\}$, where i is the synapse label. All synapses from the same input cell have the same activity; hence, in this case σ_i is identical for $i = 1, \ldots, 6$. Synaptic connections are made and broken, within retinotopic limits described by an arbor function A, to maximize $\sum_{j,i\in N_j} \sigma_i\sigma_j$, where N_j is the cortical neighborhood of synapse j, defined in the text.

(B) Model of Miller et al. The presynaptic cells are labeled by an index $\alpha, \beta, \ldots$, and the postsynaptic cells are labeled by an index $x, y, \ldots$. There is a single variable, $S(\mathbf{x}, \alpha)$, describing the total synaptic strength of connection between cells α and $\mathbf{x}$; this is equivalent to the number of synapses connecting the two cells in the model of Elliott et al. Thus, given the configuration shown in (A), the equivalent strengths in (B) in the model of Miller et al. would be $S(\mathbf{x}, \alpha) = 2$, $S(\mathbf{y}, \alpha) = 3$, $S(\mathbf{z}, \alpha) = 1$.

(C) The energy maximized by both models. The graphic shows the factors underlying the interaction energy between two weights, $S(\mathbf{x}, \alpha)$ and $S(\mathbf{y}, \beta)$. This energy is the product of the four factors shown: the two weights, the correlation between the two presynaptic cells $C(\alpha, \beta)$, and the intracortical influence between the two postsynaptic locations $I(\mathbf{x}, \mathbf{y})$. Both models maximize an energy (negative of equation 2.2, which is minimized) given by the sum over all weight pairs of this interaction energy.

serving each eye. The input layers represent cells in the lateral geniculate nucleus (LGN). Let each synapse from LGN to cortex be labeled by an index $i, j, \ldots$. Assume that all LGN inputs have two activity states, on or off. Let σ_i be a variable with value 1 if the input to synapse i is on and value -1 if that input is off. Note that multiple synapses come from the same input cell; the σ's for all synapses from a single input are always identical. Let N_j be the cortical neighborhood of synapse j, defined as the set of synapses on the same cortical cell as synapse j or on any of the four adjacent cortical cells. Elliott et al. (1996a) posit that synaptic rearrangements occur so as to minimize the energy $E = -\frac{1}{2}\sum_{j,i\in N_j} \sigma_i\sigma_j$, that is, so as to maximize the correlations between the activities of synapses on the same and neighboring

cortical cells. Arborizations are taken to be localized, so that synapses may be made only within a certain arborization radius r_A of the input cell. Rearrangements occur subject to this and other constraints on connectivity. Input activity patterns are selected from some ensemble. For each input pattern, a small number of randomly chosen allowed rearrangements are tried. Most of the model results were obtained at zero temperature, meaning that only E-reducing rearrangements are accepted; a few results were obtained at finite temperature, meaning that E-increasing rearrangements were accepted probabilistically. This discussion focuses on the zero-temperature case.

To see the equivalence to previous fixed-synapse models, let Greek letters $\alpha, \beta, \ldots$ represent retinotopic position in the input layers, and Roman letters $\mathbf{x}, \mathbf{y}, \ldots$ position in the cortical layer. Let $S^I(\mathbf{x}, \alpha)$ represent the number of synapses from the LGN neuron at position α in input layer I to the cortical cell at $\mathbf{x}$. Let $\sigma^I(\alpha)$ be the activity (± 1) of the neuron at position α in input layer I. Define an intracortical function $I(\mathbf{x}, \mathbf{y})$ to be 1 if $\mathbf{x} = \mathbf{y}$ or if $\mathbf{x}$ and $\mathbf{y}$ are cortical neighbors, and 0 otherwise. Then the energy $E = -\frac{1}{2}\sum_{j,i\in N_j} \sigma_i\sigma_j$ can be rewritten as

$$E = -\frac{1}{2}\sum_{\mathbf{x},\mathbf{y},\alpha,\beta,I,J} I(\mathbf{x}, \mathbf{y})\left[S^I(\mathbf{x}, \alpha)\sigma^I(\alpha)\right]\left[S^J(\mathbf{y}, \beta)\sigma^J(\beta)\right]. \tag{2.1}$$

Assuming that the number of synaptic rearrangements per input pattern is small, so that LGN activity patterns may be averaged over, this energy can be rewritten as (see Figure 1C)

$$E = -\frac{1}{2}\sum_{\mathbf{x},\mathbf{y},\alpha,\beta,I,J} S^I(\mathbf{x}, \alpha)I(\mathbf{x}, \mathbf{y})C^{IJ}(\alpha, \beta)S^J(\mathbf{y}, \beta), \tag{2.2}$$

where $C^{IJ}(\alpha, \beta) = \langle\sigma^I(\alpha)\sigma^J(\beta)\rangle$, the angle brackets signify average over input patterns, and the notation E has been retained for $\langle E\rangle$. Define the arbor function by $A(\mathbf{x}, \alpha) = 1$, $|\mathbf{x} - \alpha| \le r_A$; $A(\mathbf{x}, \alpha) = 0$, otherwise. Then the localization of arborizations is taken into account by minimizing E subject to $S^I(\mathbf{x}, \alpha) = 0$ whenever $A(\mathbf{x}, \alpha) = 0$.

The correlation-based models studied in Miller et al. (1989) and Miller (1990a, 1990b, 1994) begin from equations that describe (1) simple Hebbian or other correlation-based plasticity of anatomically fixed synapses and (2) cortical activity as a function of the input activity pattern. These equations are combined and averaged over the ensemble of input activation patterns, to arrive at equations for the development of synaptic connectivity of the form

$$\frac{d}{dt}S^I(\mathbf{x}, \alpha) = A(\mathbf{x}, \alpha)\sum_{\mathbf{y},\beta,I,J} I(\mathbf{x}, \mathbf{y})C^{IJ}(\alpha, \beta)S^J(\mathbf{y}, \beta). \tag{2.3}$$

Here, $S^I(\mathbf{x}, \alpha)$ represents the total synaptic strength—the average synaptic strength times the number of synapses—from α in layer I to $\mathbf{x}$. $A(\mathbf{x}, \alpha)$, the arbor function, describes the retinotopically allowed anatomical strength of connections from an input at α to a cortical cell at $\mathbf{x}$. $I(\mathbf{x}, \mathbf{y})$, the intracortical interaction function, describes the influence of synapses at cortical position $\mathbf{y}$ on the growth of simultaneously active synapses at position $\mathbf{x}$, via intracortical connectivity and/or diffusion of trophic or modulatory factors. The correlation functions $C^{IJ}(\alpha, \beta)$ describe the correlation in activity between an input at position β in LGN layer J and one at α in LGN layer I. Assuming that $I(\mathbf{x}, \mathbf{y}) = I(\mathbf{y}, \mathbf{x})$ and $C^{IJ}(\alpha, \beta) = C^{IJ}(\beta, \alpha)$, equation 2.3 can be easily shown to represent gradient descent minimization of the energy E of equation 2.2 in the variables $T^I(\mathbf{x}, \alpha) = S^I(\mathbf{x}, \alpha)/\sqrt{A(\mathbf{x}, \alpha)}$.[1] Again, this energy is minimized subject to $S^I(\mathbf{x}, \alpha) = 0$ whenever $A(\mathbf{x}, \alpha) = 0$.

Thus, an identical energy (equation 2.2; Figure 1C) is minimized by both the sprouting-and-retraction model of Elliott et al. (1996a) and the correlation-based models of Miller et al. (1989), which assume plasticity of anatomically fixed synapses. However, the energies are minimized subject to differing constraints, discussed in the appendix. The constraints imposed in the two models have a similar character, involving upper and/or lower limits on the strengths of individual connections and on the summed presynaptic and/or postsynaptic strength associated with each cell. The choices of constraints do not represent intrinsic differences between sprouting-and-retraction and fixed-synapse models. Either type of model might be run with either set of constraints.

However, a given set of constraints may be dynamically enforced in different, inequivalent ways (Miller & MacKay, 1994). I now address the form of enforcement needed to render a dynamical model equivalent to the model of Elliott et al.

3 Constraint Enforcement and Energy Minimization

Constraints on connectivity are used in developmental models to incorporate several biological facts: that biological development is competitive, so that differences between the inputs' activities, rather than the amounts

[1] Symmetrize equation 2.3 by the transformation $T^I(\mathbf{x}, \alpha) = S^I(\mathbf{x}, \alpha)/\sqrt{A(\mathbf{x}, \alpha)}$, $A(\mathbf{x}, \alpha) \neq 0$; $T^I(\mathbf{x}, \alpha) = 0$, otherwise (Miller, 1990a; Miller & Stryker, 1990; MacKay & Miller, 1990). Note that $S^I(\mathbf{x}, \alpha) = 0$ whenever $A(\mathbf{x}, \alpha) = 0$, so no information is lost by this transformation. With the stated assumptions on C and I, the resulting equation for $T^I(\mathbf{x}, \alpha)$ is symmetric under $(\mathbf{x}, \alpha) \leftrightarrow (\mathbf{y}, \beta)$, and performs gradient descent in the energy

$$E = -\frac{1}{2} \sum_{\mathbf{x},\mathbf{y},\alpha,\beta,I,J} T^I(\mathbf{x}, \alpha)\sqrt{A(\mathbf{x}, \alpha)}I(\mathbf{x}, \mathbf{y})C^{IJ}(\alpha, \beta)\sqrt{A(\mathbf{y}, \beta)}T^J(\mathbf{y}, \beta)$$

Since $\sqrt{A(\mathbf{x}, \alpha)}T^I(\mathbf{x}, \alpha) = S^I(\mathbf{x}, \alpha)$, this E is identical to that of equation 2.2.

where $\tilde{\gamma} = (\gamma_1, \ldots, \gamma_{j-1}, \tilde{\gamma}_j, \gamma_{j+1}, \ldots, \gamma_M)$. With probability $a(\gamma_j, \tilde{\gamma}_j)$ replace γ_j by the new candidate $\tilde{\gamma}_j$. Otherwise leave γ_j unchanged. Use Lemma 2.1 to evaluate $p(D|\gamma, \nu)$.
An alternative Metropolis step which updates each coordinate of γ_j separately, is described in section 3.2. For the probing distribution $g_j(\cdot)$, we use a multivariate normal $N(\gamma_j, c^2 C_\gamma)$ with $c = 0.1$. Conceptually, any probing distribution that is symmetric in its arguments, that is, $g(\tilde{\gamma}_j|\gamma_j) = g(\gamma_j|\tilde{\gamma}_j)$, would imply the desired posterior as stationary distribution of the corresponding Markov chain. For practical implementation, a probing distribution with acceptance rates not too close to zero or one is desirable. For a specialized setup, Gelman, Roberts, and Gilks (1996) showed that acceptance rates of around 25% are optimal. In the examples, we found appropriate values for c by trying a few alternative choices until we achieved acceptance rates in this range.

3. Given current values of (γ, ν), generate new values for β by a draw from the complete conditional $p(\beta|\gamma, \nu, D)$. This is a multivariate normal distribution with moments described in Lemma 2.1.

4. Given current values of (β, γ), replace the hyperparameters by a draw from the respective complete conditional posterior distributions: $p(\mu_\beta|\beta, \sigma_\beta)$ is a normal distribution, $p(\mu_\gamma|\gamma, S_\gamma)$ is multivariate normal, $p(\sigma_\beta^{-2}|\beta, \mu_\beta)$ is a Gamma distribution, $p(S_\gamma^{-1}|\gamma, \mu_\gamma)$ is Wishart, and $p(\sigma^{-2}|\beta, \gamma, y)$ is Gamma, as corresponds to a normal linear model. (See Bernardo & Smith, 1994).

The proof of the convergence of this chain follows from arguments in Tierney (1994). To judge convergence in practice, we rely on both sampled paths of parameters of interest and a convergence diagnostic proposed by Geweke (1992), as illustrated in examples 3 and 4. Once we have an approximate posterior sample $\{\theta_1, \ldots, \theta_k\}$, we may undertake various posterior and predictive tasks as usual. For example, predictive means $f(x) = E(y_{n+1}|x_{n+1} = x, D)$ can be evaluated via

$$\hat{f}(x) = \hat{E}(y_{n+1}|x_{n+1}, D) = \frac{1}{k}\sum_{t=1}^{k} E(y_{N+1}|x_{n+1}, \theta = \theta_t).$$

We illustrate some of these calculations in the examples below.

Example 1: Galaxy Data. We try to relate velocity (y_i) and radial position (x_{i1}) of galaxy NGC7531 at 323 different locations (Buta, 1987). For this example, we use only the first 80 observations. The data are shown in Figure 1. Radial positions are centered and scaled to have zero mean and unit variance, and velocities have been shifted by a constant offset of 1400. A constant

as described in section 3, this leads to a practically useful MCMC scheme for NN analyses.

The key observation in our scheme is that given the currently imputed values of the γ's, we actually have a standard hierarchical normal linear model (Lindley & Smith, 1971; Bernardo & Smith, 1994). On one hand, this will allow us to sample easily from the posterior marginals of the weights β and hyperparameters, given the γ's. On the other hand, this allows us to marginalize model represented in equations 2.2 and 2.3) with respect to β_j, $j = 1, \ldots, M$, to obtain the marginal likelihood $p(D|\gamma, \nu)$. This computation will be instrumental in the Metropolis step (step ii) in our algorithm.

The following lemma provides the marginalised likelihood, where for the sake of simplified notation, we shall omit dependence on the hyperparameters.

Lemma 2.1. *Let* $z_{ij} = z_{ij}(\gamma) = \psi(x_i'\gamma_j)$, $Z = (z_{ij})_{i=1,\ldots,N}^{j=1,\ldots,M}$, $\mathbf{1} = (1)_{i=1,\ldots,M}$, $A = Z'Z/\sigma^2$, $\rho = Z'y/\sigma^2$, $C = 1/\sigma_\beta^2 I$, $\delta = \mu_\beta/\sigma_\beta^2\mathbf{1}$. *Let* $m_b(\gamma) = (A + C)^{-1}(\rho + \delta)$ *and* $S_b(\gamma) = (A + C)^{-1}$. *Then,*

$$p(D|\gamma) = \frac{p[\beta = m_b(\gamma)]}{p[\beta = m_b(\gamma)|y, \gamma]} \prod_{i=1}^{N} p[y_i|\beta = m_b(\gamma), \gamma]$$

$$= p[\beta = m_b(\gamma)]|S_b(\gamma)|^{1/2} \prod_{i=1}^{N} p[y_i|\beta = m_b(\gamma), \gamma].$$

Proof. Conditional on γ, the model in equations 2.2 and 2.3 becomes a normal linear regression model. The posterior $p(\beta|D, \gamma)$ takes the form of a multivariate normal distribution $N[m_b(\gamma), S_b(\gamma)]$, with posterior moments $m_b(\gamma)$ and $S_b(\gamma)$, given, for example, in Bernardo and Smith (1994). By Bayes' theorem, $p(\beta|D, \gamma) = p(\beta) \prod_{i=1}^{N} p(y_i|\beta, \gamma)/p(D|\gamma)$. Substituting $\beta = m_b(\gamma)$ in the last equation, we obtain the expression for $p(D|\gamma)$.

Our hybrid, blocking, partially marginalized MCMC algorithm for inference and prediction with FFNNs is as follows:

1. Start with θ equal to some initial guess (for example, the prior means).

Until convergence is achieved, iterate through steps 2 through 4:

2. Given current values of ν only, (marginalizing over β) replace γ by Metropolis steps: For each γ_j, $j = 1, \ldots, M$, generate a proposal $\tilde{\gamma}_j \sim g_j(\gamma_j)$, with $g_j(\gamma_j)$ described below. Compute

$$a(\gamma_j, \tilde{\gamma}_j) = \min\left[1, \frac{p(D|\tilde{\gamma}, \nu)p(\tilde{\gamma}|\nu)}{p(D|\gamma, \nu)p(\gamma|\nu)}\right], \tag{2.4}$$

of their activities, determine their fate in synaptic competition (Wiesel & Hubel, 1965; Guillery, 1972; Lo & Poo, 1991; Balice-Gordon & Lichtman, 1995); that resources are limited; and that weights are single signed and finite. Because we know virtually nothing of the mechanisms by which biological competition is achieved, there is little biological guidance as to the form such constraints should take or the manner in which they should be implemented (see the discussion in Miller, 1990a, 1996b; Miller & MacKay, 1994).

In the statistical model of Elliott et al., connection patterns are searched for those that decrease the energy. The search consists of a sequence of small, random perturbations of synaptic connectivity. Constraints on connectivity are enforced by limiting the perturbations to those that obey the constraints. For example, to enforce constraints preserving total synaptic strength projected by each afferent, and retinotopically restricting the cortical locations accessible to each afferent, perturbations consist of movement of a randomly chosen synapse to a new cortical location chosen randomly from among those retinotopically accessible to the corresponding afferent.

In the dynamical model of Miller et al., a deterministic time derivative of the weights is computed at each time step. The equations must be constrained to ensure that this time derivative respects the constraints on connectivity. There are many ways to achieve this, which are not equivalent either mathematically or in developmental outcome (Miller & MacKay, 1994). Intuitively, the space of allowed connectivity patterns forms a constraint surface in the space of weights. The unconstrained time derivative may point off of this surface, in which case it must be corrected by return to the constraint surface. But to which point on the constraint surface should the weight pattern be moved? This freedom corresponds to the multiple, inequivalent ways of dynamically constraining the equation.

Equivalence to the model of Elliott et al. is achieved if constraints achieve energy minimization within the constrained space of possible weight patterns. Thus, given an unconstrained energy and a corresponding unconstrained gradient descent dynamics, the constraint implementation must achieve gradient descent in the energy *on the constraint surface*. Mathematically, this is achieved by *perpendicular projection* of the unconstrained gradient descent dynamics onto the constraint surface.[2]

In previous work (Miller & MacKay, 1994), two methods were studied of enforcing constraints limiting the sum of weights over a cell, subtractive and multiplicative. At each time step, after adding the unconstrained

[2] Beginning from a point $\mathbf{x}$, the local change in energy per unit movement in direction $\hat{\mathbf{v}}$, $\hat{\mathbf{v}} \cdot \hat{\mathbf{v}} = 1$ is given by $\nabla E(\mathbf{x}) \cdot \hat{\mathbf{v}}$, where $\nabla E(\mathbf{x})$ is the gradient of the energy at $\mathbf{x}$. Thus, the direction of maximum decrease of the energy along the constraint surface is the direction with maximum dot product with the negative of the gradient vector. This is the direction found by perpendicular projection of the negative of the gradient vector—the derivative vector under gradient descent dynamics—onto the constraint surface.

derivative to the weight vector, subtractive enforcement involves subtracting a fixed amount from each weight to return the weight vector to the constraint surface. Multiplicative enforcement involves multiplying each weight by a fixed amount, and thus subtracting an amount proportional to the weight itself. In the case of a constraint on the sum of the synaptic weights, such as $\sum_{\alpha,I} S^I(\mathbf{x}, \alpha) = S_{\text{post}}$—or, more generally, of a constraint on a linear combination of the synaptic weights, which yields a hyperplane constraint surface—perpendicular projection of the gradient vector corresponds to subtractive constraint enforcement (Miller & MacKay, 1994).[3] For a hypersphere surface, multiplicative enforcement is perpendicular projection. For other surfaces, perpendicular projection cannot be characterized as either multiplicative or subtractive.

If there are multiple constraints, one on each postsynaptic cell (or one on each presynaptic cell), perpendicular projection onto the full, multicell constraint surface is accomplished by perpendicular enforcement of each single-cell constraint.[4] These arguments apply also to restrictions to a constraint region bordered by hyperplanes, such as $S^{\min}_{\text{post}} \le \sum_{\alpha,I} S^I(\mathbf{x}, \alpha) \le S^{\max}_{\text{post}}$. Such a constraint can be imposed dynamically by allowing unconstrained development on the interior of the region and subtractive enforcement of the constraint at each border. This will result in gradient descent minimization of the energy over the allowed region.

The case of simultaneous constraints on both pre- and postsynaptic cells is similar. Perpendicular constraint enforcement of the full set of constraints minimizes energy. Assuming the single-cell constraint surfaces are hyperplanes, perpendicular enforcement corresponds to subtractive enforcement. However, the details of formulating perpendicular constraints in this case

[3] There are some technicalities involving the use of the arbor function, discussed in more detail in Miller and MacKay (1992, App. B). Briefly: one must work in the symmetric T representation discussed in footnote 1, in which the dynamics are gradient descent. There, the subtractive constraint is a perpendicular projection onto the constraint surface, although it generally does not appear perpendicular in the S representation. In the S representation, the constraint conserving $\sum_{\alpha,I} S^I(\mathbf{x}, \alpha)$ is enforced through subtraction of a multiple of $A(\mathbf{x}-\alpha)$ from $\frac{d}{dt} S^I(\mathbf{x}, \alpha)$. Transforming to the T representation, the constraint conserves $\sum_{\alpha,I} \sqrt{A(\mathbf{x}-\alpha)} T^I(\mathbf{x}, \alpha)$ and is enforced through subtraction of a multiple of $\sqrt{A(\mathbf{x}-\alpha)}$. Thus, the constraint is perpendicular in the T representation: it conserves $\mathbf{n}(\mathbf{x}) \cdot \mathbf{T}(\mathbf{x})$, and is enforced through subtraction of a multiple of $\mathbf{n}(\mathbf{x})$, where $\mathbf{n}(\mathbf{x})$ has elements $\sqrt{A(\mathbf{x}-\alpha)}$.

[4] The full constraint surface is the intersection of the constraint surfaces associated with each cell. The normals of the single-cell constraint surfaces are all normal to the full constraint surface and to one another. The sequence of perpendicular projections onto each single-cell constraint surface results in a projection onto the full constraint surface along a linear combination of the single-cell normals, and any such linear combination is perpendicular to the full constraint surface. See Miller and MacKay (1992, App. C) for mathematical formulation of the projection operators that implement constraints on a network of cells.

are somewhat more complicated. These are described briefly in Miller (1997).

Constraints on the strength of individual weights are also linear constraints. Thus, perpendicular projection onto the hyperplane bounding the constrained volume of allowed synaptic weights is achieved by subtracting off the nonallowed weight change from the gradient descent vector. This corresponds to a "hard nonlinearity": dynamics are linear within the constrained volume, but no weight changes are allowed that take a weight beyond its allowed limits. Single-synapse constraints fail to commute with constraints over summed pre- and/or postsynaptic weights. To avoid this conflict, the methods for formulating perpendicular constraints described in Miller (1997) can be used. In practice, more ad hoc methods are generally used to resolve the conflict. While this seems unlikely to alter significantly the developmental outcome from constrained energy minimization, this has not been studied.

In summary, perpendicular constraint enforcement achieves gradient descent on the multicell constraint surface. If constraints are linear in the synaptic strengths, this is achieved by subtractive enforcement. Subtractive enforcement thus realizes the intuitive picture that the biological dynamics are searching, via a correlation-based mechanism, for a "most correlated" set of inputs subject to such linear constraints. Such a search might occur by sprouting and retraction, synaptic strength modification, or some combination. While previous work (Miller & MacKay, 1994) demonstrated the differences in outcome resulting from, and in overall energies minimized by, subtractive versus multiplicative enforcement, the relationship to energy minimization or correlation maximization over the constraint surface was not previously noted.

4 Application of the Equivalence to Understand the Results of Elliott et al.

The main results of the model of Elliott et al. are to show that ocular dominance segregation will occur under their model, that the periodicity of the resulting columns corresponds to the arbor diameter, and that the degree of segregation increases with the distance over which inputs within an eye are correlated.[5] All of these results follow directly from prior quantitative analyses of the model of Miller et al. (which used subtractive constraint enforcement) (Miller et al., 1989; Miller, 1990a).

The occurrence of ocular dominance column formation was shown in

[5] Again, this describes the zero-temperature version of the model. By studying finite temperature, Elliott et al. found evidence for a phase transition that separates a high-temperature (high noise in the weight modification process), disordered regime from a low-temperature, ordered regime. Finite temperature results, of course, also apply to either model; one can construct a stochastic dynamics corresponding to the deterministic dynamics at finite temperature (e.g., van Kampen, 1992, Chap. 9).

the previous analysis to depend on two factors. First, the input activities must be such that, after averaging over activity patterns, two inputs of the same eye are always better correlated in their activities than two inputs of the opposite eye at the same retinotopic separation, at least for small retinotopic separations; and two same-eye inputs must be no worse correlated than two opposite-eye inputs for larger retinotopic separations up to an arbor radius.[6] This condition is easily met by the activity patterns used by Elliott et al.: a circle of activated cells within a single eye, centered at a random position, with all other cells inactive. Second, there must be some locally excitatory interactions between cortical cells, so that the ocular dominance of neighboring cells is coupled. This is achieved by the neighborhood function of Elliott et al., which favors the development of correlated receptive fields on neighboring cortical cells.

The spatial period of the resulting ocular dominance columns was shown previously to depend on the intracortical interaction function and the constraints, if any, on the total strength of projection of each input cell. The intracortical interaction used by Elliott et al. is purely excitatory, and one of their constraints is that the total projection from each presynaptic cell is held constant. For this case, the period of ocular dominance columns was previously shown to correspond to an arbor diameter, as found empirically by Elliott et al.

The degree of segregation was shown in the previous analysis to depend on the correlations in input activities. Two factors are involved. First, when the same-eye correlations are larger than the opposite-eye correlations only over distances small relative to an arbor radius, segregation is weak; but as the distance over which the same-eye correlations are larger than the opposite-eye correlations becomes larger, segregation becomes stronger. Second, anticorrelations between the two eyes increase the degree of segregation, and positive correlations between the two eyes reduce it. Elliott et al. find that increasing the radius of the circle of activated cells used in their activity patterns leads to sharper segregation, and for larger arbor radius, larger activity circles are needed to get a similar degree of segregation. Furthermore, they consider two models, one of which has anticorrelations between the eyes, the other of which has either reduced anticorrelations or positive correlations between the eyes, and find that the latter model requires much wider-ranging within-eye correlations (larger activity discs for a given arbor width) to achieve a similar degree of segregation, relative to the former model.[7] These results are all as predicted by the previous

[6] Mathematically: the assumption is made of symmetry between the eyes, so that $C^{LL} = C^{RR}$ and $C^{LR} = C^{RL}$. Then the condition is that $C^{D}(\alpha) \equiv C^{LL}(\alpha) - C^{LR}(\alpha)$ has the peak of its Fourier transform at frequency 0 or at a frequency that corresponds to a wavelength long relative to the arbor diameter.

[7] In the "relocation" model of Elliott et al., all synaptic changes occur by choosing an active synapse and relocating it; the change is accepted if it decreases the energy (in the

analysis. A quantitative account could be made by computing the precise within- and between-eye correlations determined by their activity patterns and sampling procedures.

Elliott et al. also model monocular deprivation. Because they conserve the total strength projected by each presynaptic arbor, they must model deprivation by hand-setting the strength of each eye's arbors. Thus, their model deals only with the arrangements of these inputs, given a fixed difference in input strength between the eyes. As found previously in the model of Miller et al., the period of the ocular dominance stripes is unchanged, but one eye's stripes become thinner and the other's thicker. In the studies of Miller et al., the final strength of each eye's projection, as well as the stripe layout, emerged dynamically.

5 Discussion

I have shown that a simple, stochastic, correlation-based sprouting-and-retraction model of synaptic development can be understood within the framework of correlation-based developmental models (Miller, 1990a, 1996a), although those models were formulated in terms of modification of anatomically fixed synapses. The various models all descend in an identical energy, corresponding to maximizing a certain measure of synaptic correlations. The analyses of the correlation-based framework accurately predict the outcomes of the sprouting-and-retraction model. The sprouting-and-retraction model emphasizes energy minimization or, equivalently, correlation maximization. This perspective in turn provides a principled reason for the use of subtractive enforcement of linear constraints on connectivity (or for multiplicative enforcement of quadratic constraints).

Thus, the hypothesis in the introduction has been confirmed and sharpened. Models of sprouting and retraction and of modification of anatomically fixed synapses can both explore the same space of possible connections and converge to the same "most-correlated" set of connections within that space. This occurs provided both maximize the same measure of correlation over the constrained set of allowed weight patterns.

The correlation-based framework was developed to allow analysis of the outcome of synaptic competition under a wide group of mechanisms. De-

zero-temperature version). Because all changes are conditional on a synapse's being active, the correlations $C^{IJ}(\alpha, \beta) = \langle \sigma(\alpha)\sigma(\beta) \rangle$ must also be computed conditional on this. Since the opposite eye is always uniformly inactive, the two eyes are completely anticorrelated, $C^{LR}(\alpha, \beta) \equiv -1$ for all α, β. In their "interchange" model, synaptic changes occur by interchanging two synapses, one active and one inactive; again, the change is accepted if it decreases the energy. Because the changes involve an inactive as well as an active synapse, the correlations of an inactive synapse with the inactive neurons in the opposite eye must also be considered in computing the correlations, so the anticorrelations between the eyes will be reduced and/or positive correlations between the eyes may be induced, depending on the sizes of the activity circle and the arbors.

velopment in this framework depends on biologically measurable functions that describe correlations in the thalamic inputs, retinotopic restrictions on arborizations, and intracortical interactions. The dependence on these functions has been characterized through both analysis and simulations (Miller et al., 1989; Miller, 1990a, 1994, 1996a). Different biological mechanisms are distinguished by the different predictions they make for the shape of these functions. For example, the model of Elliott et al. is restricted to positive intracortical interaction functions.

Previously, such correlation-based models were shown to embrace very simple models of a variety of underlying mechanisms, including activity-dependent competition for diffusible modification factors (e.g., neurotrophins) by anatomically fixed synapses (Miller et al., 1989; Miller, 1990a). Sprouting and retraction mechanisms were not previously included because it seemed likely that density-dependent and diffusive terms may arise in simplified models of their dynamics, for example, if there is a tendency for sprouting to occur from more-occupied areas into less-occupied areas; and no such terms were included in the framework studied.[8] Such terms do not arise in the model of Elliott et al. because of their assumption that all synaptic changes allowed by the constraints (see the appendix) and by the relocation or interchange rule (the effects of which can be incorporated in the correlation function; see footnote 7) are tried with equal probability, and are accepted or rejected according to their contribution to the energy. That is, synaptic densities and gradients do not affect sprouting-and-retraction probabilities other than through the energy and the constraints.

One potential problem with the fixed-synapse formalism is that it contains an initial bias in favor of weight patterns that involve uniform connectivity: if weights are initialized as small perturbations around some anatomical arbor function, then a weight pattern similar to the arbor function initially has a much larger size than weight patterns orthogonal to it. This can lead to a bias in favor of the development of such patterns, even if such patterns have less favorable energy than others (e.g., see the discussion of the effect of the DC mode's "head start" in MacKay & Miller, 1990). Although the zero-temperature model of Elliott et al. does not necessarily avoid this problem, a sprouting-and-retraction framework more generally may pro-

[8] The probability for an input from α of type I to sprout to $\mathbf{x}$ may increase with its local innervation density, $\sum_{\mathbf{x}'} f(|\mathbf{x}-\mathbf{x}'|)S^I(\mathbf{x}',\alpha)$ (here, $f(\mathbf{x})$ is some weighting function); it may decrease with total innervation density at $\mathbf{x}$, $\sum_{J,\alpha'} S^J(\mathbf{x},\alpha')$, or with the gradient of this density $\nabla_{\mathbf{x}} \sum_{J,\alpha'} S^J(\mathbf{x},\alpha')$ (if areas that are less innervated, either absolutely or relative to neighboring areas, are more likely to receive sprouts; note, however, that if total synaptic strength is conserved on each postsynaptic cell, this density is constant, and its gradient is zero). The probability might also depend on the input's total projection strength, $\sum_{\mathbf{x}'} S^I(\mathbf{x}',\alpha)$, or on the gradient of this total, $\nabla_{\alpha} \sum_{\mathbf{x}'} S^I(\mathbf{x}',\alpha)$ (e.g., smaller arbors, either in absolute terms or relative to neighboring arbors, might have a better chance of making new sprouts).

vide a natural means of avoiding such bias. The addition of noise, as in the high-temperature version of the Elliott et al. model, may also be of benefit.

It is often stated that Hebbian rules result in a set of weights corresponding to the principal component of the input data: the principal eigenvector of the input covariance matrix. This is based on the simple case of a linear covariance plasticity rule for a single, linearly activated postsynaptic cell. For this case, multiplicative constraint enforcement does lead to the principal component (e.g., Miller & MacKay, 1994), which in this case minimizes the energy of equation 2.2 over any hypersphere in weight space. However, even for this case, on the hyperplane corresponding to a linear weight constraint, the principal eigenvector is not an energy minimum, and in fact it is not in general a "special" point of any kind for the energy. That is, if Hebbian development with linear weight constraints acts to maximize correlations, then even in this simple case, the dynamics do not evolve to the principal component of the data. More complex plasticity or activity rules or network connectivity also ensure that the dynamics do not evolve to the principal component. It is therefore quite incorrect to equate Hebbian development with learning of the principal component.

In previous work (Miller et al., 1989; Miller & MacKay, 1994), it was noted that ocular dominance segregation can emerge under subtractive constraint enforcement even in the presence of partial correlation of the activities of the two eyes (as is presumably induced by vision). Under multiplicative enforcement, in contrast, ocular dominance segregation cannot emerge unless there is anticorrelation between the two eyes. Thus, subtractive enforcement seems to give a better match to the biology, at least in this respect. By connecting subtractive enforcement to correlation maximization over the allowed weight patterns, the work described in this article gives a natural biological grounding to a constraint enforcement method that also seems favored (at least relative to multiplicative enforcement) by the match of developmental outcome to biology.

Because subtractive enforcement can lead weights to saturate at their most extreme allowed values, it has been suggested (Miller & MacKay, 1994) that it may be a poor choice for models of adult plasticity, where continuous plasticity occurs in response to changing input activity distributions (Kaas, 1991). However, we demonstrate in Miller (1997) that this tendency to saturation can be eliminated by combining pre- and postsynaptic constraints. Elliott et al. (1996b) have found that their model, which does not show such saturation (discussed in the appendix), can account for some aspects of adult plasticity. Thus, these aspects can be accounted for by subtractively constrained dynamics. However, in their model, the number of synapses projected by different input classes in response to a given input activity regime is set by hand, and only the arrangement of these synapses is determined by the learning rule. Thus, it remains an open question under what conditions the energy-minimizing models with linear weight constraints discussed here can account for the reversible changes

in projection strength that occur with changes in input activities in adult cortical plasticity. Presumably the key requirement is that the perturbations of activity sufficiently shift the energy landscape on the constraint surface so that the locations of the minima are shifted.

Do biological correlation-based mechanisms rely more on sprouting and stabilization-retraction of synapses of fixed strength or on modification of synaptic strengths? There is little evidence to guide us. The two processes are obviously not mutually exclusive; for example, the decision whether to stabilize or retract newly sprouted synapses may be guided by the strengthening or weakening of their physiological strengths. In kitten visual cortex, the rapidity of monocular deprivation effects, which emerge physiologically within hours (Mioche & Singer, 1989), weakly suggests that physiological changes in synaptic efficacy may precede (and thus perhaps guide) anatomical restructuring of connections. This is also suggested by recent results showing that anatomical changes in thalamocortical afferent arbors are incomplete after 4 days of monocular deprivation (Antonini & Stryker, 1996), whereas physiological effects of monocular deprivation appear complete after just 2 days (Hensch et al., 1995). Similarly, at the neuromuscular junction, anatomical elimination of a synapse is preceded by loss of its synaptic efficacy (Colman et al., 1997), including loss of postsynaptic receptors (Balice-Gordon & Lichtman, 1993, 1995). That weight modification and axonal sprouting and retraction may be coupled is suggested by studies in several systems showing that decrease in the number of synapses onto a postsynaptic cell is associated with increase in strength of the remaining synapses (Herrera & Grinnell, 1980, 1981; Jackson & Parks, 1982; Pockett & Slack, 1982; Liu & Tsien, 1995).

Can modeling help us to distinguish between mechanisms relying on sprouting and retraction, and those relying on modification of synaptic strengths? To the extent to which the two are mathematically equivalent, of course, no theoretical distinctions can be drawn. It is conceivable that practical experimental distinctions could be drawn from simple models to which density-dependent and diffusive terms are added, as described above. However, to draw firm theoretical distinctions, more realistic knowledge of the rules governing synaptic sprouting, retraction, and modification is needed. It remains to be determined how such knowledge will modify the basic understandings of correlation-based development already achieved through study of the simple models discussed here.

Appendix: Constraints in the Two Models

The energy E of equation 2.2 is minimized in the two models subject to the following constraints:

1. Localization of connectivity: $S^I(\mathbf{x}, \alpha) = 0$ whenever $A(\mathbf{x}, \alpha) = 0$, in both models.

2. Upper and lower limits on connection strength
   - Elliott et al.: $0 \leq S^I(\mathbf{x}, \alpha)$ for all $I, \mathbf{x}, \alpha$. No explicit upper limit.
   - Miller et al.: $0 \leq S^I(\mathbf{x}, \alpha) \leq s_{\max} A(\mathbf{x}, \alpha)$ for all $I, \mathbf{x}, \alpha$, where $s_{\max}$ represents the maximum value of a synaptic weight.
3. Presynaptic total connection strength
   - Elliott et al.: The total number of synapses per presynaptic neuron is constant and conserved: $\sum_{\mathbf{x}} S^I(\mathbf{x}, \alpha) = S_{\text{pre}}$ for all I, α and for some constant S_{pre}. To model monocular deprivation, this constant is decreased for inputs from one eye.
   - Miller et al.: The effects of imposing a limitation on the sum over presynaptic weights, such as $\sum_{\mathbf{x}} S^I(\mathbf{x}, \alpha) = S_{\text{pre}}$ or $0.5S_{\text{pre}} \leq \sum_{\mathbf{x}} S^I(\mathbf{x}, \alpha) \leq 1.5S_{\text{pre}}$, are considered, but such a limitation is usually not used.
4. Postsynaptic total connection strength
   - Elliott et al.: The total number of synapses per postsynaptic neuron is constrained to remain within some range: $S_{\text{post}}^{\min} \leq \sum_{\alpha, I} S^I(\mathbf{x}, \alpha) \leq S_{\text{post}}^{\max}$ for all $\mathbf{x}$, where $S_{\text{post}}^{\min}$ (generally 1) and $S_{\text{post}}^{\max}$ are constants. A model in which the number of synapses per cortical cell is constant and conserved—that is, in which $\sum_{\alpha, I} S^I(\mathbf{x}, \alpha) = S_{\text{post}}$ for all $\mathbf{x}$—is also considered.
   - Miller et al.: The total synaptic strength per postsynaptic neuron is constant and conserved: $\sum_{\alpha, I} S^I(\mathbf{x}, \alpha) = S_{\text{post}}$ for all $\mathbf{x}$ and for some constant S_{post}.

The constraints imposed in the two models have a similar character. One notable difference is the presence of an upper limit on the strength of individual connections in the model of Miller et al., and the absence of such a limit in the model of Elliott et al. In Miller and MacKay (1994), it was shown that an upper limit on the strength of individual connections is generally needed under subtractive constraints to attain distributed receptive and/or projective fields (see the discussion in Miller, 1997). Yet Elliott et al., in the absence of an explicit upper weight limit, find apparently stable weight configurations with distributed receptive and projective fields.

There seem to be at least two explanations. First, as discussed in Miller (1997), the combination of pre- and postsynaptic constraints can eliminate the need for an upper weight limit. Second, the particular intracortical interaction used by Elliott et al. may play a role. The self-interaction L_{ii} of a connection $T_i \equiv T^I(\mathbf{x}, \alpha)$ (see footnote 1) with itself under the gradient descent dynamics is $L_{ii} = \sqrt{A(\mathbf{x}, \alpha)} I(\mathbf{x}, \mathbf{x}) C^{II}(\alpha, \alpha) \sqrt{A(\mathbf{x}, \alpha)}$. The interaction L_{ij} between two different connections from the same input cell, $T_i \equiv T^I(\mathbf{x}, \alpha)$ and $T_j \equiv T^I(\mathbf{y}, \alpha)$ is $L_{ij} = \sqrt{A(\mathbf{x}, \alpha)} I(\mathbf{x}, \mathbf{y}) C^{II}(\alpha, \alpha) \sqrt{A(\mathbf{y}, \alpha)}$. Given constraints

on only presynaptic cells, an upper weight limit is required if $L_{ii} > |L_{ij}|$ for all $j \neq i$, where j and i are labels for two weights from a single presynaptic cell (Miller & MacKay, 1994). This condition is not satisfied in the model of Elliott et al. because $I(\mathbf{x}, \mathbf{y}) = 1$ both for $\mathbf{y} = \mathbf{x}$ and for $\mathbf{y}$ the neighbor of $\mathbf{x}$, and A is constant where it is nonzero. Thus, a given input may equally well distribute its weights among a cell and its four nearest neighbors, or concentrate its inputs onto the central cell, without altering the interactions between its connections. Thus, even if only presynaptic constraints are applied, the model of Elliott et al. may not require an upper weight limit to achieve distributed receptive and projective fields.

Acknowledgments

I thank Terry Elliott for stimulating me to think about these issues and for useful discussions, and Michael Crair, Virginia De Sa, Ed Erwin, David MacKay, Michael Silver, and Todd Troyer for helpful comments on the manuscript. This work is supported by NIH grant EY11001 and grants from the Searle Scholar's Program and the Lucille P. Markey Charitable Trust.

References

Antonini, A., & Stryker, M. P. (1993a). Development of individual geniculocortical arbors in cat striate cortex and effects of binocular impulse blockade. *J. Neurosci., 13,* 3549–3573.

Antonini, A., & Stryker, M. P. (1993b). Rapid remodeling of axonal arbors in the visual cortex. *Science, 260,* 1819–1821.

Antonini, A., & Stryker, M. P. (1996). Plasticity of geniculocortical afferents following brief or prolonged monocular occlusion in the cat. *J. Comp. Neurol., 369,* 64–82.

Balice-Gordon, R. J., & Lichtman, J. W. (1993). In vivo observations of pre- and postsynaptic changes during the transition from multiple to single innervation at developing neuromuscular junctions. *J. Neurosci., 13,* 834–855.

Balice-Gordon, R. J., & Lichtman, J. W. (1995). Long-term synapse loss induced by focal blockade of postsynaptic receptors. *Nature, 372,* 519–524.

Bonhoeffer, T. (1996). Neurotrophins and activity-dependent development of the neocortex. *Curr. Opin. Neurobiol., 6,* 119–126.

Bourgeois, J.-P., Jastreboff, P. J., & Rakic, P. (1989). Synaptogenesis in visual cortex of normal and preterm monkeys: Evidence for intrinsic regulation of synaptic overproduction. *Proc. Natl. Acad. Sci. USA, 86,* 4297–4301.

Cabelli, R. J., Hohn, A., & Shatz, C. J. (1995). Inhibition of ocular dominance column formation by infusion of NT-4/5 or BDNF. *Science, 267,* 1662–1666.

Colbert, C. M., Fall, C. P., & Levy, W. B. (1994). Using adaptive synaptogenesis to model the development of ocular dominance in kitten visual cortex. In F. H. Eeckman (Ed.), *Computation in neurons and neural systems* (pp. 139–144). Boston: Kluwer.

Colman, H., Nabekura, J., & Lichtman, J. W. (1997). Alterations in synaptic

strength preceding axon withdrawal. *Science, 275,* 356–361.
Constantine-Paton, M., Cline, H. T., & Debski, E. (1990). Patterned activity, synaptic convergence and the NMDA receptor in developing visual pathways. *Ann. Rev. Neurosci., 13,* 129–154.
Cragg, B. G. (1975). The development of synapses in the visual system of the cat. *J. Comp. Neurol., 160,* 147–166.
Darian-Smith, C., & Gilbert, C. D. (1994). Axonal sprouting accompanies functional reorganization in adult cat striate cortex. *Nature, 368,* 737–740.
Debski, E. A., Cline, H. T., & Constantine-Paton, M. (1990). Activity-dependent tuning and the NMDA receptor. *J. Neurobiol., 21,* 18–32.
Elliott, T., Howarth, C. I., & Shadbolt, N. R. (1996a). Axonal processes and neural plasticity. I: Ocular dominance columns. *Cerebral Cortex, 6,* 781–788.
Elliott, T., Howarth, C. I., & Shadbolt, N. R. (1996b). Axonal processes and neural plasticity. II: Adult somatosensory maps. *Cerebral Cortex, 6,* 789–793.
Figurov, A., Pozzo-Miller, L. D., Olafsson, P., Wang, T., & Lu, B. (1996). Regulation of synaptic responses to high-frequency stimulation and LTP by neurotrophins in the hippocampus. *Nature, 381,* 706–709.
Fiorentini, A., Berardi, N., & Maffei, L. (1995). Nerve growth factor preserves behavioral visual acuity in monocularly deprived kittens. *Vis. Neurosci., 12,* 51–55.
Fraser, S. E., & Perkel, D. H. (1990). Competitive and positional cues in the patterning of nerve connections. *J. Neurobiol., 21,* 51–72.
Guillery, R. W. (1972). Binocular competition in the control of geniculate cell growth. *J. Comp. Neurol., 144,* 117–130.
Hayes, W. P., & Meyer, R. L. (1989). Impulse blockade by intraocular tetrodoxin during optic regeneration in goldfish: HRP-EM evidence that the formation of normal numbers of optic synapses and the elimination of exuberant optic fibers is activity independent. *J. Neurosci., 9,* 1414–1423.
Hensch, T. K., Crair, M. C., Ruthazer, E. S., Fagiolini, M., Gillespie, D. C., & Stryker, M. P. (1995). Robust two-day ocular dominance plasticity revealed by single-unit recording and intrinsic signal imaging of kitten area 17. *Soc. Neuro. Abs., 21,* 2023.
Herrera, A., & Grinnell, A. (1980). Transmitter release from frog motor nerve terminals depends on motor unit size. *Nature, 287,* 649–651.
Herrera, A., & Grinnell, A. (1981). Contralateral denervation causes enhanced transmitter release from frog motor nerve terminals. *Nature, 291,* 495–497.
Jackson, H., & Parks, T. (1982). Functional synapse elimination in the developing avian cochlear nucleus with simultaneous reduction in cochlear nerve axon branching. *J. Neurosci., 2,* 1736–1743.
Kaas, J. H. (1991). Plasticity of sensory and motor maps in adult mammals. *Ann. Rev. Neurosci., 14,* 137–167.
Kang, H., & Schuman, E. M. (1995). Long-lasting neurotrophin-induced enhancement of synaptic transmission in the adult hippocampus. *Science, 267,* 1658–1662.
Katz, L. C., & Constantine-Paton, M. (1988). Relationships between segregated afferents and postsynaptic neurons in the optic tectum of three-eyed frogs. *J. Neurosci., 8,* 3160–3180.

Katz, L. C., Gilbert, C. D., & Wiesel, T. N. (1989). Local circuits and ocular dominance columns in monkey striate cortex. *J. Neurosci., 9,* 1389–1399.

Korte, M., Carroll, P., Wolf, E., Brem, G., Thoenen, H., & Bonhoeffer, T. (1995). Hippocampal long-term potentiation is impaired in mice lacking brain-derived neurotrophic factor. *Proc. Natl. Acad. Sci. USA, 92,* 8856–8860.

Kossel, A., Löwel, S., & Bolz, J. (1995). Relationships between dendritic fields and functional architecture in striate cortex of normal and visually deprived cats. *J. Neurosci., 15,* 3913–3926.

LeVay, S., Stryker, M. P., & Shatz, C. J. (1978). Ocular dominance columns and their development in layer IV of the cat's visual cortex: A quantitative study. *J. Comp. Neurol., 179,* 223–244.

Liu, G., & Tsien, R. W. (1995). Properties of synaptic transmission at single hippocampal synaptic boutons. *Nature, 375,* 404–408.

Lo, Y.-J., & Poo, M.-M. (1991). Activity-dependent synaptic competition in vitro: Heterosynaptic suppression of developing synapses. *Science, 254,* 1019–1022.

MacKay, D. J. C., & Miller, K. D. (1990). Analysis of Linsker's applications of Hebbian rules to linear networks. *Network, 1,* 257–298.

McAllister, A. K., Lo, D. C., & Katz, L. C. (1995). Neurotrophins regulate dendritic growth in developing visual cortex. *Neuron, 15,* 791–803.

Miller, K. D. (1990a). Correlation-based models of neural development. In M. A. Gluck & D. E. Rumelhart (Eds.), *Neuroscience and connectionist theory* (pp. 267–353). Hillsdale, NJ: Erlbaum.

Miller, K. D. (1990b). Derivation of linear Hebbian equations from a nonlinear Hebbian model of synaptic plasticity. *Neural Comput., 2,* 321–333.

Miller, K. D. (1994). A model for the development of simple cell receptive fields and the ordered arrangement of orientation columns through activity-dependent competition between ON- and OFF-center inputs. *J. Neurosci., 14,* 409–441.

Miller, K. D. (1996a). Receptive fields and maps in the visual cortex: Models of ocular dominance and orientation columns. In E. Domany, J. L. van Hemmen, & K. Schulten (Eds.), *Models of neural networks III* (pp. 55–78). New York: Springer-Verlag. Available as ftp://ftp.keck.ucsf.edu/pub/ken/miller95.ps.

Miller, K. D. (1996b). Synaptic economics: Competition and cooperation in synaptic plasticity. *Neuron, 17,* 371–374.

Miller, K. D. (1997). Simultaneous constraints on pre- and post-synaptic cells: Effects on synaptic saturation [On-line]. Available as ftp://ftp.keck.ucsf.edu/pub/ken/pre_post.ps.gz.

Miller, K. D., Keller, J. B., & Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. *Science, 245,* 605–615.

Miller, K. D., & MacKay, D. J. C. (1992). *The role of constraints in Hebbian learning* (Tech. Rep. Memo 19). Pasadena, CA: California Institute of Technology. Available as ftp://ftp.keck.ucsf.edu/pub/ken/constraints-tr.ps.

Miller, K. D., & MacKay, D. J. C. (1994). The role of constraints in Hebbian learning. *Neural Comput., 6,* 100–126.

Miller, K. D., & Stryker, M. P. (1990). The development of ocular dominance columns: Mechanisms and models. In S. J. Hanson & C. R. Olson (Eds.), *Connectionist modeling and brain function: The developing interface* (pp. 255–350).

Cambridge, MA: MIT Press/Bradford.
Mioche, L., & Singer, W. (1989). Chronic recordings from single sites of kitten striate cortex during experience-dependent modifications of receptive-field properties. *J. Neurophysiology, 62,* 185–197.
Montague, P. R., Gally, J. A., & Edelman, G. M. (1991). Spatial signaling in the development and function of neural connections. *Cereb. Cortex, 1,* 199–220.
Movshon, J. A., & Van Sluyters, R. C. (1981). Visual neural development. *Ann. Rev. Psych., 32,* 477–522.
Pockett, S., & Slack, J. (1982). Pruning of axonal trees results in increased efficacy of surviving nerve terminals. *Brain Res., 243,* 350–353.
Purves, D., & Lichtman, J. W. (1985). *Principles of neural development.* Sunderland, MA: Sinauer Associates.
Thoenen, H. (1995). Neurotrophins and neuronal plasticity. *Science, 270,* 593–598.
van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry* (2nd Ed.). Amsterdam: North-Holland.
von der Malsburg, C. (1979). Development of ocularity domains and growth behavior of axon terminals. *Biol. Cyber., 32,* 49–62.
Weiler, I. J., Hawrylak, N., & Greenough, W. T. (1995). Morphogenesis in memory formation: synaptic and cellular mechanisms. *Behav. Brain Res., 66,* 1–6.
Wiesel, T. N., & Hubel, D. H. (1965). Comparison of the effects of unilateral and bilateral eye closure on cortical unit responses in kittens. *J. Neurophysiol., 28,* 1029–1040.

Received January 31, 1996; accepted June 12, 1997.

NOTE Communicated by Read Montague

Axonal Processes and Neural Plasticity: A Reply

T. Elliott
C. I. Howarth
N. R. Shadbolt
Department of Psychology, University of Nottingham, Nottingham, NG7 2RD, U.K.

We examine the claim that a class of sprouting-and-retraction models is mathematically equivalent to a fixed-anatomy model. We accept, subject to important caveats, a narrow mathematical equivalence of the energy functions in both classes of model. We argue that this narrow equivalence of energy functions does not, however, entail equivalence of the models. Indeed, the claim of complete model equivalence hides significant dynamical differences between the approaches, which we discuss. We also disagree that our work demonstrates that subtractive constraint enforcement is natural in fixed-anatomy models.

Miller (see "Equivalence of a Sprouting-and-Retraction Model," elsewhere in this issue) has claimed that (1) a sprouting-and-retraction framework for modeling neural plasticity (Elliott, Howarth, & Shadbolt, 1996a, 1996b) is mathematically equivalent to a fixed-anatomy model (Miller, Keller, & Stryker, 1989); (2) that this equivalence shows that subtractive normalization is natural in fixed-anatomy models; and (3) that no theoretical distinction can be drawn between sprouting-and-retraction models and fixed-anatomy models until more elaborate models are available. We briefly examine these claims.

First, modulo important caveats, we agree that our framework minimizes, subject to similar constraints, the same energy function as Miller's model. The caveats are that (1) the biological interpretation of our constraints is different, (2) our constraints do not enforce competition, and (3) competition emerges because of the range of the variables σ_i used in our framework, where letters such as i and j index axonal processes and $\sigma_i \in \{-1, 1\}$ represents the activity of axonal process i.

The first caveat is important because there is little biological support for synaptic normalization. Our constraints represent biologically plausible bounds on the minimum and maximum numbers of axonal processes supported by neurons and are not introduced as a mathematical device to enforce competition in the absence of more detailed knowledge concerning the biological mechanisms underlying competition (von der Malsburg, 1973).

The second caveat requires a little explanation. Our "relocation" and "interchange" models, in which sprouting and retraction are always cou-

pled, automatically include constraints. However, another model based on our framework, the "sprouting-and-retraction" model, uncouples sprouting and retraction, and thus no constraints are enforced, either explicitly or implicitly. Perturbations in network connectivity represent the creation of a new connection or the destruction of an old one. This third model segregates competing afferents, and in modeling anomalous peripheral activity, no hand setting of parameters is necessary (Elliott, Howarth, & Shadbolt, 1996c); in contrast, Miller adjusts afferent normalization parameters.

As for the third caveat, setting $\sigma_i = -1$ rather than $\sigma_i = 0$ to represent afferent inactivity is enough to induce competition. This is easily demonstrated by eliminating the contribution to the energy function that raises the energy when pairs of axonal processes have activities such that $\sigma_i + \sigma_j = 0$: afferent segregation then breaks down in the sprouting-and-retraction model (Elliott et al., 1996c). Curiously, elimination of this contribution from the energy function in the relocation and interchange models does not lead to a breakdown of afferent segregation, but this is not because these two models employ constraints. Rather, it is straightforward to see that the relocation and interchange rules lead automatically to the existence of a statistical force of attraction between axonal processes whose activities are positively correlated and a statistical force of repulsion between axonal processes whose activities are negatively correlated (see Elliott, Howarth, & Shadbolt, 1997, for an analysis). These noncompetitive forces are by themselves enough to induce afferent segregation, even in the absence of the competitive contribution to the energy function. Indeed, it is precisely these forces, particularly the force of repulsion, that lead to the emergence of "cortical gaps" in the relocation model (Elliott et al., 1996a). As to the justification for taking $\sigma_i = -1$ rather than $\sigma_i = 0$ to represent afferent inactivity, see the appendix to this article.

Does this equivalence of the energy functions (modulo caveats) mean that the models are equivalent? Our approach represents a set of biological assumptions and computational techniques in addition to the energy function. This is best seen by comparing the behaviors of the relocation and interchange models. Both models exhibit different results when annealed and when quenched. This is because the dynamics of minimization are different in both models. Thus, even though both minimize the same energy function, the models are not equivalent; the final outcomes are not identical. The sprouting-and-retraction model also is not equivalent to the relocation and interchange models. For example, although the energy function is identical in all three of our models, no statistical forces exist in the sprouting-and-retraction model (Elliott et al., 1997). Also, in contrast to Miller's model (Miller, 1990; Miller & MacKay, 1994), the sprouting-and-retraction model does not exhibit an organizationally determined critical period and so is able to accommodate adult plasticity in, for example, the somatosensory

cortex.[1] This significant difference between our framework and Miller's model arises because the dynamics of minimization are different. Thus, the models are different, even though the energy functions are equivalent (modulo caveats).

It may be argued that the minimization dynamics of a model are to some extent arbitrary. This is true of models that have no clear biological interpretation or for which the minimization procedure is simply a computational algorithm for finding a minimum. But our minimization procedure is constructed as an explicit model of neurons sprouting into regions of high neurotrophic support and retracting from regions of low neurotrophic support, with temperature as some measure of the noise affecting an otherwise orderly process (Elliott et al., 1996a). The dynamics of minimization in our three models represent different assumptions regarding how neurons might maximize their neurotrophic support. We do not therefore accept that the dynamics of minimization are less important than what is being minimized. Indeed, the different dynamics of our three models lead to demonstrably different behaviors and final outcomes, even though the energy functions are identical.

Second, we do not accept that our work shows that subtractive normalization is natural in fixed-anatomy models. Synaptic normalization is typically imagined to result from the decay of synaptic weights, although there is little reason to believe that global normalization will result from local decay. The decay rate will almost certainly be a nonconstant function of the concentration(s) of the decaying substance(s). Biologically, it is difficult to motivate subtractive normalization, since it assumes that the decay rate is independent of the concentration. Computationally, we agree that subtractive normalization is natural, since it leads to steepest-descent minimization. Thus, there is a tension between these two requirements, so the claim that subtractive normalization is natural in fixed-anatomy models entails that such models are not in fact biologically plausible.

A related point is that it is difficult biologically to motivate simultaneous afferent and efferent normalization in fixed-anatomy models. Because normalization is typically imagined to represent a limitation on some resource, afferent (efferent) normalization represents a limitation on a presynaptic (postsynaptic) resource. However, a synaptic weight will be some function

[1] To switch off plasticity in layer IV of the visual cortex, we invoke biochemical factors (e.g., Kasamatsu, 1983; Fox, Sato, & Daw, 1989; Gu, Liu, & Cynader, 1994), where these are controlled, presumably, by the overall level of electrical activity in the visual pathways rather than its specific pattern (e.g., Cynader, Berman, & Hein, 1976; Cynader, 1983). An organizationally determined critical period, such as that in Miller's model, could be refuted by rearing kittens under blockade of retinal activity and with direct, simultaneous stimulation of both optic nerves (e.g., Stryker & Strickland, 1984; Weliky & Katz, 1997) so that ocular dominance columns do not form, and then testing for a response to monocular deprivation to see whether the critical period has been extended. Our prediction is that it would not be.

of both pre- and postsynaptic resources. Hence, normalizing weights over both afferents and efferents is biologically problematic: afferent (efferent) normalization should be over the afferent (efferent) resource only. Thus, any attempt to suggest that simultaneously rather than consecutively implemented afferent and subtractive efferent normalization will permit adult plasticity in fixed-anatomy models runs afoul of the same biological-versus-computational dilemma raised above.

Finally, even were the models mathematically equivalent, there can be no justification for the claim that no theoretical distinction can be drawn between sprouting-and-retraction models and fixed-anatomy models. Biological models are characterized by their underlying biological assumptions, interpretations, and evidence, in addition to their mathematical structure. To emphasize the latter at the expense of the former, given that the former often legitimate procedures and techniques in one approach but not in a mathematically equivalent alternative, risks significantly limiting the power of modeling and theoretical science. Theoretical distinctions are mostly, although not always, meaningless only when experimental discrimination is impossible. Anatomical plasticity is plainly experimentally distinguishable from physiological plasticity. For a comprehensive examination of the central role of anatomical plasticity in the development of the nervous system, and an attack on the view that neural development is entirely regressive, as implicit in models such as Miller's, see, for example, Bailey and Kandel (1993), Purves (1994), and Quartz and Sejnowski (in press).

Appendix

Here we justify the use of $\sigma_i = -1$ rather than $\sigma_i = 0$ to represent afferent inactivity. To do this, first we derive a general expression for an energy function; then we manipulate it so as to extract the $\sigma_i = -1$ convention.

Let the activity of afferent process i be denoted by $a_i \in \{0, 1\}$, that is, inactivity is represented by $a_i = 0$. Notice that the σ_i variables are related to the a_i variables through $\sigma_i = 2a_i - 1$. The total input to the cell on which process i synapses is taken to be $\sum_j \mathcal{D}_{ij} a_j$, where $\mathcal{D}_{ij} = 1$ if, and only if, processes i and j synapse on the same cell, and is zero otherwise. The level of neurotrophic factor (NTF) released by the cell on which process i synapses is taken as $R_i = f_R(\sum_j \mathcal{D}_{ij} a_j)$, where f_R is some model of the overall production and release process. The NTF released is taken to diffuse rapidly through the target field, with the amount available at each target cell following diffusion assumed to be $A_i = \sum_j \tilde{\Delta}_{ij} R_j$, where $\tilde{\Delta}_{ij} = \Delta_{ij}/s_j$ with $s_j = \sum_k \mathcal{D}_{jk}$ being the total number of processes synapsing on the cell on which process j synapses. The function Δ_{ij} characterizes the diffusion process and is assumed to be appropriately normalized; we have previously used only the nearest–neighbor function. The level of NTF available at each synapse on the cell on which process i synapses is then assumed to be $\tilde{A}_i = A_i/s_i$.

Uptake of NTF is taken to be activity dependent and is assumed to promote anatomical change, by inducing either sprouting or retraction, depending on how much is taken up, and the resting level required to maintain existing terminals. This is quantified by defining $\tilde{E}_i = -a_i(\tilde{A}_i - \tau^{-1})$, where τ is some measure of either an activity-dependent requirement or a baseline, resting requirement by each process. Writing $\tilde{E}_i = E_i/(\tau \sum_j \mathcal{D}_{ij})$, we have

$$E_i = -\tau a_i \sum_j \tilde{\Delta}_{ij} f_R \left(\sum_k \mathcal{D}_{jk} a_k \right) + a_i \sum_j \mathcal{D}_{ij}. \tag{A.1}$$

This is a general expression for the energy of process i.

Setting $\tau = 2$ and with $f_R(x) = x$ as a simple model of the production and release of NTF, this reduces to

$$E_i = -2a_i \sum_j \Delta_{ij} a_j + a_i \sum_j \mathcal{D}_{ij}. \tag{A.2}$$

Replacing $\sum_j \mathcal{D}_{ij}$ by the approximate, "smeared" form $\sum_j \Delta_{ij}$, we obtain

$$E_i = -a_i \sum_j \Delta_{ij}(2a_j - 1) = -a_i \sum_j \Delta_{ij}\sigma_j. \tag{A.3}$$

Since, in our models, we consider only plasticity associated with active processes, we may replace a_i by σ_i in this expression. Summing over i, we recover the full energy function using σ_i, and not a_i, as the activity variables.

Acknowledgments

T. E. thanks the Royal Society for the support of a University Research Fellowship during the latter stages of this work.

References

Bailey, C. H., & Kandel, E. R. (1993). Structural changes accompanying memory storage. *Annu. Rev. Physiol., 55,* 397–426.

Cynader, M. S. (1983). Prolonged sensitivity to monocular deprivation in dark–reared cats: Effects of age and visual exposure. *Dev. Brain Res., 8,* 155–164.

Cynader, M. S., Berman, N., & Hein, A. (1976). Recovery of function in cat visual cortex following prolonged deprivation. *Exp. Brain Res., 25,* 139–156.

Elliott, T., Howarth, C. I., & Shadbolt, N. R. (1996a). Axonal processes and neural plasticity. I: Ocular dominance columns. *Cereb. Cortex, 6,* 781–788.

Elliott, T., Howarth, C. I., & Shadbolt, N. R. (1996b). Axonal processes and neural plasticity. II: Adult somatosensory maps. *Cereb. Cortex, 6,* 789–793.

Elliott, T., Howarth, C. I., & Shadbolt, N. R. (1996c). Neural competition and statistical mechanics. *Proc. R. Soc. Lond. Ser. B, 263,* 601–606.

Elliott, T., Howarth, C. I., & Shadbolt, N. R. (1997). Axonal processes and neural plasticity. III: Competition for dendrites. *Phil. Trans. R. Soc. Lond. Ser. B, 352,* 1975–1983.

Fox, K., Sato, H., & Daw, N. (1989). The location and function of NMDA receptors in cat and kitten visual cortex. *J. Neurosci., 9,* 2443–2454.

Gu, Q., Liu, Y., & Cynader, M. S. (1994). Nerve growth factor-induced ocular dominance plasticity in adult cat visual cortex. *Proc. Natl. Acad. Sci. U.S.A., 91,* 8408–8412.

Kasamatsu, T. (1983). Neuronal plasticity maintained by the central norepinephrine system in the cat visual cortex. *Prog. Psychobiol. Physiol. Psychol., 10,* 1–112.

Miller, K. D. (1990). Correlation-based models of neural development. In M. A. Gluck & D. E. Rumelhart (Eds.), *Neuroscience and connectionist theory* (pp. 267–353). Hillsdale, NJ: Erlbaum.

Miller, K. D., & MacKay, D. J. C. (1994). The role of constraints in Hebbian learning. *Neural Comp., 6,* 100–126.

Miller, K. D., Keller, J. B., & Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. *Science, 245,* 605–615.

Purves, D. (1994). *Neural activity and the growth of the brain.* Cambridge: Cambridge University Press.

Quartz, S. R., & Sejnowski, T. J. (in press). The neural basis of cognitive development: A constructivist manifesto. *Behav. Brain Sci.*

Stryker, M. P., & Strickland, S. L. (1984). Physiological segregation of ocular dominance columns depends on the pattern of afferent electrical activity. *Invest. Ophthalmol. Vis. Sci.* (Suppl.), *25,* 278.

von der Malsburg, C. (1973). Self-organization of orientation selective cells in the striate cortex. *Kybernetik, 14,* 85–100.

Weliky, M., & Katz, L. C. (1997). Disruption of orientation tuning in visual cortex by artificially correlated neuronal activity. *Nature, 386,* 680–685.

Received January 31, 1996; accepted July 21, 1997.

LETTER Communicated by Ad Aertsen

Synaptic Delay Learning in Pulse-Coupled Neurons

Harald Hüning*
Abteilung für Neuroinformatik, Universität Ulm, Oberer Eselsberg, D-89069 Ulm, Germany

Helmut Glünder**
Institut für Medizinische Psychologie, Ludwig-Maximilians-Universität, Goethestraße 31, D-80336 München, Germany

Günther Palm
Abteilung für Neuroinformatik, Universität Ulm, Oberer Eselsberg, D-89069 Ulm, Germany

We present rules for the unsupervised learning of coincidence between excitatory postsynaptic potentials (EPSPs) by the adjustment of postsynaptic delays between the transmitter binding and the opening of ion channels. Starting from a gradient descent scheme, we develop a robust and more biological threshold rule by which EPSPs from different synapses can be gradually pulled into coincidence. The synaptic delay changes are determined from the summed potential—at the site where the coincidence is to be established—and from postulated synaptic learning functions that accompany the individual EPSPs. According to our scheme, templates for the detection of spatiotemporal patterns of synaptic activation can be learned, which is demonstrated by computer simulation. Finally, we discuss possible relations to biological mechanisms.

1 Introduction and New Learning Scheme

The timing or coherence of a neuron's input signals determines whether the neuron behaves as an integrator or coincidence detector (Abeles, 1982). Regarding the number of impulses that are required to exceed a voltage threshold—for example, at the axon hillock or a dendritic site with voltage-dependent mechanisms—temporally incoherent signals are less effective than synchronized ones. However, if we take into account axonal and dendritic propagation times, significant coincidence cannot be expected for synchronous impulse emission (Glünder & Nischwitz, 1993). Consequently, and in contrast to the prevailing paradigm that learning manifests itself in the change of synaptic strengths, we took first steps toward a formalism for

* Present address: E/E P6, Imperial College, London SW7 2BT
** To whom correspondence should be addressed

unsupervised learning of individual synaptic delays that aims to produce coincident excitatory postsynaptic potentials (EPSPs) at a defined site. If this site differs significantly from that of the synapses (of course, within the same postsynaptic neuron), we confront the well-known communication problem associated with any form of nonlocal synaptic coincidence learning, Hebbian learning included (see, e.g., Palm, 1982, and section 5 of this article). While other authors account for delay changes by synaptic selection from a multiplicity of axonal or dendritic pathways with different propagation times (e.g., Gerstner, Ritz, & van Hemmen, 1993; Hopfield, 1995; Miller, 1989; Tank & Hopfield, 1987), we propose postsynaptic processes. Promising candidates for adjustable delays between synaptic activation (transmitter binding) and the generation of a postsynaptic potential (opening of ion channels) are experimentally demonstrated molecular messenger cascades (Hille, 1994; Wickman & Clapham, 1995) that we consider as structurally less costly than the approach noted in the previous sentence (cf. section 5).

Aside from solving timing problems in neural circuits, delay learning can serve the evaluation of spatiotemporal patterns of synaptic activation (Carr, 1993; Eggermont, 1990; Wang, 1995). For such computational purposes, the idea of adjustable delays has been considered by several authors (Baldi & Atiya, 1994; Eckmiller & Napp-Zinn, 1993; Glünder & Nischwitz, 1993; Jansen, Bluhm, Napp-Zinn, & Eckmiller, 1991; Napp-Zinn, Jansen, & Eckmiller, 1996), and recently Hopfield (1995) has suggested a neural pulse position modulation with intensity-invariant demodulation by "coordinated time delays." For nonspiking networks, learning schemes have been formulated by Baldi and Atiya (1994), Bell and Sejnowski (1995), Bodenhausen and Waibel (1991), and Tank and Hopfield (1987), but to our knowledge, no mathematical framework for unsupervised delay learning in pulse-coupled neurons has yet been published.

This article relates our threshold rule (Glünder & Hüning, 1996) for unsupervised learning of synaptic delays to the gradient descent scheme. Figure 1 shows three synapses of a neuron that are activated at times $t_{\mathrm{act},i}$ and their EPSPs delayed by τ_i. The idea is to determine the delay changes $\Delta\tau_i$ during every time interval T where the somatic or a local dendritic depolarization $u(t)$, that is, summed EPSPs, is above a learning threshold θ (see Figure 2). For their computation we must assume a secondary process that accompanies each EPSP and determines the amount and direction of the changes. With this postulated learning function $\lambda(t)$, the delay change is

$$\Delta\tau_i \sim \int_T [u(t) - \theta] \cdot \lambda(t - t_{\mathrm{act},i} - \tau_i)\, dt. \tag{1.1}$$

Thus we propose delay changes proportional to the temporal integral of the weighted learning function, where the weighting term is the suprathreshold depolarization $u(t) - \theta \geq 0$. A good choice for the learning function $\lambda(t)$ is the EPSP function's negative derivative (see section 3).

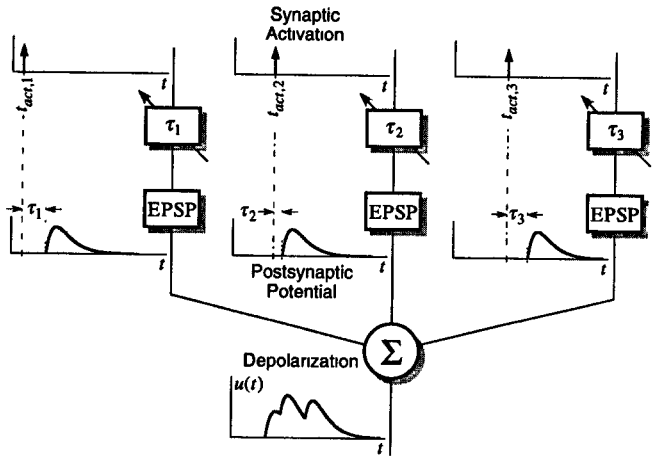


Figure 1: Differently timed ($t_{act,i}$) activation of three synapses at a cell or cell patch evokes delayed (τ_i) EPSPs that result in the net depolarization $u(t)$. The continuously adjustable delays are to be learned for coincident EPSPs.

In the next section we present a gradient descent approach to synaptic delay learning, from which a first learning function is derived, and we introduce the computation of the delay changes at the time of maximum depolarization. In section 3, we further develop this scheme to our threshold rule and generalize the concept of the learning function. We proceed with the simulated formation of a detector that becomes tuned to a spatiotemporal pattern of synaptic activation, and we consider the issue of stability. In the conclusion, we briefly relate our theoretical investigations to known and expected biophysical and neurobiological mechanisms.

2 Relation to Unsupervised Gradient Descent Learning

We relate the unsupervised learning of synaptic delays to schemes of error minimization by using the mathematically convenient parabolic EPSP function (see Figure 2A),

$$h_p(t) = \begin{cases} 1 - t^2 & \text{for } -1 \le t \le 1 \\ 0 & \text{else} \end{cases}$$

For reasons that will soon become evident, we define the activation onset (beginning of the transmitter binding) of synapse i as $t_{act,i} = t_{ref} - t_i$. With a relative activation time $t_i > 0$, it then precedes the reference time t_{ref} at which the delay changes are computed. We assume that N excitatory and

linearly transmitting synapses contribute—each with a single EPSP—to a neuron's depolarization $u(t = t_{\text{ref}})$. For this to happen, the synaptic delays τ_i must obey the relation $|t_i - \tau_i| < 1$. Then the summed potential is

$$u(t = t_{\text{ref}}) = N - \sum_i (t_i - \tau_i)^2.$$

Clearly, $u(t = t_{\text{ref}})$ becomes maximum for N coincident EPSPs, which can be achieved by minimizing—through gradient descent—the squared Euclidean distance between the relative activation times t_i and the associated delays τ_i. This leads to the well-known learning rule, here for synaptic delays $\Delta\tau_i \sim t_i - \tau_i$. Unfortunately, the relative activation times t_i are unknown to the neuron. However, formally $t_i - \tau_i$ can be expressed by the derivative of the EPSP function as $-\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}h_p(t + t_i - \tau_i)|_{t=0}$, using $t_{\text{ref}} = 0$ for simplicity. An essential point of this article is that such a secondary and clearly hypothetical signal is indispensable. It accompanies each EPSP (see similar ideas in Gerstner et al., 1993), and we refer to it as a synaptic learning function $\lambda(t)$. For the parabolic EPSP $h_p(t)$, the learning function resulting from gradient descent is $\lambda_p(t) \sim -\frac{\mathrm{d}}{\mathrm{d}t}h_p(t) = 2t$ for $-1 \le t \le 1$ and zero where the EPSP is zero as well. At the reference time, all EPSPs' learning functions are sampled to give the delay changes (learning increments) of the corresponding synapses.

Although specific signals may exist that define a reference time, we now propose to consider the time at which the depolarization $u(t)$ is maximum. If the sum of N parabolic EPSPs exhibits a single maximum, then the hereby defined reference time becomes $t_{\max} = \frac{1}{N}\sum_i(\tau_i - t_i)$ and the maximum potential is $u(t_{\max}) = N - \|\vec{e}\|^2$, with the components of the error vector

$$e_i = (t_i - \tau_i) - \frac{1}{N}\sum_j (t_j - \tau_j).$$

Hence, if we keep to the learning function $\lambda_p(t)$, we arrive at the learning rule $\Delta\tau_i \sim e_i$ (Hüning, 1995). Here, the sampling of the learning function takes place at the maximum of the depolarization. Although this signal-defined reference time is less ad hoc, maximum detection is difficult to implement, highly sensitive to noise, and thus biologically quite implausible.

3 Temporally Distributed Delay Learning (Threshold Rule)

As a scheme for the unsupervised learning of synaptic delays that is more robust with respect to noisy potentials we finally propose the depolarization-dependent threshold rule (see equation 1.1). With this scheme, delay changes are executed either continuously during or at the end of learning intervals T_μ for which the net depolarization remains above a learning threshold θ (see Figures 2A and 2B, bottom). Although learning defined by equation 1.1 appears functional also without the suprathreshold function

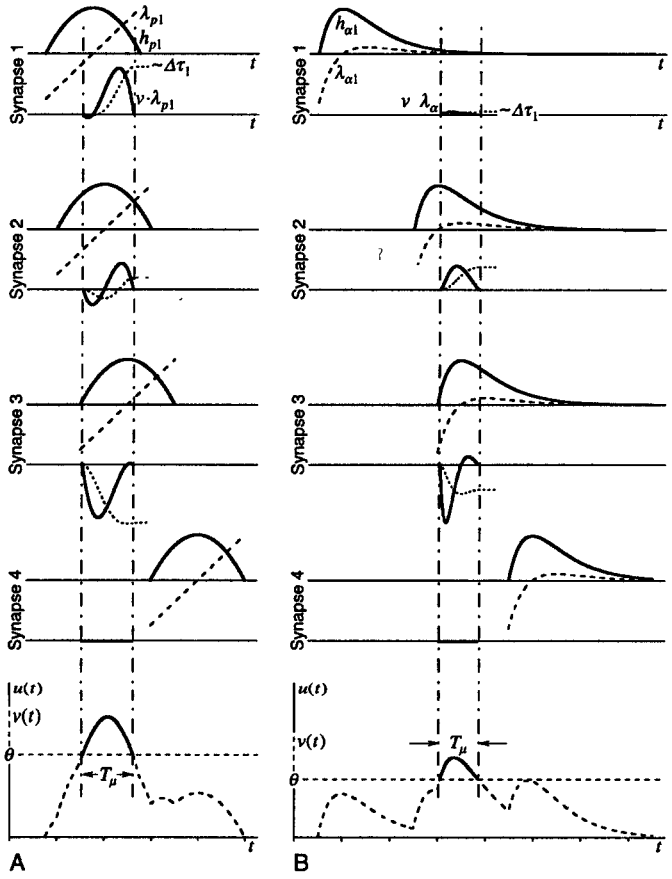


Figure 2: Evaluation of individual synaptic delay changes $\Delta\tau_i$ from the net depolarization $u(t)$ (bottom) of a cell patch with four active synapses. For every synapse, we show an EPSP, its learning function λ (dashed line, except bottom), the weighted learning function $v \cdot \lambda$ (with $v(t) = u(t) - \theta \geq 0$), and its integral (dotted line) that is proportional to the delay change. EPSP shape $h(t)$: (A) parabolic, $h_p(t) = 1 - t^2$ for $-1 \leq t \leq 1$; (B) α-function, $h_\alpha(t) = \alpha^2 t \cdot e^{-\alpha t}$ for $t \geq 0$.

$v(t) = u(t) - \theta \geq 0$, we include this weighting to avoid conflicting and sometimes stable oscillatory delay changes that otherwise can result from concurring activation patterns at successive learning intervals (see section 4).

If every parabolic EPSP and therefore the associated synaptic learning function $\lambda_p(t) \sim -\frac{d}{dt}h_p(t)$ entirely cover the learning interval, we obtain from equation 1.1 the learning rule $\Delta\tau_i \sim V_\mu \cdot e_i$ with the suprathreshold area $V_\mu = \int_{T_\mu}[u(t) - \theta]\,dt$. Figure 2A depicts a situation where three parabolic EPSPs cover the learning interval, while a fourth EPSP comes later and is not captured. In contrast to schemes relying on reference times, where EPSPs are not captured if they do not contribute to the sampled depolarization, our threshold rule shows a gradual coupling of EPSPs that only partly reach into the learning interval. In the latter case, the delay changes increase with every presentation of a spatiotemporal activation pattern until an EPSP's maximum enters the learning interval. This behavior becomes pronounced with more realistic EPSPs—that is, with unimodal functions that steeply rise and slowly decay. (For asymptotically decaying EPSPs, we reasonably assume learning functions of finite duration $\lambda(t > t_\varepsilon) = 0$, with $h(t > t_\varepsilon) < \varepsilon$, where ε may depend on the noise level.) With this kind of asymmetric EPSP function and $\lambda(t) \sim -\frac{d}{dt}h(t)$, early EPSPs will be captured long before their maxima enter the learning interval, whereas late EPSPs, which rise after the interval, either fail to be captured (the fourth synapse in Figure 2B) or create a separate learning interval (for a lower threshold than in Figure 2B).

Because various synaptic learning functions are feasible for a given unimodal EPSP function $h(t)$, we have investigated general requirements. Evidently learning functions must change sign from minus to plus in order to give the direction of the delay changes. Formally, we have found that all learning functions $\lambda(t) = -\frac{d}{dt}f\{h(t)\}$ with any monotone increasing function f comply with the demand that learning must stop, that is, the integral (see equation 1.1) must vanish, if coincidence of the EPSPs is reached (Hüning, 1995). This holds for all threshold settings. Function f permits one to tailor the properties of the learning process. For instance, it may serve the smoothing of a learning function's otherwise discontinuous onset and the restriction of its duration. Furthermore, we can conclude that coincidence learning still works with EPSPs of various amplitudes.

4 Simulation of Spatiotemporal Template Learning

We demonstrate unsupervised synaptic delay learning by the simulated formation of detectors for spatiotemporal patterns of synaptic activation. As an example, we consider the time courses of activation at 10 synapses of a formal neuron (see Figure 3A). Each of the two distinct patterns lasts longer than a single EPSP. Before the repeated presentations of the pattern pair, the 10 synaptic delays are randomly distributed in the interval $0.5\Delta \leq \tau_i \leq 3\Delta$, where Δ is the duration of the parabolic EPSP. Therefore, and because both patterns are well separated in time, the neuron, at best, will be tuned to one of them. A steady time course of the neuron's depolarization (see Figure 3B) is reached after 21 presentations of the pattern pair. Owing to

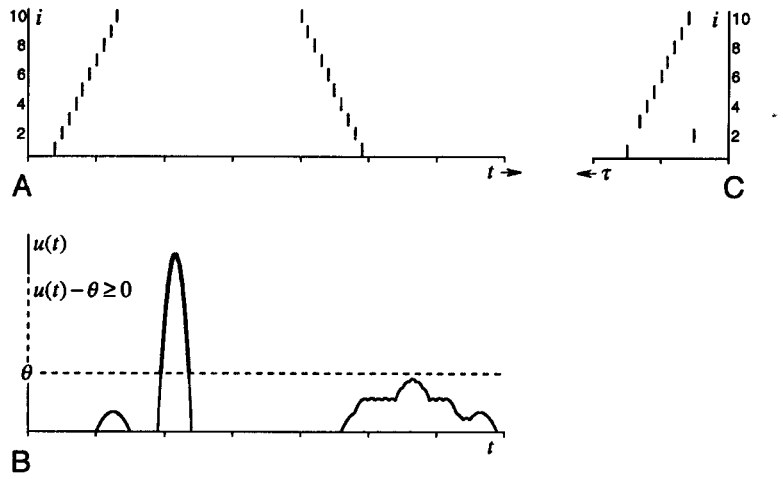


Figure 3: Formation of templates for spatiotemporal patterns. (A) Sample of the stimulation patterns at 10 synapses of a neuron. (B) Steady time course of the net depolarization after delay learning. (C) Final delay configuration $\tau(i)$.

the greater similarity of the first pattern to the specific random initialization, the threshold rule has produced a detector for this pattern which is obvious from the final synaptic delay configuration (slightly imperfect template) depicted in Figure 3C.

Apart from the functionality of delay learning with the threshold rule, our simulation illustrates the effect of temporally limited EPSPs and learning functions, as well as the competition of different patterns. As long as an EPSP contributes to the suprathreshold depolarization, the corresponding synaptic delay becomes adjusted, which in our example is not fulfilled for the second synapse. Furthermore, if both patterns produce suprathreshold depolarizations, we then obtain opposing delay changes. However, oscillations are avoided by the weighting term of equation 1.1, which drives the learning process toward the pattern that evoked the largest initial suprathreshold voltage.

Because a neuron's delay tuning is not changed by patterns that remain subthreshold, a sufficiently high learning threshold retains a tuning even without any further occurrences of the pattern that gave rise to it. Therefore, a threshold that adapts toward the peaks of the depolarization provides a stable delay tuning. Intermediate thresholds cause an adaptive averaging behavior. Accordingly, the delay tuning can follow a slowly changing and repeatedly presented pattern of synaptic activation (Napp-Zinn et al., 1996),

provided the threshold is crossed at every occurrence. For small, random fluctuations of a pattern, the delays are expected to be tuned to the temporal mean, because the integral (see equation 1.1) behaves approximately linear around the zero crossing for realistic EPSP functions.

5 Conclusion

In summarizing our results, in particular concerning their neurobiological implications, we have to speculate about possible biological mechanisms, not an easy task for theoreticians. However, our general impression of the recent progress in the investigation of synaptic mechanisms gives us a good confidence that suitable biological mechanisms for anything that is logically possible will be found eventually. So one should not worry too much about the concrete mechanisms proposed below.

We have presented a systems view of unsupervised and robust coincidence learning in pulse-coupled neurons that essentially relies on three assumptions.

1. **Only a sum of EPSPs is accessible at a defined measuring site at which the EPSP coincidence is to be established.** Similar to long-term changes of synaptic strengths (Brown, Kairiss, & Keenan, 1990), synaptic delay changes also are assumed to depend on pre- and postsynaptic potentials. With respect to postsynaptic potentials, local dendritic learning is based on dendritic depolarization, whereas more global neural learning relies on the potential at a neuron's axon hillock. The process of delay learning will lead to coinciding EPSPs at these sites.

2. **The time course of the voltage above a learning threshold at the measuring site is available to the individual synapses.** Voltage thresholds are biologically plausible (Artola & Singer, 1993), and their adaptation according to the long-term mean of the depolarization was proposed earlier (Bienenstock, Cooper, & Munro, 1982). In the case of local dendritic learning schemes, the suprathreshold depolarization can easily be sensed by synapse-related molecular mechanisms. Rules that are nonlocal within the postsynaptic cell require the suprathreshold depolarization to be instantaneously signaled, for example, from a neuron's axon hillock, back to all its synapses, which appears more involved. This well-known and indeed fundamental communication problem exists with any form of nonlocal synaptic coincidence learning, Hebbian learning of synaptic strengths included. Except for the work reported by Stuart and Sakmann (1994), to date we have to rely more on speculations than on direct experimental evidence for possible communication mechanisms. Interestingly, Hebbian learning today generally is assumed local (Brown et al., 1990), although Hebb (1958) described a global scheme: "When an axon of a neuron x is near enough to fire a neuron y and does so, some change takes place such that x becomes more effective at exciting y. What is this change and how does it work? This is a question to which we have no final answer." In case of passive or active dendritic

propagation of action potentials (e.g., back from the soma to the synapses), these potentials will act in the same way as the dendritic depolarizations do in our scheme; they will define the learning intervals and the weighting of the learning function.

3. **A uniform learning function is attributed to every synapse and is triggered at the opening of its ion channels (EPSP onset).** Delay learning, as proposed in this article, requires that the synapse contributes to the postsynaptic depolarization and that a postulated differentiating (biochemical) process parallels its individual contribution. This kind of process could be realized by the interaction of an intracellular messenger, such as an activated G protein, and channel proteins (Destexhe, Mainen, & Sejnowski, 1995).

Under these circumstances, we have shown how to compute the delay change of an active synapse from the values of its learning function in conjunction with its suprathreshold depolarization. Biologically speaking, we assume the suprathreshold depolarization to have a nonlinear influence on the differentiating (biochemical) process. During the periods of suprathreshold depolarization, this process could, for example, modify the temporal behavior of intracellular messengers that determine the delay between the transmitter binding to a transmembrane receptor and the intracellular opening of ion channels. This modification could be similar to changes of presynaptic messenger cascades, as initiated by retrograde diffusion of nitric oxide (Montague, 1993), that are hypothesized to cause long-term changes of synaptic strengths. In this respect, we do not rule out alternative mechanisms for delay changes, such as modifications of the temporal behavior of presynaptic molecular processes. Currently there is increasing interest in membrane-delimited mechanisms of rather direct and thus comparatively fast (within a second) interaction between activated G proteins and ion channels (Hille, 1994; Wickman & Clapham, 1995), but to our knowledge, on a millisecond time scale, the properties and their modifiability of these interactions have not been investigated yet.

Although the concrete biophysical or biochemical realization of delay learning is still unclear, we have demonstrated that this simple learning mechanism is well within the possibilities of our current neurobiological knowledge and would provide a useful addition to the commonly accepted plasticity of synaptic efficacy.

Acknowledgments

We thank A. Bibbig and T. Wennekers for helpful comments and R. Hudson for improving the text.

References

Abeles, M. (1982). Role of cortical neuron: Integrator or coincidence detector? *Isr. J. Med. Sci., 18,* 83–92.

Artola, A., & Singer, W. (1993). Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. *Trends in Neurosciences, 16*, 480–487.

Baldi, P., & Atiya, A. F. (1994). How delays affect neural dynamics and learning. *IEEE Transactions on Neural Networks, 5*, 612–621.

Bell, A. J., & Sejnowski, T. J. S. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*, 1129–1159.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience, 2*, 23–48.

Bodenhausen, U., & Waibel, A. (1991). The tempo 2 algorithm: Adjusting time-delays by supervised learning. In R. P. Lippman, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in neural information processing systems 3* (pp. 155–161). San Mateo, CA: Morgan Kaufmann.

Brown, T. H., Kairiss, E. W., & Keenan, C. L. (1990). Hebbian synapses: Biophysical mechanisms and algorithms. *Annual Review of Neuroscience, 13*, 475–511.

Carr, C. E. (1993). Processing of temporal information in the brain. *Annual Review of Neuroscience, 16*, 223–243.

Destexhe, A., Mainen, Z. F., & Sejnowski, T. J. (1995). Synaptic currents, neuromodulation, and kinetic models. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 956–959). Cambridge, MA: MIT Press.

Eckmiller, R., & Napp-Zinn, H. (1993). Information processing in biology-inspired pulse coded neural networks. In Y. Okabe (Ed.), *Proc. Int. Joint Conference on Neural Networks '93* (pp. 643–648). Nagoya, Japan.

Eggermont, J. J. (1990). *The correlative brain: Theory and experiment in neural interaction*. Berlin: Springer-Verlag.

Gerstner, W., Ritz, R., & van Hemmen, J. L. (1993). Why spikes? Hebbian learning and retrieval of time-resolved excitation patterns. *Biological Cybernetics, 69*, 503–515.

Glünder, H., & Hüning, H. (1996). Detection of spatio-temporal spike patterns by unsupervised synaptic delay learning. In N. Elsner & H.-U. Schnitzler (Eds.), *Brain and evolution* (Vol. 2). Stuttgart: Thieme.

Glünder, H., & Nischwitz, A. (1993). On spike synchronization. In A. Aertsen (Ed.), *Brain theory: Spatio-temporal aspects of brain function* (pp. 251–258). Amsterdam: Elsevier.

Hebb, D. O. (1958). *Textbook of psychology*. Philadelphia: Saunders.

Hille, B. (1994). Modulation of ion-channel function by G-protein-coupled receptors. *Trends in Neurosciences, 17*, 531–536.

Hopfield, J. J. (1995). Pattern recognition computation using action potential timing for stimulus representation. *Nature, 376*, 33–36.

Hüning, H. (1995). *Lernen durch Laufzeitvariation in Netzwerken impulsgekoppelter Neurone*. Diploma thesis, Rheinisch-Westfälische Technische Hochschule, Aachen.

Jansen, M., Bluhm, M., Napp-Zinn, H., & Eckmiller, R. (1991). Asynchronous pulse-processing neural net hardware for dynamic functions based on frequency and phase information. In U. Ramacher, U. Rückert, & J. Nossek

(Eds.), *Proc. 2. Int. Conf. Microelectronics and Neural Networks* (pp. 359–365). Munich: Kyrill & Method.

Miller, R. (1989). Cortico-hippocampal interplay: Self-organizing phase-locked loops for indexing memory. *Psychobiology, 17,* 115–128.

Montague, P. R. (1993). The NO hypothesis. In B. Smith & G. Adelman (Eds.), *Neuroscience year: Supplement 3 to the encyclopedia of neuroscience* (pp. 100–102). Boston, MA: Birkhäuser.

Napp-Zinn, H., Jansen, M., & Eckmiller, R. (1996). Recognition and tracking of impulse patterns with delay adaptation in biology-inspired pulse processing neural net (BPN) hardware. *Biological Cybernetics, 74,* 449–453.

Palm, G. (1982). *Neural assemblies: An alternative approach to artificial intelligence.* Berlin: Springer-Verlag.

Stuart, G. J., & Sakmann, B. (1994). Active propagation of somatic action potentials into neocortical pyramidal cell dendrites. *Nature, 367,* 69–72.

Tank, D. W., & Hopfield, J. J. (1987). Neural computation by concentrating information in time. *Proceedings of the National Academy of Science of the USA, 84,* 1896–1900.

Wang, D. (1995). Temporal pattern processing. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 967–971). Cambridge, MA: MIT Press.

Wickman, K. D., & Clapham, D. E. (1995). G-protein regulation of ion channels. *Current Opinion in Neurobiology, 5,* 278–285.

Received January 29, 1996; accepted September 9, 1997.

LETTER Communicated by Ad Aertsen

Neural Processing in the Subsecond Time Range in the Temporal Cortex

Kiyohiko Nakamura
Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama 226, Japan

The hypothesis that cortical processing of the millisecond time range is performed by latency competition between the first spikes produced by neuronal populations is analyzed. First, theorems that describe how the mechanism of latency competition works in a model cortex are presented. The model is a sequence of cortical areas, each of which is an array of neuronal populations that laterally inhibit each other. Model neurons are integrate-and-fire neurons. Second, the model is applied to the ventral pathway of the temporal lobe, and neuronal activity of the superior temporal sulcus of the monkey is reproduced with the model pathway. It consists of seven areas: V1, V2/V3, V4, PIT, CIT, AIT, and STPa. Neural activity predicted with the model is compared with empirical data. There are four main results: (1) Neural responses of the area STPa of the model showed the same fast discrimination between stimuli that the corresponding responses of the monkey did: both were significant within 5 ms of the response onset. (2) The hypothesis requires that the response latency of cortical neurons should be shorter for stronger responses. This requirement was verified by both the model simulation and the empirical data. (3) The model reproduced fast discrimination even when spontaneous random firing of 9 Hz was introduced to all the cells. This suggests that the latency competition performed by neuronal populations is robust. (4) After the first few competitions, the mechanism of latency competition always detected the strongest of input activations with different latencies.

1 Introduction

Primates recognize and respond to complex visual stimuli within half a second. Signals from the retina take 25–50 ms to reach the primary visual cortex V1, and the motor system requires about 150 ms to produce a response (Kalaska & Crammond, 1992). This means that 200–300 ms is available for cortical processing. Not only is the processing time restricted to less than a third of a second, the firing rate of single cortical neurons is usually less than 50 Hz. Each neuron thus produces as few as 10 spikes during the processing.

The temporal constraints also limit the number of processing steps (Oram & Perrett, 1994). The time between spike generation in one cell and its in-

Neural Computation **10, 567–595** (1998)

fluence on the firing rate of a postsynaptic cell is about 5 ms. Latency of the earliest cell response in the superior temporal sulcus (STS) is 70–100 ms, and cells in the STS can discriminate forms of visual stimuli within 5 ms of response onset (Oram & Perrett, 1992). This suggests that the processing from the cortex V1 to the STS could take place within about 50 ms and that as few as 10 synaptic relays occur during the processing. The ventral pathway of form recognition passes through seven areas—V1, V2/V3, V4, PIT, CIT, AIT, and STPa—so only one or two synaptic relays can occur within each area.

This article analyzes a hypothetical mechanism that performs the cortical processing within that time range (Nakamura, 1992, 1993a). The mechanism was intended to account for the following characteristics. First, each cell produces as few as 10 spikes during the few hundred milliseconds available for this processing. Second, cortical processing of that time range usually becomes possible only after training. For example, children need to experience repeated stimulus-response cycles before they learn to discriminate cats from other animals within a few hundred milliseconds. Third, neural mechanisms of the brain are robust against noise because brain cells are subject not only to sensory signals but also to many kinds of noise.

The mechanism works through two processes: latency competition and reinforcement of cortical connections. The competition is between neuronal populations that laterally inhibit each other (Nakamura & Ichikawa, 1989). When the populations are activated simultaneously, those receiving the strongest activations fire first and inhibit the others from firing. This mechanism detects the most strongly activated populations in the time range of firing latency, because the first spikes indicate that the populations producing them have been most strongly activated. Suppose that the competition functions as neural processing of every cortical area. Then the cortical processing by way of multiple areas could be done within a few hundred milliseconds. The process of reinforcement is to increase the synaptic efficacy of cortical pathways that lead the first spikes to relevant cells of the motor cortex. Repeated learning selectively reinforces the pathways along which neuronal excitation producing relevant motor responses travels first (Nakamura, 1993a).

A number of network models using lateral inhibition have been proposed. In many, modules producing the strongest outputs win the competition by inhibiting the others, and stimulus selectivity emerges gradually through computational cycles of connectivity change (Rumelhart & Zipser, 1986; Grossberg, 1987; Fukushima, 1980). A main difference from the mechanism presented here is that these earlier models did not take account of temporal aspects of the competition in each cycle. As a result, they failed to provide an estimate of time taken by each competitive cycle. Since they have not dealt specifically with time for each cycle, their response time is usually measured by time taken for the emergence of the stimulus selectivity. Consequently, the earlier models have been considered to describe slow

mechanisms. The model presented here has pointed out that winners of every neural competition are determined by firing latency and that the cycles of connectivity change correspond to learning process to form reinforced pathways producing the rapid and relevant responses.

Coultrip, Granger, and Lynch (1992) presented a network model of the latency competition, in which single neurons compete and a single inhibitory neuron delivers lateral inhibition. This network architecture may not work in noisy environments because damage to single neurons immediately changes the results of the competition. Biological mechanisms need to be robust and to function appropriately despite noise. The circuit of the model presented here (Nakamura & Ichikawa, 1989) is composed of neuronal populations and encodes signals in ratios of firing neurons in the populations. Its behavior is hardly varied under noisy conditions.

Another network model using a latency mechanism, presented by Opara and Wörgötter (1996), has demonstrated that a delay mechanism in the early stage of visual pathway speeds up synchronization for cell assemblies representing different objects. Although they did not take neural plasticity into consideration, a number of studies have shown cortical plasticity in the visual pathway of the temporal lobe. Miyashita and colleagues (Miyashita & Chang, 1988; Sakai & Miyashita, 1991; Miyashita, 1990) have shown that extensive training affects the selectivity of cells in the AIT cortex. Kobatake, Tanaka, Wang, and Tamori (1993) also have reported that training makes cells of the inferotemporal cortex responsive to the presented stimuli. These observations suggest that the stimulus selectivity of cells might result from the adjustment of cortical connectivity. The model here accounts for the stimulus selectivity produced by the connectivity adjustment.

Oram and Perrett (1992) have shown that cells of the monkey STS discriminate visual stimuli within 5 ms of response onset. We expect the hypothetical mechanism of latency competition and connectivity reinforcement to account for this discrimination. To see if it actually does, we developed a model of the ventral pathway of the temporal lobe and used a computer simulation of their experiment to see whether the model reproduced the neuronal activity of the monkey.

2 Model

2.1 Architecture of the Model Cortex. The model cortex consists of the sensory cortex, association areas, and motor cortex. These areas are connected in series, and each is represented by an array of neuronal populations (see Figure 1A). Every population has three types of neurons: excitatory stellate cells, pyramidal cells, and inhibitory interneurons (see Figure 1B). Stellate cells transmit signals to pyramidal cells, which project to inhibitory interneurons and also send fibers to the next areas. Inhibitory neurons extend lateral projections to nearby pyramidal cells. Each cell in Figure 1B represents not a single cell but a population of a type of cells. Because each

as one

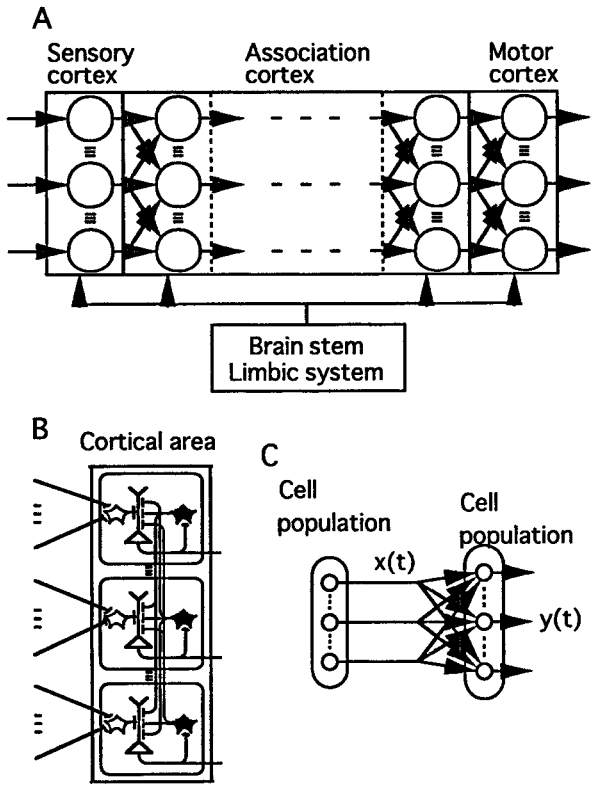


Figure 1: Model cortex. (A) Arrays of neuronal populations and connections between them. Projections from the brain stem and limbic system pervade the whole cortex. (B) Neuronal circuitry of a single cortical area. (C) Populations of neurons of two types and connections between them.

cell in Figure 1B is modeled by a cell population as shown in Figure 1C, every connection in Figure 1B represents a bundle of fibers connecting cell populations.

Fig 1 * The neuronal populations need not be "columns," but they may spatially overlap. What is necessary for our model is the neuronal connectivity shown in Figure 1. If the populations overlap, inhibitory collaterals (Kawaguchi, 1995; Kang, Kaneko, Ohnishi, Endo, & Araki, 1994) pervade many populations.

2.2 Model Neuron. Each neuron is represented by a single compartment, and its change in membrane potential V is given by

$$C(dV/dt) = -\sum_k G_k V + \sum_k G_k E_k, \tag{2.1}$$

where C is membrane capacitance and G_k and E_k are the channel conductance and electromotive force of ion k, respectively. Conductances G_k change when the neuron receives spikes, and we simplify the process of spike generation: When V exceeds a threshold h, the neuron fires and delivers spikes of duration τ_s. After that, the neuron enters a refractory state, which lasts τ_r and during which the neuron does not respond to any spikes that may arrive. By the end of the refractory period, V has returned to the resting potential V_o. Consecutive firing reduces the firing rate, and this is modeled by increasing the threshold h:

$$dh/dt = \Delta_h \delta(t) - C_h(h - h_o), \tag{2.2}$$

where $\delta(t) = 1$ at time t when the neuron begins to deliver a spike and 0 otherwise and where Δ_h, C_h, and h_o are, respectively, the increment per spike, the coefficient of recovery, and the initial value of h. This equation implies that firing threshold h rises by Δ_h every firing and returns to the initial value h_o at the rate C_h.

2.3 Input-Output Variables of Neuronal Populations. The nervous system is considered to encode information in ratios of firing neurons in neuronal populations. These ratios are averages across neurons, not over time. Let $x(t)$ denote the ratio of firing neurons to all the neurons of a population at time t. Neurons producing spikes are counted as the firing neurons. The conductance G_k of a neuron receiving projections from the population is written in the form

$$G_k(t) = [W_k x(t) + 1]G_k^o, \tag{2.3}$$

where $G_k(t)$ and G_k^o denote values of G_k at time t and at $x(t) = 0$, respectively, and W_k is a coefficient representing the effectiveness of synaptic transmission on change of conductance G_k. The value of W_k is derived from conductance changes produced by single spikes (for details, see Nakamura, 1993b).

2.4 Synaptic Plasticity. Synaptic plasticity of corticocortical connections is assumed, and the synapses of connections are reinforced when the following three conditions are satisfied simultaneously: (1) the connections deliver spikes, (2) the postsynaptic neurons fire, and (3) the projections from the brain stem and limbic system are activated. The first and third

conditions were introduced according to findings concerning the plasticity of ocular dominance in the cat visual cortex (Karil, Dubin, Scott, & Stark, 1986; Gordon, Allen, & Trombley, 1988). Although this plasticity was found during development and in a specific part of the cortex, it is assumed here to function even in adults and all over the neocortex. The second condition is based on characteristics of Ca^{2+}, which is involved in switching biochemical processes of plasticity. Sufficient Ca^{2+} influx is necessary for the processes. Since Ca^{2+} channels are voltage dependent, a large increase in membrane potential by firing evokes the influx. The projections from the brain stem and limbic system may convey signals related to attention and drive. Their activation instructs the nervous system that the current sensory stimuli should be memorized.

3 Model Analysis of Latency Competition Mechanism

The model describes how cortical processing that occurs in the millisecond time range is performed and is based on the work of Nakamura and Ichikawa (1989) and Nakamura (1993a). A sensory stimulus activates certain neuronal populations of the sensory cortex, and these activated populations deliver spikes to the next area, where, as shown in Figure 2A, the populations receiving the strongest connections are the first to fire. The first spikes excite populations of inhibitory interneurons (see Figure 1B) that laterally inhibit the other populations before they fire. Consequently, only the populations receiving the strongest connections are allowed to produce spikes. This mechanism is described by the following theorems. First, response latency of neuronal populations decreases with strong activation:

Theorem 1. *When $x(t)$ is a step input, response latency of neuronal populations decreases with large synaptic input $W_k x(t) G_k^o$ under the following conditions: (1) Synaptic input is excitatory and acts only on a single ion conductance G_{k_a}, and (2) increments in $W_{k_a} x(t)$ are nearly the same for all the neurons of the population.*

For the proof, see appendix A. The assumption of step input $x(t)$ will be discussed in section 6.3. The conditions were introduced for analytical simplicity, but they also have some biological plausibility. Condition 1 is satisfied if the corticocortical connections deliver the same transmitter, and it acts on a single ion conductance. Condition 2 holds if the input increments are produced by an increase in $x(t)$.

The lateral inhibition restricts firing to the most strongly activated populations:

Theorem 2. *Assume the same conditions as in Theorem 1. Then only the most strongly activated populations fire if the projections of pyramidal cells are strong enough to fire the inhibitory interneurons and if the resultant inhibition is strong*

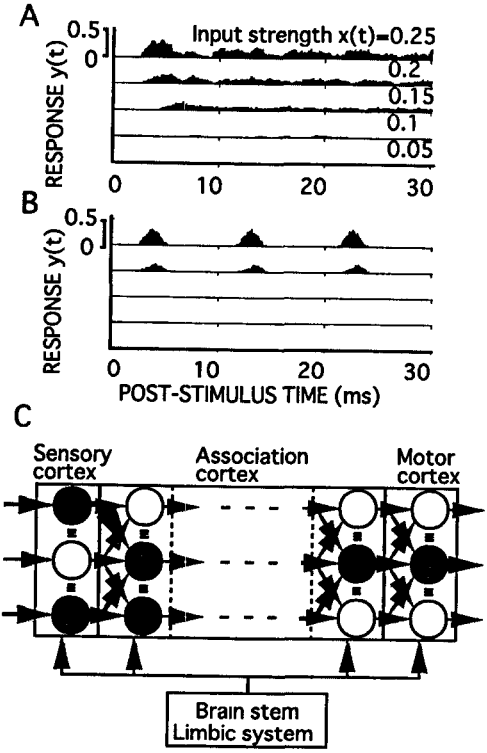


Figure 2: Neural mechanism of latency competition. (A) Responses of model neuronal populations receiving stepwise activations of different strengths. The abscissa and the ordinate, respectively, indicate the time after onset of the activations and the ratios of firing neurons in the populations. (B) Responses of neuronal populations in a single cortical area of the model (see Figure 1B). Five rows show changes in ratios of firing neurons in the populations receiving the same activations as in A. (C) Schematic illustration of connectivity reinforcement. Bold arrows denote reinforced corticocortical connections. Filled circles show activated neuronal populations, and open circles show unactivated neuronal populations.

enough to cancel out excitation of pyramidal cells and lasts longer than the refractory period of pyramidal cells.

Quantitative description and proof are given in appendix B. Figure 2B shows a simulation result. Since the interneurons inhibit pyramidal cells of their own populations as well, excitation of the inhibitory neurons is sup-

pressed. When the circuit is released from the inhibition, it repeats the same competitive process. As a result, the most strongly activated populations make oscillatory firing. Note that the competitive mechanism detects the strongest inputs with the first spikes, and their latency is as few as milliseconds. If processing in each cortical area is performed by the mechanism, the cortical processing by way of multiple areas can be done in a few hundred milliseconds. And no more than one spike is needed for each neuron during the cortical processing. The short processing time and the small number of spikes required for the processing meet the biological constraints already noted.

Consider how the cortical processing by way of many areas is performed in reinforced circuits. Assume that synaptic weights of corticocortical fibers are initially weak and set at random values. Suppose a sensory stimulus excites some populations of the sensory cortex. It follows from Theorem 2 that only the populations receiving the strongest connections produce spikes in the first area of association cortex. While the sensory stimulus is presented, every activation of projections from the brain stem and limbic system selectively reinforces the same connections between the firing populations and makes the initially strong connections stronger. This makes the specific populations of the area strongly responsive to the sensory stimulus (see Figure 2C).[1] Connections to subsequent areas are reinforced in the same way, and strong pathways running through those reinforced connections are formed. Consequently, we have the following theorem:

Theorem 3. *Assume the same conditions as in Theorem 2 and that corticocortical connections are initially set at random strength and are so weak that only strong activation of some populations in the projecting areas may weakly excite populations in the target areas. When the projections from the brain stem and limbic system are repeatedly activated under a sensory stimulus, certain pathways of the model cortex are selectively reinforced, and neuronal populations along these pathways become strongly responsive to the sensory stimulus.*

If the projections from the brain stem and limbic system convey reward signals, connections with the motor cortex are reinforced only if activation of their postsynaptic neurons produces reward motor responses. This is because firing of those neurons leads to activation of the projections and evokes the synaptic change (Nakamura, 1993). After reinforced pathways to the motor cortex are established, sensory signals travel first along them, and animals respond with rewarded actions.

[1] Reward signals from the brain stem and limbic system are delivered after the motor response, and sensory stimuli may therefore change before reward signals arrive. This time delay can be offset by reverberation in the corticohippocampal circuits. A detailed discussion of this has been published in Nakamura (1993b), which also discusses how the neural circuit starts or stops learning.

4 Simulation

The model was used to simulate the experiment done by Oram and Perrett (1992).

4.1 Model of Temporal Lobe. The experimental data were obtained in response to the presentation of visual form (head views), so the model cortex was that of the ventral pathway of the temporal lobe. It consists of seven areas: V1, V2/V3, V4, PIT, CIT, AIT, and STPa. They are connected in series, and the V1, V4, and PIT areas also send direct projections to the V4, CIT, and AIT areas (see Figure 3A). Each cortical area is a 5 × 5 array of neuronal populations (see Figure 3B), and each cell population contains 100 stellate cells, 100 pyramidal cells, and 100 inhibitory interneurons. Those neurons are connected as shown in Figure 1. The simulator includes $100 \times 3 \times 5 \times 5 \times 7 (= 52,500)$ neurons in total. The strength of corticocortical connections becomes weaker as the length of the connections increases, as shown in Figure 3C. Lateral inhibition also becomes exponentially weaker with distance. The synaptic plasticity is formulated as

$$dW_k/dt = C_r x(t) I(t) r(t) W_k - C_d W_k, \tag{4.1}$$

where $I(t)$ is the firing function of the postsynaptic neuron ($I(t) = 1$ when the neuron fires and 0 otherwise), $r(t)$ is the firing ratio of the projections from the brain stem and limbic system, and C_r and C_d are coefficients of reinforcement and decay, respectively. Values of W_k are assumed not to exceed an upper limit $\hat{W}_k$. The parameter values of equations 1 to 4 are given in appendix C.

4.2 Stimuli. Stimuli to the model cortex were activations of neuronal populations of the area V1. They were activated in the six patterns shown in Figure 4. For the populations indicated by filled circles, firing ratios of the afferent connections were increased from 0.009 (spontaneous activity) to 0.03. Since duration of spikes was set to be 1 ms, the ratio of 0.03 corresponds to a firing rate of 30 Hz of single neurons.

Spontaneous activity was also introduced to all neurons: Potential V was randomly raised above the firing threshold h_0 at the rate of spontaneous activity (S/A). The rate was set at 9 Hz according to the empirical data (Oram & Perrett, 1992).

4.3 Procedure. This study assumed that the monkeys Oram and Perrett (1992) used had learned the head views before the experiment, and the learning process was therefore simulated first. Initially, values W_k of corticocortical connections were set at small random values, and the neurons of STPa made no response to any stimulus. Learning was conducted. The projections from the brain stem and limbic system were kept

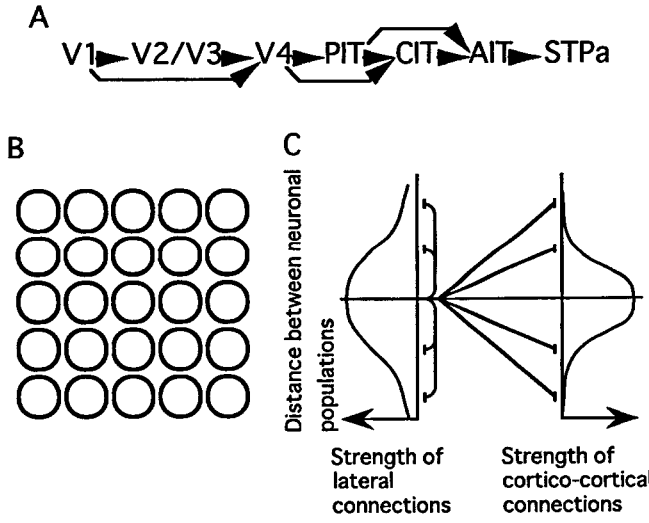


Figure 3: (A) Corticocortical connections of the ventral pathway of the temporal lobe. (B) A 5 × 5 array of neuronal populations in a single cortical area of the model simulator. (C) Change in synaptic efficacy of lateral and corticocortical connections with increasing distance between neuronal populations. The efficacy decreases exponentially as the distance increases.

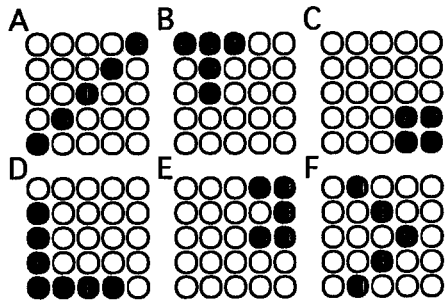


Figure 4: Six patterns of activation of the area V1. Filled circles indicate activated neuronal populations, and open circles are unactivated populations.

activated—that is, $r(t)$ was increased to 0.03—and the six stimuli were sequentially provided. The learning was continued until the activity of neurons in the STPa was significantly different from the S/A. The time required for the learning depended on the values of reinforcement and decay parameters C_r and C_d. A large value of C_r reduces computational time but makes learning unstable.

After learning, cell responses of the STPa were examined, just as in the experiment. The six test stimuli and a control stimulus (no activations of the area V1) were provided five times. Timing of spikes of the S/A was different from stimulus to stimulus. During the examination, it was assumed that reinforcement was negligible, that is, that $C_r = 0$.

4.4 Data Analysis. First, the same analysis that Oram and Perrett (1992) had done was conducted. Pyramidal cells were classified according to the mean cell response (spikes per second) over a period of 250 ms starting at stimulus onset. In the simulation, the stimulus onset was the activation of area V1, whereas in the experiment, it was the presentation of visual stimuli (Oram & Perrett, 1992). Therefore, the former follows the latter by a few tens of milliseconds. Cells were selected when analysis of variance (ANOVA) showed that at least one test stimulus elicited activity significantly ($p < 0.025$) greater than that observed after the control stimulus. For each cell, responses were categorized into three groups: Best (response to the most effective stimuli), Worst (response to the least effective stimuli), and Mid (response between the former two). The three categories were defined as 20% ranges of the full range of responses to different stimuli: Worst, 0–20%; Mid, 40–60%; and Best, 80–100%. Of the selected cells, those that had responses of all three categories were picked out.

The responses obtained in five trials were averaged within each response category for each cell on a bin-by-bin basis (bin size was 5 ms). This yielded a peristimulus time histogram (PSTH) with 120 bins. The latency was taken as the first of three consecutive time bins in which the mean response was in excess of the 95% confidence interval of the S/A. If different categories gave different latencies, the shortest was taken as the cell response latency. The mean firing rate of the cells during the first, second, and fifth 100 ms of the response after the estimated cell response latency was also calculated.

Two response measurements introduced by Oram and Perrett (1992) were calculated: population response and average cell response. Population response is PSTH profiles calculated in the three response categories for the entire population of selected cells. They were obtained by three procedures: (1) normalizing the response magnitude of each cell to the magnitude of the difference between the S/A and the peak response of the Best category, then (2) averaging the response rate in each time bin across all cells, and finally (3) renormalizing as procedure 1.

Average cell response was calculated from PSTHs, each of which was shifted by the cell's estimated response latency. The shifted PSTHs were

processed according to the three procedures just described. Oram and Perrett introduced this synchronization of response onset because response onsets of cells of monkeys were fairly widely distributed, and it was necessary to examine the difference between cell responses following response onset.

To estimate the efficiency of discrimination between different response categories, the responses in each time bin were subjected to a two-way ANOVA with response category being a fixed factor and cell a random factor. Firing rates above S/A were used without normalizing the magnitude of responses.

The simulation provided data concerning concurrent activity in multiple cells, which was not recorded in the experiment. Concurrent cell responses to one activation pattern were examined at the model STPa. The activation pattern was that of Figure 4B. The concurrent cell response was calculated as follows: For each of the three response categories, 50 cells that made responses of the corresponding category were selected at random. The responses of the 50 cells were averaged within each response category for each trial on a bin-by-bin basis. The difference in average concurrent responses between the response categories was estimated on each bin for each the five trials with the use of ANOVA. For each category, mean firing rates of the corresponding 50 cells were used.

Synaptic coefficient W_k change caused by learning was calculated. After the learning period, values of W_k were averaged within each postsynaptic neuronal population to obtain the average synaptic efficacy between neuronal populations. Increments in the average synaptic efficacy from their initial values were calculated as percentages of the initial values. The same calculation was also performed after the first half of the learning period.

5 Results

At the end of learning, 2469 cells showed statistically significant responses to some stimuli; they were 98.7% of the pyramidal cells ($5 \times 5 \times 100$ cells) of the model STPa. Numbers of cells responding to the six activation patterns are listed in Table 1 for each the three response categories. Among them, 399 cells made responses of all of the Best, Mid, and Worst categories. In the experiment (Oram & Perrett, 1992), 44 cells were suitable for the data analysis. For comparison, 50 of the 399 cells were sampled at random.

Change in firing rates of those cells is listed in Table 2. The PSTHs of the cells were also calculated. A PSTH of a cell for the Best response category is illustrated in Figure 5A. The response showed a very rapid rise followed by a slower decline in firing rate, as did the monkey cells. The rate of rise depends on membrane parameters C, G_k, and E_k, which were set according to physiological data, and on synaptic efficacy W_k, which was adjusted by learning. The decline was produced by increasing the firing threshold for the model neurons. The rate of this decline was fitted to the empirical data

Table 1: Distribution of Cells Responding to the Six Input Activations.

Activation Pattern	Response Category		
	Best	Mid	Worst
A	1	0	398
B	210	99	90
C	0	89	310
D	139	160	49
E	49	0	325
F	0	112	250

Note: Activation pattern labels correspond to those in Figure 4.

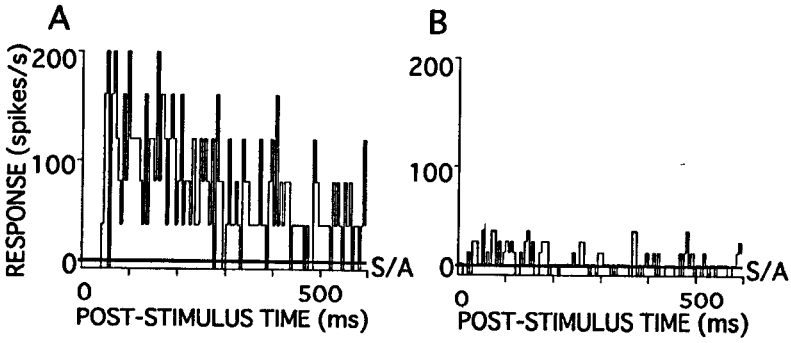


Figure 5: Peristimulus time histograms of the Best (A) and Worst (B) response categories of one cell.

by adjusting parameters Δ_h and C_h.[2] A PSTH of a cell for the Worst response category is shown in Figure 5B. The firing rate showed a small rise followed by a rapid and complete decay.

Means of the estimated response latency for the Best, Mid, and Worst response categories were, respectively, 41.2, 49.6, and 62.4 ms. Increments in the mean latency from that of the Best category are shown in Figure 6.

[2] There was still some disparity. Firing rates for the Best and Mid categories in the model were larger than those in the experiment, and for the Worst category they were smaller. Reduction of firing rates for the former categories made response for the Worst category statistically indistinguishable from the S/A. Since this model was intended to analyze latency competition that was related to response onset, only the decline in firing rate was modeled.

Table 2: Modeled and Measured Firing Rates (spikes per second).

	A. Model			B. Experiment		
Period	Best	Mid	Worst	Best	Mid	Worst
Peak	187.3	136.8	36.3	115.8	82.4	51.2
First 100 ms	94.6	59.8	17.1	66.1	43.6	24.3
Second 100 ms	82.7	34.1	12.0	45.1	31.8	16.0
Fifth 100 ms	32.9	14.4	9 6	27.7	19.7	13.1

Source: Oram & Perrett, 1992.

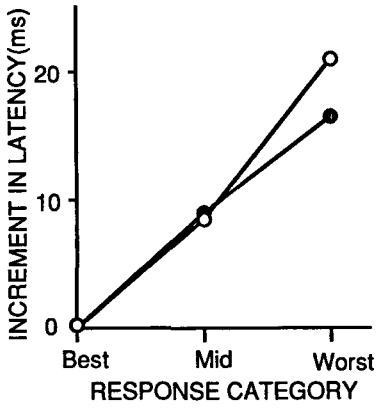


Figure 6: Increment in mean response latency versus the response category. Stimuli of the Best category fire cells at the greatest mean rate during the post-stimulus period of 250 ms, and those of the Worst category fire cells at the smallest mean rate. Open circles indicate data of the model simulation, and filled circles indicate the empirical data of Oram and Perrett (1992).

Weaker responses occur in longer latencies for both the model simulation and experiment. The mean response latency of cells of the monkey was 111.2 ms for the Best response category. Assuming that visual signals took 70 (111.2 – 41.2) ms to reach area V1, Figure 6 shows that response latencies of the model STPa are in good accordance with those of the monkey STPa.

5.1 Population Response. The amplitude-normalized responses (S/A, 0%; Peak from Best, 100%) were averaged by response category to produce the population PSTH profiles (see Figure 7A). They reproduce those of the experiment (see Figure 7B). Although the firing rates of model cells tend to

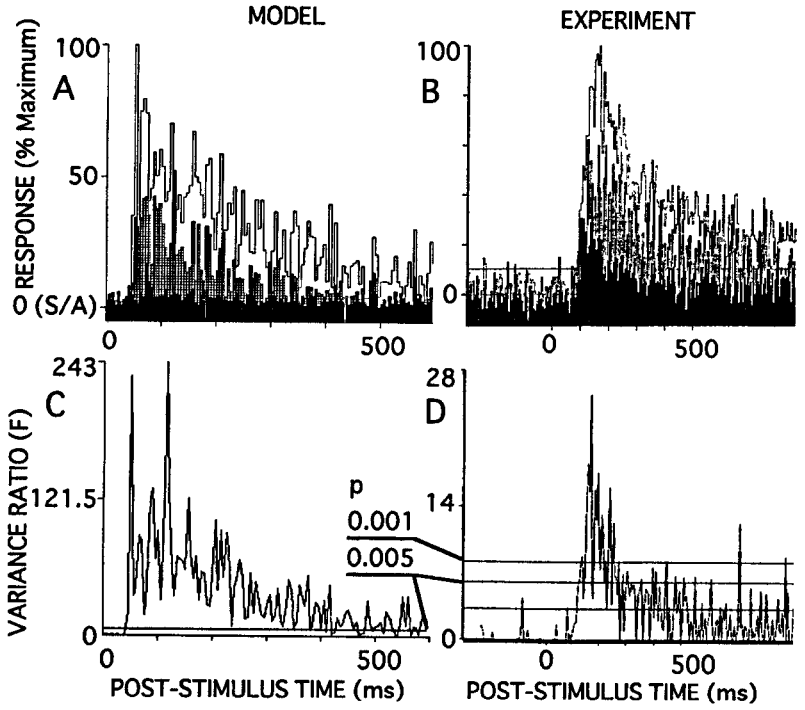


Figure 7: Cell population responses and discrimination between the responses. (A, B) PSTH response profiles in the model simulation and in the experiment (Oram & Perrett, 1992). Clear bars show the Best category, hatched bars show the Mid category, and solid bars show the Worst category. Firing rate is expressed as a percentage of peak response in the Best category. (C, D) The results of statistical evaluation of response discrimination in the simulation and the experiment: *F*-ratio computed for each time bin across the three categories. Discrimination reaches significance ($P < 0.05$) 45 ms after stimulus onset (activation of V1) in the simulation and 116 ms after stimulus onset (presentation of visual stimuli) in the experiment.

be higher than those of the monkey (see Table 2), the simulation results are fairly close to the empirical data.

A two-way ANOVA (fixed factor, response category; random factor, the 50 cells) was performed between the three response categories for each of the 120 time bins. Values of the *F*-ratio are plotted against time in Figure 7C. The discrimination between stimuli reaches a statistically significant level

45 ms after stimulus onset (activation of V1) and is very reliable, as in the experiment (see Figure 7D). The mean response latency after stimulus onset of the selected cells is 37.6 ms. In the experiment, the time needed for the discrimination measure to reach a significant level was 116 ms, and the response latency was 90 ms. Note that stimuli in the experiment were stimuli to the retina and that those times should therefore be longer in the empirical data than in the simulation. Significant discrimination in the model simulation occurred 7.4 ms after response onset and in the experiment 26 ms.

5.2 Average Cell Response. Some cells did not have a clear response onset for some categories. Nineteen cells had the defined response onset for all the response categories 20 ms after stimulus onset. Responses of those cells were synchronized at 20 ms after stimulus onset. That is, cell responses of the three categories were shifted by the difference between the cell's estimated response latency and 20 ms of poststimulus time. For those 19 cells, PSTH profiles were calculated after the synchronization of response onsets (see Figure 8A). The synchronization makes the Best, Mid, and Worst responses begin at the same latency, though responses of different categories could occur at different latency for each cell.

Discrimination between stimuli was again analyzed by ANOVA of firing rate in different time bins across the 19 cells. Changes in the *F*-ratio values with time are shown in Figure 8C. The discrimination reaches a statistically significant level within the first 5 ms of response onset, as it does in the empirical data (see Figure 8D).

5.3 Concurrent Cell Response. For each of five trials, the concurrent responses of the cells were averaged within each response category (see Figure 9A). Responses of the Best category rose first every trial. They were followed by responses of the Mid category. Some trials did not produce clear responses of the Worst category. All the responses were considerably oscillatory, and cells of the same response categories fired synchronously. This would correspond to the synfire patterns proposed by Abeles (Abeles et al., 1993). The oscillation did not appear in the population PSTH profiles because the responses there were averaged across the trials.

Discrimination among responses of the Best, Mid, and Worst categories was analyzed by ANOVA of firing rate in different time bins. The statistical evaluation is illustrated in Figure 9B. Statistically significant discrimination ($p < 0.005$) occurred at 50 ms for all the five trials, and the timing agreed with the onsets of the concurrent responses for the Best category.

5.4 Reinforcement of Synaptic Connectivity. Increments in the average synaptic efficacy were calculated after the learning period and after the first half of the learning period. Figure 10 shows the values of connections between neuronal populations in the second rows of the population matrices

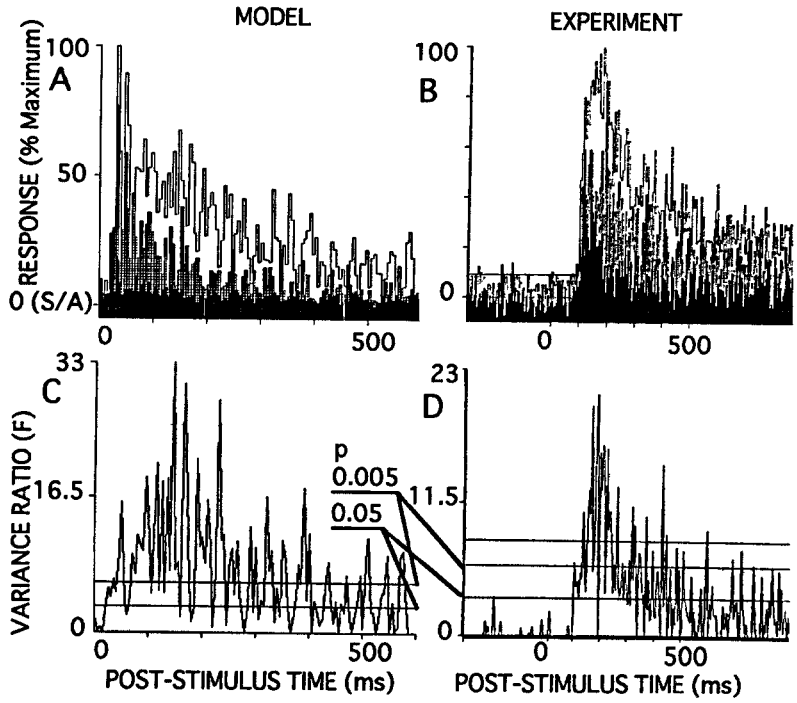


Figure 8: Average cell responses and discrimination between the responses. (A, B) PSTH response profiles in the model simulation and in the experiment (Oram & Perrett, 1992). Response latencies of contributing cells were synchronized to 20 ms (poststimulus). Best, Mid, and Worst categories are, respectively, denoted by clear, hatched, and solid bars. (C, D) The results of statistical evaluation of response discrimination in the simulation and the experiment: *F*-ratio computed for each time bin across the three categories. Discrimination reaches significance ($P < 0.05$) within 5 ms of response onset in both the simulation and the experiment.

of cortical areas. Dashed arrows indicate connections whose increments in average synaptic efficacy were between 3% and 30% of their initial values. Solid arrows indicate connections whose average synaptic efficacy increased by more than 30% of the initial values. As learning proceeded, pathways of reinforced connections extended from area V1 to area STPa.

as two

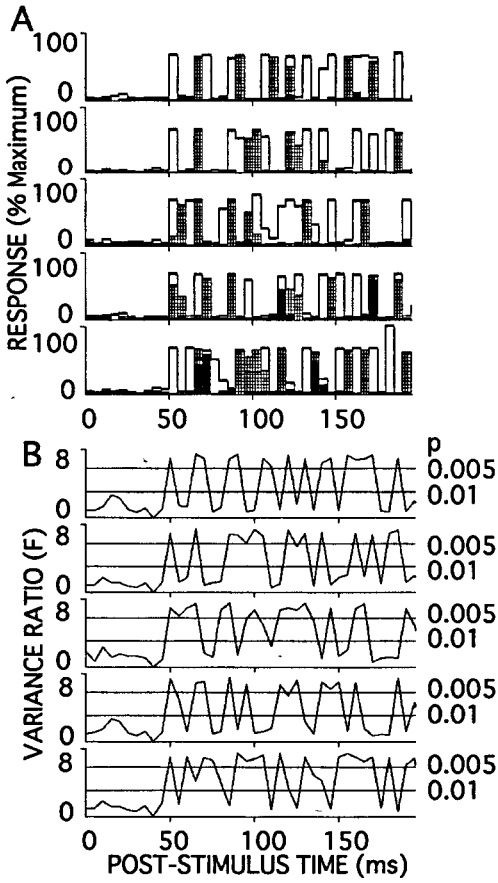


Figure 9: Average concurrent responses of multiple cells and discrimination between responses. (A) Responses averaged across 50 cells for five trials. Clear bars show the Best category, hatched bars show the Mid category, and solid bars show the Worst category. Firing rate is expressed as a percentage of peak response in the Best category (300 spikes/s). (B) Statistical evaluation of response discrimination: *F*-ratio computed for each time bin across the three categories.

as one

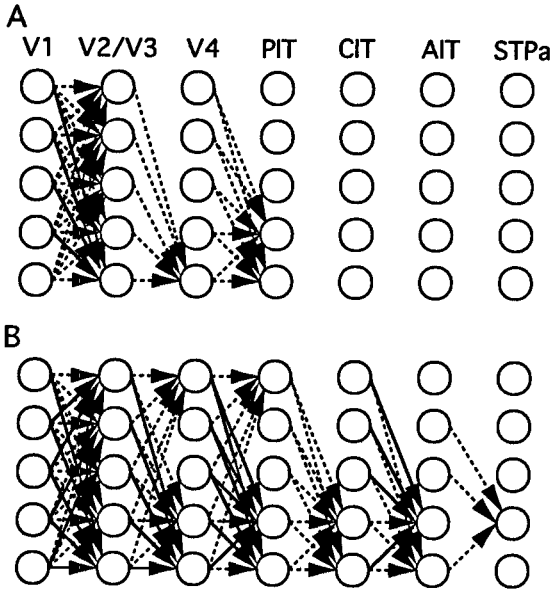


Figure 10: Reinforcement of corticocortical connections. Increments in the average synaptic efficacy after the first (A) and second (B) halves of the learning period. Circles represent neuronal populations of the second rows in the population matrices of cortical areas. Dashed arrows indicate connections whose increment in average synaptic efficacy was between 3% and 30% of their initial values. Solid arrows indicate connections whose average synapyic efficacy increased by more than 30% of the initial values.

6 Discussion

6.1 Latency and Response Discrimination. The mechanism of latency competition transmits the strongest activation first, so strong responses should occur at short latency. Unless the latency of strong responses is shorter, the hypothesis presented here does not hold. This requirement was verified in both the simulation results and the empirical data (see Figure 6). Differences of latency between response categories in the simulation were close to those in the empirical data. This suggests that the number of synaptic relays in the shortest pathways for the discrimination in the monkey STPa is nearly the same as that in the model circuit and therefore that the pathways of the monkey might be included in the ventral pathway of the temporal lobe of the model cortex.

Oram and Perrett (1992) calculated a measure of discrimination between the Best and Worst categories. The discrimination was expressed as $100 \times (R_B - R_W)/R_B$, where R_B and R_W are the mean firing rate levels above S/A in the Best and Worst categories during the first 100 ms of the response after the estimated cell response latency. They found no correlation between the discrimination measure and cell's response latency. The reason might be that the cell's response latency principally depends on the length of pathways transmitting the first spikes and is not related to the above discrimination measure. Instead, our model predicts that for each cell, responses of the Best category should come earlier than those of the other categories, no matter how long the cell's response latency is. This was verified by the data shown in Figure 6.

6.2 Robustness of Populational Behavior Against Noise. Neural mechanisms must work well in the noisy environment of the brain. Our model assumes that cortical circuits encode information in firing ratios of neuronal populations. The simulation has shown that in all the trials, the average concurrent response of neuronal populations began at the shortest latency to stimuli of the Best response category even though spontaneous activity (9 Hz) provided random spikes (see Figure 9A). This enables the mechanism of latency competition to work under the random spontaneous activity. In every trial, discrimination between stimuli of the three response categories reached a statistically significant level within 5 ms of response onset (see Figure 9B).

6.3 Robustness Against Latency Variability of Input Signals. The mechanism of latency competition may be suspected not to work correctly if input activations start after different latencies. The firing latency of retinal ganglion cells varied up to 30 ms with changes in contrast. Robustness of the latency competition against this variability was examined for two typical cases. The first is that every neuronal population of area V1 receives projections from cells with different firing latencies. Figures 11A and B show a result of simulation, where the circuit of Figure 1B receives inputs, each of which is produced by 100 cells with different firing latencies. The latencies of the first spikes distribute in a gaussian fashion: mean $\pm$ SD $= 15 \pm 5$ ms. Firing ratios of afferents increase not stepwise but gradually (see Figure 11A). Even in response to gradually rising inputs, the circuit allowed only the most strongly activated populations to fire (see Figure 11B). That is, the latency competition correctly detected the strongest activations.

The second case is that each population of area V1 receives projections from cells of the same latency and the latency varies from population to population. Neuronal populations of area V1 receive stepwise inputs of different latencies (see Figure 11C). If strong inputs start early, they certainly win the latency competition. The simulation investigated the opposite case: Weak inputs start early. In the simulation, every interval between input on-

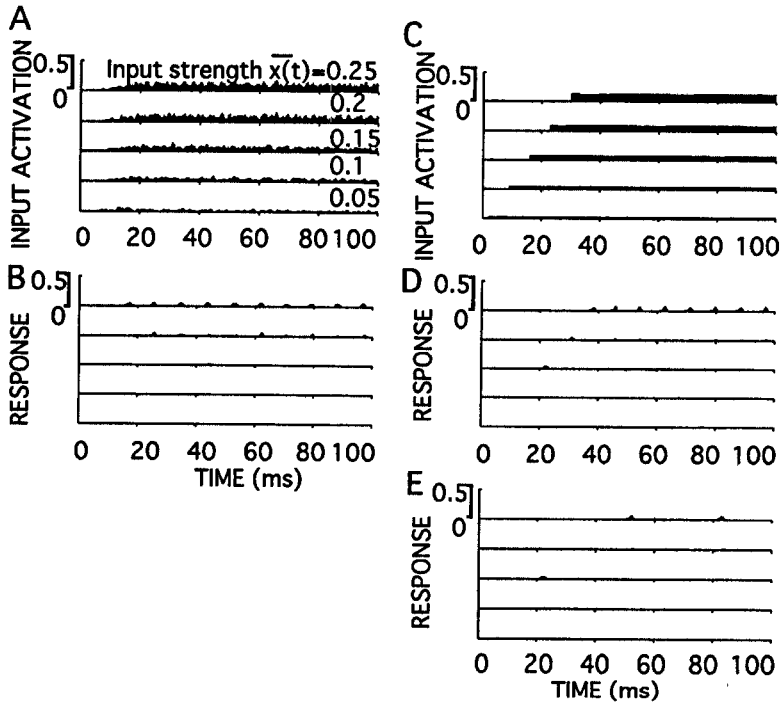


Figure 11: Robustness of latency competition against input latency variability. The abscissa and the ordinate, respectively, indicate time and the ratios of firing neurons in the populations. (A) Input activations, each of which is produced by 100 cells with different firing latencies. Mean firing ratio tends to 0.05 for the bottom row, and limit values increase to 0.25 for the top row. (B) Responses of the circuit of cortical area shown in Figure 1B to inputs of A. (C) Input activations, each of which is produced by 100 cells with the same firing latency. The latency is the shortest for the bottom row, increases by 7 ms, and is the longest for the top row. Firing ratio is 0.05 for the bottom row and increases to 0.25 for the top row. (D) Responses of the circuit of cortical area shown in Figure 1B to inputs of C. Duration of inhibition is 4 ms. (E) Responses of the circuit of cortical area shown in Figure 1B to inputs of C. Duration of inhibition is 25 ms.

sets was 7 ms, and the weakest input began 28 ms earlier than the strongest input. Figure 11D shows the responses. The weakest input was too weak to fire any cell. The second and third weakest inputs activated some cells and laterally inhibited all the populations. At release from the inhibition,

the fourth weakest (second strongest) input had begun, and the four inputs competed. The two strongest excited their target populations and again inhibited all the populations. When the second inhibition ended, the strongest input had been on, and all the inputs participated in the competition. The result was that the two strongest inputs produced oscillatory responses. As above, the latency competition resulted in the same response that was produced by simultaneous inputs after all the inputs had been on. In the early stage, weaker inputs fired some cells. The irrelevant response would be reduced, however, if the inhibition were of long duration. If the oscillation of 40 Hz at area V1 was produced by lateral inhibition, its estimated duration would be 25 ms. Figure 11E shows a simulation result where inhibition lasts 25 ms. Although small, irrelevant responses were made in the first competition, the circuit correctly detected the strongest activations after the second competition. We see that except in the first competition, the latency competition makes relevant responses even to inputs with an onset variability of nearly 30 ms.

In the discussion, input activations were not oscillatory. This is plausible for inputs to area V1, but activation of the subsequent areas may be oscillatory. We will see here that neuronal populations receive nonoscillatory inputs even if neuronal populations of the preceding area deliver oscillatory activations. Every neuronal population receives projections from many populations in the preceding area (see Figure 12A). Therefore, the activation the population receives is a superposition of activations from the multiple populations. Suppose each of the activations is oscillatory. Intervals of the oscillation are determined by the interval of inhibition. If we assume that the interval differs in the inhibitory neuronal populations, the superposition of oscillatory activations becomes less oscillatory. Figure 12B shows a superposition of three oscillatory activations. It is less oscillatory than the original three. A superposition of more activations is expected to be close to a step input.

The discussion also suggests a possible way of integrating signals conveyed by different pathways. The signals could arrive after different latencies at areas where they are integrated. For example, auditory signals arrive at the STS cells about 100 ms earlier than visual signals (Bruce, Desimone, & Gross, 1981). Despite the latency difference, animals can integrate those signals to respond in a few hundred milliseconds. Suppose that a cortical area receives input along two cortical pathways (see Figure 12C) and that the first pathway delivers activation earlier than the second pathway. Also assume that the first pathway activates the two neuronal populations more strongly than the other three populations, but activation of both the pathways provides more synaptic input to the latter three populations than to the former two. Figure 12D shows the behavior of the model cortical area. The responses of the former and the latter populations are, respectively, shown in the top two rows and the bottom three rows. Before activation of the second pathway, the two neuronal populations strongly activated by

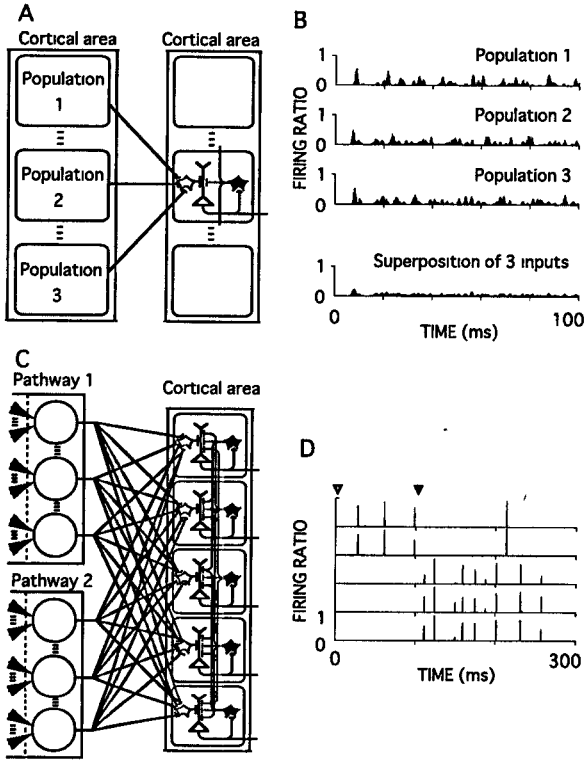


Figure 12: Integration of signals from different cortical pathways. (A) Corticocortical connections in a single pathway. Three activated neuronal populations project to stellate cells of a population in the next area. (B) Oscillatory activations of the three corticocortical projections in A (top three rows) and the superposition of them (bottom row). The refractory periods of inhibitory neurons in the model cortical areas differ from neuron to neuron; the distribution of lengths of the refractory periods is gaussian (mean $\pm$ SD $= 10.0 \pm 3.33$ ms). (C) Two cortical pathways converging at a cortical area. The first pathway activates the two neuronal populations more strongly than the other three populations, but activation of both the pathways provides more synaptic input to the latter three populations than to the former two. (D) Responses of neuronal populations of a cortical area that receives activations of two pathways at different onset times. Open and solid triangles, respectively, indicate onset times of activations of the first and the second pathways. The responses of the populations strongly activated by the first pathway are shown in the top two rows, and the responses of the populations strongly activated under activation of both the pathways are shown in the bottom three rows.

the first pathway won the competition and made oscillatory responses. After the second pathway also began to fire, the other three populations won the competition and produced oscillatory firing. Again, the mechanism of latency competition correctly detected the strong activation from the integrated signals, except in the first few competitions. If the motor system needs to accumulate some neural activation in order to move, irrelevant responses of the first competitions could be neglected, and relevant motor response would be made.

From the above consideration, we see that the latency competition is not only performed by the first spikes but is repeated in oscillation by the subsequent spikes. This makes the competition mechanism robust against some latency variability of neural inputs.

6.4 Feedback Connections. There are many feedback connections in the cortical circuitry, though the model presented here does not include them. This study focused on processing of the millisecond time range, which biological time constraints suggest should be performed without using feedback connections (Oram & Perrett, 1994). They might, however, accelerate processing. Auditory signals activate cells of the STS earlier than visual signals, and feedback connections could transmit the activation to earlier stages of visual pathways such as the AIT and MST. This might increase the excitability of those areas and make them respond faster to visual stimuli. It is expected that feedback connections play different roles in different aspects of cortical processing. How the mechanism of latency competition might cooperate with other cortical mechanisms is an issue for future research.

Appendix A

Suppose $x(t) = \mathbf{x}_0$ for $t < 0$ and $x(t) = \mathbf{x}$ for $t \geq 0$. Let $A(x(t)) = \sum_k (W_k x(t) + 1) G_k^o / C$ and $V(x(t)) = \sum_k (W_k x(t) + 1) G_k^o E_k / (CA(x(t)))$. Assuming V is constant for $t < 0$, we obtain from equation 2.1

$$V = \begin{cases} V(\mathbf{x}_0) & \text{if } t < 0 \\ V(\mathbf{x}) + (V(\mathbf{x}_0) - V(\mathbf{x})) \exp(-A(\mathbf{x})t) & \text{if } t \geq 0 \end{cases}. \tag{A.1}$$

Let T_0 be a time interval required for the neuron to fire the first time. Since the synaptic input is excitatory, $V(\mathbf{x}) > h$. Setting $V = h$ in equation A.1 yields

$$T_0 = A(\mathbf{x})^{-1} \log \frac{V(\mathbf{x}) - V(\mathbf{x}_0)}{V(\mathbf{x}) - h}. \tag{A.2}$$

Let T be the time intervals required to fire after the first firing. It is given by

$$T = A(\mathbf{x})^{-1} \log \frac{V(\mathbf{x}) - V_o}{V(\mathbf{x}) - h}. \tag{A.3}$$

It follows that the uth firing begins at

$$t_u = T_0 + (u-1)(\tau_h + \tau_r + T), \quad u = 1, 2, \ldots \tag{A.4}$$

Suppose that the neuron is producing the uth spike at time t; that is, $t_u \le t$ and $t_u > max[0, t - \tau_h]$. Equations A.2, A.3, and A.4 indicate that t_u is a function of (W_k). Condition 1 reduces $\sum_k W_k x(t)$ to $W_{k_a} x(t)$ in $V(x(t))$ and $A(x(t))$. It follows that t_u depends only on W_{k_a} among W_k. By counting neurons that are producing the uth spikes at time t, we can calculate the ratio of firing neurons in the population, which is denoted by $y(t)$:

$$y(t) = \sum_{u=1}^{u_e - 1} \int_{W_{k_a} \in \{W_{k_a} | t_u \le t, t_u \ge t - \tau_h\}} f(W_{k_a}) dW_{k_a} + \int_{W_{k_a} \in \{W_{k_a} | t_{u_e} \le t, t_{u_e} \ge max[0, t - \tau_h]\}} f(W_{k_a}) dW_{k_a}, \tag{A.5}$$

where $f(W_{k_a})$ denotes the probability density of neurons receiving connections with coefficient W_{k_a} in the populations, and u_e is the maximum integer not more than $t/(\tau_h + \tau_r + T) + 1$. Let $y_1(t)$ and $f_1(W_{k_a})$ denote values of $y(t)$ and $f(W_{k_a})$ when W_{k_a} is larger by ΔW_{k_a}. Condition 2 implies that $f_1(W_{k_a}) = f(W_{k_a} - \Delta W_{k_a})$. Assuming that ΔW_{k_a} is small, we have

$$\begin{aligned} y_1(t) &\simeq y(t) - \Delta W_{k_a} \frac{dy(t)}{dW_{k_a}} \\ &= y(t) - \Delta W_{k_a} \frac{dy(t)}{dt} \frac{dt}{dt_u} \frac{dt_u}{dW_{k_a}}. \end{aligned} \tag{A.6}$$

Let us define the response latency of neuronal populations, which is denoted by t_1^*, as the time when the value of $y(t)$ reaches a certain value y^* for the first time. Since $y(t)$ is not decreasing at t_1^*, $dy(t)/dt \ge 0$ at t_1^*. $dt/dt_u = 1$. From equation A.4, $dt_u/dW_{k_a} = dT_0/dW_{k_a} + (u-1)dT/dW_{k_a}$.

$$\frac{dT_0}{dW_{k_a}} = \frac{-1}{G_{k_a}^0 A^2(\mathbf{x})} \left[\frac{G_{k_a}^0 (h - V(\mathbf{x}_0))(E_{k_a} - V(\mathbf{x}))}{C(V(\mathbf{x}) - V(\mathbf{x}_0))(V(\mathbf{x}) - h)} + \frac{G_{k_a}^0}{CA(\mathbf{x}_0)} \frac{A(\mathbf{x})(E_{k_a} - V(\mathbf{x}_0))}{V(\mathbf{x}) - V(\mathbf{x}_0)} + \frac{G_{k_a}^0}{C} \log \frac{V(\mathbf{x}) - V(\mathbf{x}_0)}{V(\mathbf{x}) - h} \right]. \tag{A.7}$$

Since the afferents are excitatory, $E_{k_a} \ge V(\mathbf{x}) \ge h > V(\mathbf{x}_0)$. It follows from equation A.7 that $dT_0/dW_{k_a} < 0$. Similarly, $dT/dW_{k_a} < 0$ because $V(\mathbf{x}_0) \ge V_o$. From these, we have $dt_u/dW_{k_a} < 0$. It follows from equation A.6 that $y_1(t_1^*) \ge y(t_1^*)$. Consequently, the value of $y_1(t)$ reaches y^* not later than t_1^*. Let t_2^* denote the time when the value of $y_1(t)$ reaches y^* for the first time.

$t_2^* \leq t_1^*$. Let $y_2(t)$ and $f_2(W_{k_a})$ denote values of $y(t)$ and $f(W_{k_a})$ when W_{k_a} is larger by $2\Delta W_{k_a}$. The same consideration results in $y_2(t)$ reaching y^* not later than t_2^* and so not later than t_1^*. Similarly, whenever W_{k_a} is larger by any value, the value of $y(t)$ with the larger W_{k_a} reaches y^* not later than t_1^*.

Appendix B

Theorem 1 ensures that neuronal populations receiving the strongest activation produce spikes first. The first spikes excite the inhibitory interneurons if W_{k_a} of the connections with the inhibitory neurons is larger than the following value. Consider the weakest activation, that is, that only one spike is delivered. Then input $\mathbf{x} = 1/n$, where n denotes the number of neurons of the projecting population. If T given by equation A.3 is shorter than spike duration τ_s, the neuron is certain to fire. Let $T(1/n) = A(1/n)^{-1} \log[V(1/n) - V_o]/[V(1/n) - h]$. $T(1/n)$ decreases as W_{k_a} increases (see appendix A). It follows that there is some value of W_{k_a} that makes $T(1/n) = \tau_s$. If W_{k_a} is larger than that value, the inhibitory neuron is certain to fire. If the input stays larger than $\mathbf{x}$ for a period of τ_e, the condition $T(1/n) = \tau_s$ is replaced by $T(\mathbf{x}) = \tau_e$.

Assume the inhibitory spikes act only on the k_ith ion conductance of pyramidal cells. This appendix will show that firing of the inhibitory neurons reduces the membrane potential of pyramidal cells close to the inhibitory electromotive force E_{k_i} ($<V_o$) if the inhibitory connections are strong enough. Suppose only one inhibitory neuron fires, and assume that inhibitory neurons produce a train of spikes every time they fire (Kawaguchi, 1995). Let n_t and τ_t, respectively, be the number of spikes in single trains and an interval between spike discharges. Then the strength of inhibitory input is $1/n_t$ during spikes and 0 between spikes, where n_t denotes the number of inhibitory neurons in the population. Let $\hat{x}_{k_a}$ denote the maximum strength of excitatory input x_{k_a} during the inhibition. It follows from equation A.1 that the value of V at the end of the first spike, which is denoted by V_1, is in the range defined by

$$V_1 \leq V(\hat{x}_{k_a}, 1/n_t)(1 - \exp(-A(\hat{x}_{k_a}, 1/n_t)\tau_s)) + V_1^* \exp(-A(\hat{x}_{k_a}, 1/n_t)\tau_s), \tag{B.1}$$

where $V(\hat{x}_{k_a}, 1/n_t)$ and $A(\hat{x}_{k_a}, 1/n_i)$, respectively, denote $\sum_k (W_k x_k + 1) G_k^o E_k / (CA(x_{k_a}, x_{k_i}))$ and $\sum_k (W_k x_k + 1) G_k^o / C$, where $x_{k_a} = \hat{x}_{k_a}$, $x_{k_i} = 1/n_i$, and $x_k = 0$ for the other k, and where V_1^* denotes the value of V at the onset of the first spike:

$$\frac{dV_1}{dW_{k_i}} = \frac{G_{k_i}^o (E_{k_i} - V(\hat{x}_{k_a}, 1/n_t))}{n_i CA(\hat{x}_{k_a}, 1/n_i)} (1 - \exp(-A(\hat{x}_{k_a}, 1/n_t)\tau_s)) + (V(\hat{x}_{k_a}, 1/n_t) - V_1)(G_{k_i}^o / n_t C) A(\hat{x}_{k_a}, 1/n_i) \times \tau_s \exp(-A(\hat{x}_{k_a}, 1/n_t)\tau_s). \tag{B.2}$$

The k_ith input being inhibitory means that $V(\hat{x}_{k_a}, 1/n_i)$ decreases to E_{k_i} as W_{k_i} increases and therefore $E_{k_i} < V(\hat{x}_{k_a}, 1/n_i)$. If $V_1^* > E_{k_i}$, there is some value of W_{k_i} such that $V(\hat{x}_{k_a}, 1/n_i) < V_1^*$. It follows from these that there is some value of W_{k_i} such that $dV_1/dW_{k_i} < 0$. Consequently, if W_{k_i} is above that value, V_1 decreases as W_{k_i} increases. Let V_2^* denote the value of V at the onset of the second spike. From equation A.1, we have

$$V_2^* \leq V(\hat{x}_{k_a}, 0) + (V_1 - V(\hat{x}_{k_a}, 0)) \exp(-A(\hat{x}_{k_a}, 0)(\tau_t - \tau_s)). \tag{B.3}$$

Only V_1 depends on W_{k_i} in equation B.3 and decreases as W_{k_i} increases. There is some value of W_{k_i} such that $V_2^* < V_1^*$. The value ensures that the neuron never fires before the onset of the second spike. Similarly, let V_j and V_j^*, respectively, denote the values of V at the ends and the onsets of the jth spikes, $j = 1, 2, \ldots, n_t$. There are some values of W_{k_i} such that V_j decreases as W_{k_i} increases and $V_j > V_{j+1}^*$. And from equation B.1, we know that V_j tends to E_{k_i} as W_{k_i} increases. It follows that there is some value of W_{k_i} such that the neuron does not fire during the spike train and V_{n_t} is close to $E_{k_i} (< V_o)$. From the assumption that the length of the spike trains, $n_t \times \tau_t + \tau_s$, is longer than the refractory period of pyramidal cells, τ_r, it follows that at release from the inhibition, pyramidal cells that produced spikes have recovered from the refractory state. Since those cells receive inhibition after the recovery, their membrane potentials are not more than the resting potential V_o. And they are larger than those of the unfired cells because the fired cells are assumed to neglect any input (including inhibitory input) during the refractory period and so do not receive inhibition any longer than the unfired cells. After release from the inhibition, the second competition takes place, in which the previously fired cells fire more quickly than the previously unfired cells. Again, only the most strongly activated populations are fired.

Appendix C

The model neuron is assumed to have Na^+, K^+, and Cl^- channels. Values of G_k^o are 0.888888 $(M\Omega cm^2)^{-1}$ for Na^+, 44.4444 $(M\Omega cm^2)^{-1}$ for K^+, and 154.666 $(M\Omega cm^2)^{-1}$ for Cl^-. Values of E_k are 60.0 mV for Na^+, −90.0 mV for K^+, and −65.0 mV for Cl^-. The initial threshold potential for firing, h_o, is −40.0 mV. $\Delta_h = 0.1$ mV/spike. $C_h = 6.0 \times 10^{-4}$. $V_o = -70.0$ mV. $C = 1.0\mu$F. $\tau_s = 1.0$ ms. At every firing, excitatory neurons produce a single spike, and inhibitory neurons produce a train of spikes (three spikes in this simulation). The interspike interval in the train is 1.5 ms. $\tau_r = 4.0$ ms for excitatory neurons, and the τ_r of inhibitory neurons varies with a normal distribution, from neuron to neuron. The mean is 10.0 ms, and the standard deviation (SD) is one-third of the mean. The probability that the variable is below the mean $-3 \times$ SD is 0.2% and each neuronal population has 100 neurons. It follows that the neuronal populations rarely include neurons with negative τ_r.

Synaptic efficacy should differ from neuron to neuron, so values of W_k were assumed to be normally distributed. The mean values were chosen so that neurons may fire at physiologically plausible rates. The means were 2500 for projections of stellate cells to pyramidal cells and for projections of pyramidal cells to inhibitory cells. The projections with synaptic efficacy of those mean values fire about 3% of the neurons in their target populations when the projections are activated at the ratio of $x(t) = 0.03$. Lateral connections of inhibitory neurons diverge exponentially. Means of W_k decline at a rate of $\exp(C_i(l_x^2 + l_y^2))$ where C_i is a coefficient, l_x is the distance between the projecting population and target population on the x-axis, and l_y is that distance on the y-axis. The mean of W_k of the lateral connections is 2 at $l_x = l_y = 0$ and $C_i = 0.5$. Corticocortical connections between areas were initially so weak that no cortical cells might respond to any activation of area V1. They also diverge exponentially, at a rate given by $\exp(C_c(l_x^2 + l_y^2))$, where $C_c = 1.0$ and the mean of W_k of the corticocortical connections is 250 at $l_x = l_y = 0$. Projections to area V1 are strong enough to fire stellate cells of area V1: the means of W_k are 3000, and all SDs are one-third of the means. $\hat{W}_k$ are two times as large as the means. The time needed for spikes to travel along the connections is 0.2 ms inside the cortical areas, 6.0 ms between the sequential cortical areas, and 9.0 ms along the shortcut connections.

The coefficient of reinforcement $C_r = 10.0$ and the coefficient of decay $C_d = 2.43 \times 10^{-5}$. The upper limit of synaptic efficacy $\hat{W}_k$ is twice the initial mean value.

Acknowledgments

I thank M. W. Oram and D. I. Perrett for sending copies of their empirical data. This research was supported in part by Grants in Aid for Scientific Research, Ministry of Education, Science and Culture of Japan, and Research for the Future Program, Japan Society for the Promotion of Science.

References

Abeles, M., Vaadia, E., Bergman, H., Prut, Y., Haalman, I., & Slovin, H. (1993). Dynamics of neuronal interactions in the frontal cortex of behaving monkeys. *Concepts in Neurosci., 4,* 131–158.

Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the Macaque. *J. Neurophysiol., 46,* 369–384.

Coultrip, R., Granger, R., & Lynch, G. (1992). A cortical model of winner-take-all competition via lateral inhibition. *Neural Networks, 5,* 47–54.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern., 36,* 193–202.

Gordon, B., Allen, E. E., & Trombley, P. Q. (1988). The role of norepinephrine in

plasticity of visual cortex. *Prog. Neurobiol., 30,* 171–191.
Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognition, 11,* 23–63.
Kalaska, J. F., & Crammond, D. J. (1992). Cerebral cortical mechanisms of reaching movements. *Science, 255,* 1517–1523.
Kang, Y., Kaneko, T., Ohnishi, H., Endo, K., & Araki, T. (1994). Spatiotemporally differential inhibition of pyramidal cells in the cat motor cortex. *J. Neurophysiol., 71,* 280–293.
Karil, R. E., Dubin, M. W., Scott, G., & Stark, L. A. (1986). Elimination of action potentials blocks the structural development of retino-geniculate synapses. *Nature, 323,* 156–158.
Kawaguchi, Y. (1995). Physiological subgroups of nonpyramidal cells with specific morphological characteristics in layer II/III of rat frontal cortex. *J. Neurosci., 15,* 2638–2655.
Kobatake, E., Tanaka, K., Wang, G., & Tamori, Y. (1993). Effects of adult learning on the stimulus selectivity of cells in the inferotemporal cortex. *Soc. Neurosci. Abstr., 19,* 975.
Miyashita, Y. (1990). Associative representation of the visual long-term memory in the neurons of the primate temporal cortex. In E. Iwai & M. Mishkin (Eds.), *Vision, memory and the temporal lobe* (pp. 75–87). New York: Elsevier.
Miyashita, Y., & Chang, H.-S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature, 331,* 68–70.
Nakamura, K. (1992). Neuropopulational mechanism of parallel cortical processing in the millisecond range. *Soc. Neurosci. Abstr., 18,* 1210.
Nakamura, K. (1993a). Temporal competition as an optimal parallel processing of the cerebrohypothalamic system. *Proc. IEEE Int. Conf. on Neural Networks, 1,* 64–69.
Nakamura, K. (1993b). A theory of cerebral learning regulated by the reward system I: Hypotheses and mathematical description. *Biol. Cybern., 68,* 491–498.
Nakamura, K., & Ichikawa, A. (1989). Timing information in transient behavior of neuropopulations. *IEEE Trans. on Syst., Man, and Cybern., 19,* 1188–1193.
Opara, R., & Wörgötter, F. (1996). Using visual latencies to improve image segmentation. *Neural Computation, 8,* 1493–1520.
Oram, M. P., & Perrett, D. I. (1992). Time course of neural responses discriminating different views of the face and head. *J. Neurophysiol., 68,* 70–84.
Oram, M. P., & Perrett, D. I. (1994). Modeling visual recognition from neurobiological constraints. *Neural Networks, 7,* 945–972.
Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition fundations* (Vol. 1, pp. 151–193). Cambridge, MA: MIT Press.
Sakai, K., & Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. *Nature, 354,* 152–155.

Received October 15, 1996; accepted July 15, 1997.

LETTER Communicated by Sebastian Seung and Miguel Nicolelis

Temporal-Code to Rate-Code Conversion by Neuronal Phase-Locked Loops

Ehud Ahissar
Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel

Peripheral sensory activity follows the temporal structure of input signals. Central sensory processing uses also rate coding, and motor outputs appear to be primarily encoded by rate. I propose here a simple, efficient structure, converting temporal coding to rate coding by neuronal phase-locked loops (PLL). The simplest form of a PLL includes a phase detector (that is, a neuronal-plausible version of an ideal coincidence detector) and a controllable local oscillator that are connected in a negative feedback loop. The phase detector compares the firing times of the local oscillator and the input and provides an output whose firing rate is monotonically related to the time difference. The output rate is fed back to the local oscillator and forces it to phase-lock to the input. Every temporal interval at the input is associated with a specific pair of output rate and time difference values; the higher the output rate, the further the local oscillator is driven from its intrinsic frequency. Sequences of input intervals, which by definition encode input information, are thus represented by sequences of firing rates at the PLL's output. The most plausible implementation of PLL circuits is by thalamocortical loops in which populations of thalamic "relay" neurons function as phase detectors that compare the timings of cortical oscillators and sensory signals. The output in this case is encoded by the thalamic population rate. This article presents and analyzes the algorithmic and the implementation levels of the proposed PLL model and describes the implementation of the PLL model to the primate tactile system.

1 Introduction

The distinction between rate and temporal coding is not always clear (Theunissen & Miller, 1995). For example, temporal coding is sometimes regarded as rate coding with a fine time resolution. In this article, temporal coding will refer to coding in which the exact time of every spike is informative. Rate coding will be associated here with a temporal window, the rate bin, within which the exact temporal information is not informative and the information is carried by the average firing rate over the entire temporal window. The rate bin is usually determined by the integration times of the readout mechanisms. A rate encoded signal can thus be described by a series of numbers, each of which represents the average firing rate in a

single rate bin (see appendix A.1). Fluctuations in the average firing rate of a neuron over different rate bins are considered here as fluctuations of rate-encoded information, and not as temporal coding, as has been considered previously (Richmond & Optican, 1987). A temporally encoded signal is described by a series of numbers, each of which represents either the timing of a single spike or a single interspike interval (ISI; see appendix A.1). The information contained in the spiking times can be presented in different ways; two of them are depicted in Figure 1: $M(n)$ describes the deviations of the actual train from an imaginary, ideally periodic, "carrier" train and $I(n)$ describes the ISIs. Figure 1 also demonstrates the distinction between temporal and rate coding; the spike train in this example carries a significant amount of information if a temporal coding is assumed (see Figure 1a), but almost no information if a rate coding with a particular rate bin is assumed (see Figure 1b). Practically, this distinction is important for reading out the information of the spike train. A readout mechanism based on rate will lose more and more information as its integration time increases. To read out temporally encoded information, a rate-based mechanism needs to employ integration times shorter than half of the input temporal resolution, an implementation that is both nonefficient and, with fine input resolution, not practical for neurons. The other alternative is to utilize preprocessing by time-sensitive mechanisms—mechanisms that produce populations of spikes, where the number of spikes in a population directly represents the ISI at the input and the exact times of these output spikes is not important.

Another important distinction is between peripheral and central temporal encodings (Perkel & Bullock, 1968). When a stimulus is temporally encoded at the periphery, the peripheral ISIs directly describe stimulus features such as spatial periods (Darian-Smith & Oke, 1980) whereas when stimuli are temporally encoded centrally, the centrally generated temporal structures are usually not directly related to the stimulus (Engel, Konig, Kreiter, Schillen, & Singer, 1992; Abeles, Bergman, Margalit, & Vaadia, 1993). This article deals with central decoding of peripherally generated temporal encodings. Using the term *decoding* in neuronal contexts should not imply that the original signals are reconstructed, but rather that the encoded information is extracted (Perkel & Bullock, 1968).

In mammals, sensory information is encoded by both rate and temporal coding (Sejnowski, 1995; Carr, 1993; Middlebrooks & Green, 1991; Wang, Merzenich, Beitel, & Schreiner, 1995; Johansson & Vallbo, 1983; Berkley, 1978). Whereas spatial static information is usually encoded by rate, dynamic information, generated during movements of either the stimulus or the sensory organ, is encoded also by temporal cues (see, for example, encoding of spatial intervals by ISIs of tactile [Darian-Smith & Oke, 1980] and visual [Shadlen & Newsome, 1994] neurons). In contrast, motor control is assumed to utilize rate coding predominantly (Georgopoulos, 1986; Fetz, 1993; Wise, 1993), even at the early stage of motor planning (Boussaoud & Wise, 1993). Thus, information carried by the sensory temporal

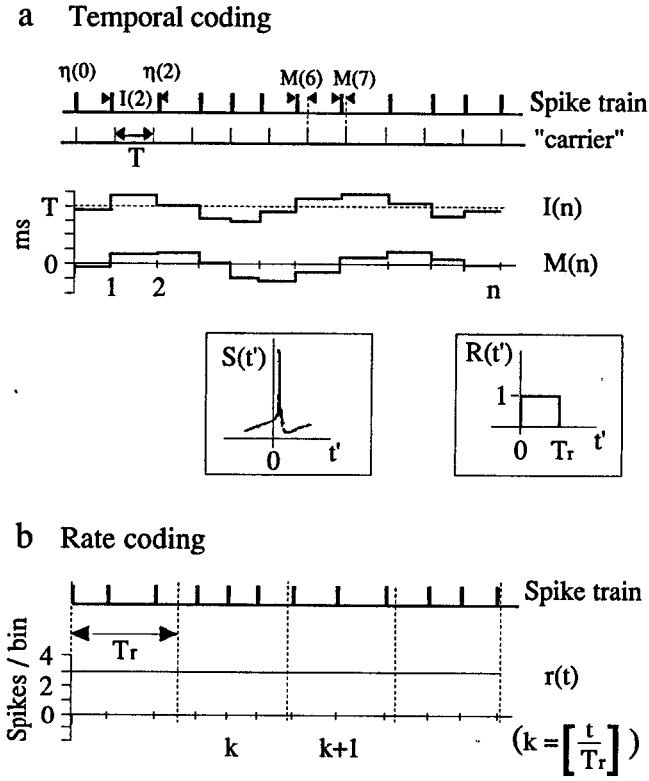


Figure 1: Encoding by spike trains. (a) Temporal encoding. Two possible presentations of the information contained in the spiking times: a series of ISIs $[I(n)]$ and a series of absolute modulations $[M(n)]$. $M(n)$ is the difference between the actual timing of the nth spike and the timing expected by an imaginary, ideally periodic, "carrier" train that has the same average periodicity and no modulation (see appendix A1). (b) Rate encoding. The spike train is divided into several rate bins (four in this case), and the total number of spikes in each bin determines the value of the rate function $r(t)$ for that bin. Insets: $S(t')$, a spike triggered at $t' = 0$; $R(t')$, a step function with a unity gain and a duration of a rate bin, T_r.

components is probably translated, by neuronal circuits in the brain, to rate-encoded signals that are "readable" by the motor system. If such a translation occurs early in a sensory pathway, the translation would also facilitate integration of temporally encoded information with other, rate-encoded sensory information. This necessity for translation was elegantly demonstrated by Mountcastle and his colleagues (Talbot, Darian-Smith, Ko-

rnhuber, & Mountcastle, 1968; Mountcastle, 1993) over the past few decades.

A mechanism that utilizes neuronal delay lines to transform temporal coding to rate coding has been suggested by Jeffress (1948). Such delay lines exist in the electric sensory system of electric fishes and in the subcortical auditory systems of birds and mammals (reviewed in Carr, 1993). These delay lines are probably utilized to decode temporal disparities, which in the submillisecond and low millisecond ranges can determine interaural time differences and echo delays, respectively. As the delay increases above a few hundred microseconds, implementations of delay lines require multiple neuronal elements, and the accuracy decreases (Carr, 1993). A mechanism that uses synaptic time constants appears more suitable to decode temporally encoded information in the millisecond range (Buanomuno & Merzenich, 1995). Both of these mechanisms describe "passive," open-loop decoding schemes that are based on classification of different ISIs according to their interaction with predetermined neuronal temporal features. In this article, I suggest an "active," closed-loop decoding mechanism, which dynamically adapts its working parameters to match the incoming signal. This phase-locked loop (PLL) model is based on a local oscillator "measuring" the instantaneous temporal period of the input by comparing it to its own period. During decoding, the local oscillator updates its period according to the result of the comparison, such that it remains tuned to the changing input. The PLL is a well-known mechanism in electrical engineering where it is often used for the decoding of phase-modulated signals. The algorithm presented here was adopted from that of continuous-time electronic PLLs (Gardner, 1979) and modified to fit discrete-time neuronal PLLs. This approach uses a small neuronal network as a PLL, unlike a previous approach that described a single-neuron as a PLL (Hoppensteadt, 1986).

Neuronal PLLs appear suitable to decode temporally encoded information in the range of a few to a few hundred milliseconds. However, decoding by a single neuronal PLL is usually limited to phase modulations that are in the order of its intrinsic period. Thus, decoding in different frequency ranges requires different PLL circuits, and decoding of large modulations requires an ensemble of several PLLs. In this article, the mechanism of a single PLL is described in detail, whereas only the principles of operation are described for the postulated ensemble.

Depending on the parameters of a stimulus, sensory firing could engage different temporal forms. For example, the peripheral tactile response to a moving grating can be one spike per bar or a burst of a variable length per bar (Darian-Smith & Oke, 1980; Morley and Goodwin, 1987). For clarity, simple temporal forms will be assumed here. Sensory firing with bursts does not affect the principles of decoding described here (S. Serulnik and E. Ahissar, unpublished observations), although it affects the decoding details.

2 The PLL Model

2.1 The PLL Algorithm. The simplest version of a first-order PLL (Gardner, 1979) is adopted here. A first-order neuronal PLL is composed of two elements (see Figure 2a): a phase detector (PD) and a rate-controlled oscillator (RCO). The RCO is a local oscillator whose output frequency, and thus the timing of its output spikes, is controlled by the firing rate of its input. If the input is zero, the RCO will fire at its intrinsic frequency. The more excitatory the input, the higher the RCO's output frequency, and the more inhibitory the input, the lower is the RCO's frequency. The PD compares the phase—the time of arrival—of each of the spikes of a repetitive input against the phase of the RCO spikes and produces an output that is a "measure" of (i.e., its firing rate is proportional to) that phase difference. The RCO can be regarded as a rate-to-temporal code converter and the PD as a temporal-to-rate converter. The PD's output (R_d) is fed into the RCO's input and changes the RCO's firing phase in the direction that will cancel the phase difference (in fact, cancel any deviation from some constant phase difference), establishing a negative feedback loop (see section A.2). Note that in the following description *phase difference* and *temporal difference* are interchangeable terms, both expressed in time units.

The PLL is considered *locked* when the RCO's instantaneous frequency equals the input's instantaneous frequency. The phase difference, in the *locked* state, depends on the difference between the input frequency and the RCO's intrinsic frequency (see appendices A.3 and A.4). While locked, the RCO generates one and only one oscillation cycle for each input cycle. For simplicity, assume that a single spike represents a single cycle, even though a short burst or an ensemble of single spikes over a cell population is also possible. In the absence of noise and with ideal PLLs (see appendix A.3), the RCO's output spike train is a perfect replica of the input spike train, but with a delay of one cycle plus a constant phase shift. This is because, with ideal PLLs, any deviation of the input from the expected ISI is followed by an identical deviation of the RCO's ISI at the next cycle.

The decoding (or recoding) of the input information is based on the delayed internal replica of the input spike train. As long as the PLL is locked, the RCO's ISI has to be modulated by the same information that modulates the input ISI. Thus, the input information is represented by the rate-encoded signal that drives the RCO (see appendix A.4). This signal is the PLL's output. The same decoding mechanism can be described differently, at least for ideal PLLs: every input ISI is "stored" as the next RCO's ISI. Thus, at each cycle, an input ISI is compared with the input's previous ISI, and the change, which is the encoded information, is detected by the PD and presented as a rate-encoded signal (see appendix A.4). While the PD's (=PLL's) output is affecting the RCO at every cycle, it can be integrated over several cycles by a potential readout mechanism. The readout integration time, or rate bin, determines the maximal rate of information that can be represented inter-

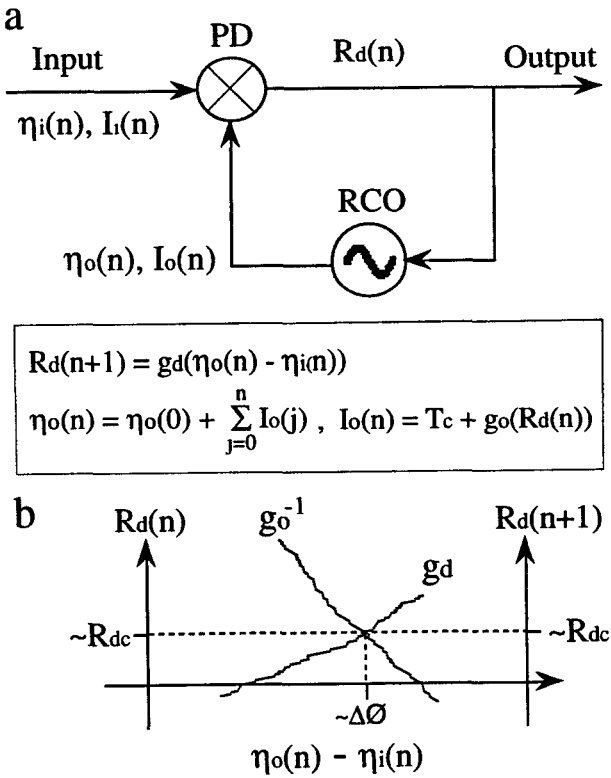


Figure 2: The first-order neuronal PLL algorithm. (a) Schematic diagram. The output of the phase detector (PD) is proportional to the difference between the timing of its two inputs. The output frequency of the rate-controlled oscillator (RCO) is modulated by the firing rate of its input (see loop equations at bottom and appendix A.2). (b) Schematic examples of transfer functions. For every I_i within the working range, $R_d(n+1)$ increases as a function of the timing difference $[\eta_o(n) - \eta_i(n)]$ and the timing difference decreases, via decrement of $I_o(n)$, and thus of $\eta_o(n)$, as a function of $R_d(n)$. The crossing (working) point is approximately (exactly for linear systems) at $(\Delta\varnothing, R_{dc})$.

nally; higher rates require shorter integration periods. Note that the phase of the input is not lost but rather preserved by the firing phase of the PLL's output, which is phase locked to the input (see, for example, Figure 6b).

Thus, the output of the PLL is a rate-encoded signal proportional to the difference between the RCO's intrinsic period and the input instantaneous ISI. This signal can be decomposed to two components: a DC component

(R_{dc}), which represents the difference between the RCO's intrinsic period and the average input ISI, and an AC component (R_{ac}), which represents the dynamic input information (see appendices A.2 and A.4). An ideal PLL should be able to track any change in the input ISI within one cycle. Practical PLLs, however, are limited in both their working ranges—the ranges of trackable input frequencies (see appendix A.2.3)—and lock-in times—the time required for moving into a new locked state. The lock-in dynamics, which is mainly determined by the loop gain (see appendix A.2.3), limits the maximal rate of change in the input frequency that a given PLL can track and decode.

PLLs of higher than first order have low-pass filters between the PD and RCO. Such filters improve the loop performance, especially in noisy conditions (Gardner, 1979; Viterbi, 1966). Low-pass filtering, also referred to as input integration, is an elementary feature of nerve cells. It is assumed that the RCO uses such filtering in its input stage. However, for simplicity, higher-order circuits will not be discussed here, since the first-order version is sufficient for code translation.

2.2 Implementations of PLLs. There are two main families of PLL implementations: excitatory PLLs (ePLLs) are those implementations in which the PD excites the RCO (see Fig. 3a, dashed lines), and inhibitory PLLs (iPLLs) are those implementations in which the PD inhibits the RCO (see Figure 3a, dotted lines). Here, only two specific implementations of these families—the AND-like and NAND-like implementations—are described in detail. Thus, ePLL will refer to an AND-like and iPLL to a NAND-like implementation, unless otherwise noted. In the following descriptions, only essential components are included, a case that probably does not occur in the brain. The described implementations should thus be regarded as building blocks that can be used separately or in combination in the brain. Accordingly, the principles of, not the exact, operation of such hypothetical PLL circuits are discussed.

The basic ePLL is a straightforward implementation of the PLL algorithm (see Figure 2) and involves two sets of neurons: the PD and RCO sets (see Figure 3a). The basic iPLL, in addition to these two sets of neurons, involves a set of inhibitory cells (INH). In both ePLL and iPLL, every component is implemented by a set of neurons similar to each other. These sets of neurons are interconnected by "diverging/converging" pathways; every cell in the projecting set sends axons to many cells in the target set, and every cell in the target set receives synapses from many cells in the projecting set. The set of RCOs of a given PLL is regarded as a set of coupled oscillators that oscillate at the same frequency. The redundancy of the RCO and INH cells has no specific role in the presented implementations beyond improving robustness. However, the efficiency of phase detection by a PD composed of a population of cells is significantly better than the efficiency of a single coincidence detector. The number of coincidence-detecting neurons that

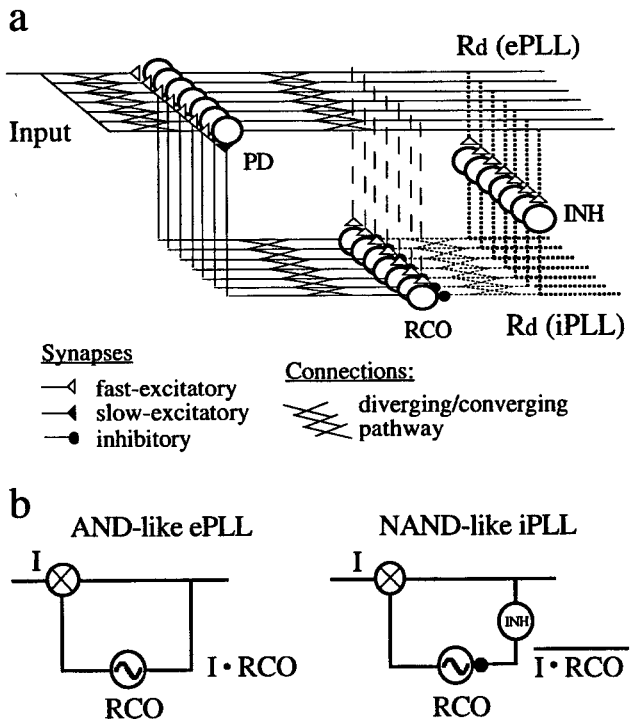


Figure 3: Possible implementations of a single PLL. (a) Connection schemes. For clarity, the width (w) was arbitrarily set at 7. Dotted lines indicate inhibitory PLL (iPLL); dashed lines, excitatory PLL (ePLL); and INH, inhibitory neurons. (b) Schematics of two possible PLL implementations where the PD operates in an AND-like manner.

compose the PD set defines the "width" (w) of a single PLL. Arbitrarily, it is assumed that the other neuronal sets (INH and RCO) have the same width. A reasonable estimation for the minimal value of w can be derived from the number of peripheral fibers activated by a "point" stimulus. In the tactile case, for example, this number is around 20 (Johansson & Vallbo, 1980).

2.2.1 Implementations of PDs. In principle, each neuron can operate as a degenerated PD. When functioning as a "coincidence detector" (Goldberg & Brown, 1969; Abeles, 1982), a cell will fire only if a certain number of its inputs will be synchronously active—that is, a single neuron detects a zero- or near-zero-phase difference among its inputs. Thus, to serve as a PD, the neuron's inputs should be predominantly organized into two groups, with

the inputs always being temporally coordinated within each group. The neuron will function as an AND-like zero-phase detector if its threshold is set such that neither of these two groups of inputs is able by itself to activate the cell, but there is a high probability that synchronous operation of both inputs will activate it. Coincidence detection, however, is not sufficient for a PD. A usable PD should have a range within which its output is a monotonic increasing or a monotonic decreasing function of the phase difference (see appendix A.2).

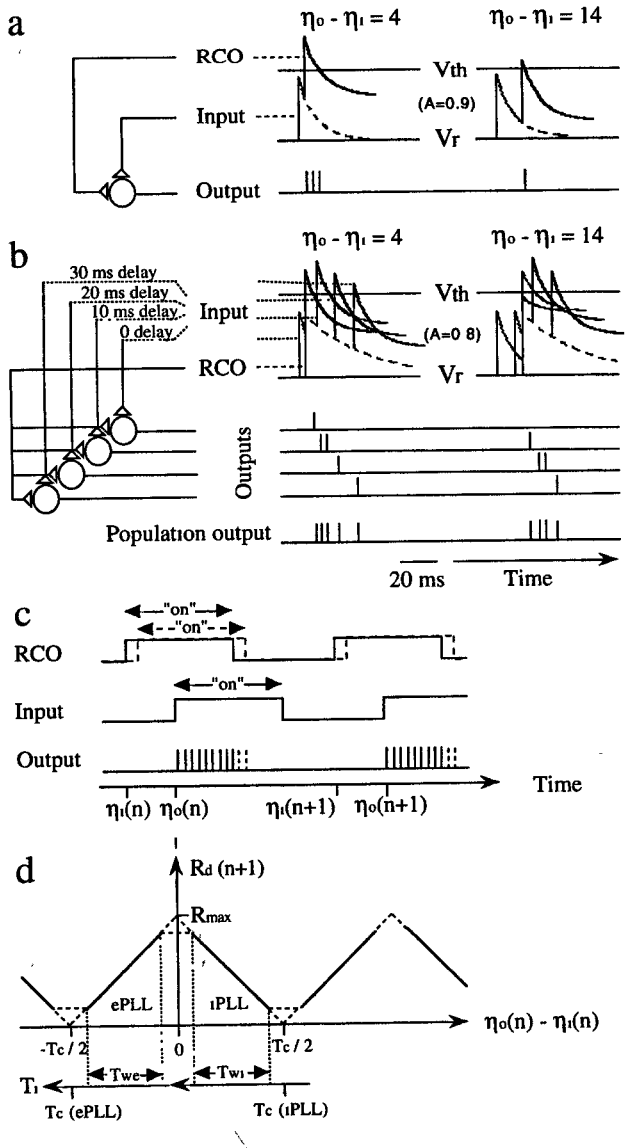
2.2.1.1 A single neuron AND-like PD. The PD capacity of a single neuron is due to the nonzero time constants of its inputs. If a neuron receives two major inputs, the range of delays that it will be able to resolve (i.e., its working range) will depend on the amplitude and time constants of the two inputs. For example, suppose the excitatory postsynaptic potentials (EPSPs) of both inputs, when measured at the axon hillock, exhibit short rising times, long decays, and relatively strong amplitudes (see Figure 4a). The longer the delay between the inputs—the phase difference—the shorter the period in which the membrane potential will be above threshold, and thus, the fewer the spikes that will be emitted. Thus, the output rate of a single neuronal PD is generally a monotonic decreasing function of the input phase difference.

For frequencies near 30 Hz, for example, the working range of an appropriately tuned neuron covers about half a cycle (see appendix A.5), which is satisfactory for a PD (Gardner, 1979). However, the refractoriness of a single neuron results in a poor output resolution—usually poorer than 2 ms. For example, tracking a frequency of 30 Hz with an error less than 1 Hz requires the RCO to be informed about deviations as small as 1 ms in the input spike train. A single neuron with a refractory period of 2 ms or more cannot provide this resolution. In addition, single-cell reliability is limited, and noise will significantly influence the single neuron PD's performance.

2.2.1.2 A population AND-like PD. To increase a PD's resolution, a number of single cells, say w, can be arranged in parallel such that all receive the same input, but with different delays (see Figure 4b). Let T_{wo} denote the effective width (see appendix A.5) of the RCO's output and T_{wi} the effective width of the input. The most efficient phase detection occurs with $T_{wo} = T_{wi}$. In this case, every phase difference between 0 and T_{wo} produces a different population sum (see appendix A.1) at the PD's output. Since the population sum is directly related to the overlap period, this dependency is monotonic. If the input delays are generated by constant and reliable delay lines, the phase differences will be also coded by the PD's population vector (see appendix A.1). Both "sum PD" and "vector PD" are valid PD implementations.

Schematically, the two input signals to the PD can be described as square waves (see Figure 4c) whose duty cycles are determined by their effective

widths; an input will be considered as "on" at all time points in which, had the other input been considered "on," summation of the generated EPSPs would be suprathreshold in at least one of the PD neurons. If the PD's output is a linear function of the inputs' overlap time, then the transfer function would take the form described in Figure 4d (see appendix A.5).



Since g_d is required to be monotonous, it is clear that a PLL can function only in limited ranges of phase differences: either within one of the increasing monotonic ranges (see Figure 4d, ePLL) or within one of the decreasing monotonic range (see Figure 4d, iPLL). In the AND-like ePLL, the RCO's output leads the input ($\eta_o - \eta_i < 0$), and in the AND-like iPLL, the RCO's output lags the input (see Figure 4d). Each of these implementations requires a different circuit to achieve the negative feedback (see appendix A.5). In the ePLL, the PD excites directly the RCO, and in the iPLL the PD excites inhibitory interneurons (INH) which, in turn, inhibit the RCO (see Figure 3 and section A.5.2). Note that the PD transfer function is periodic. Thus, large, instantaneous input modulations can move the PLL from one working range to another, producing only instantaneous tracking errors—that is, losing or "filling in" one or more input cycles.

Figure 4: *Facing page.* Implementations of neuronal AND-like PDs. (a) A single-cell PD. The two inputs, which are massive, generate two giant EPSPs with exponential decays ($A = 0.9$, $\tau = 10$ ms). More spikes are emitted when the overlapping time is larger (i.e., when there is a smaller time difference between the inputs). V_r, resting voltage; V_{th}, threshold voltage. (b) A population PD. A population of cells in which each cell receives a delayed version of the Input, each after a different delay. The RCO signal decays more slowly ($\tau = 33$ ms) than the Input and arrives simultaneously at all cells. As the time difference between the arrival of the RCO signal and the Input increases, fewer cells will be activated. (c) A schematic description of a population AND-like PD. The population signal of the Input is a pulse function, which is the "envelope" of all the delayed versions of the input, ignoring fluctuations due to EPSP shapes. At any time when both the Input and the RCO signal are "on," it is assumed that at least one of the PD cells will be activated (see section 2.2.1.2). The RCO signal is described as a pulse function, where the pulse duration is defined by the part of the RCO-driven EPSP in which adding an EPSP of the input (in any of the PD cells) will drive the membrane voltage above threshold. If the time difference between the two inputs decreases (dashed line of the RCO), the PD's output becomes stronger (two additional, dashed, spikes). (d) An example of a linear AND-like PD transfer function (g_d). The output is stronger for time delays $[\eta_o(n) - \eta_i(n)]$ that have smaller absolute values (larger overlap) and monotonically decreases in response to larger time differences. The exact form of the periodic transfer function depends on the input parameters (dashed lines; see section 2.2.1.2). The working range of the iPLL includes ISIs that are longer than the intrinsic period (T_c), while the working range of the ePLL includes ISIs that are shorter than the intrinsic period.

2.2.1.3 Other PD implementations. The AND-like implementation adopted here is not the only possible one. PDs could implement an OR function, in which the PD fires when either of its two inputs is active, or an ANDNOT function, in which the PD fires only when the input is active and the RCO is silent (E. Ahissar and M. Zacksenhouse, unpublished observations). Variations of these three basic mechanisms are also possible. For example, each input can activate the PD by itself, whereas a synchronous activation augments the PD's output (an AOR operation). All of these mechanisms can be implemented in either ePLL or iPLL configurations. Since the transfer function of the RCO is probably always a decreasing one, the potential working ranges for each implementation are those in which the PD function is increasing (see appendix A.2).

2.2.2 Implementations of RCOs. Almost any single neuron can be regarded as a voltage-controlled oscillator (VCO or VCON; Hoppensteadt, 1986). However, the PLL circuit presented here requires that the RCO exhibit an explicit periodic output activity. One possible model for a neuronal intrinsic oscillator is Calvin's regenerative firing mode (Calvin, 1975). According to this model, each spike is followed by a strong afterhyperpolarization (AHP), which recovers at some rate until it reaches the threshold again, generates a new spike, and then restarts the process. The average ISI, T_o, is determined by the depth of the AHP and the average input to the neuron. Modulations of the input produce modulations of the RCO's ISI (inhibition extends $I_o(t)$, whereas excitation shortens it (Calvin, 1975; Perkel, Schulman, Bullock, Moore, & Segundo, 1964; Hoppensteadt, 1986; see Figure 5). In another possible model, the RCO has intrinsically generated subthreshold oscillations that become suprathreshold with an appropriate DC input. The frequency of such oscillations is often controlled by the input (Llinas, Grace, & Yarom, 1991). Both subthreshold and suprathreshold intrinsic oscillations often present close-to-linear input-output (current to frequency) transfer functions (Calvin, 1975; Llinas et al., 1991; Silva, Amitai, & Connors, 1991). This implies that the input-rate to output-frequency transfer function of these oscillators is close to linear, since the amount of input current accumulated during a cycle is directly related to the rate of synaptic activation.

Three different frequencies are associated with an RCO. The *intrinsic frequency* ($f_c = 1/T_c$) is the RCO's frequency when the input to the RCO is quiescent. The *local frequency* is the RCO's frequency when the input to the PLL is quiescent, a situation that may include spontaneous activity within the loop. The *working frequency* ($f_o = 1/T_o$) is the RCO's average frequency during the decoding of a specific input.

2.3 Simulation. Validation of the basic idea of a neuronal PLL circuit and a demonstration of such a circuit's operation is provided by a simulation of a simple circuit that includes only the essential elements of the iPLL (see

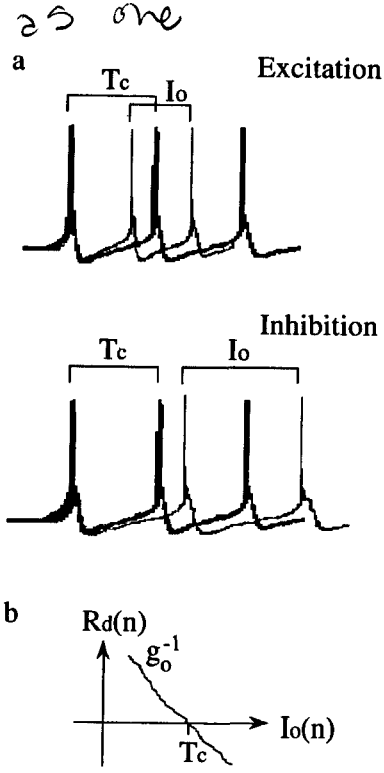


Figure 5: A neuronal RCO mechanism. (a) Output signal. The thick traces describe the membrane voltage of an RCO with no modulating input—when the RCO oscillates at its intrinsic frequency. Additional excitation or less inhibition will increase the depolarization slope, and thus increase the frequency of the RCO's output (top, thin trace). Additional inhibition or less excitation will decrease the slope, and hence decrease the frequency (bottom, thin trace). (b) A schematic transfer function of the RCO plotted as in Figure 2b. As the input (R_d) increases, the ISI (I_o) decreases.

Figure 6a). The simulation was performed on a DEC 3100 workstation using Genesis, a general-purpose neuronal simulator (Wilson & Bower, 1989). Neurons were represented by two compartments: one that represented an excitable soma that obeyed Hodgkin-Huxley kinetics and another that represented the dendrites. Three types of synapses were simulated: (1) *fast-excitatory*, non-NMDA-like synapses with a conductance time constant (τ) of 1 ms; (2) *slow-excitatory*, NMDA-like synapses with $\tau = 20$ ms, and (3) *slow-inhibitory*, $GABA_B$-like synapses with $\tau = 20$ ms. Axons were simulated as delay lines that conducted action potentials. Intrinsic oscillations were simulated by increasing the maximal sodium conductance by approximately

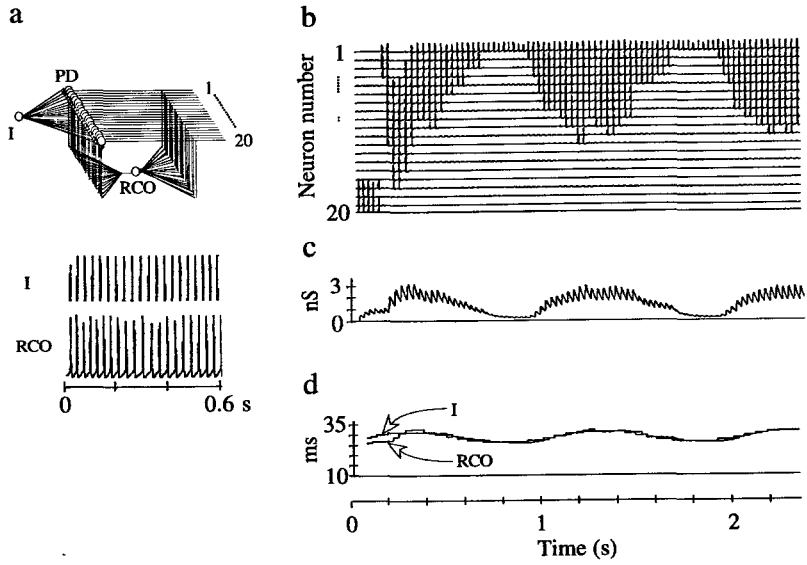


Figure 6: Computer simulation of a neuronal PLL circuit. (a) Simulated circuit and spike trains. The circuit was composed of one input cell (I), 20 PD neurons (PD), 20 different delay lines from the input to the PD neurons, and one RCO neuron (RCO) receiving an inhibitory input from each of the 20 PD neurons. The timings of the input spikes and the membrane voltage of the RCO are presented at the bottom. (b) The PLL's output, which is the population output of the PD. The spike trains of the 20 PD neurons are depicted. Each line represents, as a function of time, the membrane voltage of one PD neuron. (c) The RCO's integrated input—the total inhibitory conductance caused by synaptic input to the RCO neuron. (d) The instantaneous ISIs of the input I and the RCO are described as a function of time. After a lock-in stage, the two curves essentially merge.

50% (Alonso & Llinas, 1989; Llinas et al., 1991; Silva et al., 1991). The intrinsic oscillating frequency of the simulation was set by tuning the membrane capacitance.

The width of the loop was set to $w = 20$. The input was simulated by a single input cell (I) whose output was conveyed to the PD neurons via 20 axons, whose delays to the 20 PD neurons were uniformly distributed between 14 and 20 ms and which formed fast-excitatory synapses on PD neurons. All 20 PD neurons converged on a single RCO neuron via slow-inhibitory synapses. For simplicity, the INH neurons were discarded and replaced by direct inhibitory connections from the PD to the RCO. A single RCO neuron represented the hypothesized 20 RCO neurons. This RCO neuron fed back, by slow-excitatory synapses, each of the PD neurons.

The ability of this simplistic simulated PLL circuit to decode periodic modulations of periodic input signals was tested by "injecting" excitatory intracellular currents into the input cell's soma. Figures 6b–d depict the results of one simulation. The input signal was a 1 Hz modulation of a carrier frequency (35 Hz), with a modulation depth of 20%. The RCO frequency locks to the input frequency (see Figure 6d); the instantaneous ISI, of both the input and the RCO, are described by the two curves. After a lock-in stage, the two curves merge, which indicates the frequency locking. In the locked condition, the input modulation of 1 Hz is decoded by the PLL and approximated by a 1 Hz population signal (both population vector and population sum) at the PD's output (see Figure 6b). At any given time, both the population vector (the actual firing neurons) and the population sum (total spikes across the population) represent the input ISI (within the PD resolution limits). The integrated inhibition (see Figure 6c) modulates the RCO's frequency. This integrated signal is an integration of the PLL's output, and it provides a measure of the population sum.

This simple simulated circuit was able to decode modulations of up to 2 Hz with a 20% modulation depth. However, one cannot learn about the decoding limitations of the PLL from this simulation, since only a specific, limited circuit was simulated. For example, the resolution of PLL decoding depends on the number of elements, and the range of decodable modulation depths and frequencies, as well as lock-in dynamics, depends on the loop gain. This simulation mainly demonstrates how PLL neuronal signals should look in principle. To demonstrate the dependency of lock-in dynamics on the loop gain, I performed a MATLAB simulation of the iPLL, using equations A.4, A.9, and A.15 (see Figure 2) and a periodic PD function with the profile depicted in Figure 4d. The results are shown in Figure 7. All time variables are expressed in T_c (the RCO's intrinsic period; see appendix A.2) units. For an input period (T_i) of 1.2 T_c and an initial phase difference $[\eta_o(0) - \eta_i(0)]$ of 0.3 T_c, lock-in time was one cycle for loop gain $(G) = -1$ (see Figure 7a). When G was too small in absolute value (lowermost trace) the RCO could not approach the input period. The reason was that with such gains, the phase difference that was required to follow T_i exceeded the PD's working range ($T_c/2$; see Figure 4d). Thus, with this specific PD function, the working range of the iPLL was $T_c < T_i \leq T_c(1 + |G|/2)$. iPLLs with $G < -2$ (upper-most oscillating trace) were not stable (see appendix A.2, equation A.17). Between these two limits of G, iPLLs could lock in to the input, where lock-in times increased with increased deviation of G from -1. However, lock-in times also depended on the initial phase difference (see Figure 7b). Thus, even with ideal PLLs, having $G = -1$, lock-in to the onset of an input train might take more than one cycle, due to the phase difference. A single cycle lock-in is guaranteed only when the PLL is already locked to the input and a sudden change in the input periodicity is introduced, as demonstrated in Figure 7c. Here, after four cycles of

1.1 T_c the input ISI was changed to 1.4 T_c and from then on was modulated around 1.25 T_c with a modulation period of 10 T_i and a modulation depth of 0.4 (peak-to-peak). Six iPLLs with $0.5 \leq G \leq 1.75$ were simulated. Lock-in to the input onset was not immediate, due to the nonoptimal initial phase difference (0.3 T_c). However, after the PLLs were locked, those who could track the maximal input period (those with $G \leq -1$, see working ranges above) tracked it more or less smoothly. The tracking errors are plotted in Figure 7d. It can be seen that the iPLL with $G = -1$ (x's) tracked the input modulations with no errors, while the other iPLLs exhibit tracking errors, as expected (see section 2.1 and appendix A.3). The output rate produced by the iPLL with $G = -1$ is depicted in Figure 7e. Finally, the PLLs that could reach the maximal T_i (those with $G \leq -1$ in this case) were able to follow the highest possible rate of input modulations: ($2T_i$) (see Figure 7f).

Note that the maximal rate of input modulations trackable by PLLs does not indicate the maximal resolution of temporal decoding by PLLs. The latter is determined by the smallest deviation of input ISI that can be detected by a PLL and is determined by the resolution of the PD's population output (e.g., the resolution of the y-axis in Figure 7e). Thus, PLLs can detect (and represent by rate) temporal changes with a resolution that can be much higher than the maximal rate of temporal modulations that they can track. For example, a PLL whose working range is 100 ms $< T_i \leq$ 150 ms can, with enough PD resolution, distinguish between inputs of 110 ms and 111 ms, although it cannot track 1 kHz modulation.

3 Tactile PLLs

The mammalian tactile system contains the neuronal elements required for the function of thalamocortical PLLs. Following is a proposal for a plausible implementation of PLLs by the primate tactile system. The peripheral tactile system, which acquires sensations during exploration of textures, has been described in detail over the past three decades. The tactile system includes the three following subsystems, which are classified according to the temporal nature of their responses: slowly adapting (SA) receptors and neurons, which respond optimally over the low range (~0–20 Hz) of stimulus frequencies; rapidly adapting (RA) receptors and neurons, which respond best over frequencies of medium range (~20–40 Hz); and Pacinian (PC) receptors and neurons, which mainly transfer information at high frequencies (>80 Hz) (Talbot et al., 1968; Freeman & Johnson, 1982; Johansson, Landstrom, & Lundstrom, 1982; Goodwin, John, Sathian, & Darian-Smith, 1989). The glabrous fingertip is innervated mainly by RA receptors, by lower numbers of SA receptors, and by only a small number of PC receptors (Johansson & Vallbo, 1979; Darian-Smith & Kenins, 1980).

Less is known about the central mechanisms underlying tactile decoding and processing. The tactile pathways from the periphery to the cortex preserve the phase of the stimulus (Darian-Smith & Oke, 1980; Mountcas-

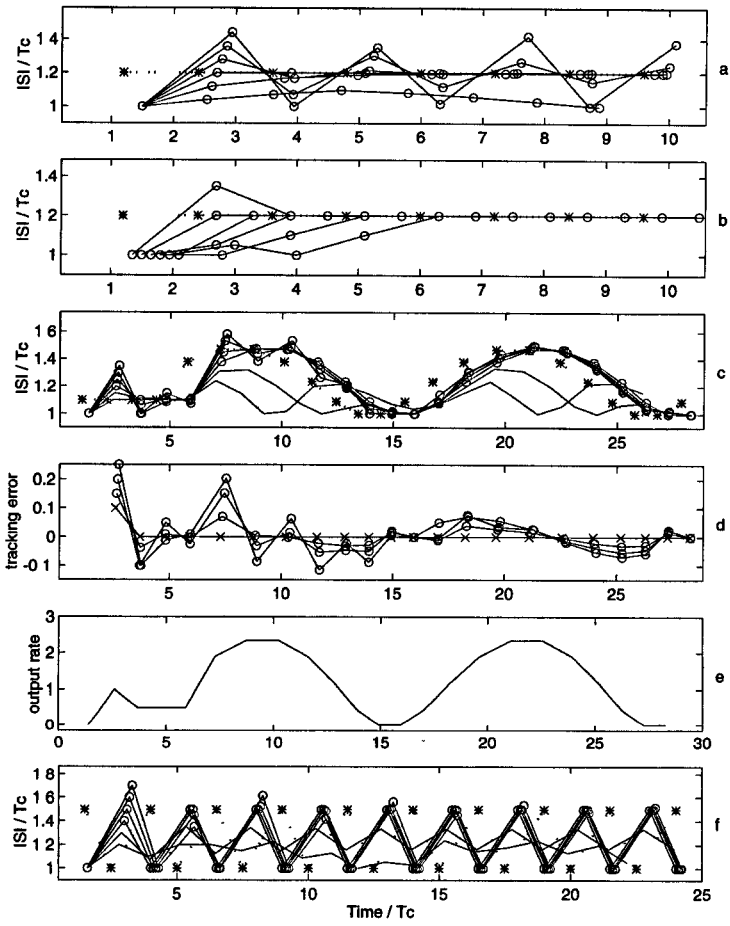


Figure 7: Dependence of lock-in dynamics on input parameters. The iPLL algorithm was simulated in MATLAB (see text). Simulations steps (n) were counted from 1. (a) Input (stars and dotted line): A steady input period at $T_i = 1.2T_c$. G values (traces from bottom up at $n = 2$): 0.2, 0.6, 1.0, 1.4, 1.8, 2.2. Initial phase difference $[\Delta\eta(1)] = 0.3T_c$ (b) Input: As in a. $G = -1$. $\Delta\eta(1) = 0.15, 0.30, 0.45, 0.60, 0.75, 0.90\ T_c$. (c) Input: Four cycles of $1.1T_c$ followed by $I_i(n) = (1.25 + 0.25\sin(2\pi(n-4)/10))T_c$, for $n > 4$. G values (traces from bottom up at $n = 2$): 0.5, 0.75, 1.0, 1.25, 1.5, 1.75. iPLLs with $G > -1$ are plotted without symbols. $\Delta\eta(1) = 0.3T_c$. (d) Tracking errors for the simulation in c were computed as $(I_o(n) - I_i(n-1))/T_c$, for $n > 1$. Only $G \leq -1$ are shown. $G = -1$ is plotted with X's. (e) The PLL's output rate (R_d), in arbitrary units, for the simulation in c. (f) Input: $T_i = 1.25\ T_c$, modulation rate ($2T_i$), modulation depth 40%. G values as in c. $\Delta\eta(1) = 0.1T_c$

tle, Talbot, Sakata, & Hyvärinen, 1969; Ferrington & Rowe, 1980; Burton & Sinclair, 1991; Gardner, Palmer, Hamalainen, & Warren, 1992). However, the degree of phase locking gradually decreases along the afferent pathways, with the largest reduction probably occurring at the transition from the thalamus to the cortex, a transition that is also accompanied by an increased complexity of response (Darian-Smith, Sugitani, Heywood, Karita, & Goodwin, 1982; Sinclair & Burton, 1988; Burton & Sinclair, 1994). This increased complexity could be due to significant processing that occurs already at the thalamocortical level (Gottschaldt, Vahle-Hinz, & Hicks, 1983).

3.1 Temporal Encoding of Textures. I consider here only textures (of variable patterns and heights) on flat surfaces (e.g., textures of sandpapers, clothes, woods, artificial gratings, or braille pages). A finger traversing these types of surfaces usually moves along sections of approximately straight lines (see Figure 8, top). The information contained in these textures consisting of a collection of ridges is expressed by three variables: amplitude, average spatial period (where period is 1/frequency), and local spatial modulations. The information carried by the average spatial period (X_t) is called here the *roughness* of the surface and the information carried by local spatial modulations (p_t) the *pattern* of the surface. I use *italics* to distinguish this specific stimulus-defined *roughness* from the more general roughness percept. As I will show below, decoding *roughness* information can contribute to the roughness percept.

When a surface is transversed by fingertips, the spatial information is encoded in two ways:

1. Spatial encoding: Across the contact area (~0.6 cm^2 in humans), at any given moment, the spatial features are reflected by the corresponding skin deformations leading to a spatially encoded response of the relevant receptor population.

2. Temporal encoding: At any given skin location, receptors are responding to the fluctuations of the indentation amplitude produced by the movement (see section A.6).

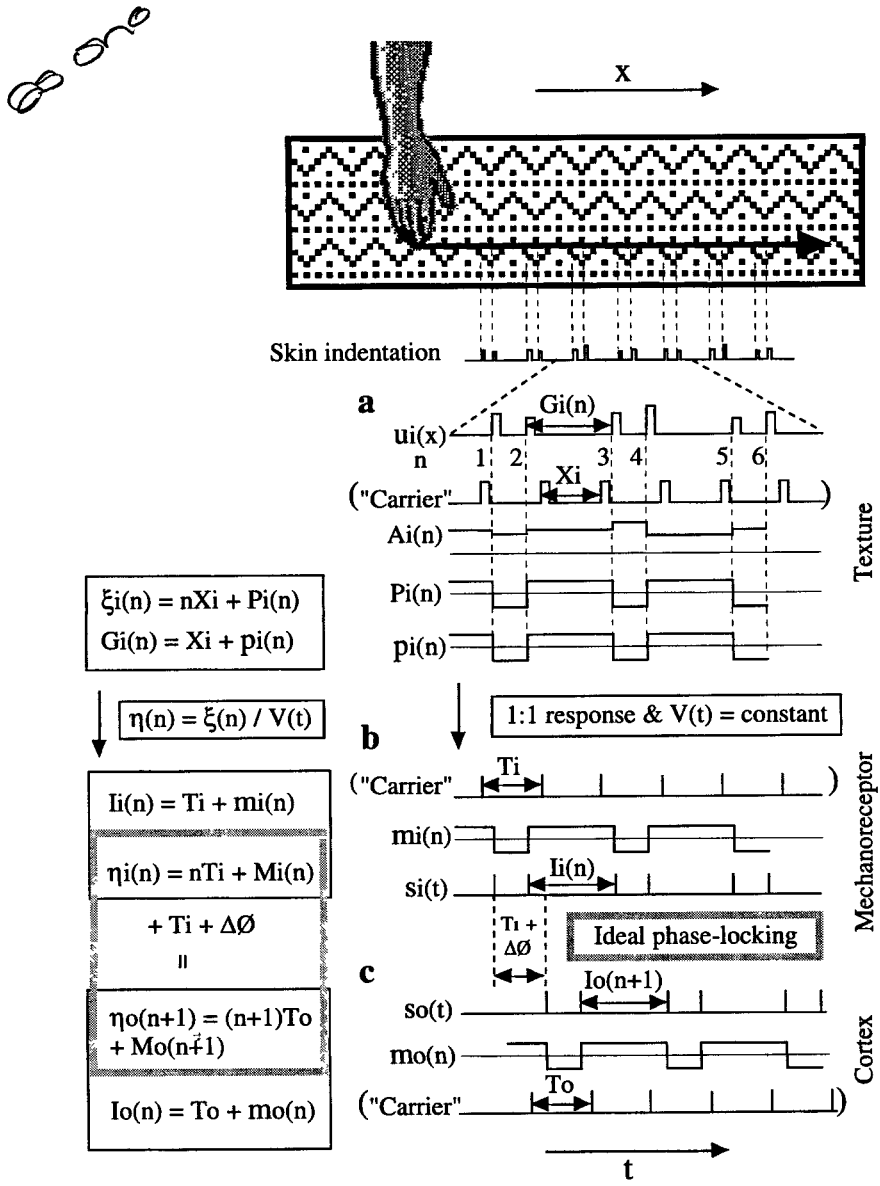
Spatial encoding is probably best mediated by the SA receptors (Phillips, Johansson, & Johnson, 1990) and needs to be decoded by mechanisms utilizing spatial comparisons. Temporal encoding is probably best mediated by RA (and, to a lesser degree, PC) receptors, which respond reliably to temporal modulations (Darian-Smith & Oke, 1980; Morley & Goodwin, 1987). Such temporally encoded signals could be efficiently decoded by thalamocortical PLLs. However, since point skin indentations are modulated by both amplitude (due to vertical surface fluctuations) and time (due to horizontal interval fluctuations; see Figure 8, top), the interpretation of the decoded signals could be ambiguous. Electronic implementations of PLL, facing similar problems, always include an amplitude limiter at the input

stage (Gardner, 1979). Interestingly, the RA peripheral system employs a similar mechanism. Responses of RA mechanoreceptive fibers to vibratory stimuli, or moving spatial gratings, having amplitudes between fewer than 10 to hundreds of microns, are often of a 1:1 type; they fire one and only one spike per vibratory or grating cycle, regardless of the amplitude (Talbot et al., 1968; Darian-Smith & Oke, 1980; Goodwin & Morley, 1987; Gardner & Palmer, 1989). Every RA fiber exhibits a 1:1 response within a specific range of amplitudes (termed the "plateau" range; Talbot et al., 1968) and temporal frequencies (Darian-Smith & Oke, 1980). Outside these ranges, fibers respond with bursts of variable lengths per cycle (Darian-Smith & Oke, 1980; Morley & Goodwin, 1987), depending on the force and frequency of the stimulus (Darian-Smith & Oke, 1980; Goodwin et al., 1989).

Thus, in the case of the RA system and with a constant finger velocity, the encoding of the horizontal features of textures is straightforward. The horizontal (or temporal) modulations of the periodic indentation profile, as a function of x (or t), can be described by the two methods used above to describe temporal periodic signals (see Figure 1 and appendix A.1): either with respect to an imaginary "carrier" signal (see Figure 8a, $P_i(n)$) or with respect to the spatial intervals themselves (see Figure 8a, $p_i(n)$) (see appendix A.6). During scanning, the timing of the nth mechanoreceptive spike is uniquely determined by the location of the nth ridge (see Figure 8b). If the response type is 1:1, the RA mechanoreceptive fibers should fire one and only one spike per every ridge in the surface, and the sensory transformation takes a simple form: the horizontal spatial structure is directly represented by the temporal structure of the RA spike trains (see appendix A.6). With different ranges of finger forces, the 1:1 response becomes a 1:n response, and the transformation is more complex. However, as long as the duration of the bursts is small relative to the input average cycle (which is usually the case; Darian-Smith & Oke, 1980; Morley & Goodwin, 1987), decoding efficiency should hardly be affected since input onset times, which are the important parameters for the decoding, are not affected. Yet the increased input intensity and duration caused by the bursts should be compensated by a proper tuning of the PLL's loop parameters (S. Serulnik & E. Ahissar, unpublished observations). Thus, for optimal performance, PLL parameters should be tuned according to the expected form of input bursts.

3.2 Decoding by Thalamocortical PLLs. As I will show, the decoding of tactile signals by PLLs requires an additional feedback loop. Thus, the postulated temporal tactile decoder, as one module within a global tactile texture decoder, includes many parallel PLLs embedded within a sensorimotor feedback loop (see Figure 9). According to the model, the movement of fingers across a surface activates skin mechanoreceptors (MR) which convert the spatial details into temporal signals. The RA and PC mechanoreceptors at the fingertip include amplitude limiters (L), which eliminate amplitude modulations. The parallel array of input filters (IF; mechanoreceptors and

their fibers) transfers the filtered signal to an array of somatotopic organized PLLs, each specifically tuned to a particular frequency restricted to one of the tactile submodalities (SA, RA, or PC). Therefore, every point on the skin is driving a set of PLLs, each tuned to a particular frequency (see Dykes,



1983). The output of all the PLLs is fed to two readout networks, IP and IR, for *pattern* and *roughness* evaluation, respectively. The IR's output drives the velocity controller (VC), which closes the loop by controlling the finger velocity.

Each PLL thus processes information about different spatial frequencies of the explored surface. How does the brain know which PLLs provide relevant information about the actual surface being explored and how can it focus on these PLLs? If the PDs employ nonperiodic, sigmoid-like transfer functions, the answer to the first question would be simple. Only PLLs that are tuned to the relevant (i.e., informative) temporal frequencies should present modulated output signals. All other PLLs should produce outputs that are saturated at either the highest or the lowest possible values. Thus, the modulation depth of the AC output component, or a related measure such as $|R_{ac}(t)|$, should provide a reliable measure of the amount of information contained in each PLL's output. This criterion appears to be valid also for periodic PDs (like the one in Figure 4d). As the input average frequency moves away from the center of the working range, the probability of the instantaneous frequencies to exceed the bounds of this range increases. Once the input frequency exceeds one of these bounds, the PLL's output is closer to its average value, and its modulation depth decreases. Thus, local maxima of $|R_{ac}(t)|$ represent maximal information. Furthermore, it is most likely that these local maxima will be graded among different PLLs, and a global maximum will also be available. The reason is that each PLL can obtain a larger

Figure 8: *Facing page.* Illustration of a temporal encoding of spatial features. The movement of the hand (arrow) across a surface generates skin displacements at the zone of contact. This series of displacements can be described as a spatial signal $[u_i(x)]$ that represents the texture in this one-dimensional direction of movement. (a) Decomposition of the spatial signal. $u_i(x)$ can be decomposed into vertical $[A_i(n)]$ and horizontal [either an imaginary "carrier" $+P_i(n)$, or $p_i(n)$] components. The similarity between $P_i(n)$ and $p_i(n)$ in this example is due to the regularity of the *pattern*. $G_i(n)$ is the interridge interval. (b) Receptor transformation. Assuming a 1:1 response of mechanoreceptive fibers and a constant velocity, $u_i(x)$ is converted to a temporal signal described by $s_i(t)$. The temporal signal, which is carried by the mechanoreceptive fibers, can be also decomposed into subcomponents. However, due to the 1:1 response, which neglects amplitude changes, the amplitude component is constant and equal to 1, and thus is ignored. $I_i(n)$ is the ISI. Refer to appendix A.1 for the definitions of other terms. (c) Decoding by an ideal PLL. The phase-locking mechanism of the PLL forces the RCO's output $[s_o(t)]$ to track the peripheral input $[s_i(t)]$ with a delay of one cycle (T_i) and a constant phase difference $(\Delta\varnothing)$. As a result, the brain can extract the modulation $(M_o(n)$ or $m_o(n))$ that describes the *pattern* and the average interval (T_o) that describes the *roughness*. See appendices A.1, A.3, and A.6 for explanation of other symbols.

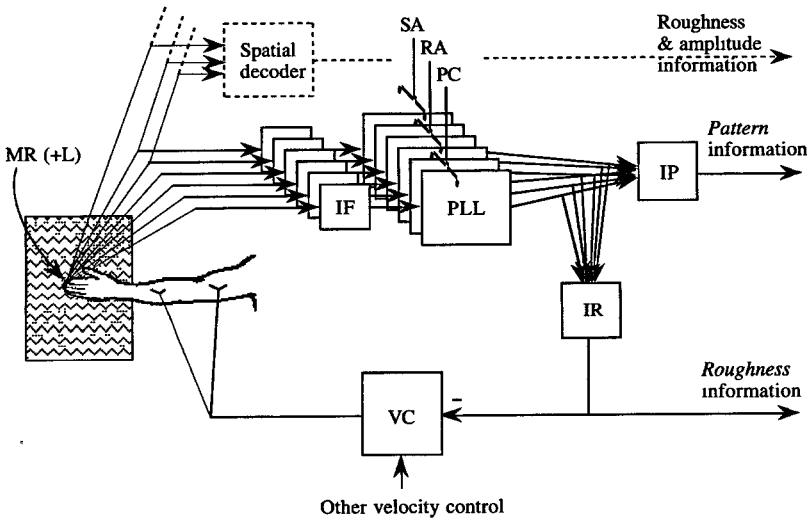


Figure 9: PLLs within a global texture decoder. Many topographically-organized PLLs reside in parallel; only six are shown in the figure. Each PLL is tuned to a specific frequency range within a specific submodality: SA, RA, or PC. The input of each PLL is received from a set of mechanoreceptive fibers through a set of dorsal column nuclei relay neurons, which together comprise the input filter (IF). Most of the mechanoreceptors (MR) include amplitude limiters (L). IP and IR are readout networks that produce *pattern-* and *roughness*-related information, respectively. A hypothetical spatial decoder illustrates the operation of additional mechanisms in parallel.

working range if it tunes the slope of its PD function according to its average frequency—smaller slopes for lower frequencies. In this case, the global maximum will indicate the PLL whose working range is fully exploited.

Since neuronal excitation is often sensitive to the variability at the input (Aertsen, Erb, & Palm, 1994), circuits that detect maximal variabilities can be implemented. If such circuits are included in the PLLs' readout networks (e.g., IP in Figure 9), they can assist the selection of one of the submodalities (SA, RA, or PC) and the specific PLLs within that submodality that are most informative. Other factors affecting this selection probably include visual, cognitive, and additional tactile information, such as that obtained by spatial decoders (see Figure 9). According to this selection, the finger velocity is deliberately determined to be in the range that will generate temporal frequencies in the appropriate range for the chosen PLLs. By setting the finger velocity, the system focuses on the selected PLLs, since they will generate the most informative output. To keep this focus steady, an automatic feed-

back system is required to carry out the fine-tuning of the velocity. Such a feedback system could be tuned to maximize the amount of output information from the selected set of PLLs, using $|R_{ac}(t)|$ as a measure. Although such an operation makes sense, I propose that if it is implemented, it is implemented as a higher-order feedback loop. For maintaining the input frequencies around a selected PLLs' working point, such a feedback system could simply operate on fluctuations of the averaged PLL's output, R_{dc}.

Let us represent each selected group of PLLs by a single PLL. Once a particular PLL is selected, the sensorimotor circuit responsible for temporal decoding can be described by two loops (see Figure 10a): the inner loop is the selected PLL, which extracts the input temporally encoded information (see Figure 8) and recodes it by rate (see appendix A.7), and the outer loop is an automatic velocity control (AVC), which keeps the input frequency of the PLL centered around the PLL's working frequency. IP and IR are reduced in this description to single filters, assumed to produce outputs related mainly to the selected PLL. The other inputs to IP and IR are assumed to be averaged out. The general case of processing, in which the average input frequency can change (albeit slowly) over time, even beyond the working range of the PLL, is assumed here. Thus, both the average input ISI and the average PLL's output are functions of time ($T_i(t)$ and $R_{dc}(t)$, respectively).

3.2.1 The automatic velocity control (AVC). When the PLL is locked, the average RCO's ISI is approximately equal to the average input ISI ($T_o(t) \approx T_i(t)$; see appendix A.3). If either the average input spatial period ($X_i(t)$) or the finger velocity ($V(t)$) is changed, the PLL will move to a new working point in which $T_o(t) \approx T_i(t)$. Such a new working point will be associated with a new average output rate ($R_{dc}(t)$) of the PLL. However, if the working range of the PLL is limited, as is the case for any practical implementation, this adaptive process is also limited, and consistent drifts in the input average frequency can eventually lead to a loss of locking as the PLL leaves its working range. From the point of view of the sensorimotor system, there are two possible solutions to this problem: it can either have many PLL circuits, each tuned to a different working range (the open-loop approach), or it can actively maintain the input temporal frequency within a working window (the closed-loop approach). The closed-loop approach, whose operation is postulated here, requires that, while operating near the center of the PLL's working range, if $T_o(t)$ is driven toward the limits of the working range, an action will be taken to bring $T_i(t)$ back to its original value via the control of the finger velocity, $V(t)$ (see appendix A.8).

The algorithm for the tactile AVC is composed of five elements (see Figure 10a): (1) a multiplier (MR), which multiplies the finger velocity by the spatial frequency of the texture; (2) a PLL circuit, which converts the *roughness* and *pattern* information to the DC and AC components of a rate signal; (3) a base-band filter (BBF), which transfers only the frequencies related to the *pattern*; (4) a low-pass filter (LPF), which transfers only the frequencies

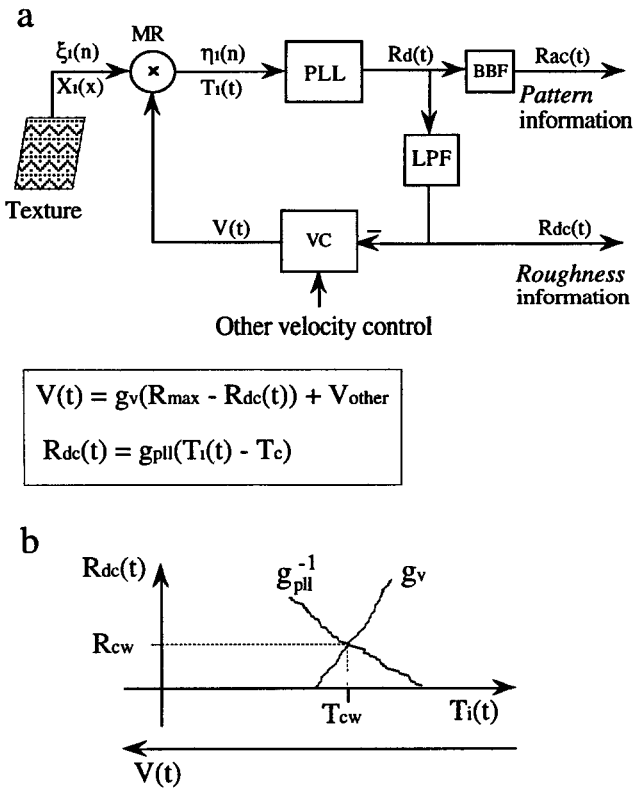


Figure 10: The algorithm of the AVC loop. (a) Loop components. MR, multiplier that includes an amplitude limiter; PLL, one of the PLL circuits in Figure 9 that is selected for optimization; BBF, base-band filter; LPF, low-pass filter; VC, velocity controller. $\xi i(n)$, location of the nth input ridge; $X_i(x)$, average interridge interval; $\eta_i(n)$, timing of the nth input spike; $T_i(t)$, average input ISI; $R_d(t)$, firing rate of the PLL's output; $R_{ac}(t)$, the integrated signal representing the *pattern*; $R_{dc}(t)$, the integrated signal representing changes in *roughness*; $V(t)$, the finger velocity. The loop equations (inset) are explained in appendix A.8. (b) Schematic examples of transfer functions. The crossing point (T_{cw}, R_{cw}) is the working point of the AVC, which, optimally, fits the desired working point of the selected PLL.

related to changes in the *roughness*; and (5) a velocity controller (VC), which controls the finger velocity.

The negative feedback nature of the AVC maintains the PLL's mean input ISI close to the PLL's desired working point (T_{cw}, R_{cw}), which is the center of the working range of the selected PLL. An increase in either the average spatial frequency or the finger velocity will result in the input's average ISI decreasing, $R_{dc}(t)$ increasing, and $V(t)$ decreasing (see Figure 10 and sec-

tion A.8). As a result, $T_l(t)$ will be driven back toward T_{cw} with a dynamics that depends on the actual l transfer functions. An opposite reaction occurs when either the average spatial frequency or the finger velocity decreases. Note that this servo operation holds for any given transfer functions, provided that they establish a negative feedback. Thus, dependence of tactile inputs on motor outputs (Chapman, 1994; Nelson, 1996) should affect the details of the AVC operation but not its principles.

3.3 Implementations of Tactile PLLs.

3.3.1 Implementations of tactile PDs. The tactile RA system appears to have evolved such that thalamic RA "relay" cells can be used as efficient phase detectors. The main features of the RA system contributing to this efficiency are (1) a rectangular-like distribution of the conduction velocities of RA fibers (Talbot et al., 1968; Darian-Smith & Kenins, 1980); (2) a close-to-uniform receptor sensitivity across the receptive field (Johansson & Vallbo, 1983; Gardner & Palmer, 1989); and (3) slow (long duration) cortical-to-relay neurons excitatory postsynaptic potentials (EPSPs) (Deschenes, Paradis, Roy, & Steriade, 1984). In general, the first two of these features are also typical of the PC, but not of the SA, subsystem; the third is probably common to all three of the tactile subsystems.

If organized correctly, the lemniscal input to the thalamic relay neurons can implement a square-wave-like signal, like in Figure 4c. Given features 1 and 2 of the RA system, the lemniscal input contains subpopulations in which, for a given point stimulus, different subsets of the input are active at different times. A "point" stimulus—an abrupt indentation at a single skin locatio—will generate a uniform response across all RA receptors that include that point in their receptive field, due to the uniform sensitivity of each receptor across its receptive field. When light touch is used, the skin indentation is assumed to be within a plateau range of amplitudes where the response has the form of one spike per one point stimulus (Talbot et al., 1968; Darian-Smith & Oke, 1980). If each subpopulation of fibers that share a skin location contains fibers with different conduction velocities, these activations will arrive at the thalamic relay neurons at different times for each fiber, like in Figure 4b. In this case, each of the fibers can be considered as a delay line generating a specific delay from skin activation to the firing of a lemniscal fiber. For the conductance velocities (Talbot et al., 1968; Darian-Smith & Kenins, 1980) and hand length (~50 cm) of monkeys, the spread of lemniscal firings probably contains mainly latencies between 7 and 14 ms, not including the duration of input bursts. This range corresponds to about one-fourth of a cycle of 30 Hz oscillations and is a reasonable range for a PD (see section 2.2.1). However, different spreads of the afferent signal are optimal for different PLL working frequencies. Thus, it is expected that channels conveying lower frequencies will employ larger temporal spreads.

3.3.2 Implementations of tactile RCOs. Obvious candidates for RCOs are the posterior SII local oscillators (Ahissar & Vaadia, 1990). Many of the neurons in this area display oscillatory patterns; however, not all of them can be considered local oscillators. At least 15% of the neurons in that area probably oscillate due to local mechanisms (Ahissar & Vaadia, 1990). The rest of the oscillating neurons (about 30% of the population) are either externally driven by the local oscillators, or their local oscillations are masked by a significant amount of noncorrelated input. The local oscillators can either directly drive thalamocortical neurons, if they project to the thalamus, or drive corticothalamic neurons. Note that these single-cell oscillations do not merely reflect sleeplike thalamocortical spindles (Steriade, McCormick, & Sejnowski, 1993) since they appear in wakefulness, include mainly gamma frequencies, and are not correlated among neighboring neurons.

There is no direct evidence yet that indicates an RCO-like operation of the SII oscillating neurons. However, these neurons lose their oscillatory patterns when stimulated with nonperiodic tactile stimuli (Ahissar & Vaadia, 1990). This finding is consistent with the cortical oscillators trying to track the nonperiodic input. More important, the distribution of oscillating frequencies of these oscillators matches the peripheral distribution of best frequencies (see Figure 11). A more direct evidence was obtained for SI oscillators in monkeys, employing RA frequencies (~30 Hz; Lebedev & Nelson, 1995), and in rodents, employing whisking frequencies (~10 Hz; Ahissar, Alkon, Zacksenhouse, & Haidarliu, 1996). These oscillators can be entrained by tactile periodic stimuli near their spontaneous frequencies, but usually not with significantly higher or lower frequencies. Thus, PLL circuits might exist in parallel in thalamocortical loops involving SI and SII cortices.

3.3.3 Implementation of Readout Networks. Each of the two readout networks, IP and IR, should implement at least two functions. The simple one, which is required for the AVC operation, is filtering out the unnecessary information. Both low-pass and bandpass filters are easy to implement by neuronal networks, utilizing synaptic integrations and decays. In addition, these networks should probably include circuits that compute and compare input variabilities (IP) and input averages (IR). Detailed implementations of these filters are beyond the scope of this article. Although the PLLs' outputs are described as converging to the readout networks (see Figure 9), they do not necessarily have to converge. The readout networks can utilize parallel processing and produce population outputs. Accordingly, the single lines standing for the outputs of the two filters in Figure 9 denote the unity of information conveyed by their outputs rather than the outputs' physical widths.

4 Discussion

4.1 Advantages and Limitations of PLLs. The PLL algorithm is used extensively in electrical engineering for decoding of phase and frequency

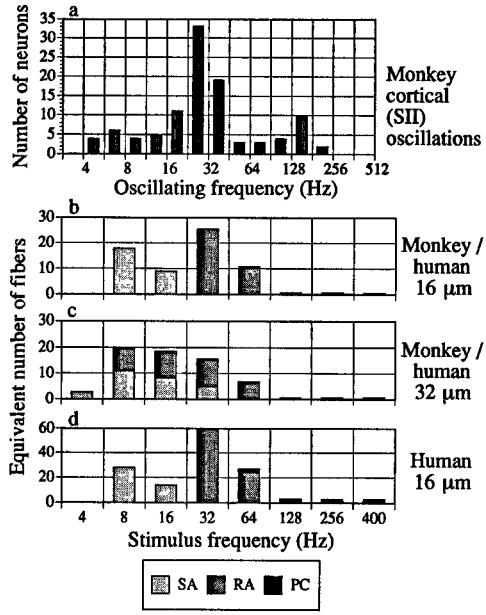


Figure 11: Cortical oscillating frequencies and peripheral frequency tunings. (a) Distribution of oscillating frequencies of cortical (posterior SII) single-cell oscillators (104 frequencies observed in 76 neurons of which 18 exhibited more than a single frequency). Only frequencies larger than 2.8 Hz were included for compatibility with the peripheral data. (modified from Ahissar and Vaadia, 1990). (b–d) Estimated distributions of peripheral tuning to the frequency of sinusoidal skin displacements. Ordinates depict estimations, for each input frequency, of the number of mechanoreceptive fibers that innervate the skin contact area and are tuned to that frequency. The distribution of tuning to vibratory frequencies among the input fibers was estimated here by calculating the "equivalent number of fibers" tuned to each frequency. The equivalent number of fibers per submodality is the fraction of the average response of that submodality at each frequency multiplied by the average number of fibers of the same submodality that innervates the stimulated area of skin. During light touch, the contact areas of skin for humans and monkeys are ~0.6 cm^2 (Lamb, 1983) and ~0.2 cm^2 (Goodwin & Morley, 1987), respectively. (b) Mean responses of mechanoreceptive fibers were obtained from data published for humans (Johansson et al., 1982) and innervation densities from data published for monkeys (Darian-Smith & Kenins, 1980). Skin contact area was assumed to be 0.2 cm^2. Peak-to-peak indentation amplitude is 16 μm. (c) Same as b, except that the indentation amplitude is 32 μm. (d) Same as b, except that the innervation densities were obtained from data published for humans (Johansson & Vallbo, 1979) and skin contact area was assumed to be 0.6 cm^2.

modulated signals, frequency synthesis, and pulse synchronization. When utilized as a phase demodulator, a PLL exhibits an excellent noise immunity due to its adaptive narrowband filtering (Gardner, 1979). This narrowband filtering is achieved by comparing the input against a specific internal frequency and becomes adaptive because of the feedback control of the internal frequency.

In principle, temporally encoded neuronal signals (see Figure 1) are phase-modulated signals; therefore, utilization by the brain of a PLL mechanism to decode temporally-encoded signals should be advantageous. However, there are limitations inherent in the PLL mechanism that the brain would have to compensate for. One limitation arises from the adaptive behavior of the PLL, which limits the PLL's capacity to track rapid changes in the input. As with any other negative feedback loop, a few input cycles may be needed before the PLL can lock in to a new input and efficient decoding can commence. Nevertheless, learning-induced fine tunings of the loop parameters can reduce to a minimum (down to 1 cycle) the number of lock-in cycles.

Another limitation of PLLs stems from the limited working ranges of their implementations. A PLL cannot track, although it can detect input modulations that are larger than its working range. The working ranges of basic neuronal PLLs are usually around half a cycle, with the upper limit probably being one cycle (see section 2.2). Thus, a typical, "nonsophisticated," neuronal PLL is limited to inputs with modulation depths of less than 50%. If an RCO cannot produce the required frequencies, the PLL's working range will be even more limited. This limitation can be circumvented by having several PLL circuits in parallel, each tuned to a different frequency range and decoding a different segment of the input information. In addition, "sophisticated" implementations of PDs can extend working ranges and reduce lock-in times.

A significant advantage of neuronal PDs is that transitions from one implementation to another can occur within a given anatomical circuit by changing cellular parameters. For example, at low excitability levels, a PD neuron can implement an AND-like function, at high excitability levels an OR-like function, and at intermediate excitability levels an AOR-like function (see section 2.2.1.3). Thus, neuronal PLLs can dynamically change their loop parameters, including gain and working range, to accommodate to global sensory changes or requirements. For example, a full-cycle working range can be implemented by asymmetrical PDs that employ an AND-like function for negative phase differences [$\eta_o(n) - \eta_i(n) < 0$] and an OR-like function for positive phase differences. In such an asymmetrical AOR-like PD, the order of input activation determines the sensitivity of the PD neurons. The periodic PD transfer function of such asymmetrical PDs has the shape of a sawtooth instead of the triangular shape of the symmetrical PDs (see Figure 4d). The advantages of a sawtooth PD function are that the

working ranges are larger, and with very large input modulations, the PLL immediately shifts to another valid working range.

Other options that are probably available for neuronal PLLs are dynamic tuning of the RCO's local frequency, asymmetric RCO transfer functions, and combined excitatory-inhibitory implementations of PLLs (see Figure 3a) with asymmetric or dynamically shifted relative weights.

4.2 Plausible Sites for PLLs. Neuronal circuits that contain local oscillators probably can function as PLLs without any specific tuning. In principle, the feedback connections and the basic phase detection function of any neuron (see section 2.2.1.1) establish the essential requirements of the loop. Nevertheless, efficient operation at a specific frequency range requires additional tuning of the cellular and circuit parameters (see section 2.2). Some neuronal systems have at least some of the required parameters for efficient PLLs. Following is a summary of the requirements from the circuitry and local oscillators and a review of plausible sites.

4.2.1 Potential Circuits. The sensory thalamocortical loops are attractive candidates for PLL circuits, since PLL performance improves when the RCO-to-PD connections are massive, and having PLLs early in a processing stream would be advantageous for facilitating sensory-sensory integration. Nevertheless, feedback circuits within or between cortical areas could function as PLLs as well. Within thalamocortical loops, the natural implementation would be that the thalamic relay neurons function as PDs and corticothalamic neurons at deep cortical layers function as RCOs. Natural candidates for INH neurons in iPLLs are the cortical inhibitory interneurons in layer 4 (White & Keller, 1987; Agmon & Connors, 1992; Swadlow, 1995). However, other combinations, including inhibitory neurons of the reticular nucleus of the thalamus, or oscillatory neurons in superficial layers that drive the corticothalamic neurons, are also possible, as long as the loop transfer functions establish a stable negative feedback loop (see appendix A.2). Within thalamocortical systems, many PLLs are expected to function in parallel, each tuned to a different combination of receptive field and working range.

4.2.2 Local oscillators. In a PLL, a local oscillator should function as an RCO, that is, its output frequency should be controllable by the input. The RCO's oscillations can be sub- or suprathreshold during spontaneous activity, as long as when decoding starts, the oscillations become suprathreshold. With single or groups of cells, which are oscillating due to intrinsic mechanisms, control of the frequency of oscillations by the input is expected to obey the simple neuronal rules required by the PLL: excitatory inputs should increase the frequency of oscillations, whereas inhibitory inputs should decrease the frequency (see section 2.2.2). Thus, single-cell oscillators are excellent candidates for RCOs. In contrast, oscillations generated outside the

processing network are not expected to be affected by the excitation levels of the network and thus cannot function as RCOs. The effect of neuronal input on a network that oscillates due to specific connectivity of excitatory and inhibitory cells is not obvious and depends on the exact connectivity and the exact input. Thus, neuronal ensembles that exhibit network oscillations will not be considered candidates for RCOs but rather as circuits that can be driven by RCOs.

Brain activity contains oscillations in a wide range of frequencies, from circadian to millisecond ranges. However, only frequency ranges compatible with perceptual time scales, during which sensory temporal codes should be transferred to motor rate codes, will be discussed here. Emphasis will be placed on the 10–100 Hz range, although lower and higher frequencies can probably also be used for perceptual processing by PLLs.

4.2.3 The primate somatosensory system. Candidates for somatosensory RCOs were presented in section 3.3.2. If PLLs indeed operate in somatosensory thalamocortical circuits, network oscillations that occur in the primary somatosensory and motor areas (reviewed in Fetz, 1993) during the performance of tactile tasks or during periods with global excitation could be due to propagation of the working frequencies from the PD neurons to the sensorimotor areas. If input modulations are not overly strong, the working frequency is preserved in the synchronous PD firings, even though the firing rate of the whole population might be modulated (see Figure 6).

4.2.4 The rodent vibrissal system. Many rodents achieve tactile sensory acquisition with an active process in which their whiskers move back and forth in a sinusoidal-like manner with frequencies near 10 Hz (Welker, 1964; Simons, 1995). Rodents use such whisking to localize (Welker, 1964) and identify objects with strategies and resolution capabilities comparable to those of primates achieved by applying manual active touch (Simons, 1995). During active whisking, the vibrissal pathway of the rat oscillates synchronously at around 10 Hz (Nicolelis, Baccala, Lin, & Chapin, 1995). These oscillations probably originate in the cortex, but in every cycle, the peripheral neurons fire first, followed by the firing of cortical ones, which in turn is followed by a firing of thalamic neurons. These observations are fully consistent with PLL circuits of 10 Hz operating in the thalamocortical vibrissal system of the rat. Recently we observed that cortical oscillators in the somatosensory cortices of anesthetized rats and guinea pigs exhibit three modalities of oscillating frequencies, at roughly 1, 10, and 100 Hz (Ahissar et al., 1996). It is possible that the ~10 Hz oscillators are utilized in PLLs that detect the location of external objects and that the ~100 Hz oscillators are utilized in PLLs that decode the texture of these objects (see Carvell & Simons, 1995).

4.2.5 The auditory system. Under normal conditions, most of the single-cell oscillations in the auditory cortex have frequencies below 14 Hz (Ahissar & Vaadia, 1990). Thus, if auditory PLLs exist, they probably decode low-frequency information derived from relatively slow processes such as speech or movements of sound sources (Ahissar, Ahissar, Bergman, & Vaadia, 1992).

4.2.6 The olfactory system. Network oscillations occurring in the olfactory system (Freeman, 1975) are probably utilized to enhance cortical processing or to encode sensory information (Hopfield, 1995), but not to decode temporally encoded sensory information which is not conjectured in this sense.

4.2.7 The visual system. During stimulations, the visual pathway often exhibits synchronous, wide-band oscillations (Neuenschwander & Singer, 1996; Engel et al., 1992; Eckhorn, 1994). Whether these oscillations are utilized for the decoding of temporally encoded information or for recoding spatially encoded information is not yet clear. The fact that the internal frequencies are usually much higher than the so-called temporal frequencies of the stimulus (i.e., the frequencies at which single receptors are stimulated) is not indicative in this case. The temporal structure of the retinal output depends also on the frequency of the sequential activation of neighboring receptors since, at least in cats, several receptors usually converge onto single ganglion cells. The direct dependency of cortical frequencies on stimulus velocities (Gray, Engel, Konig, & Singer, 1990; Eckhorn, Frien, Bauer, Woelbern, & Kehr, 1993) supports the direct coupling between peripheral and cortical oscillations. Visual temporal decoding by PLLs could rely on single-cell oscillators in the thalamus or the cortex. Neurons in the lateral geniculate nucleus (LGN) exhibit spontaneous stable oscillations (Ghose & Freeman, 1992) that are disturbed once visual patterns are presented, which is consistent with these oscillators trying to track modulated temporal structures. Neurons in superficial layers of the cortex exhibit intrinsic oscillations during stimulations (Gray & McCormick, 1996). These neurons could function as RCOs probably only after initial sensory or internal preparatory excitation.

4.2.8 Summary. The perceptual mechanisms of two sensory (tactile and visual) systems that apparently can use PLL circuits involve motion of the sense organs during sensory acquisition. Such movements result in encoding of spatial information in temporal firing patterns, information that can be efficiently decoded by PLL circuits. Hand movements can easily be measured with a resolution higher than that of the tactile receptive fields, while such measurements with eye movements are difficult (Carpenter, 1988). Thus, accurate testing of PLL's predictions in visual systems is limited. PLLs might be implemented differently in these two systems. For example, tactile RCOs are expected to be cortical, whereas visual ones could be thalamic.

According to the PLL model, the network oscillations observed in the visual and sensorimotor cortices reflect oscillatory activities in either the output or the readout stages of PLL circuits.

After the cessation of oscillatory sensory stimulations in both the visual and somatosensory modalities, the brain persists in emitting synchronized oscillations having the same frequency of the stimulus (Narici et al., 1987). This "oscillatory memory" requires a closed-loop operation, at either the cellular or the circuit level, as suggested by the PLL model. Testing of this phenomenon at low frequencies revealed that best resonating frequencies for the somatosensory modality were 6 and 8 Hz and for the visual modality 10 Hz. It will be interesting to see whether similar phenomena occur with oscillations around 30 Hz in the somatosensory system and 40 to 100 Hz in the visual system.

4.3 Experimental Evidence for Thalamocortical Tactile PLLs. Current physiological and anatomical data are compatible with PLL's being implemented within and across the thalamic ventrobasal nuclei (VB), SI, and SII areas. Neurons that can be considered as local oscillators in the SA, PC, and mostly RA ranges exist in SI and SII areas of primates (see section 3.3.2). The high percentage of posterior SII oscillators and the grouping of oscillators with frequencies that correspond to the three submodalities (see Figure 11) suggests an important role for SII in temporal decoding of textures. In fact, lesions in SII of primates significantly impair tactile texture decoding (Murray & Mishkin, 1984; Carlson, 1990). Furthermore, the direct motor connections of SII to the primary motor cortex (MI) (Jones, 1986; Burton, 1986) would facilitate participation of SII in a basic sensorimotor loop, such as the one described by the AVC loop.

The input tactile channels are evidently not fully segregated; sensory information is probably shared by different frequency channels and even between different submodalities. Thus, the decoding details cannot be as simple as described here. However, if channel segregation holds to a certain degree, PLL-like decoding could occur, and in this case the decoding principles outlined in this article should hold. Note that although anatomical continuity within input channels is required for input pathways running up to the cortex and back to the thalamus, physiological consistency of response type is required only up to the thalamus. In fact, the PLL model suggests that a significant code transformation occurs at the thalamocortical level. Thus, the findings that cortical response types are not correlated with peripheral ones (e.g., Tremblay, Ageranioti-Belanger, & Chapman, 1996) are not in conflict with the PLL model.

Below are presented data that are consistent with (i.e., can be explained by) the PLL model and data that support the model (i.e., that are more consistent with the PLL model than with other models). Since no other specific mechanism has yet been suggested for texture decoding at the circuit

level, the PLL will be compared with the open-loop model of local oscillators (Ahissar, 1995) and with "non-PLL" mechanisms in general.

4.3.1 Data consistent with PLLs functioning in the tactile thalamocortical system.

- The RA system employs amplitude limiting, uniform receptive fields, and temporal dispersions, all required for efficient thalamic phase detection (see sections 3.1 and 3.3.1).
- The RA pathway exhibits a high degree of phase locking that preserves the temporal information up to the thalamus (see the introduction to section 3).
- The mechanoreceptors and their fibers act as bandpass filters; they emphasize a certain range of the input frequency spectrum (Johansson et al., 1982; Freeman & Johnson, 1982; Goodwin et al., 1989), as required for efficient PLL decoding.
- The reciprocal connections between the thalamus and cortex are modality and somatic specific (Jones, 1986; Doetsch, Standage, Johnston, & Lin, 1988; Hoogland, Welker, & Van der Loos, 1987).
- The circuitry required for the function of thalamocortical ePLLs and iPLLs exists in mammals (Jones, 1986; White & Keller, 1987; Agmon & Connors, 1992).
- Thalamic relay neurons are activated with short EPSPs from afferent projections and long EPSPs from cortical inputs (Deschenes et al., 1984), an arrangement that facilitates thalamic PD operation (see section 3.3.1).
- Oscillating frequencies of SI neurons can be controlled locally (Silva et al., 1991; Amitai, 1994; Ahissar et al., 1996).
- Two successive stimuli to the same location on the skin are not differentiable for delays between 0 and 15–40 ms (Rosner, 1961), consistent with a PLL-like mechanism that "samples" the input using RA-range frequencies.
- Talbot et al. (1968) suggested the existence of a central mechanism that "alters its own activity [which "measures"] the dominant period in the input train of impulses." The PLL, by altering its own activity (the RCO's frequency), can "measure" the dominant input period.

4.3.2 Data that support PLLs in the tactile thalamocortical system.

- Local oscillators in SI of monkeys (Lebedev & Nelson, 1995) and of anesthetized rats and guinea pigs (Ahissar et al., 1996) can be entrained by oscillatory tactile stimuli when the input frequency is close to the local frequency.

- When vibratory stimuli are applied within series of decreasing intensities, minimal detection thresholds were consistently slightly lower than during series of increasing intensities (Talbot et al., 1968). Whereas this result cannot be explained by neural mechanisms involving adaptation or habituation, it is easily explained by PLL-like mechanisms: A minimal input intensity is required to lock in the PLL (increasing series), but once the PLL is locked (decreasing series), less input intensity is necessary to keep it locked, since the local oscillators already fire in phase with the input.
- A qualitative coding transformation, from temporally oriented at the thalamus (Sinclair, Sathian, & Burton, 1991) to rate oriented at the cortex (Sinclair & Burton, 1991), appears to occur at the thalamocortical level of monkeys performing a texture discrimination task. The gradual nature of the cortical responses is more consistent with the PLL than with alternative open-loop mechanisms producing labeled-line coding (Ahissar, 1995).

4.3.3 Data consistent with inhibitory PLLs.

- There is evidence "that presumed inhibitory interneurons in the cat SI could be activated first by thalamic inputs among cortical neurons and set to inhibit the output cells" (Yamamoto, Samejima, & Oka, 1988, p. 199).
- Activities of local oscillators in SI of the behaving monkey are often inhibited by vibrotactile stimuli (Lebedev & Nelson, 1995).
- Local oscillators in SI of the rat receive strong inhibitory input (Chagnac-Amitai & Connors, 1989).
- With SII neurons of cats, firing in phase with a vibratory stimulus is impaired when GABA receptors are blocked (Alloway, Sinclair, & Burton, 1988). This is consistent with cortical phase locking being achieved by inhibitory PLL circuits.
- In response to thalamic (VB) stimulation, corticothalamic neurons in SI of cats exhibit inhibitory postsynaptic potentials (Landry & Dykes, 1985). Synaptic excitation is also observed in some of these neurons, which suggests a combination of ePLLs and iPLLs.
- Somatosensory cortical neurons of rats have been classified according to whether they are coactivated with fast (~20 Hz) electroencephalogram waves (CoE cells) or not (CoI cells) (Angel, 1983). CoE neurons exhibit rhythmic firing around 20 Hz, dominate the electroencephalogram when CoI neurons are quiet, respond to peripheral inputs with longer latencies then CoI neurons, and activate thalamic (reticular) cells with a shorter latency then CoI neurons do. All of these

phenomena are consistent with CoE functioning as RCO neurons and with CoI functioning as inhibitory interneurons (INH).

- About 25% of the SI neurons of the monkey exhibit a sharp, positive sigmoidal dependency on the spatial period (Sinclair & Burton, 1991; see also Darian-Smith et al., 1982), as expected by iPLLs (see Figure 6). Negative sigmoidal dependency, as predicted by ePLLs, was not observed.

4.3.4 Data that seem inconsistent with PLLs in tactile thalamocortical systems.

- Although the primary function of the proposed PLLs in tactile thalamocortical systems would be in perception of patterns, these PLLs should, using only temporal input information, be able to convey information about the roughness of the scanned texture. However, in some cases, estimation of roughness is independent of the temporal parameters of the peripheral input (Lederman, 1981). This would imply that even if PLL circuits exist in the somatosensory system, their contribution to the perception of roughness is negligible. However, so far, only a partial range of possible conditions has been studied—for example, only spatial frequencies of relatively high frequencies (groove widths of 0.175–1 mm; Lederman, 1981). Also, the involvement of temporal information was tested only after subjects were trained to estimate the roughness of different textures at different velocities. Since the nervous system is capable of developing perceptual constancies over many parameters, one of which is probably finger velocity, naive rather than trained subjects should have been used. When naive subjects are required to identify forms or discriminate gratings, perception indeed depends on finger velocity (Vega-Bermudez, Johnson, & Hsiao, 1991; Ahissar & Gamzu, 1995). Furthermore, during training with a difficult discrimination task, subjects developed scanning strategies that were based on maximizing differences between temporal frequencies by controlling the scanning hand velocities (Gamzu, Haidarliu, & Ahissar, 1994).

- As the spatial frequency of the stimulus decreases, the SA and RA mechanoreceptive fibers in the hand of the monkey fire more spikes per second, even if the peak temporal frequency of the stimulus is kept constant (Goodwin & Morley, 1987). This seems to contradict a basic assumption of the tactile PLL that peripheral firing depicts in a 1:1 manner the existence of texture ridges. These experiments were conducted with indentations (1 mm) well above the peripheral threshold (tens of microns; Talbot et al., 1968), which probably forced the peripheral fibers to function outside their plateau range (see Goodwin et al., 1989). Nevertheless, the spatial features were still represented by the peripheral temporal structure, though with a 1:n ratio (Mor-

ley & Goodwin, 1987; Goodwin et al., 1989). In principle, PLLs can decode 1:n input ratios, and the decoding efficiency depends on the parameters. For example, if the bursts increase the input spread (T_{wi}; see Figure 4) beyond $T_i/2$, performance could be impaired. However, if they bring T_{wi} closer to $T_i/2$, performance should improve. Since the length of these bursts increases as the spatial frequency decreases (Goodwin et al., 1989), they might indeed improve the efficiency of the putative PLLs. Note, however, that the peripheral burst lengths are not necessarily preserved at the outputs of the dorsal column nuclei.

- The PLL model for the tactile thalamocortical system predicts that the temporal information of the input will be represented by the firing rates of cortical populations. In contrast, Burton and Sinclair (1994) concluded that the cortex probably encodes spatial features of the surface independent of temporal factors. However, since only correlations between *average* values of cortical rates and hand velocities were computed, velocity was not systematically varied, and "velocity was not well controlled" (Sinclair & Burton, 1991, p. 165), these results cannot rule out representations of input temporal structures in cortical rates. Indeed, Chapmann and colleagues found recently, applying systematic variations of input velocities, that the firing rates of 66% of SI neurons are directly related to the stimulus velocity (Tremblay et al., 1996).

- Johnson and Lamb (1981) argued that the temporal dispersion caused by a wide distribution of conduction velocities of mechanoreceptive fibers, especially those of RA fibers, can contribute to a spatial dispersion (blurring) of the image of the scanned texture. This would indeed be the case with a central mechanism that blindly integrates input signals from all input fibers. However, a PLL-like mechanism can actually benefit from such dispersions, which convert input "click trains" to lemniscal square waves (see Figure 4c) and enable the PD operation (see sections 2.2.1 and 3.3.1).

- Connor and Johnson (1992) have compared spatial and temporal encoding schemes as possible candidates to underlie tactile roughness estimation and showed that spatial variations have the closest correlation with roughness estimations. However, in some spacing ranges, temporal variations are better than spatial variations in predicting the subjects' reports (compare Figures 9 and 10 in Connor & Johnson, 1992). A view consistent with this and other (e.g., Ahissar & Gamzu, 1995) studies is that different perceptual mechanisms are emphasized as a function of the task at hand and the range of the spatial frequencies being explored. Roughness estimation tasks and low-spatial-frequency pattern discrimination tasks are probably primarily dealt with by spatial mechanisms (Connor & Johnson, 1992, and Ahissar & Gamzu,

1995, respectively) while high-spatial-frequency pattern discrimination tasks are primarily dealt with by temporal mechanisms (Ahissar & Gamzu, 1995), such as the PLL. This is consistent with the finding that superimposed vibrations improve stationary grating resolution in a range of high spatial frequencies and decrease performance in a range of lower frequencies (Johnson & Phillips, 1981).

4.3.5 Summary of experimental evidence. Experimental data indicate that under certain conditions, operation of a PLL-like mechanism is feasible at the thalamocortical level of mammalian tactile systems. In this system, current data favor the existence of iPLLs over ePLLs, although combined operation of both implementations has been indicated. If such PLLs exist, they should function in parallel to other, nontemporal, decoding mechanisms.

4.4 Interactions with other tactile mechanisms. If PLL circuits do exist in the brain, it is likely that they do not exist as isolated circuits and they operate in parallel with other temporal and nontemporal decoding mechanisms. In fact, Johnson, Phillips, and colleagues have shown that some spatial features are most efficiently resolved by the SA system (Johnson & Lamb, 1981; Phillips & Johnson, 1981; Phillips et al., 1990), and this resolution appears to occur without significant dependency on temporal parameters (Phillips, Johnson, & Hsiao, 1988; Phillips, Johansson, & Johnson, 1992). They suggested that both "spatial" (involving the SA system) and "nonspatial" (involving the RA system) mechanisms underlie texture perception; the RA system probably encodes the microscopic dimensions and the SA the macroscopic dimensions of the texture (Johnson & Phillips, 1984). Similarly, I suggest that PLL circuits are embedded in and intermingled with other circuits and that, as a whole, these circuits function as a texture analyzer (Taylor, Lederman, & Gibson, 1973). Within such embedded and intermingled circuits, operations that obey PLL principles should occur in parallel to other operations that obey other principles. Other possible operations could be purely spatial, such as spatial variation detection (Connor & Johnson, 1992), or spatiotemporal, such as cross-coincidence detection among parallel phase-preserved input signals. In real time, the adaptive brain can emphasize one or another operation, according to the task at hand and previous experience. Thus, PLL circuits, which probably occur predominantly in the RA system, can decode temporal information related to the *pattern*—to the fine details of the surface—while spatial mechanisms (e.g., Bankman, Hsiao, & Johnson, 1990), which predominantly use the SA system, can decode rate-encoded information related to the macroscopic details (e.g., the roughness or shape) of textures. In addition, SA-based intensity mechanisms can refine *pattern* perception by using detailed spatial information, and PLL circuits can refine roughness perception by using fine temporal information.

4.5 Predictions of the Tactile PLL Model. The predictions derived directly from the algorithm are considered critical; a rejection of any one of them results in a rejection of the model or, at least, a major modification of it. A rejection of an implementation-specific prediction results in the rejection of only that specific implementation. The electrophysiological predictions require a distinction between two types of neurons: RCO neurons, which are local oscillators, and PD (or PD-driven) neurons, whose oscillating activity is externally driven. A partial list of both types of predictions follows. The implementation-specific predictions are detailed only for the AND-like, vector PDs (see section 2.2.1.2). For these implementations, the thalamocortical relay neurons can function as PDs only if the lemniscal input is subthreshold; therefore, the related predictions apply only for light touch, such as that used for texture discrimination.

Algorithm-Derived ("critical") Predictions.

AVC predictions.

a1. During a difficult identification of a patterned texture, the exploring velocities are expected to maintain the average temporal frequency of the input within one of the three ranges that correspond to the trimodal distribution of cortical oscillating frequencies (see Figure 11), with the RA range being preferred.

PLL predictions.

a2. RCO neurons are expected to track, within a range around their spontaneous oscillating frequency, variations in the frequency of a vibratory stimulus.

a3. When a periodic stimulus is applied at a frequency that matches the frequency of the RCO, the PD neurons are expected to be phase locked with both the stimulus and the RCO neurons, and, during phase-locking, the spikes of the PD neurons should usually (and in AND-like implementations always) lag those of the RCO neurons.

a4. When the loop is locked, the net excitatory input to the RCO should be a monotonic increasing function of the input frequency. This is because in order to follow a higher frequency, a neuronal RCO needs to be excited further.

a5. While the PLL is locked, as the frequency of the stimulus is increased, the delay between the input and the RCO neurons, $(\eta_0 - \eta_i)$, becomes more positive (see Figure 4d).

a6. Within the PLL's working range, the response of the PD population should be monotonic with the input frequency. The polarity of this relationship depends on the implementation (see Figure 3) and the measurement point (e.g., before or after an inhibitory stage).

Implementation-Specific ("Noncritical") Predictions for AND-Like Vector PDs.

Thalamocortical Implementations.

t1. Within groups (or "rods"; Jones, 1986) of thalamocortical relay neurons that share the same receptor modality, receptive field location, and cortical projecting area, different neurons will have different response latencies (phase shifts) that preferentially cover a range of several ms.

Excitatory PLLs (ePLLs).

e1. Usually an abrupt and strong peripheral stimulus should increase the instantaneous frequency of an RCO.

e2. The higher the input frequency, the higher the fraction of PD neurons that should respond to the input.

e3. Entrainment of single PD neurons should exhibit a steplike or sigmoidal dependency on input frequency. They should not be entrained to low frequencies and should start to respond once the input frequency becomes higher than a certain threshold.

e4. At low input frequencies, only PD neurons that respond with short latencies should respond. As the input frequency increases, additional PD neurons, which have incrementally longer response latencies, should be recruited. Thus, although all PD neurons can maintain phase locking at high-input frequencies, PD neurons with shorter response latencies should maintain phase locking to stimuli of lower frequencies.

Inhibitory PLLs (iPLLs).

i1. Usually an abrupt and strong peripheral stimulus should decrease the instantaneous frequency of an RCO.

i2. The higher the input frequency, the lower the fraction of PD neurons that should respond to the input.

i3. Entrainment of single PD neurons should exhibit a steplike or sigmoidal dependency on input frequency. They should be entrained to low frequencies and should stop responding once the input frequency becomes higher than a certain threshold.

i4. At high-input frequencies, only PD neurons that respond with long latencies should respond. As the input frequency decreases, additional PD neurons, which have decrementally shorter response latencies, should be recruited. Thus, although all PD neurons can maintain phase locking at low input frequencies, a PD neuron with a longer response latency should maintain phase locking to stimuli of higher frequencies.

Appendix

A.1 Temporally- and Rate-Encoded Neuronal Signals. Any spike train that consists of N spikes of a single neuron can be described as (see Figure 1a)

$$s(t) = \sum_{n=0}^{N-1} S(t - \eta(n)) \tag{A.1}$$

where $S(t')$ describes a single spike triggered at $t' = 0$ (see Figure 1a inset), and $\eta(n)$ describes the series of spike timings,

$$\eta(n) = \eta(0) + nT + M(n) = \eta(0) + nT + \sum_{j=1}^{n} m(j), \qquad n > 0 \tag{A.2}$$

where T is the average ISI; $M(n)$ is the "absolute" modulation of T for the nth spike and represents the deviation of the timing of the nth spike from the expected timing of the nth spike of the equivalent ideal oscillator having the same T; and $m(n)$ is the "cycle modulation" and represents the deviation of the nth ISI from T. By definition, the total modulation over the whole spike train should be zero

$$\left[M(N) = \sum_{j=1}^{n} m(j) = 0 \right].$$

For an ideal oscillator, $M(n) = m(n) = 0$ for every n. The instantaneous ISI is (see Figure 1a):

$$I(n) = \eta(n) - \eta(n-1) = T + m(n), \qquad n > 0 \tag{A.3}$$

$$\eta(n) = \eta(0) + \sum_{j=1}^{n} I(j), \qquad n > 0. \tag{A.4}$$

It is assumed, as a convention, that the spike train was not modulated prior to $n = 0$; therefore:

$$I(0) = T; \qquad m(o) = 0; \qquad m(j) = 0, \quad j < 0 \tag{A.5}$$

The information carried by the spike train is described by T and m. Generally the information carried by T and m could be referred to as rate-encoded and temporally-encoded, respectively, since T is a measure of the average firing rate over the whole period and m is a measure of the fine temporal modulations within that period.

A description of a signal by rate requires the division of a spike train into rate bins, with each rate bin being represented by a single number. Each of these single numbers can be evaluated by a variety of functions, ranging from a simple spike count to a weighted average that uses a postsynaptic filter function. Here, rate-encoded signals will be represented by simple spike counts over each rate bin. If other measurements are required, the spike count terms (e.g., $A(k)$ in equation A.6) should simply be replaced with other terms. Thus, a spike train can be described by a series of counts of spikes, where each count corresponds to a single rate bin (see Figure 1b):

$$R_x(t) = \sum_{k=0}^{Nr-1} A(k)R(t - kT_r), \tag{A.6}$$

where T_r is the rate bin, $R(t')$ is a pulse function that equals 1 for $0 \leq t' < T_r$ and 0 otherwise (see Figure 1b, inset), $A(k)$ is the spike count of the neuron at the kth rate bin, and N_r is the number of rate bins in the spike train.

For simplicity, negative "firing rates" will be ascribed to inhibitory inputs. Thus, a rate signal is defined as the difference between the count of spikes leading to EPSPs and the count of spikes leading IPSPs. For example, a single excitatory cell can produce only positive rate values, and a single inhibitory cell can produce only negative rate values.

Two kinds of population rate coding are considered here: population sum, which, per each rate bin k, is the sum of all $A_i(k)$, and population vector, which, per each rate bin k, is the array of all $A_i(k)$.

A.2 Phase-Locked Loop.

A.2.1 Rate-controlled oscillator. The RCO's output signal is:

$$s_o(t) = \sum_{n=0}^{N-1} S(t - \eta_o(n)) \tag{A.7}$$

where

$$\eta_o(n) = \eta_o(0) + nT_c + M_c(n). \tag{A.8}$$

T_c is the RCO's intrinsic period—its ISI when it receives no input—and $M_c(n)$ is the nth spike's absolute modulation (see equations A.2 and A.3 for other related definitions). The ISI (the "cycle") of the RCO is controlled by its input in the following way:

$$I_o(n) = T_c + g_o(R_d(n)), \tag{A.9}$$

where g_o, in the general case, is a monotonic decreasing or a monotonic increasing function, $g_o(0) = 0$, and $R_d(n)$ is the input to the RCO integrated over the interval preceding spike n, during $I_o(n)$ (see Figure 1 and

section 2.2.2). In neuronal implementations, g_o will probably always be a decreasing function in which the more positive (excitatory) the oscillator's input is, the sooner the oscillator will fire its next spike, and vice versa for more negative (inhibitory) inputs.

The average value of $R_d(n)$ is not necessarily, and usually will not be, 0. Therefore, T_c will not necessarily equal the average ISI of the RCO. To be consistent with equation A.3, for any given decoding period, $R_d(n)$ will be described as being composed of two components: a DC component (the average value, R_{dc}) and an AC component (the residual modulations, $R_{ac}(n)$):

$$R_d(n) = R_{dc} + R_{ac}(n), \tag{A.10}$$

and the average ISI of the RCO, T_o, will be:

$$T_o = T_c + \langle g_o(R_d(n)) \rangle, \tag{A.11}$$

where $\langle x \rangle$ is the average value of x over the described decoding period. Thus, the output timings of the RCO can be rewritten as:

$$\begin{aligned} \eta_o(n) &= \eta_o(0) + nT_o + M_o(n) \\ &= \eta_o(0) + nT_o + \sum_{j=1}^{n} m_o(j), \quad n > 0. \end{aligned} \tag{A.12}$$

For a linear g_o we get:

$$\langle g_o(R_d(n)) \rangle = g_o(R_{dc}), \tag{A.13}$$

and the instantaneous ISI is (from equations A.3, A.9, and A.11),

$$I_o(n) = T_o + g_o(R_{ac}(n)). \tag{A.14}$$

A.2.2 Phase detector. The PD's output, $R_d(n)$, is a rate-encoded signal, which is a function of the difference between the arrival times of the PD's two inputs,

$$R_d(n+1) = g_d(\eta_o(n) - \eta_i(n)), \tag{A.15}$$

where g_d is a monotonic increasing or a monotonic decreasing function. The difference $\eta_o(n) - \eta_i(n)$ is simply the difference between the times of appearance of the nth spikes of the RCO and the PLL's input, where n is counted only within a locked state, when the RCO's and the input's spikes are paired. With neuronal implementations, g_d probably cannot achieve a strict monotonic shape but rather will assume a staircase-like form. There will be ranges of phase difference within which the PD will produce a constant output. The size of these ranges determines the PD's resolution (see section 2.2.1) and, hence, also the PLL's resolution.

A.2.3 The loop gain. The ability of the PLL to be locked to the input depends on the loop's functioning as a negative feedback loop. In such a negative feedback loop, any deviation of the input from its expected frequency will produce an error signal (R_{ac}) that will drive the RCO's frequency in the direction that will reduce the error—in the same direction as the input's deviation. To provide a negative feedback, the gain along the loop during one cycle, referred to as the *loop gain*, should be negative. The loop gain, G, is computed per a *working point* (e.g., the crossing point in Figure 2b), assuming a constant input. It is equal to the gain of a small perturbation from the working point that is obtained after one cycle and is approximately (exactly for a linear system)

$$G \approx g_o' g_d', \tag{A.16}$$

where, with continuous g_o and g_d, g_o' and g_d' are the derivatives of g_o and g_d, respectively, at the working point. With discrete g_o or g_d, g_x' equals $\Delta y/\Delta x$, where Δy is the output change generated by a minimal input change (Δx). The PLL will be stable only if any deviation from a working point, generated within the loop while the input is constant will be attenuated at the next cycle. A perturbation will be canceled at the next cycle if $G = -1$, will be attenuated if $-1 < G < 0$, and will be inverted and attenuated if $-2 < G < -1$. Outside this range, any perturbation will increase in absolute magnitude with each successive cycle. Thus, a necessary, although not sufficient, requirement for a stable PLL is

$$-2 < G < 0. \tag{A.17}$$

Therefore, to keep a PLL stable g_d and g_o must have opposite slopes around the working point (see, for example, Figure 2b). A range within which input modulations can be decoded is defined as a *working range* of the PLL. This range is determined by equation A.17, the dynamic range of the RCO, and the input average frequency.

A.3 Tracking. For clarity, let us consider an *ideal PLL*, in which the derivatives of g_d and g_o are constant (equal to k_d and k_o, respectively), $G = -1$, the RCO fires single spikes per cycle, and there is no noise. Suppose the input to the PLL (hereafter "the Input") is:

$$s_i(t) = \sum_{n=0}^{N-1} S(t - \eta_i(n)), \qquad I_i(n) = T_i + m_i(n), \qquad n > 0 \tag{A.18}$$

and

$$m_i(n) = 0, \qquad n \le 0.$$

When the loop is locked, there is one and only one RCO spike per each Input spike. If the Input is not modulated (i.e., it is perfectly periodic), the timings of the RCO's spikes will differ from the Input spikes only by a constant time delay (phase shift),

$$\eta_o(n) = \eta_i(n) + \Delta\varnothing \tag{A.19}$$

where $\Delta\varnothing$ is a constant time difference and n is the index of the Input cycle. And

$$I_o(n) = I_i(n). \tag{A.20}$$

When the Input is temporally modulated, the modulation is detected by the PD, which detects the difference between the expected $\eta_i(n)$ ("stored" as $\eta_o(n)$) and the actual timing. The detected difference corrects the RCO's frequency so that the latter matches the Input frequency. However, this correction will take place only at the following cycle $(n+1)$:

$$I_o(n) = I_i(n-1) \tag{A.21}$$

and, from equation A.3,

$$T_o + m_o(n) = T_i + m_i(n-1). \tag{A.22}$$

By definition, T_o is the average ISI of the RCO, and as long as the loop is locked, it is equal to the average Input ISI, T_i,

$$T_o = T_i \tag{A.23}$$

and, therefore,

$$m_o(n) = m_i(n-1). \tag{A.24}$$

Thus, the Input modulation is replicated by the RCO's modulation, with one cycle lag.

The average periods, T_o and T_i, are defined de facto for every decoding period. Thus, the PLL does not "know" the exact values for these averages during the decoding, and a decomposition of its output signal to the different components will fit the above definitions only at the end of the decoding period. Note, however, that this non-causal process relates only to the observer's interpretation of the decoding process and does not relate to the process itself, since the decoding utilizes the actual timings [$\eta_o(n)$ and $\eta_i(n)$] of the signals (equations A.9 and A.15).

A.4 Decoding. It can be shown that with ideal PLLs,

$$R_{dc} = g_d(\Delta\varnothing) \tag{A.25}$$

$$R_{ac}(n+1) = -g_d(m_i(n)) = g_o^{-1}(m_i(n)) \tag{A.26}$$

$$\Delta\varnothing = T_c - T_i + \varnothing_m. \tag{A.27}$$

$\Delta\varnothing$ is the average phase difference and, together with R_{dc}, determines the PLL's working point (see Figure 2b). $\varnothing_m$ is a constant delay that depends on the implementation (see section A.5.2).

If g_d or g_o is not linear, the AC component will depend on the DC component, that is, on the working point. However, since g_d (and g_o^{-1}) is monotonic, $R_d(n)$ is unique (within the resolution limits) for every input. When $G \neq -1$, the above solutions are the steady-state solutions that are obtained after a variable number of cycles, depending on G.

Using minimal rate bin ($= T_i$), the output of the ideal PLL is (see equation A.6),

$$R_d(t) = g_d(T_c - T_i + \varnothing_m) - \sum_{n=0}^{N-1}[g_d(m_i(n-1))R(t - nT_i)]. \tag{A.28}$$

Readout mechanism that employ longer rate-bins should sample or integrate the PLL's output.

A.5 PD Implementations.

A.5.1 A single neuron PD. The working range of such a PD is determined by the effective widths of its inputs—the maximal delay from an onset of an EPSP in which, if an EPSP from the other input is added, the membrane voltage will cross the threshold. For similar inputs whose (EPSP amplitude)/(threshold distance) $= A$

$$T_w = \tau \ln(A/(1-A)). \tag{A.29}$$

Assuming $\tau = 10$ ms is the decay time constant, if $A = 0.8$, the working range (T_w) is almost 14 ms, and if $A = 0.9$, it is about 22 ms.

A.5.2 Population PDs. Within the working range of the ePLL (T_{we}; see Figure 4d),

$$R_d(n+1) = R_{\max} + k_d[\eta_o(n) - \eta_i(n)], \qquad k_d > 0. \tag{A.30}$$

Since g_o is monotonic decreasing, $G < 0$ and the basic algorithm can be implemented straightforwardly by the circuit denoted by the dashed lines in Figure 3a. With the PD implementation of Figure 4d, the average delay is

$$\Delta\varnothing = T_c - T_i - R_{\max}/k_d. \tag{A.31}$$

Within the working range of the iPLL (T_{wi}, see Figure 4d),

$$R_d(n+1) = R_{\max} - k_d[\eta_o(n) - \eta_i(n)], \qquad k_d > 0. \tag{A.32}$$

In this case, $G < 0$ because of the INH that are added to the loop (see Figure 3a, solid lines). The average delay for the Figure 4d iPLL implementation is

$$\Delta\varnothing = T_c - T_i + R_{\max}/k_d. \tag{A.33}$$

The PLL's output, in both the excitatory and inhibitory implementations, is a population output.

A.6 Tactile Signals. It is assumed, for simplicity, that for any given scanning direction, all ridges have negligible widths. Textures consisting of a collection of such ridges can be described, along any one-dimensional direction, in a discrete form by

$$u_i(x) = \sum_{n=0}^{N} A_i(n)U[x - \xi_i(n)], \tag{A.34}$$

where $U(x')$ describes a single ridge at $x' = 0$ with a unit height, $A_i(n)$ is the height of the nth ridge, and $\xi_i(n)$ describes the location of the nth ridge,

$$\begin{aligned} \xi_i(n) &= \xi_i(0) + nX_i + P_i(n) = \xi_i(0) + nX_i + \sum_{j=0}^{n} p_i(j); \\ G_i(n) &= X_i + p_i(n), \end{aligned} \tag{A.35}$$

where X_i is the average spatial period, $P_i(n)$ is the absolute modulation of this period for the nth ridge, $p_i(n)$ is the cycle modulation, and $G_i(n)$ is the inter-ridge-interval (see section A.1). If a mechanoreceptor response is 1:1, then for a constant finger velocity, V, the sensory transformation is simply

$$\eta_i(n) = \xi_i(n)/V, \qquad \eta_i(0) = \xi_i(0) = 0, \tag{A.36}$$

and similar relationships hold for the signals' components:

$$T_i = X_i/V; \quad I_i(n) = G_i(n)/V; \quad m_i(n) = p_i(n)/V. \tag{A.37}$$

A.7 Tactile Decoding. With ideal PLLs, minimal rate bins ($= T_i$) and constant finger velocity, the two output components (see equations A.28 and A.37) are,

$$R_{dc}(t) = g_d(T_c - X_i(t)/V(t) + \varnothing_m) \tag{A.38}$$

$$R_{ac}(t) = -\sum_{n=0}^{N-1}[g_d(p_i(n-1)/V(t))R(t - nT_i)] \tag{A.39}$$

where $\varnothing_m$ is an implementation-specific delay.

A.8 Automatic Velocity Control. Suppose the desired working point is T_{cw}; then the requirement is that

$$X_i(t)/V(t) = T_{cw}, \qquad dx = V(t)dt, \tag{A.40}$$

and

$$V(t) = X_i(t)/T_{cw} \tag{A.41}$$

is the finger velocity that the AVC should produce. The AVC is a servo-regulating loop (see Figure 10) whose controlled variable, $T_i(t)$, should be kept constant. The loop equations are:

$$V(t) = g_v(R_{\max} - R_{dc}(t)) + V_{\text{other}} \tag{A.42}$$

$$R_{dc}(t) = g_{pll}(T_i(t) - T_c) \tag{A.43}$$

where g_{pll} and g_v are the PLL's and VC's transfer functions, respectively (see Figure 10b), $R_{\max}$ is the maximal possible $R_{dc}(t)$, V_{other} is the velocity additive component caused by the "other velocity control," and T_c is the PLL's intrinsic frequency (see appendix A.2).

Acknowledgments

I thank S. Serulnik and M. Zacksenhouse for extensive and illuminating discussions on phase-locked loops; A. Aertsen, M. Ahissar, D. Blake, T. Flash, K. O. Johnson, D. Shoham, A. Treves, S. Ullman, and several anonymous referees for their helpful comments on previous versions of this article; and B. Schick for reviewing the manuscript. This work was supported by the Alon Foundation, Israel; the Minna-James-Heineman Foundation, Germany; and grant 93-198 from the United States–Israel Binational Science Foundation, Jerusalem, Israel.

Note added in proof:
Recently we found that the behavior of cortical oscillators in the barrel cortices of anesthetized rats and guinea pigs confirms predictions a2 and a5. Also, the behavior of multi-units at the thalamic recipient layers of these cortices is consistent with predictions a5 (under the assumption of AND-like PD operation) and a6 (Ahissar, E., Haidarliu, S., & Zacksenhouse, M. (1997) "Decoding temporally encoded sensory input by cortical oscillations and thalamic phase comparators." *Proc. Natl. Acad. Sci. USA. 94*:11633–11638). Note also that the results of Nicolelis et al. (1995) confirm prediction a3.

References

Abeles, M. (1982). Role of the cortical neuron: Integrator or coincidence detector? *Isr. J. Med. Sci., 18,* 83–92.

Abeles, M., Bergman, H., Margalit, E., & Vaadia, E. (1993). Spatiotemporal firing patterns in the frontal cortex of behaving monkeys. *J. Neurophysiol., 70,* 1629–1638.

Aertsen, A., Erb, M., & Palm, G. (1994). Dynamics of functional coupling in the cerebral cortex: An attempt at a model-based interpretation. *Physica-D, 75,* 103–128.

Agmon, A., & Connors, B. W. (1992). Correlation between intrinsic firing patterns and thalamocortical synaptic responses of neurons in mouse barrel cortex. *J. Neurosci., 12,* 319–329.

Ahissar, E. (1995). *Conversion from temporal-coding to rate-coding by neuronal phase-locked loops* (Tech. Rep. GC-EA/95-4). Rehovot, Israel: Weizmann Institute of Science.

Ahissar, M., Ahissar, E., Bergman, H., & Vaadia, E. (1992). Encoding of sound source and movement: The activity of single neurons and interactions between adjacent neurons in the primary auditory cortex of monkeys. *J. Neurophysiol., 67,* 203–215.

Ahissar, E., Alkon, G., Zacksenhouse, M., & Haidarliu, S. (1996). Cortical somatosensory oscillators and the decoding of vibrissal touch. *Soc. Neurosci. Abstr., 22,* 18.

Ahissar, E., & Gamzu, E. (1995). Utilization of temporally-encoded versus spatially-encoded information during the performance of a tactile discrimination task. *Soc. Neurosci. Abstr., 21,* 1018.

Ahissar, E., & Vaadia, E. (1990). Oscillatory activity of single units in a somatosensory cortex of an awake monkey and their possible role in texture analysis. *Proc. Natl. Acad. Sci. USA, 87,* 8935–8939.

Alloway, K. D., Sinclair, R. J., & Burton, H. (1988). Responses of neurons in somatosensory cortical area II of cats to high-frequency vibratory stimuli during iontophoresis of a GABA antagonist and glutamate. *Somatosens. Mot. Res. 6*(2), 109–140.

Alonso, A., & Llinas, R. R. (1989). Subthreshold Na^+-dependent theta-like rhythmicity in stellate cells of entorhinal cortex layer II. *Nature, 342,* 175–177.

Amitai, Y. (1994). Membrane potential oscillations underlying firing patterns in

neocortical neurons. *Neuroscience, 63,* 151–161.

Angel, A. (1983). The functional interrelations between the somatosensory cortex and the thalamic reticular nucleus: Their role in the control of information transfer across the specific somatosensory thalamic relay nucleus. In G. Macchi, A. Rustioni, and R. Spreafico (Eds.), *Somatosensory integration in the thalamus* (pp. 222–239). Amsterdam: Elsevier.

Bankman, I. N., Hsiao, S. S., & Johnson, K. O. (1990). Neural image transformation in the somatosensory system of the monkey: Comparison of neurophysiological observations with responses in a neural network model. In *Cold Spring Harbor Symposia on Quantitative Biology* (Vol. 55, pp. 611–620). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Berkley, M. (1978). Vision: Geniculocortical system. In R. B. Masterton (Ed.), *Handbook of behavioral neurobiology, Vol 1: Sensory integration* (pp. 165–207). New York: Plenum Press.

Boussaoud, D., & Wise, S. P. (1993). Primate frontal cortex: Effects of stimulus and movement. *Exp. Brain Res., 95,* 28–40.

Buanomuno, D., & Merzenich, M. M. (1995). Temporal information transformation into a spatial code by a neural network with realistic properties. *Science, 267,* 1028–1030.

Burton, H. (1986). Second somatosensory cortex and related areas. In E. G. Jones and A. Peters (Eds.), *Cerebral cortex* (Vol. 5, pp. 31–98). New York: Plenum Press.

Burton, H., & Sinclair, R. J. (1991). Second somatosensory cortical area in Macaque monkeys: 2. Neuronal responses to punctate vibrotactile stimulation of glabrous skin on the hand. *Brain Res., 538,* 127–135.

Burton, H., & Sinclair, R. J. (1994). Representation of tactile roughness in thalamus and somatosensory cortex. *Can. J. Physiol. Pharmacol., 72,* 546–557.

Calvin, W. H. (1975). Generation of spike trains in CNS neurons. *Brain Res., 84,* 1–22.

Carlson, M. (1990). The role of somatic sensory cortex in tactile discrimination in primates. In E. G. Jones and A. Peters (Eds.), *Cerebral Cortex* (Vol. 8B, pp. 451–486). New York: Plenum Press.

Carpenter, R. H. S. (1988). *Movements of the eyes.* 2nd ed. London: Pion.

Carr, C. E. (1993). Processing of temporal information in the brain. *Annu. Rev. Neurosci., 16,* 223–243.

Carvell, G. E., & Simons, D. J. (1995). Task- and subject-related differences in sensorimotor behavior during active touch. *Somatosens. Mot. Res., 12,* 1–9.

Chagnac-Amitai, Y., & Connors, B. W. (1989). Synchronized excitation and inhibition driven by intrinsically bursting neurons in neocortex. *J. Neurophysiol., 62,* 1149–1162.

Chapman, C. E. (1994). Active versus passive touch: Factors influencing the transmission of somatosensory signals to primary somatosensory cortex. *Can. J. Physiol. Pharmacol., 72,* 558–570.

Connor, C. E., & Johnson, K. O. (1992). Neural coding of tactile texture: Comparison of spatial and temporal mechanisms for roughness perception. *J. Neurosci., 12*(9), 3414–3426.

Darian-Smith, I., & Kenins, P. (1980). Innervation density of mechanoreceptive fibres supplying glabrous skin of the monkey's index finger. *J. Physiol., 309,* 147–155.

Darian-Smith, I., & Oke, L. E. (1980). Peripheral neural representation of the spatial frequency of a grating moving at different velocities across the monkey's finger pad. *J. Physiol., 309,* 117–133.

Darian-Smith, I., Sugitani, M., Heywood, J., Karita, K., & Goodwin, A. (1982). Touching textured surfaces: Cells in somatosensory cortex respond both to finger movement and to surface features. *Science, 218,* 906–909.

Deschenes, M., Paradis, M., Roy, J. P., & Steriade, M. (1984). Electrophysiology of neurons of lateral thalamic nuclei in cat: Resting properties and burst discharges. *J. Neurophysiol., 51,* 1196–1219.

Doetsch, G. S., Standage, G. P., Johnston, K. W., & Lin, C. S. (1988). Thalamic connections of two functional subdivisions of the somatosensory forepaw cerebral cortex of the raccoon. *J. Neurosci., 8*(6), 1873–1886.

Dykes, R. W. (1983). Parallel processing of somatosensory information: A theory. *Brain Res. Rev., 6,* 47–115.

Eckhorn, R. (1994). Oscillatory and non-oscillatory synchronizations in the visual cortex and their possible roles in associations of visual features. *Prog. Brain Res., 102,* 405–426.

Eckhorn, R., Frien, A., Bauer, R., Woelbern, T., & Kehr, H. (1993). High frequency (60–90 Hz) oscillations in primary visual cortex of awake monkey. *NeuroReport, 4,* 243–246.

Engel, A. K., Konig, P., Kreiter, A. K., Schillen, T. B., & Singer, W. (1992). Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends. Neurosci., 15,* 218–226.

Ferrington, D. G., & Rowe, M. (1980). Differential contributions to coding of cutaneous vibratory information by cortical somatosensory areas I and II. *J. Neurophysiol., 43,* 310–331.

Fetz, E. E. (1993). Cortical mechanisms controlling limb movement. *Curr. Opin. Neurobiol., 3,* 932–939.

Freeman, A. W., & Johnson, K. O. (1982). Cutaneous mechanoreceptors in Macaque monkey: Temporal discharge patterns evoked by vibration, and a receptor model. *J. Physiol., 323,* 21–41.

Freeman, W. J. (1975). *Mass action in the nervous system.* New York: Academic Press.

Gamzu, E., Haidarliu, S., & Ahissar, E. (1994). Sensorimotor control: Dependence of finger velocity on the scanned spatial frequency during performance and learning of a tactile task. *Isr. Soc. Neurosci., 3,* 57.

Gardner, F. M. (1979). *Phaselock techniques.* New York: Wiley.

Gardner, E. P., & Palmer, C. I. (1989). Simulation of motion on the skin. I. Receptive fields and temporal frequency coding by cutaneous mechanoreceptors of OPTACON pulses delivered to the hand. *J. Neurophysiol., 62,* 1410–1435.

Gardner, E. P., Palmer, C. I., Hamalainen, H. A., & Warren, S. (1992). Simulation of motion on the skin. V. Effect of stimulus temporal frequency on the representation of moving bar patterns in primary somatosensory cortex of monkeys. *J. Neurophysiol., 67,* 37–63.

Georgopoulos, A. P. (1986). On reaching. *Annu. Rev. Neurosci., 9,* 147–170.

Ghose, G. M., & Freeman, R. D. (1992). Oscillatory discharge in the visual system: Does it have a functional role? *J. Neurophysiol., 68,* 1558–1574.

Goldberg, J. M., & Brown, P. B. (1969). Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: Some physiological mechanisms of sound localization. *J. Neurophysiol., 32,* 613–636.

Goodwin, A. W., John, K. T., Sathian, K., & Darian-Smith, I. (1989). Spatial and temporal factors determining afferent fibre responses to a grating moving sinusoidally over the monkey's fingerpad. *J. Neurosci., 9*(4), 1280–1293.

Goodwin, A. W., & Morley, J. W. (1987). Sinusoidal movement of a grating across the monkey's fingerpad: Effect of contact angle and force of the grating on afferent fiber responses. *J. Neurosci., 7*(7), 2192–2202.

Gottschaldt, K.-M., Vahle-Hinz, C., & Hicks, T. P. (1983). Electrophysiological and micropharmacological studies on mechanisms of input-output transformation in single neurones of the somatosensory thalamus. In G. Macchi, A. Rustioni, & R. Spreafico (Eds.), *Somatosensory integration in the thalamus* (pp. 199–216). Amsterdam: Elsevier.

Gray, C. M., Engel, A. K., Konig, P., & Singer, W. (1990). Stimulus-dependent neuronal oscillations in cat visual cortex: Receptive field properties and feature dependence. *Eur. J. Neurosci., 2,* 607–619.

Gray, C. M., & McCormick, D. A. (1996). Chattering cells: Superficial pyramidal neurons contributing to the generation of synchronous oscillations in the visual cortex. *Science, 274,* 109–113.

Hoogland, P. V., Welker, E., & Van der Loos, H. (1987). Organization of the projections from barrel cortex to thalamus in mice studied with Phaseolus vulgaris-leucoagglutinin and HRP. *Exp. Brain Res., 68,* 73–87.

Hopfield, J. J. (1995). Pattern recognition computation using action potential timing for stimulus representation. *Nature, 376,* 33–36.

Hoppensteadt, F. C. (1986). *An introduction to the mathematics of neurons.* Cambridge: Cambridge University Press.

Jeffress, L. A. (1948). A place theory of sound localization. *J. Comp. Physiol. Psychol., 41,* 35–39.

Johansson, R. S., Landstrom, U., & Lundstrom, R. (1982). Responses of mechanoreceptive afferent units in the glabrous skin of the human hand to sinusoidal skin displacements. *Brain Res., 244,* 17–25.

Johansson, R. S., & Vallbo, A. B. (1979). Tactile sensibility in the human hand: Relative and absolute densities of four types of mechanoreceptive units in glabrous skin. *J. Physiol., 286,* 283–300.

Johansson, R. S., & Vallbo, A. B. (1980). Spatial properties of the population of mechanoreceptive units in the glabrous skin of the human hand. *Brain Res., 184,* 353–366.

Johansson, R. S., & Vallbo, A. B. (1983). Tactile sensory coding in the glabrous skin of the human hand. *Trends. Neurosci., 6,* 27–32.

Johnson, K. O., & Lamb, G. D. (1981). Neural mechanisms of spatial tactile discrimination: Neural patterns evoked by braille-like dot patterns in the monkey. *J. Physiol., 310,* 117–144.

Johnson, K. O., & Phillips, J. R. (1981). Tactile spatial resolution. I. Two-point

discrimination, gap detection, grating resolution, and letter recognition. *J. Neurophysiol., 46,* 1177–1191.

Johnson, K. O., & Phillips, J. R. (1984). Spatial and nonspatial neural mechanisms underlying tactile spatial discrimination. In C. von Euler, O. Franzen, V. Lindblom, and D. Ottoson (Eds.), *Somatosensory mechanisms* (pp. 237–248). London: Macmillan.

Jones, E. G. (1986). Connectivity of the primate sensory-motor cortex. In E. G. Jones & A. Peters (Eds.), *Cerebral cortex* (Vol. 5, pp. 113–184). New York: Plenum Press.

Lamb, G. D. (1983). Tactile discrimination of textured surfaces: Psychophysical performance measurements in humans. *J. Physiol., 338,* 551–565.

Landry, P., & Dykes, R. W. (1985). Identification of two populations of corticothalamic neurons in cat primary somatosensory cortex. *Exp. Brain Res., 60,* 289–298.

Lebedev, M. A., & Nelson, R. J. (1995). Rhythmically firing (20–50 Hz) neurons in monkey primary somatosensory cortex: Activity patterns during initiation of vibratory-cued hand movements. *J. Comp. Neurosci., 2,* 313–334.

Lederman, S. J. (1981). The perception of surface roughness by active and passive touch. *Bulletin of the Psychonomic Society, 18,* 253–255.

Llinas, R. R., Grace, A. A., & Yarom, Y. (1991). *In vitro* neurons in mammalian cortical layer 4 exhibit intrinsic oscillatory activity in the 10- to 50-Hz frequency range. *Proc. Natl. Acad. Sci. USA, 88,* 897–901.

Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. *Annu. Rev. Psychol., 42,* 135–159.

Morley, J. W., & Goodwin, A. W. (1987). Sinusoidal movement of a grating across the monkey's fingerpad: Temporal patterns of afferent fiber responses. *J. Neurosci., 7,* 2181–2191.

Mountcastle, V. B. (1993). Temporal-order determinants in a somatesthetic frequency discrimination—Sequential order coding. *Ann. N.Y. Acad. Sci., 682,* 150–170.

Mountcastle, V. B., Talbot, W. H., Sakata, H., & Hyvärinen, J. (1969). Cortical neuronal mechanisms in flutter-vibration studied in unanesthetized monkeys: Neuronal periodicity and frequency discrimination. *J. Neurophysiol., 38,* 452–484.

Murray, E. A., & Mishkin, M. (1984). Relative contributions of SII and area 5 to tactile discrimination in monkeys. *Behav. Brain Res., 11,* 67–83.

Narici, L., Romani, G. L., Salustri, C., Pizzella, V., Modena, I., & Papanicolaou, A. C. (1987). Neuromagnetic evidence of synchronized spontaneous activity in the brain following repetitive sensory stimulation. *Int. J. Neurosci., 32,* 831–836.

Nelson, R. J. (1996). Interactions between motor commands and somatic perception in sensorimotor cortex. *Curr. Opin. Neurobiol., 6,* 801–810.

Neuenschwander, S., & Singer, W. (1996). Long-range synchronization of oscillatory light responses in the cat retina and lateral geniculate nucleus. *Nature, 379,* 728–732.

Nicolelis, M. A. L., Baccala, L. A., Lin, R. C. S., & Chapin, J. K. (1995). Sensori-

motor encoding by synchronous neural ensemble activity at multiple levels of the somatosensory system. *Science, 268,* 1353–1358.

Perkel, D. H., & Bullock, T. H. (1968). Neural coding. *Neurosci. Res. Prog. Bull., 6,* 221–248.

Perkel, D. H., Schulman, J. H., Bullock, T. H., Moore, G. P., & Segundo, J. P. (1964). Pacemaker neurons: Effects of regularly spaced synaptic input. *Science, 145,* 61–63.

Phillips, J. R., & Johnson, K. O. (1981). Tactile spatial resolution. II. Neural representation of bars, edges, and gratings in monkey primary afferents. *J. Neurophysiol., 46,* 1192–1203.

Phillips, J. R., Johansson, R. S., & Johnson, K. O. (1990). Representation of braille characters in human nerve fibers. *Exp. Brain Res., 81,* 589–592.

Phillips, J. R., Johansson, R. S., & Johnson, K. O. (1992). Responses of human mechanoreceptive afferents to embossed dot arrays scanned across fingerpad skin. *J. Neurosci., 12,* 827–839.

Phillips, J. R., Johnson, K. O., & Hsiao, S. S. (1988). Spatial pattern representation and transformation in monkey somatosensory cortex. *Proc. Natl. Acad. Sci. USA, 85,* 1317–1321.

Richmond, B. J., & Optican, L. M. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. II. Quantification of response waveform. *J. Neurophysiol., 57,* 147–161.

Rosner, B. S. (1961). Neural factors limiting cutaneous spatiotemporal discrimination. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 725–737). Cambridge, MA: MIT Press.

Sejnowski, T. J. (1995). Pattern recognition: Time for a new neural code? *Nature, 376,* 21–22.

Shadlen, M. N., & Newsome, W. T. (1994). Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol., 4,* 569–579.

Silva, L. R., Amitai, Y., and Connors, B. W. (1991). Intrinsic oscillations of neocortex generated by layer 5 pyramidal neurons. *Science, 251,* 432–435.

Simons, D. J. (1995). Neuronal integration in the somatosensory whisker/barrel cortex. In E. G. Jones and I. T. Diamond (Eds.), *Cerebral cortex* (Vol. 11, pp. 263–297). New York: Plenum Press.

Sinclair, R., & Burton, H. (1988). Responses from area 3b of somatosensory cortex to textured surfaces during active touch in primate. *Somatosens. Res., 5,* 283–310.

Sinclair, R. J., & Burton, H. (1991). Neuronal activity in the primary somatosensory cortex in monkeys (Macaca mulatta) during active touch of textured surface gratings: Responses to groove width, applied force, and velocity of motion. *J. Neurophysiol., 66,* 153–169.

Sinclair, R. J., Sathian, K., & Burton, H. (1991). Neuronal responses in ventroposterolateral nucleus of thalamus in monkeys (Macaca mulatta) during active touch of gratings. *Somatosens. Mot. Res., 8,* 293–300.

Steriade, M., McCormick, D. A., & Sejnowski, T. J. (1993). Thalamocortical oscillations in the sleeping and aroused brain. *Science, 262,* 679–685.

Swadlow, H. A. (1995). Influence of VPM afferents on putative inhibitory interneurons in S1 of the awake rabbit—Evidence from cross-correlation, mi-

crostimulation, and latencies to peripheral sensory stimulation. *J. Neurophysiol., 73,* 1584–1599.

Talbot, W. H., Darian-Smith, I., Kornhuber, H. H., & Mountcastle, V. B. (1968). The sense of flutter-vibration: Comparison of the human capacity with response patterns of mechanoreceptive afferents from the monkey hand. *J. Neurophysiol., 31,* 301–334.

Taylor, M. M., Lederman, S. J., & Gibson, R. H. (1973). Tactual perception of texture. In E. Carterette & M. Friedman (Eds.), *Handbook of perception* (Vol. 3, pp. 251–272). New York: Academic Press.

Theunissen, F., & Miller, J. P. (1995). Temporal encoding in nervous systems: A rigorous definition. *J. Comp. Neurosci., 2,* 149–162.

Tremblay, F., Ageranioti-Belanger, S. A., & Chapman, C. E. (1996). Cortical mechanisms underlying tactile discrimination in the monkey. I. Role of primary somatosensory cortex in passive texture discrimination. *J. Neurophysiol., 76,* 3382–3403.

Vega-Bermudez, F., Johnson, K. O., & Hsiao, S. S. (1991). Human tactile pattern recognition: Active versus passive touch, velocity effects, and patterns of confusion. *J. Neurophysiol., 65,* 531–546.

Viterbi, A. J. (1966). *Principles of coherent communication.* New York: McGraw-Hill.

Wang, X., Merzenich, M. M., Beitel, R., & Schreiner, C. E. (1995). Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: Temporal and spatial characteristics. *J. Neurophysiol., 74,* 2685–2706.

Welker, W. I. (1964). Analysis of sniffing of the albino rat. *Behaviour, 22,* 223–244.

White, E. L., & Keller, A. (1987). Intrinsic circuitry involving the local axon collaterals of corticothalamic projection cells in mouse SmI cortex. *J. Comp. Neurol., 262,* 13–26.

Wilson, M. A., & Bower, J. M. (1989). The simulation of large-scale neural networks. In C. Koch and I. Segev (Eds.), *Methods in neuronal modeling: From synapses to networks* (pp. 291–333). Cambridge, MA: MIT Press.

Wise, S. P. (1993). Monkey motor cortex: Movements, muscles, motoneurons and metrics. *Trends. Neurosci., 16,* 46–49.

Yamamoto, T., Samejima, A., & Oka, H. (1988). Short latency activation of local circuit neurons in the cat somatosensory cortex. *Brain Res., 461,* 199–203.

Received August 1, 1996; accepted June 3, 1997.

LETTER Communicated by Christoph von der Malsburg

Deformation Theory of Dynamic Link Matching

Toru Aonishi
Department of Biophysical Engineering, Faculty of Engineering Science, Osaka University, Osaka 560, Japan

Koji Kurata
Department of Systems and Human Science, Graduate School of Engineering Science, Osaka University, Osaka 560, Japan

Dynamic link matching is a self-organizing topographic mapping between a template image and a data image. The mapping tends to be continuous, linking two points sharing similar local features, which, as a result, can lead to its deformation to some degree. In analyzing such deformation mathematically, we reduced the model equation to a phase equation, which enabled us to clarify the principles of the deformation process and the relationship between high-dimensional models and low-dimensional ones. We also elucidated the characteristics of the model in the context of the standard regularization theory.

1 Introduction

Pattern recognition invariant to deformation or transformation can be performed by dynamic link matching (DLM) (Konen & von der Malsburg, 1993). DLM specifies a flexible match between a template pattern and a data pattern, where local features in the data pattern have to be matched with their counterparts in the template pattern. DLM is based on the model of a self-organizing topographic map with a weak tendency to link two points sharing similar features. Since the self-organizing map has a kind of elasticity, the map tends to resemble closely the identity map, resisting disturbances. DLM shares this characteristic with the self-organizing map but will also generally link two points of similar local features, causing map deformation to some extent.

DLM has been applied effectively to some engineering examples (Lades et al., 1993; Bienenstock & Doursat, 1994), but no mathematical analysis has yet been made. Therefore, we propose a mathematically tractable model based on a system of local excitation. We reduce a model equation to a phase equation (Ermentrout, 1981), which gives us a mathematical understanding of the principle of the flexible matching process.

DLM can be classified into two types. One, a high-dimensional representation (Bienenstock & von der Malsburg, 1987; Konen & von der Malsburg,

1993; Konen, Maurer, & von der Malsburg, 1994), is the map expressed by the synaptic weight distribution on the product space of the two image spaces. The other, a low-dimensional representation, corresponds directly to the graph matching (Lades et al., 1993; Bienenstock & Doursat, 1994). In this case, the map is expressed by reference vectors as in Kohonen's model (Kohonen, 1982).

Matching two images is, in general, an ill-posed problem. Information is insufficient when matching local features in a data image with their counterparts in a template image. By introducing topography constraints, however, this deficiency can be addressed in the definition of the problem. The potential of the phase equation in our model is equivalent to the cost function in standard regularization theories (Poggio, Torre, & Koch, 1985), since it consists of a stabilizer (a topography constraint) and a data-fitting term (to match local features in the data image with their counterparts in the template image). Our theory bridges the gap between low-dimensional representations and high-dimensional representations, because we start from a high-dimensional model equation and derive a low-dimensional phase equation.

2 Model Description

Figure 1a shows a schematic diagram of our model. The problem is how to build a flexible match between a one-dimensional data image and a one-dimensional template image. The input patterns are feature vector functions defined on F_1 and F_2. Here, we simply assume $F_1 = F_2 = \mathbf{R}$ to avoid boundary effects. However, our analysis is useful in understanding models with boundary conditions, because at points far from the boundary the boundary effect is negligible and is not important for the deformation process of the mapping, which is the theme of this article. The feature vectors are defined as $I_1(r_1)$, $r_1 \in F_1$ and $I_2(r_2)$, $r_2 \in F_2$. I_1 and I_2 are expected to be equal up to a certain deformed topographic transformation. The system has to match local features in the data image with their counterparts in the template image, and thus produce a topographic transformation from F_1 to F_2. The map is expressed by the synaptic weight distribution $w(r_1, r_2)$ on $F_1 \times F_2$.

The model equation is

$$\frac{\partial}{\partial t} w(r_1, r_2, t) = -w + \mathcal{K} f(w) + \varepsilon^2 s(r_1, r_2), \tag{2.1}$$

$$\mathcal{K} f(w) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dr_1' dr_2' k(r_1', r_2') f\left(w(r_1 - r_1', r_2 - r_2')\right),$$

where $|\varepsilon| << 1$ and $s(r_1, r_2)$ is the local similarity between $I_1(r_1)$ and $I_2(r_2)$,

$$s(r_1, r_2) = v(I_1(r_1), I_2(r_2)). \tag{2.2}$$

This equation consists of two parts. The first part, $-\frac{\partial}{\partial t} w - w + \mathcal{K} f(w)$, is

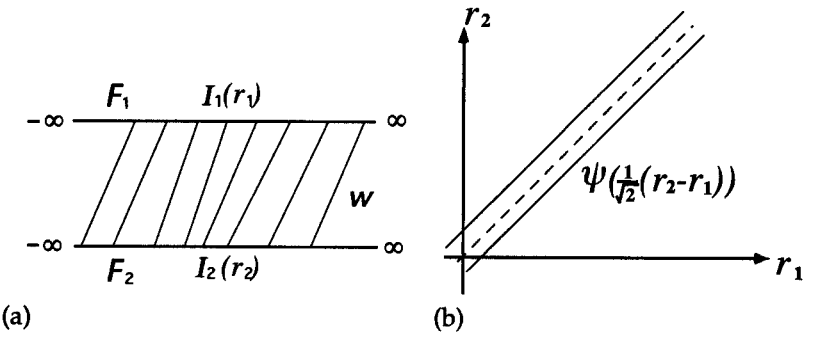


Figure 1: (a) Schematic diagram of our model. I_1 and I_2 are equal up to a certain deformed topographic transformation. The system has to produce a topographic mapping from F_1 to F_2. The mapping is expressed by the synaptic weight distribution $w(r_1, r_2)$ on $F_1 \times F_2$. (b) Schematic diagram of a topographic mapping (an equilibrium solution). This solution is a bank on a diagonal line.

the self-organizing part, which conserves the topography. The second part, $\varepsilon^2 s(r_1, r_2)$, is a perturbation term. This part links two points assigned similar features in I_1 and I_2. Here we examine the case where the perturbation is very weak.

3 Self-Organizing Part

Let us begin with the self-organizing part,

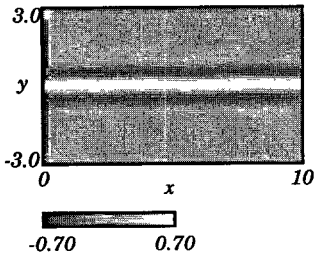
$$\frac{\partial}{\partial t} w(r_1, r_2, t) = -w + \mathcal{K} f(w). \tag{3.1}$$

The integration kernel k in the linear operator $\mathcal{K}$ is of a two-dimensional isotropic Mexican hat type, and f is a sigmoidal function with $f'(x) > 0$ (monotonically increasing function).

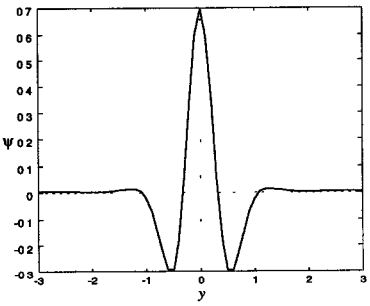
If all of the parameters are set appropriately, we can make the dynamics have a stable equilibrium solution of the following form:

$$w(r_1, r_2, t) = \psi\left(\frac{1}{\sqrt{2}}(r_2 - r_1)\right).$$

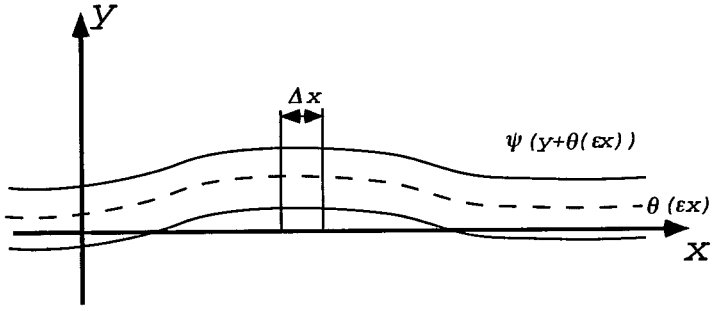
This solution is a bank on a diagonal line as shown in Figure 1b. Figures 2a and 2b shows an example of a bank and its one-dimensional cross-section obtained by computer simulation with another coordinate system (see equation 3.3). Some work (Takeuchi & Amari, 1979; Amari, 1980; Häussler & von



(a)



(b)



(c)

Figure 2: (a) Rotated equilibrium solution obtained by computer simulation. (b) One-dimensional cross-section of the equilibrium solution. (c) Schematic diagram of a deformed solution by perturbation. We can assume that this solution does not vary along the x-axis within a small region.

der Malsburg, 1983) has analytically shown the stability of equilibrium solutions of topographic connections. In section 6, we discuss the necessary condition for the stability of the solution when $f(x)$ is a step function. It is not easy to analyze its stability mathematically. However, in this article, the stability is verified by computer simulation.

If $r_1 \in F_1$ is fixed, w is a function in the domain F_2. In such a case, we can define the topographic mapping p as

$$p: \begin{array}{l} F_1 \to F_2 \\ r_1 \to r_2 | \arg\max_{r_2} w. \end{array}$$

In our system, the basin of attraction for a single bank solution (see Figure 2a) is small because k is localized and isotropic. Thus, we set a single

bank pattern as the initial state; otherwise, a spotty pattern would be formed. This single bank solution is stable against small perturbations, as verified by computer simulation. This is because of the sigmoidal function, f, which suppresses all of the perturbations under its threshold h and makes our system stable around $w = 0$.

In DLM, it is essential that the self-organizing dynamics have a bank-shaped stable equilibrium solution on a diagonal line (a topographic map). It is possible to complicate the model to avoid the formation of stripe or hexagon patterns with long-range synaptic competition. More complicated models, however, would lose the mathematical simplicity, and some could not be analyzed mathematically.

Another merit of a difference of gaussian (DOG) kernel is that its rotation symmetricity makes the mapping solution very flexible and much more expedient with magnification, contraction, and deformation. Many DLM models are not so strong against magnification and contraction. We have compared commonly used anisotropic kernels and DOG kernels by computer simulation and have found that a DOG kernel can make a model much more tolerant of deformation (deformation involving a local magnification and contraction). A theory on deformation for DLM therefore needs to be formed and verified by computer simulation; the use of a DOG kernel is very appropriate.

Equation 3.1 is popular as a system of local excitation in 2D neural fields. Furthermore, this form of dynamics can be represented by a simplified model for Hebbian learning between two 1D neural fields. The lateral connection of each 1D neural field consists of short-range excitatory and long-range inhibitory connections. Interlayer synaptic weight w increases between excited cells as in Hebbian learning. When some synapse is enforced, other synaptic connections nearby decrease due to synaptic competition, since both neurons absorb the limited nutrition available from nearby cells. f represents the nonlinearity of the synapse.

Through simple mathematical manipulation, equation 3.1 can be expressed in terms of another coordinate system,

$$\frac{\partial}{\partial t} w(x, y, t) = -w + \mathcal{K} f(w), \tag{3.2}$$

$$\mathcal{K} f(w) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dx' dy' k(x', y') f\left(w(x - x', y - y')\right),$$

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}, \tag{3.3}$$

where kernel k is not changed by the transformation of a coordinate because of its isotropy. The equilibrium solution is rewritten as:

$$w(x, y, t) = \psi(y). \tag{3.4}$$

Figure 2a shows an example of an equilibrium solution obtained by numerical calculation with equation 3.2.

Equation 3.2 is invariant to spatial shifts along the y-axis. Therefore, $\psi(y-\theta)$ is also an equilibrium solution with any constant θ. θ stands for the phase of the solution. If we define $\psi(0)$ as the maximum value of ψ, θ represents the center of the receptive field, which can be considered as the reference vector, for example, as in Kohonen's (1982) model.

Here, we estimate the stability against the fluctuation constant about the x-axis. Substituting $w(x,y,t) = \psi(y-\theta) + \varepsilon u(y,t)$ into equation 3.2 and linearizing around $\varepsilon = 0$, we obtain the following equation:

$$\frac{\partial u}{\partial t} = \mathcal{L}_\theta u,$$
$$\mathcal{L}_\theta = -1 + \mathcal{K}_y(f'(\psi(y-\theta))\cdot). \tag{3.5}$$

Here, $\mathcal{L}_\theta$ denotes a linear operator, and $\mathcal{K}_y$ is the following 1D convolution operator in the direction y:

$$\mathcal{K}_y v(x,y) = \int_{-\infty}^{\infty} dy' \bar{k}(y') v(x, y-y'),$$
$$\bar{k}(y) = \int_{-\infty}^{\infty} dx k(x,y).$$

All of the eigenvalues of $\mathcal{L}_\theta$ are nonpositive since the equilibrium solution $\psi(y-\theta)$ is stable. However, there is an eigenvalue of 0 with the eigenfunction $\psi'(y-\theta)$ because $-\psi + \mathcal{K}f(\psi) = 0$. This eigenfunction corresponds to a small spatial shift of ψ in the direction of y, because $\psi(y-\theta+\varepsilon) \approx \psi(y-\theta) + \varepsilon\psi'(y-\theta)$. We assume no other eigenfunctions exist for the eigenvalue of 0, that is,

$$\ker\mathcal{L}_\theta = \mathrm{span}\{\psi'(y-\theta)\}. \tag{3.6}$$

This assumption is equivalent to the waveform stability of $\psi(y-\theta)$. This means that although the equilibrium solution is irresistant to a spatial shift, it can maintain its waveform. The waveform stability of our solution is unproved, but it can be easily verified by computer simulation. The stability of modes with eigenvalue 0 is dramatically changed by a small perturbation. If other modes with eigenvalue 0 were to exist, the waveform would be broken.

Even with a boundary, the solution maintains an approximate waveform stability as long as the solution decreases toward zero in both directions and is sufficiently small at the boundary points.

4 Perturbation

Here, we consider the perturbed model,

$$\frac{\partial}{\partial t}w(x,y,t) = -w + \mathcal{K}f(w) + \varepsilon^2 s(x,y), \tag{4.1}$$

where the local similarity in equation 2.2 is transformed into $s(x,y)$ using equation 3.3. Since ε is very small, the self-organizing part is only weakly affected by the image data. The equilibrium solution 3.4 is gently deformed by the effect of the perturbation. We assume that the deformed solution can be expressed as

$$w(x,y,t) = \psi(y - \theta(\xi,\tau)) + \varepsilon u_1(\xi,y,\tau) + \varepsilon^2 u_2(\xi,y,\tau), \tag{4.2}$$

where $\xi = \varepsilon x$, $\tau = \varepsilon^2 t$. ξ is a large space variable, and τ denotes a slow time variable. This solution can be regarded as constant with respect to x within a tight range about the x-axis as shown in Figure 2c. εu_1 and $\varepsilon^2 u_2$ are fluctuations caused by the effects of the perturbation.

Since the deformation is caused by $\varepsilon^2 s$ in the neighborhood of an equilibrium solution for an unperturbed system, the order of this deformation process is equal to the order of the perturbation, that is, $O(\varepsilon^2)$. In order to treat this small phenomenon in the theory, it is necessary to shrink time t and space x. If we set $\xi = \varepsilon x$ and $\tau = \varepsilon^2 t$, we can derive the following $O(\varepsilon^2)$ terms:

$$\frac{\partial}{\partial t} = \varepsilon^2 \frac{\partial}{\partial \tau} \quad \text{and} \quad \frac{\partial^2}{\partial x^2} = \varepsilon^2 \frac{\partial^2}{\partial \xi^2}.$$

Thus, we can derive a diffusion term to describe the resistance against deformation. However, if x and t shrink with the same order, we cannot derive a spatial derivatives term (the $\partial/\partial x$ term is erased in the case of our system; see appendix A), that is, the deformation process is neglected in the theory, which is equivalent to a previous analysis (Aonishi, Kurata, & Mito, 1997).

From this assumption, we can derive the following equations (see appendix A):

$$0 = -\psi + \mathcal{K}_y f(\psi) + \varepsilon \mathcal{L}_\theta u_1 + \varepsilon^2 \left(\mathcal{L}_\theta u_2 - m_2\right), \tag{4.3}$$

$$\begin{aligned} m_2 = &-\psi'(y-\theta)\frac{\partial \theta}{\partial \tau} \\ &- \frac{1}{2}G_1(y-\theta)\left(\frac{\partial \theta}{\partial \xi}\right)^2 + \frac{1}{2}G_2(y-\theta)\frac{\partial^2 \theta}{\partial \xi^2} \\ &- \frac{1}{2}\mathcal{K}_y f''(\psi(y-\theta))u_1{}^2 - s(x,y), \end{aligned}$$

$$G_1(y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy'k(x',y')x'^2 \times \left(f'(\psi(y-y'))\psi''(y-y') + f''(\psi(y-y'))\psi'(y-y')^2\right),$$

$$G_2(y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy'k(x',y')x'^2 f'(\psi(y-y'))\psi'(y-y').$$

Here, we cannot neglect the $O(\varepsilon^2)$ term, since the order of the perturbation is $O(\varepsilon^2)$. The coefficient of ε^0 satisfies 0 from the equilibrium condition. Therefore, equation 4.3 can be reduced as follows:

$$0 = \mathcal{L}_\theta u_1 + \varepsilon\left(\mathcal{L}_\theta u_2 - m_2\right). \tag{4.4}$$

This equation demonstrates that $\mathcal{L}_\theta u_1 = O(\varepsilon)$. Therefore, from equation 3.6, we can obtain u_1 in the following form,

$$u_1 = \alpha\psi'(y-\theta) + \varepsilon\overline{u}_1, \tag{4.5}$$

where $\varepsilon\overline{u}_1$ is a higher-order fluctuation in u_1, and α is an arbitrary constant, since $\mathcal{L}_\theta u_1 = \alpha\mathcal{L}_\theta\psi'(y-\theta) + \varepsilon\mathcal{L}_\theta\overline{u}_1 = \varepsilon\mathcal{L}_\theta\overline{u}_1 = O(\varepsilon)$. Substituting this into equation 4.4 and neglecting higher-order terms, we obtain

$$\mathcal{L}_\theta\left(\overline{u}_1 + u_2\right) - m_2 = 0. \tag{4.6}$$

$$m_2 = -\psi'(y-\theta)\frac{\partial\theta}{\partial\tau} - \frac{1}{2}G_1(y-\theta)\left(\frac{\partial\theta}{\partial\xi}\right)^2 + \frac{1}{2}G_2(y-\theta)\frac{\partial^2\theta}{\partial\xi^2} - \frac{1}{2}\alpha^2\mathcal{K}_y f''(\psi(y-\theta))\psi'(y-\theta)^2. - s(x,y).$$

Here, we can erase fluctuation term $\mathcal{L}_\theta\left(\overline{u}_1 + u_2\right)$ in equation 4.6 using the following averaging manipulation. We average equation 4.6 using the weight $f'(\psi(y-\theta))\psi'(y-\theta)$,

$$\begin{aligned}\int_{-\infty}^{\infty} dyf'(\psi(y-\theta))\psi'(y-\theta)m_2 &= \int_{-\infty}^{\infty} dyf'(\psi(y-\theta))\psi'(y-\theta)\mathcal{L}_\theta\left(\overline{u}_1 + u_2\right)\\ &= \int_{-\infty}^{\infty} dyf'(\psi(y-\theta))\left(-\psi'(y-\theta) + \mathcal{K}_y\psi'(y-\theta)\right)\left(\overline{u}_1 + u_2\right)\\ &= \int_{-\infty}^{\infty} dyf'(\psi(y-\theta))\mathcal{L}_\theta\psi'(y-\theta)\left(\overline{u}_1 + u_2\right).\end{aligned}$$

From equation 3.6, $\mathcal{L}_\theta \psi'(y-\theta) = 0$. Thus, we can obtain

$$\int_{-\infty}^{\infty} dy f'(\psi(y-\theta))\psi'(y-\theta)m_2 = 0. \tag{4.7}$$

Therefore, we can erase the fluctuation term.

Since $k(x, y)$ is an even function, the model is symmetrical about the y-axis. Thus, we can expect stationary solution $\psi(y)$ to be an even function. If we have an uneven stationary solution in the symmetric model, almost all small symmetrical perturbation on the model will transform it into a traveling wave solution (see appendix B), for there is no generic reason for it to stand still. In other words, uneven stationary solutions in symmetric models are structurally unstable. By computer simulation, we could verify that $\psi(y)$ is an even stationary solution, as shown in Figure 2b. Thus $\psi'(y)$ and $f'(\psi(y))\psi'(y)$ are odd functions, $G_1(y)$ is an even function, and $G_2(y)$ is an odd function. Eliminating some terms by averaging with the weight $f'(\psi(y-\theta))\psi'(y-\theta)$, we obtain the following phase equation:

$$c\frac{\partial\theta(\xi,\tau)}{\partial\tau} = \frac{d}{2}\frac{\partial^2\theta(\xi,\tau)}{\partial\xi^2} + \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} dy f(\psi(y-\theta))s(\xi/\varepsilon, y), \tag{4.8}$$

where

$$c = \int_{-\infty}^{\infty} dy f'(\psi(y))\psi'(y)\psi'(y), \qquad d = \int_{-\infty}^{\infty} dy f'(\psi(y))\psi'(y)G_2(y).$$

From variational principles, phase equation 4.8 has the following potential V:

$$V = \frac{d}{4c}\int_{-\infty}^{\infty} d\xi \left(\frac{\partial\theta(\xi,\tau)}{\partial\xi}\right)^2 - \frac{1}{c}\int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} dy f(\psi(y-\theta))s(\xi/\varepsilon, y). \tag{4.9}$$

This equation consists of two parts: one for smoothing and the other for data fitting. Thus, it is obvious how it is analogous to the standard regularization theory. The data-fitting term derives the mapping so that $f(\psi(y-\theta))$ and $s(x, y)$ have maximal covariance, giving the mapping the tendency to connect two points sharing similar local features. It should also be noted that θ is a low-dimensional expression of weak deformation. The low-dimensional representation of mapping p is obtained from θ as the solution of the following self-consistent equation:

$$p:\ F_1 \to F_2$$
$$r_1 \to r_2 \left| \frac{1}{\sqrt{2}}(r_1 - r_2) = \theta\left(\frac{\varepsilon}{\sqrt{2}}(r_1 + r_2)\right)\right. .$$

5 Parameters

We used the following convolution kernel,

$$k(x,y) = \frac{1}{\sigma_1^2}\exp\left(-\frac{x^2+y^2}{2\sigma_1^2}\right) - \frac{1}{\sigma_2^2}\exp\left(-\frac{x^2+y^2}{2\sigma_2^2}\right),$$

and the sigmoidal function

$$f(x) = \frac{1}{1+\exp(-\beta(x-h))},$$

where $\sigma_1 = 0.25$, $\sigma_2 = 0.33$, $\beta = 10$, and $h = 0.36$. In following simulations, we set $\varepsilon = 0.2$.

6 Stability of the Solution

With the $\beta \to \infty$ limit, $f(x)$ tends to form a step function:

$$f(x) = \begin{cases} 1 & x > h \\ 0 & x \le h \end{cases}.$$

In this case, we can simply derive the following two conditions for the stability of $\psi(y)$. One is a one-dimensional stability condition and is equivalent to the stability condition for local excitation in one-dimensional neural fields (Amari, 1977). The other is a two-dimensional stability condition and is needed to keep the solution straight, as in Figure 2a. A stable solution must satisfy both conditions.

One-dimensional stability is a necessary condition for preventing perturbations that are constant with respect to x, that is, $w(x,y,t) = \psi(y-\theta) + \varepsilon u(y,t)$. Let $\psi(y) > h$, that is, $f(\psi(y)) = 1$ in the region $[-\infty,\infty] \times [\frac{-a}{2}, \frac{a}{2}]$. $\psi(y)$ exists if a satisfies

$$K(a) = h, \tag{6.1}$$

$$\begin{aligned} K(a) &= \int_{-\infty}^{\infty} dx \int_0^a dy k(x,y) \\ &= \sqrt{2\pi}\int_0^a dy \left(\frac{1}{\sigma_1}\exp\left(-\frac{y^2}{2\sigma_1^2}\right) - \frac{1}{\sigma_2}\exp\left(-\frac{y^2}{2\sigma_2^2}\right)\right). \end{aligned}$$

One-dimensional stability in this case,

$$K'(a) < 0, \tag{6.2}$$

is a necessary condition for the stability of the solution $\psi(y)$. The function $K(a)$ is depicted in Figure 3a. If $0 < h < K(a_{\max})$, condition 6.1 allows two

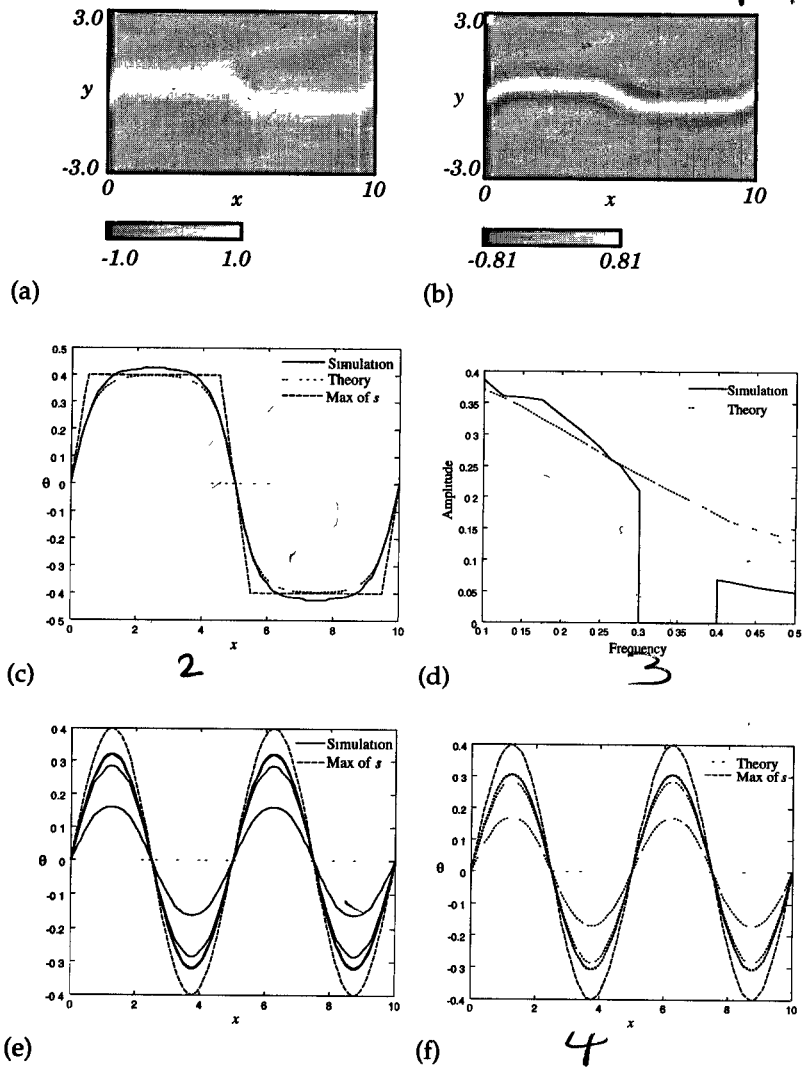


Figure 4: (a) Example of $s(x, y)$. (b) Deformed equilibrium solution by perturbation with the local similarity function shown in a. (c) Ridge of $s(x, y)$ and that of deformed equilibrium solution (b), together with a theoretical curve from equation 4.8. (d) Frequency responses to a sinusoidal ridge of $s(x, y)$ (the amplitude is 0.4). (e,f) The time-dependent behavior of θ. Time interval between two curves is 5.6 in τ. $d = 0.019453$, $c = 5.034132$.

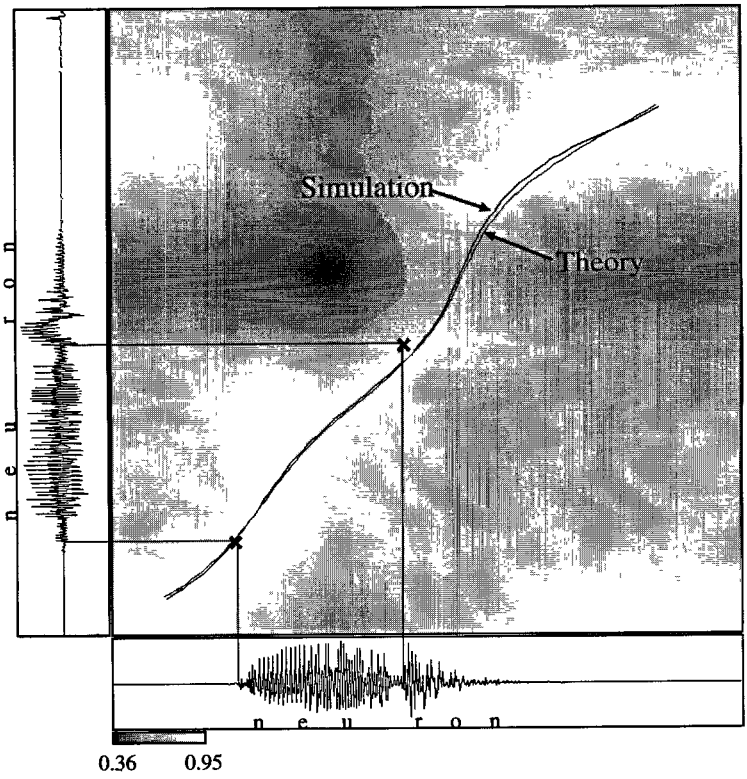


Figure 5: Two pieces of real data and the local similarity $s(r_1, r_2)$ between the data, together with a simulation result and a theoretical curve. We also show syllables for speech data.

Figure 5 shows two pieces of real data and the local similarity $s(r_1, r_2)$ between the data. We performed computer simulations by numerical calculation with equation 4.1 with the periodic boundary condition, as in the former simulations. On the local similarity $s(r_1, r_2)$ in Figure 5, we superimposed the simulation result (the ridge of the deformed equilibrium solution) and a theoretical curve, together with the starting points of the speech data syllables. Our system could match between the points in I_1 and their counterparts in I_2. The theory strongly correlated with a behavior of the system.

expect the stability conditions with infinite β to be applied in establishing stability at large values of β. In the case of $h = 0.36$, the solution satisfies both stability conditions.

7 Computer Simulation

In numerical simulations to verify our theoretical results, computer simulation was performed for equation 4.1 with the periodic boundary condition. Here, the artificial local similarity $s(x, y)$ was set as shown in Figure 4a. We set the initial state $w(x, y, 0) = \psi(y)$ in Figure 2a. In the following simulations, $d = 0.019453$ and $c = 5.034132$. Figure 4b shows the deformed equilibrium solution by perturbation with the local similarity in Figure 4a. In Figure 4c we show $\theta(x)$, the ridge of the deformed equilibrium solution 4b, together with a theoretical curve from equation 4.8.

We also investigated spatial frequency responses of the system. The similarity was assumed to be a function with a sinusoidal ridge. The amplitude of the ridge was fixed to 0.4. Figure 4d shows the amplitude of the solution wave versus various frequencies of similarity waves. The dotted curve in the same figure indicates the theory from phase equation 4.8. According to Figure 4d, the theory strongly conforms to the simulation results when $f < 0.3$, but when $f > 0.3$ our theory breaks down. The solutions (which initially have a continuous ridge) in the range of $0.3 < f < 0.4$ break into blobs, and thus our theory ceases to be valid. The solutions when $f > 4.0$ keep the continuous ridge, but the assumption that the solution varies gently along the x-axis is no longer satisfied.

Figures 4e and 4f shows the time-dependent behavior of θ. Our theory provides a good description of the simulation data in the time scale.

8 Application to Real Data

Our design is very simple so as to be mathematically tractable; however, this creates limitations in its application. Even so, we have provided one example of an application to real data. From real speech data, we extracted power spectra at each time with wavelet filters for 20 different frequencies corresponding to Fourier transforms limited by a gaussian window in the time domain. Then the 20-dimensional power spectrum data for each time were assigned to a vector $I_i(r_i)$. In this simulation, I_1 and I_2 were extracted from two different pieces of data, which were pronounced "neuron" by the same Japanese speaker. Here, r_i denotes the time. All parameters were set as in the former simulations, except $\varepsilon^2 = 0.045$. We used the following local similarity function:

$$v(I_1(r_1), I_2(r_2)) = 1.0 - |I_1(r_1) - I_2(r_2)|. \tag{8.1}$$

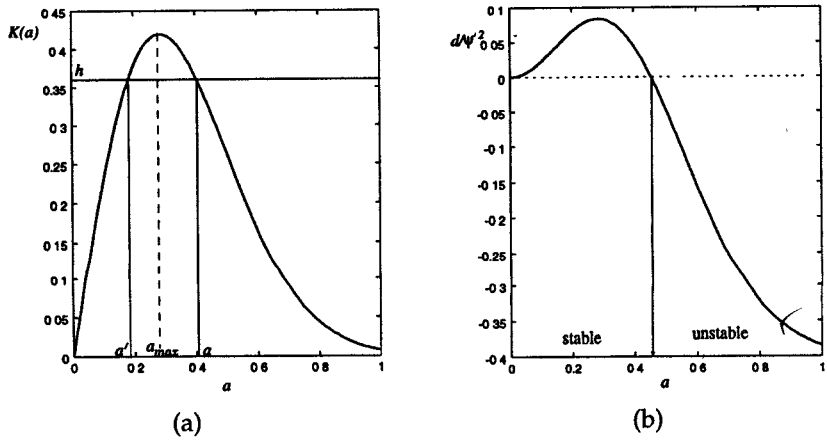


Figure 3: (a) Schematic diagram of $K(a)$. Only a is a stable solution; a' is an unstable fixed point. (b) Diffusing coefficient d with the $\beta \to \infty$ limit. The vertical axis corresponds to d/ψ'^2, and the horizontal axis corresponds to a range of solutions. We can find a bifurcation point of the stability parameterized by a. Here, all parameters are set as in section 5.

solutions: a and a' in Figure 3a. However, condition 6.2 permits only a as a stable solution, while a' becomes an unstable fixed point.

Obviously, $d > 0$ in equation 4.8 is a necessary condition for the stability of the solution. This is a two-dimensional stability condition. In the unstable case, our phase equation is no longer valid.

In the case of a step function, we can explicitly derive the diffusing coefficient,

$$d = 2\sqrt{2\pi}\psi'(a/2)^2 \left(\sigma_1 \left(1 - \exp\left(-\frac{a^2}{2{\sigma_1}^2}\right)\right) - \sigma_2 \left(1 - \exp\left(-\frac{a^2}{2{\sigma_2}^2}\right)\right)\right), \tag{6.3}$$

where a is the size of the connected region in which the condition $f(\psi(y)) = 1$ holds, as previously discussed for one-dimensional stability. Figure 3b shows the diffusion coefficient $d/\psi'(a/2)^2$ versus the range a. We can find a critical point where the sign of d changes, that is, a bifurcation point of the stability parameterized by a.

Figures 3a and 3b shows that a region a exists in which both conditions are satisfied. Here, all parameters are set as in section 5, except for β. We can

9 Conclusion

In this article, we proposed mathematically tractable DLM, a topographic mapping formed between a template image and a data image. The mapping is continuous but tends to link two points sharing similar local features, which can result in some degree of deformation. To analyze such deformation mathematically, we derived a phase equation from a model equation. We demonstrated that the theory complies with the behavior of the system using computer simulations.

DLM can be classified into two types. One is a high-dimensional representation. Equation 2.1 corresponds to this type since the map is expressed by a synaptic weight distribution on the product space of two image spaces. The other is a low-dimensional representation. The potential (see equation 4.9) is the same as in the cost function of standard regularization theories, since it consists of a stabilizer (a topography constraint) and a data-fitting term (to match local features in the data image with their counterparts in the template image). Our theory bridges the gap between these two representations and gives us a mathematical understanding of the principle behind the flexible matching process.

Low-dimensional versions of DLM require both a much lower computational time and less memory than do high-dimensional versions of DLM. We used an example to demonstrate that these two versions are equivalent in the neighborhood of the identity map. However, some high-dimensional models (Konen & von der Malsburg, 1993; Konen et al., 1994) have a very wide basin for the successful matchings they make. With such models, we can start from constantly distributed connections, but we have to set low-dimensional models to some mapping to start them. If this "prejudice" is too far from correct matching, then we cannot reach it.

Appendix A

Substituting equation 4.2 into equation 4.1, we obtain

$$\begin{aligned}\varepsilon^2\psi'(y-\theta(\xi,\tau))\frac{\partial\theta}{\partial\tau} = &-\psi(y-\theta)-\varepsilon u_1-\varepsilon^2 u_2\\ &+\mathcal{K}f(\psi(y-\theta(\xi,\tau)))\\ &+\varepsilon\mathcal{K}\left(f'(\psi(y-\theta(\xi,\tau)))u_1(\xi,y,\tau)\right)\\ &+\varepsilon^2\mathcal{K}\left(f'(\psi(y-\theta(\xi,\tau)))u_2(\xi,y,\tau)\right)\\ &+\frac{\varepsilon^2}{2}\mathcal{K}\left(f''(\psi(y-\theta(\xi,\tau)))u_1(\xi,y,\tau)^2\right)\\ &+\varepsilon^2 s(x,y)\\ &+O(\varepsilon^3). \qquad \text{(A.1)}\end{aligned}$$

Here, we can represent the $\mathcal{K}(\cdots)$ term in equation A.1 as follows:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy' k(x',y')u(\varepsilon(x-x'),y-y') = \int_{-\infty}^{\infty} dx' \frac{1}{\varepsilon} k\left(\frac{x'}{\varepsilon},y'\right) u(\varepsilon x - x', y-y'). \tag{A.2}$$

As $\varepsilon \to 0$, $\frac{1}{\varepsilon}k(\frac{x}{\varepsilon},y)$ tends to the following function,

$$\frac{1}{\varepsilon}k(\frac{x}{\varepsilon},y) = \sqrt{2\pi}\delta(x)\left(\frac{1}{\sigma_1}\exp\left(-\frac{y^2}{2\sigma_1^2}\right) - \frac{1}{\sigma_2}\exp\left(-\frac{y^2}{2\sigma_2^2}\right)\right). \tag{A.3}$$

Thus, we can expand $u(\varepsilon x - x', y-y')$ into a polynomial around $x' = 0$ as follows:

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy' \frac{1}{\varepsilon} k\left(\frac{x'}{\varepsilon},y'\right) u(\varepsilon x - x', y-y') \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy' \frac{1}{\varepsilon} k\left(\frac{x'}{\varepsilon},y'\right) \\
&\quad \times \left[u(\varepsilon x, y-y') - u_x(\varepsilon x, y-y')x' + u_{xx}(\varepsilon x, y-y')\frac{x'^2}{2} + \cdots\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy' k(x',y') \\
&\quad \times \left[u(\varepsilon x, y-y') - u_x(\varepsilon x, y-y')\varepsilon x' + u_{xx}(\varepsilon x, y-y')\frac{\varepsilon^2 x'^2}{2} + \cdots\right] \\
&= \int_{-\infty}^{\infty} dy' \left[M_0(y')u(\varepsilon x, y-y') - \varepsilon M_1(y')u_x(\varepsilon x, y-y') \right. \\
&\qquad \left. + \frac{\varepsilon^2}{2}M_2(y')u_{xx}(\varepsilon x, y-y') + \cdots\right],
\end{aligned} \tag{A.4}$$

where $M_n(y)$ is the moment of $k(x,y)$,

$$M_n(y) = \int_{-\infty}^{\infty} dx' k(x',y) x'^n. \tag{A.5}$$

If n is odd, $M_n(y) = 0$, since k is even w.r.t. x.

Thus, the $\mathcal{K}(\cdots)$ term in equation A.1 can be expanded as follows:

$$\begin{aligned}
&\mathcal{K}\psi(y-\theta(\xi,\tau)) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy' k(x',y') f(\psi(y-y'-\theta(\varepsilon x,\tau))
\end{aligned}$$

$$
\begin{aligned}
&+ \varepsilon \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy'k(x',y')x' \\
&\times f'(\psi(y-y'-\theta(\varepsilon x,\tau)))\psi'(y-y'-\theta(\varepsilon x,\tau)) \\
&\times \frac{\partial\theta(\varepsilon x,\tau)}{\partial\varepsilon x} + \frac{\varepsilon^2}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy'k(x',y')x'^2 \\
&\times \big(f'(\psi(y-y'-\theta(\varepsilon x,\tau)))\psi''(y-y'-\theta(\varepsilon x,\tau)) \\
&+ f''(\psi(y-y'-\theta(\varepsilon x,\tau)))\psi'(y-y'-\theta(\varepsilon x,\tau))^2\big)\left(\frac{\partial\theta(\varepsilon x,\tau)}{\partial\varepsilon x}\right)^2 \\
&- \frac{\varepsilon^2}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy'k(x',y')x'^2 \\
&\times f'(\psi(y-y'-\theta(\varepsilon x,\tau)))\psi'(y-y'-\theta(\varepsilon x,\tau)) \\
&\times \frac{\partial^2\theta(\varepsilon x,\tau)}{\partial(\varepsilon x)^2} + O(\varepsilon^3) \qquad \text{(A.6)}
\end{aligned}
$$

$$
\begin{aligned}
&\varepsilon\mathcal{K}\big(f'(\psi(y-\theta(\xi,\tau)))u_1(\xi,y,\tau)\big) \\
&\quad = \varepsilon\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy'k(x',y')f'(\psi(y-y'-\theta(\varepsilon x,\tau))u_1(\varepsilon x,y-y',\tau) \\
&\quad - \varepsilon^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy'k(x',y')x'\big(f'(\psi(y-y'-\theta(\varepsilon x,\tau)) \\
&\quad \times \frac{\partial}{\partial\varepsilon x}u_1(\varepsilon x,y-y',\tau) - f''(\psi(y-y'-\theta(\varepsilon x,\tau))) \\
&\quad \times \psi'(y-y'-\theta(\varepsilon x,\tau))u_1(\varepsilon x,y-y',\tau)\frac{\partial\theta(\varepsilon x,\tau)}{\partial\varepsilon x}\big) \\
&\quad + O(\varepsilon^3) \qquad \text{(A.7)}
\end{aligned}
$$

$$
\begin{aligned}
&\varepsilon^2\mathcal{K}\big(f'(\psi(y-\theta(\xi,\tau)))u_2(\xi,y,\tau)\big) \\
&\quad \varepsilon^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy'k(x',y')f'(\psi(y-y'-\theta(\varepsilon x,\tau)))u_2(\varepsilon x,y-y',\tau) \\
&\quad + O(\varepsilon^3) \qquad \text{(A.8)}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\varepsilon^2}{2}\mathcal{K}\big(f''(\psi(y-\theta(\xi,\tau)))u_1(\xi,y,\tau)^2\big) \\
&\quad = \frac{\varepsilon^2}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy'k(x',y') \\
&\quad \times f''(\psi(y-y'-\theta(\varepsilon x,\tau)))u_1(\varepsilon x,y-y',\tau)^2 + O(\varepsilon^3) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(A.9)}
\end{aligned}
$$

Erasing odd moments and neglecting higher-order terms, we derive equation 4.3.

Appendix B

We show that an uneven stationary solution in the symmetric model (see equation 3.2) is transformed into a traveling wave solution by some small symmetrical perturbation. Here, we add small constant perturbation εs to equation 3.2 as follows:

$$\frac{\partial}{\partial t}w(x,y,t) = -w + \mathcal{K}f(w) + \varepsilon s, \tag{B.1}$$

Solutions in the neighborhood of an equilibrium solution for an unperturbed system can be represented as

$$w(x,y,t) = \psi(y-\theta(\tau)) + \varepsilon u(y,\tau) \tag{B.2}$$

where $\tau = \varepsilon t$. Substituting equation B.2 into B.1, expanding a polynomial around $\varepsilon = 0$, and neglecting higher-order terms, we obtain

$$\psi'(y-\theta(\tau))\frac{d\theta}{d\tau} = \mathcal{L}_\theta u + s. \tag{B.3}$$

Averaging equation B.3 using the weight $f'(\psi(y-\theta))\psi'(y-\theta)$ as in equation 4.7, we obtain the following phase equation:

$$c\frac{d\theta}{d\tau} = d, \tag{B.4}$$

$$c = \int_{-\infty}^{\infty} dy f'(\psi(y))\psi'(y)\psi'(y), \qquad d = s\int_{-\infty}^{\infty} dy f'(\psi(y))\psi'(y). \tag{B.5}$$

If $\psi(y)$ is uneven, $d \neq 0$. Therefore, an uneven stationary solution is transformed into a traveling wave solution by a constant perturbation.

Acknowledgments

This work was partially supported by Grants-in-Aid for Scientific Research in Priority Area (2) No. 07252219, Grants-in-Aid for the Encouragement of Young Scientists No. 2871, and JSPS Research Fellowships for Young Scientists.

References

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

Amari, S. (1980). Topographic organization of nerve fields. *Bulletin of Mathematical Biology*, *42*, 339–364.

Aonishi, T., Kurata, K., & Mito, T. (1997). *A phase locking theory of matching between rotated images by a dynamic link matching*. Unpublished manuscript.

Bienenstock, E., & Doursat, R. (1994). A shape-recognition model using dynamical links. *Network, 5*, 241–258.

Bienenstock, E., & von der Malsburg, C. (1987). A neural network for invariant pattern recognition. *Europhysics Letters, 4*, 121–126.

Ermentrout, G. B. (1981). $n : m$ Phase-locking of weakly coupled oscillators. *Journal of Mathematical Biology, 6*, 327–342.

Häussler, A. F., & von der Malsburg, C. (1983). Development of retinotopic projections—An analytical treatment. *J. Theor. Neurobio., 2*, 47–73.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern., 43*, 59–69.

Konen, W. K., Maurer, T., & von der Malsburg, C. (1994). A fast dynamic link matching algorithm for invariant pattern recognition. *Neural Networks, 7*(6/7), 1019–1030.

Konen, W. K., & von der Malsburg, C. (1993). Learning to generalize from single examples in the dynamic link architecture. *Neural Computation, 5*, 719–735.

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computer, 42*(3), 300–311.

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature, 317*, 314–319.

Takeuchi, A., & Amari, S. (1979). Formation of topographic maps and columnar microstructures in nerve fields. *Biol. Cybern., 35*, 63–72.

Received August 26, 1996; accepted June 7, 1997.

LETTER Communicated by Klaus Obermayer

Constrained Optimization for Neural Map Formation: A Unifying Framework for Weight Growth and Normalization

Laurenz Wiskott
Computational Neurobiology Laboratory, Salk Institute for Biological Studies, San Diego, CA 92186-5800, U.S.A. http://www.cnl.salk.edu/CNL/

Terrence Sejnowski
Computational Neurobiology Laboratory, Howard Hughes Medical Institute, Salk Institute for Biological Studies, San Diego, CA 92186-5800, U.S.A.
Department of Biology, University of California, San Diego, La Jolla, CA 92093, U.S.A.

Computational models of neural map formation can be considered on at least three different levels of abstraction: detailed models including neural activity dynamics, weight dynamics that abstract from the neural activity dynamics by an adiabatic approximation, and constrained optimization from which equations governing weight dynamics can be derived. Constrained optimization uses an objective function, from which a weight growth rule can be derived as a gradient flow, and some constraints, from which normalization rules are derived. In this article, we present an example of how an optimization problem can be derived from detailed nonlinear neural dynamics. A systematic investigation reveals how different weight dynamics introduced previously can be derived from two types of objective function terms and two types of constraints. This includes dynamic link matching as a special case of neural map formation. We focus in particular on the role of coordinate transformations to derive different weight dynamics from the same optimization problem. Several examples illustrate how the constrained optimization framework can help in understanding, generating, and comparing different models of neural map formation. The techniques used in this analysis may also be useful in investigating other types of neural dynamics.

1 Introduction

Neural maps are an important motif in the structural organization of the brain. The best-studied maps are those in the early visual system. For example, the retinotectal map connects a two-dimensional array of ganglion cells in the retina to a corresponding map of the visual field in the optic tectum of vertebrates in a neighborhood-preserving fashion. These are called topographic maps. The map from the lateral geniculate nucleus (LGN) to the primary visual cortex (V1) is more complex because the inputs coming from

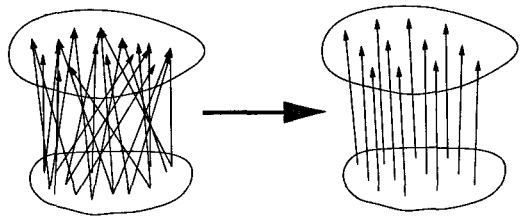


Figure 1: Goal of neural map formation: The initially random all-to-all connectivity self-organizes into an orderly connectivity that appropriately reflects the correlations within the input stimuli and the induced correlations within the output layer. The output correlations also depend on the connectivity within the output layer.

LGN include signals from both eyes and are unoriented, but most cells in V1 are tuned for orientation, an emergent property. Neurons with preferred orientation and ocular dominance in area V1 form a columnar structure, where neurons responding to the same eye or the same orientation tend to be neighbors. Other neural maps are formed in the somatosensory, the auditory, and the motor systems. All neural maps connect an input layer, possibly divided into different parts (e.g., left and right eye), to an output layer. Each neuron in the output layer can potentially receive input from all neurons in the input layer (here we ignore the limits imposed by restricted axonal arborization and dendritic extension). However, particular receptive fields develop due to a combination of genetically determined and activity-driven mechanisms for self-organization. Although cortical maps have many feedback projections (for example, from area V1 back to the LGN), these are disregarded in most models of map formation and will not be considered here.

The goal of neural map formation is to self-organize from an initial random all-to-all connectivity a regular pattern of connectivity, as in Figure 1, for the purpose of producing a representation of the input on the output layer that is of further use to the system. The development of the structure depends on the architecture, the lateral connectivity, the initial conditions, and the weight dynamics, including growth rule and normalization rules.

Fig 1 The first model of map formation, introduced by von der Malsburg (1973), was for a small patch of retina stimulated with bars of different orientation. The model self-organized orientation columns, with neighboring neurons having receptive fields tuned to similar orientation. This model already included all the crucial ingredients important for map formation: (1) characteristic correlations within the stimulus patterns, (2) lateral interactions within the output layer, inducing characteristic correlations there

as well, (3) Hebbian weight modification, and (4) competition between synapses by weight normalization. Many similar models have been proposed since then for different types of map formation (see Erwin, Obermayer, & Schulten, 1995; Swindale, 1996; and Table 2 for examples). We do not consider models that are based on chemical markers (e.g., von der Malsburg & Willshaw, 1977). Although they may be conceptionally similar to those based on neural activities, they can differ significantly in the detailed mathematical formulation. Nor do we consider in detail models that treat the input layer as a low-dimensional space, say two-dimensional for the retina, from which input vectors are drawn (e.g., Kohonen, 1982, but see section 6.8). The output neurons then receive only two synapses per neuron, one for each input dimension.

The dynamic link matching model (e.g., Bienenstock & von der Malsburg, 1987; Konen, Maurer, & von der Malsburg, 1994) is a form of neural map formation that has been developed for pattern recognition. It is mathematically similar to the self-organization of retinotectal projections; in addition, each neuron has a visual feature attached, so that a neural layer can be considered as a labeled graph representing a visual pattern. Each synapse has associated with it an individual value, which affects the dynamics and expresses the similarity between the features of connected neurons. The self-organization process then not only tends to generate a neighborhood preserving map, it also tends to connect neurons having similar features. If the two layers represent similar patterns, the map formation dynamics finds the correct feature correspondences and connects the corresponding neurons.

Models of map formation have been investigated by analysis (e.g., Amari, 1980; Häussler & von der Malsburg, 1983) and computer simulations. An important tool for both methods is the objective function (or energy function) from which the dynamics can be generated as a gradient flow. The objective value (or energy) can be used to estimate which weight configurations would be more likely to arise from the dynamics (e.g., MacKay & Miller, 1990). In computer simulations, the objective function is maximized (or the energy function is minimized) numerically in order to find stable solutions of the dynamics (e.g., Linsker, 1986; Bienenstock & von der Malsburg, 1987).

Objective functions, which can also serve as a Lyapunov function, have many advantages. First, the existence of an objective function guarantees that the dynamics does not have limit cycles or chaotic attractors as solutions. Second, an objective function often provides more direct and intuitive insight into the behavior of a dynamics, and the effects of each term can be understood more easily. Third, an objective function allows additional mathematical tools to be used to analyze the system, such as methods from statistical physics. Finally, an objective function provides connections to more abstract models, such as spin systems, which have been studied in depth.

Although objective functions have been used before in the context of neural map formation, they have not yet been investigated systematically. The goal of this article is to derive objective functions for a wide variety of models. Although growth rules can be derived from objective functions as gradient flows, normalization rules are derived from constraints by various methods. Thus, objective functions and constraints have to be considered in conjunction and form a constrained optimization problem. We show that although two models may differ in the formulation of their dynamics, they may be derived from the same constrained optimization problem, thus providing a unifying framework for the two models. The equivalence between different dynamics is revealed by coordinate transformations. A major focus of this article is therefore on the effects of coordinate transformations on weight growth rules and normalization rules.

1.1 Model Architecture. The general architecture considered here consists of two layers of neurons, an input and an output layer, as in Figure 2. (We use the term *layer* for a population of neurons without assuming a particular geometry.) Input neurons are indicated by ρ (retina) and output neurons by τ (tectum); the index ν can indicate a neuron in either layer. Neural activities are indicated by a. Input neurons are connected all-to-all to output neurons, but there are no connections back to the input layer. Thus, the dynamics in the input layer is completely independent of the output layer and can be described by mean activities $\langle a_\rho \rangle$ and correlations $\langle a_\rho, a_{\rho'} \rangle$. Effective lateral connections within a layer are denoted by $D_{\rho\rho'}$ and $D_{\tau\tau'}$; connections projecting from the input to the output layer are denoted by $w_{\tau\rho}$. The second index always indicates the presynaptic neuron and the first index the postsynaptic neuron. The lateral connections defined here are called *effective*, because they need not correspond to physical connections. For example, in the input layer, the effective lateral connections represent the correlations between input neurons regardless of what induced the correlations, $D_{\rho\rho'} = \langle a_\rho, a_{\rho'} \rangle$. In the example below, the output layer has short-term excitatory and long-term inhibitory connections; the effective lateral connections, however, are only excitatory. The effective lateral connections thus represent functional properties of the lateral interactions and not the anatomical connectivity itself.

To make the notation simpler, we use the definitions $i = \{\rho, \tau\}$, $j = \{\rho', \tau'\}$, $A_{ij} = D_{\tau\tau'} A_{\rho'} = D_{\tau\tau'} \langle a_{\rho'} \rangle$, and $D_{ij} = D_{\tau\tau'} D_{\rho\rho'} = D_{\tau\tau'} \langle a_\rho, a_{\rho'} \rangle$ in section 3 and later. We assume symmetric matrices $A_{ij} = A_{ji}$ and $D_{ij} = D_{ji}$, which requires some homogeneity of the architecture, that is, $\langle a_\rho \rangle = \langle a_{\rho'} \rangle$, $\langle a_\rho, a_{\rho'} \rangle = \langle a_{\rho'}, a_\rho \rangle$, and $D_{\tau\tau'} = D_{\tau'\tau}$.

In the next section, a simple model is used to demonstrate the basic procedure for deriving a constrained optimization problem from detailed neural dynamics. This procedure has three steps. First, the neural dynamics is transformed into a weight dynamics, where the induced correlations are expressed directly in terms of the synaptic weights, thus eliminating neu-

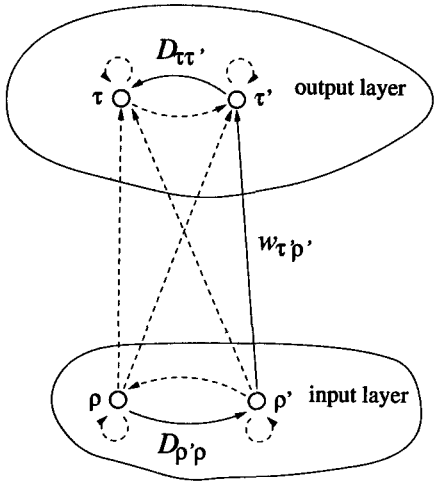


Figure 2: General architecture: Neurons in the input layer are connected all-to-all to neurons in the output layer. Each layer has effective lateral connections D representing functional aspects of the lateral connectivity (e.g., characteristic correlations). As an example, a path through which activity can propagate from neuron ρ to neuron τ is shown by solid arrows. Other connections are shown as dashed arrows.

ral activities from the dynamics by an adiabatic approximation. Second, an objective function is constructed, which can generate the dynamics of the growth rule as a gradient flow. Third, the normalization rules need to be considered and, if possible, derived from constraint functions. The last two steps depend on each other insofar as growth rule, as well as normalization rules, must be inferred under the same coordinate transformation. The three important aspects of this example—deriving correlations, constructing objective functions, and considering the constraints—are then discussed in greater detail in the following three sections, respectively. Readers may skip section 2 and continue directly with these more abstract considerations beginning in section 3. In section 6, several examples are given for how the constrained optimization framework can be used to understand, generate, and compare models of neural map formation.

2 Prototypical System

As a concrete example, consider a slightly modified version of the dynamics proposed by Willshaw and von der Malsburg (1976) for the self-organization

of a retinotectal map, where the input and output layer correspond to retina and tectum, respectively. The dynamics is qualitatively described by the following set of differential equations:

Neural activity dynamics

$$\dot{m}_\rho = -m_\rho + (k * a_{\rho'})_\rho \tag{2.1}$$

$$\dot{m}_\tau = -m_\tau + (k * a_{\tau'})_\tau + \sum_{\rho'} w_{\tau\rho'} a_{\rho'} \tag{2.2}$$

Weight growth rule

$$\dot{w}_{\tau\rho} = a_\tau a_\rho \tag{2.3}$$

Weight normalization rules

$$\text{if } w_{\tau\rho} < 0 : w_{\tau\rho} = 0 \tag{2.4}$$

$$\text{if } \sum_{\rho'} w_{\tau\rho'} > 1 : w_{\tau\rho} = \tilde{w}_{\tau\rho} + \frac{1}{M_\tau}\left(1 - \sum_{\rho'} \tilde{w}_{\tau\rho'}\right) \quad \text{for all } \rho \tag{2.5}$$

$$\text{if } \sum_{\tau'} w_{\tau'\rho} > 1 : w_{\tau\rho} = \tilde{w}_{\tau\rho} + \frac{1}{M_\rho}\left(1 - \sum_{\tau'} \tilde{w}_{\tau'\rho}\right) \quad \text{for all } \tau \tag{2.6}$$

where m denotes the membrane potential, $a_\nu = \sigma(m_\nu)$ is the mean firing rate determined by a nonlinear input-output function σ, $(k * a_{\nu'})$ indicates a convolution of the neural activities with the kernel k representing lateral connections with local excitation and global inhibition, $\tilde{w}_{\tau\rho}$ indicates weights as obtained by integrating the differential equations for one time step, that is, $\tilde{w}_{\tau\rho}(t+\Delta t) = w_{\tau\rho}(t) + \Delta t\, \dot{w}_{\tau\rho}(t)$, M_τ is the number of links terminating on output neuron τ, and M_ρ is the number of links originating from input neuron ρ. Equations 2.1 and 2.2 govern the neural activity dynamics on the two layers, equation 2.3 is the growth rule for the synaptic weights, and equations 2.4–2.6 are the normalization rules that keep the sums over synaptic weights originating from an input neuron or terminating on an output neuron equal to 1 and prevent the weights from becoming negative. Notice that since the discussion is qualitative, we included only the basic terms and discarded some parameters required to make the system work properly. One difference from the original model is that subtractive instead of multiplicative normalization rules are used.

2.1 Correlations. The dynamics within the neural layers is well understood (Amari, 1977; Konen et al., 1994). Local excitation and global inhibition lead to the development of a local patch of activity, called a *blob*. The shape and size of the blob depend on the kernel k and other parameters of the

system and can be described by $B_{\rho'\rho_0}$ if centered on input neuron ρ_0 and $B_{\tau'\tau_0}$ if centered on output neuron τ_0. The location of the blob depends on the input, which is assumed to be weak enough that it does not change the shape of the blob. Assume the input layer receives noise such that the blob arises with equal probability $p(\rho_0) = 1/R$ centered on any of the input neurons, where R is the number of input neurons. For simplicity we assume cyclic boundary conditions to avoid boundary effects. The location of the blob in the output layer, on the other hand, is affected by the input,

$$i_{\tau'}(\rho_0) = \sum_{\rho'} w_{\tau'\rho'} B_{\rho'\rho_0}, \tag{2.7}$$

received from the input layer and therefore depends on the position ρ_0 of the blob in the input layer. Only one blob can occur in each layer, and the two layers need to be reset before new blobs can arise. A sequence of blobs is required to induce the appropriate correlations.

Konen et al. (1994) have shown that without noise, blobs in the output layer will arise at location τ_0 with the largest overlap between input $i_{\tau'}(\rho_0)$ and the final blob profile $B_{\tau'\tau_0}$, that is, the location for which $\sum_{\tau'} B_{\tau'\tau_0} i_{\tau'}(\rho_0)$ is maximal. This winner-take-all behavior makes it difficult to analyze the system. We therefore make the assumption that in contrast to this deterministic dynamics, the blob arises at location τ_0 with a probability equal to the overlap between the input and blob activity,

$$p(\tau_0|\rho_0) = \sum_{\tau'} B_{\tau'\tau_0} i_{\tau'}(\rho_0) = \sum_{\tau'\rho'} B_{\tau'\tau_0} w_{\tau'\rho'} B_{\rho'\rho_0}. \tag{2.8}$$

Assume the blobs are normalized such that $\sum_{\rho'} B_{\rho'\rho_0} = 1$ and $\sum_{\tau_0} B_{\tau'\tau_0} = 1$ and that the connectivity is normalized such that $\sum_{\tau'} w_{\tau'\rho'} = 1$, which is the case for the system above if the input layer does not have more neurons than the output layer. This implies $\sum_{\tau'} i_{\tau'}(\rho_0) = 1$ and $\sum_{\tau_0} p(\tau_0|\rho_0) = 1$ and justifies the interpretation of $p(\tau_0|\rho_0)$ as a probability.

Although it is plausible that such a probabilistic blob location could be approximated by noise in the output layer, it is difficult to develop a concrete model. For a similar but more algorithmic activity model (Obermayer, Ritter, & Schulten, 1990), an exact noise model for the probabilistic blob location can be formulated (see the appendix). With equation 3.8 the probability for a particular combination of blob locations is

$$p(\tau_0, \rho_0) = p(\tau_0|\rho_0)p(\rho_0) = \sum_{\tau'\rho'} B_{\tau'\tau_0} w_{\tau'\rho'} B_{\rho'\rho_0} \frac{1}{R}, \tag{2.9}$$

and the correlation between two neurons defined as the average product of their activities is

$$\langle a_\tau a_\rho \rangle = \sum_{\tau_0\rho_0} p(\tau_0, \rho_0) B_{\tau\tau_0} B_{\rho\rho_0} \tag{2.10}$$

$$= \sum_{\tau_0\rho_0} \sum_{\tau'\rho'} B_{\tau'\tau_0} w_{\tau'\rho'} B_{\rho'\rho_0} \frac{1}{R} B_{\tau\tau_0} B_{\rho\rho_0} \tag{2.11}$$

$$= \frac{1}{R} \sum_{\tau'\rho'} \left(\sum_{\tau_0} B_{\tau'\tau_0} B_{\tau\tau_0} \right) w_{\tau'\rho'} \left(\sum_{\rho_0} B_{\rho'\rho_0} B_{\rho\rho_0} \right) \tag{2.12}$$

$$= \frac{1}{R} \sum_{\tau'\rho'} \bar{B}_{\tau\tau'} w_{\tau'\rho'} \bar{B}_{\rho'\rho}, \qquad \text{with } \bar{B}_{\nu'\nu} = \sum_{\nu_0} B_{\nu'\nu_0} B_{\nu\nu_0}, \tag{2.13}$$

where the brackets $\langle \cdot \rangle$ indicate the ensemble average over a large number of blob presentations. $\frac{1}{R}\bar{B}_{\rho'\rho}$ and $\bar{B}_{\tau\tau'}$ are the effective lateral connectivities of the input and the output layer, respectively, and are symmetrical even if the individual blobs $B_{\rho\rho_0}$ and $B_{\tau\tau_0}$ are not, that is, $D_{\rho'\rho} = \frac{1}{R}\bar{B}_{\rho'\rho}$, $D_{\tau\tau'} = \bar{B}_{\tau\tau'}$, and $D_{ij} = D_{ji} = D_{\tau\tau'} D_{\rho'\rho} = \frac{1}{R}\bar{B}_{\tau\tau'}\bar{B}_{\rho'\rho}$. Notice the linear relation between the weights $w_{\tau'\rho'}$ and the correlations $\langle a_\tau a_\rho \rangle$ in the probabilistic blob model (see equation 2.13).

Substituting the correlation into equation 2.3 for the weight dynamics leads to:

$$\langle \dot{w}_{\tau\rho} \rangle = \langle a_\tau a_\rho \rangle = \frac{1}{R} \sum_{\tau'\rho'} \bar{B}_{\tau\tau'} w_{\tau'\rho'} \bar{B}_{\rho'\rho}. \tag{2.14}$$

The same normalization rules given above (equations 2.4–2.6) apply to this dynamics. Since there is little danger of confusion, we neglect the averaging brackets for $\langle \dot{w}_{\tau\rho} \rangle$ in subsequent equations and simply write $\dot{w}_{\tau\rho} = \langle a_\tau, a_\rho \rangle$.

Although we did not give a mathematical model of the mechanism by which the probabilistic blob location as given in equation 2.8 could be implemented, it may be interesting to note that the probabilistic approach can be generalized to other activity patterns, such as stripe patterns or hexagons, which can be generated by Mexican hat interaction functions (local excitation, finite-range inhibition) (von der Malsburg, 1973; Ermentrout & Cowan, 1979). If the probability for a stripe pattern's arising in the output layer is linear in its overlap with the input, the same derivation follows, though the indices ρ_0 and τ_0 will then refer to phase and orientation of the patterns rather than location of the blobs.

Using the probabilistic blob location in the output layer instead of the deterministic one is analogous to the soft competitive learning proposed by Nowlan (1990) as an alternative to hard (or winner-take-all) competitive learning. Nowlan demonstrated superior performance of soft competition over hard competition for a radial basis function network tested on recognition of handwritten characters and spoken vowels, and suggested there might be a similar advantage for neural map formation. The probabilistic blob location induced by noise might help improve neural map formation by avoiding local optima.

2.2 Objective Function. The next step is to find an objective function that generates the dynamics as a gradient flow. For the above example, a suitable objective function is

$$H(\mathbf{w}) = \frac{1}{2R} \sum_{\tau\rho\tau'\rho'} w_{\tau\rho} \bar{B}_{\rho\rho'} \bar{B}_{\tau\tau'} w_{\tau'\rho'}, \tag{2.15}$$

since it yields equation 2.14 from $\dot{w}_{\tau\rho} = \frac{\partial H(\mathbf{w})}{\partial w_{\tau\rho}}$, taking into account that $\bar{B}_{\nu\nu'} = \bar{B}_{\nu'\nu}$.

2.3 Constraints. The normalization rules given above ensure that synaptic weights do not become negative and that the sums over synaptic weights originating from an input neuron or terminating on an output neuron do not become larger than 1. This can be written in the form of inequalities for constraint functions g:

$$g_{\tau\rho}(\mathbf{w}) = w_{\tau\rho} \geq 0, \tag{2.16}$$

$$g_{\tau}(\mathbf{w}) = 1 - \sum_{\rho'} w_{\tau\rho'} \geq 0, \tag{2.17}$$

$$g_{\rho}(\mathbf{w}) = 1 - \sum_{\tau'} w_{\tau'\rho} \geq 0. \tag{2.18}$$

These constraints define a region within which the objective function is to be maximized by steepest ascent. While the constraints follow uniquely from the normalization rules, the converse is not true. In general, there are various normalization rules that would enforce or at least approximate the constraints, but only some of them are compatible with the constrained optimization framework. As shown in section 5.2.1, compatible normalization rules can be obtained by the method of Lagrangian multipliers. If a constraint g_x, $x \in \{\tau\rho, \tau, \rho\}$ is violated, a normalization rule of the form

$$\text{if } g_x(\tilde{\mathbf{w}}) < 0: \qquad w_{\tau\rho} = \tilde{w}_{\tau\rho} + \lambda_x \frac{\partial g_x}{\partial \tilde{w}_{\tau\rho}} \qquad \text{for all } \tau\rho, \tag{2.19}$$

has to be applied, where λ_x is a Lagrangian multiplier and determined such that $g_x(\mathbf{w}) = 0$. This method actually leads to equations 2.4–2.6, which are therefore a compatible set of normalization rules for the constraints above. This is necessary to make the formulation as a constrained optimization problem (see equations 2.15–2.18) an appropriate description of the original dynamics (see equations 2.3–2.6).

This example illustrates the general scheme by which a detailed model dynamics for neural map formation can be transformed into a constrained optimization problem. The correlations, objective functions, and constraints are discussed in greater detail and for a wide variety of models below.

3 Correlations

In the above example, correlations in a highly nonlinear dynamics led to a linear relationship between synaptic weights and the induced correlations. We derived effective lateral connections in the input as well as the output layer mediating these correlations. Corresponding equations for the correlations have been derived for other, mostly linear activity models (e.g., Linsker, 1986; Miller, 1990; von der Malsburg, 1995), as summarized here.

Assume the dynamics in the input layer is described by neural activities $a_\rho(t) \in \mathbb{R}$, which yield mean activities $\langle a_\rho \rangle$ and correlations $\langle a_\rho, a_{\rho'} \rangle$. The input received by the output layer is assumed to be a linear superposition of the activities of the input neurons:

$$i_{\tau'} = \sum_{\rho'} w_{\tau'\rho'} a_{\rho'}. \tag{3.1}$$

This input then produces activity in the output layer through effective lateral connections in a linear fashion:

$$a_\tau = \sum_{\tau'} D_{\tau\tau'} i_{\tau'} = \sum_{\tau'\rho'} D_{\tau\tau'} w_{\tau'\rho'} a_{\rho'}. \tag{3.2}$$

As seen in the above example, this linear behavior could be generated by a nonlinear model. Thus, the neurons need not be linear, only the effective behavior of the correlations (cf. Sejnowski, 1976; Ginzburg & Sompolinsky, 1994). The mean activity of output neurons is

$$\langle a_\tau \rangle = \sum_{\tau'\rho'} D_{\tau\tau'} w_{\tau'\rho'} \langle a_{\rho'} \rangle = \sum_j A_{ij} w_j. \tag{3.3}$$

Assuming a linear correlation function ($\langle a_\rho, \alpha(a_{\rho'} + a_{\rho''}) \rangle = \alpha \langle a_\rho, a_{\rho'} \rangle + \alpha \langle a_\rho, a_{\rho''} \rangle$ with a real constant α) such as the average product or the covariance (Sejnowski, 1977), the correlation between input and output neurons is

$$\langle a_\tau, a_\rho \rangle = \sum_{\tau'\rho'} D_{\tau\tau'} w_{\tau'\rho'} \langle a_{\rho'}, a_\rho \rangle = \sum_j D_{ij} w_j. \tag{3.4}$$

Note that $i = \{\rho, \tau\}$, $j = \{\rho', \tau'\}$, $A_{ij} = A_{ji} = D_{\tau\tau'} A_{\rho'} = D_{\tau\tau'} \langle a_{\rho'} \rangle$, and $D_{ij} = D_{ji} = D_{\tau\tau'} D_{\rho'\rho} = D_{\tau\tau'} \langle a_{\rho'}, a_\rho \rangle$. Since the right-hand sides of equations 3.3 and 3.4 are formally equivalent, we will consider only the latter one in the further analysis, bearing in mind that equation 3.3 is included as a special case.

In this linear correlation model, all variables may assume negative values. This may not be plausible for the neural activities a_ρ and a_τ. However,

equation 3.4 can be derived also for nonnegative activities, and a similar equation as equation 3.3 can be derived if the mean activities $\langle a_\rho \rangle$ are positive. The difference for the latter would be an additional constant, which can always be compensated for in the growth rule.

The correlation model in Linsker (1986) differs from the linear one introduced here in two respects. The input (see equation 3.1) has an additional constant term, and correlations are defined by subtracting positive constants from the activities. However, it can be shown that correlations in the model in Linsker (1986) are a linear combination of a constant and the terms of equations 3.3 and 3.4.

4 Objective Functions

In general, there is no systematic way of finding an objective function for a particular dynamical system, but it is possible to determine whether there exists an objective function. The necessary and sufficient condition is that the flow field of the dynamics be curl free. If there exists an objective function $H(\mathbf{w})$ with continuous partial derivatives of order two that generates the dynamics $\dot{w}_i = \partial H(\mathbf{w})/\partial w_i$, then

$$\frac{\partial \dot{w}_i}{\partial w_j} = \frac{\partial^2 H(\mathbf{w})}{\partial w_j \partial w_i} = \frac{\partial^2 H(\mathbf{w})}{\partial w_i \partial w_j} = \frac{\partial \dot{w}_j}{\partial w_i}. \tag{4.1}$$

The existence of an objective function is thus equivalent to $\partial \dot{w}_i/\partial w_j = \partial \dot{w}_j/\partial w_i$, which can be checked easily. For the dynamics given by

$$\dot{w}_i = \sum_j D_{ij} w_j \tag{4.2}$$

(cf. equation 2.14), for example, $\partial \dot{w}_i/\partial w_j = D_{ij} = \partial \dot{w}_j/\partial w_i$, which shows that it can be generated as a gradient flow. A suitable objective function is

$$H(\mathbf{w}) = \frac{1}{2} \sum_{ij} w_i D_{ij} w_j \tag{4.3}$$

(cf. equation 2.15), since it yields $\dot{w}_i = \partial H(\mathbf{w})/\partial w_i$.

A dynamics that cannot be generated by an objective function directly is

$$\dot{w}_i = w_i \sum_j D_{ij} w_j, \tag{4.4}$$

as used in Häussler and von der Malsburg (1983), since for $i \neq j$ we obtain $\partial \dot{w}_i/\partial w_j = w_i D_{ij} \neq w_j D_{ji} = \partial \dot{w}_j/\partial w_i$, and $\dot{w}_i$ is not curl free. However, it is

sometimes possible to convert a dynamics with curl into a curl-free dynamics by a coordinate transformation. Applying the transformation $w_i = \frac{1}{4}v_i^2$ ($\mathcal{C}^w$) to equation 4.4 yields

$$\dot{v}_i = \frac{\mathrm{d}v_i}{\mathrm{d}w_i}\dot{w}_i = \sqrt{w_i}\sum_j D_{ij}w_j = \frac{1}{2}v_i\sum_j D_{ij}\frac{1}{4}v_j^2, \tag{4.5}$$

which is curl free, since $\partial\dot{v}_i/\partial v_j = \frac{1}{2}v_i D_{ij}\frac{1}{2}v_j = \partial\dot{v}_j/\partial v_i$. Thus, the dynamics of $\dot{v}_i$ in the new coordinate system $\mathcal{V}^w$ can be generated as a gradient flow. A suitable objective function is

$$H(\mathbf{v}) = \frac{1}{2}\sum_{ij}\frac{1}{4}v_i^2 D_{ij}\frac{1}{4}v_j^2, \tag{4.6}$$

since it yields $\dot{v}_i = \partial H(\mathbf{v})/\partial v_i$. Transforming the dynamics of $\mathbf{v}$ back into the original coordinate system $\mathcal{W}$, of course, yields the original dynamics in equation 4.4:

$$\dot{w}_i = \frac{\mathrm{d}w_i}{\mathrm{d}v_i}\dot{v}_i = \frac{1}{4}v_i^2\sum_j D_{ij}\frac{1}{4}v_j^2 = w_i\sum_j D_{ij}w_j. \tag{4.7}$$

Coordinate transformations thus can provide objective functions for dynamics that are not curl free. Notice that $H(\mathbf{v})$ is the same objective function as $H(\mathbf{w})$ (see equation 4.3) evaluated in $\mathcal{V}^w$ instead of $\mathcal{W}$. Thus $H(\mathbf{v}) = H(\mathbf{w}(\mathbf{v}))$ and H is a Lyapunov function for both dynamics.

More generally, for an objective function H and a coordinate transformation $w_i = w_i(v_i)$,

$$\dot{w}_i = \frac{\mathrm{d}}{\mathrm{d}t}[w_i(v_i)] = \frac{\mathrm{d}w_i}{\mathrm{d}v_i}\dot{v}_i = \frac{\mathrm{d}w_i}{\mathrm{d}v_i}\frac{\partial H}{\partial v_i} = \left(\frac{\mathrm{d}w_i}{\mathrm{d}v_i}\right)^2\frac{\partial H}{\partial w_i}, \tag{4.8}$$

which implies that the coordinate transformation simply adds a factor $(\mathrm{d}w_i/\mathrm{d}v_i)^2$ to the original growth term obtained in the original coordinate system $\mathcal{W}$. For the dynamics in equation 4.4 derived under the coordinate transformation $w_i = \frac{1}{4}v_i^2$ ($\mathcal{C}^w$) relative to the dynamics of equation 4.2, we verify that $(\mathrm{d}w_i/\mathrm{d}v_i)^2 = w_i$. Equation 4.8 also shows that fixed points are preserved under the coordinate transformation in the region where $\mathrm{d}w_i/\mathrm{d}v_i$ is defined and finite but that additional fixed points may be introduced if $\mathrm{d}w_i/\mathrm{d}v_i = 0$.

This effect of coordinate transformations is known from the general theory of relativity and tensor analysis (e.g., Dirac, 1996). The gradient of a potential (or objective function) is a covariant vector, which adds the factor

put neurons, $\langle a_\tau \rangle = \sum_j A_{\tau j} w_j$. This term is, for instance, important to form topographic maps. Functional aspects of term Q are discussed in section 6.3.

5 Constraints

A constraint is either an inequality describing a surface (of dimensionality $RT - 1$ if RT is the number of weights) between valid and invalid region or an equality describing the valid region as a surface. A normalization rule is a particular prescription for how the constraint has to be enforced. Thus, constraints can be uniquely derived from normalization rules but not vice versa.

5.1 Orthogonal Versus Nonorthogonal Normalization Rules. Normalization rules can be divided into two classes: those that enforce the constraints orthogonal to the constraint surface, that is, along the gradient of the constraint function, and those that also have a component tangential to the constraint surface (see Figure 4). We refer to the former ones as *orthogonal* and to the latter ones as *nonorthogonal*.

Only the orthogonal normalization rules are compatible with an objective function, as is illustrated in Figure 5. For a dynamics induced as an ascending gradient flow of an objective function, the value of the objective function constantly increases as long as the weights change. If the weights cross a constraint surface, a normalization rule has to be applied iteratively to the growth rule. Starting from the constraint surface at point $\mathbf{w}'$, the gradient ascent causes a step to point $\tilde{\mathbf{w}}$ in the invalid region, where $\tilde{\mathbf{w}} - \mathbf{w}'$ is in general nonorthogonal to the constraint surface. A normalization rule causes a step back to $\mathbf{w}$ on the constraint surface. If the normalization rule is orthogonal, that is, $\mathbf{w} - \tilde{\mathbf{w}}$ is orthogonal to the constraint surface, $\mathbf{w} - \tilde{\mathbf{w}}$ is shorter than or equal to $\tilde{\mathbf{w}} - \mathbf{w}'$ and the cosine of the angle between the combined step $\mathbf{w} - \mathbf{w}'$ and the gradient $\tilde{\mathbf{w}} - \mathbf{w}'$ is nonnegative, that is, the value of the objective function does not decrease. This cannot be guaranteed for nonorthogonal normalization rules, in which case the objective function of the unconstrained dynamics may not even be a Lyapunov function for the combined system, including weight dynamics and normalization rules. Thus, only orthogonal normalization rules can be used in the constrained optimization framework.

The term *orthogonal* is not well defined away from the constraint surface. However, the constraints used in this article are rather simple, and a natural orthogonal direction is usually available for all weight vectors. Thus, the term *orthogonal* will also be used for normalization rules that do not project back exactly onto the constraint surface but keep the weights close to the surface and affect the weights orthogonal to it. For more complicated constraint surfaces, more careful considerations may be required.

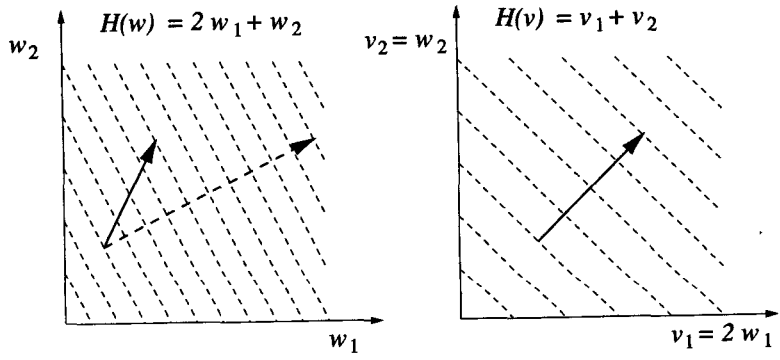


Figure 3: The effect of coordinate transformations on the induced dynamics. The figure shows a simple objective function H in the original coordinate system $\mathcal{W}$ (left) and the new coordinate system $\mathcal{V}$ (right) with $w_1 = v_1/2$ and $w_2 = v_2$. The gradient induced in $\mathcal{W}$ (dashed arrow) and the gradient induced in $\mathcal{V}$ and then backtransformed into $\mathcal{W}$ (solid arrows) have the same component in the w_2 direction but differ by a factor of four in the w_1 direction (cf. equation 4.8). Notice that the two dynamics differ in amplitude and direction, but that H is a Lyapunov function for both.

dw_i/dv_i through the transformation from $\mathcal{W}$ to $\mathcal{V}$. Since $\dot{\mathbf{v}}$ as a kinematic description of the trajectory is a contravariant vector, this adds another factor dw_i/dv_i through the transformation back from $\mathcal{V}$ to $\mathcal{W}$. If both vectors were either covariant or contravariant, the back-and-forth transformation between the different coordinate systems would have no effect. The same argument holds for the constraints in section 5.2. In some cases, it may also be useful to consider more general coordinate transformations $w_i = w_i(\mathbf{v})$ where each weight w_i may depend on all variables v_j, as is common in the general theory of relativity and tensor analysis. Equation 4.8 would have to be modified correspondingly. In Figure 3, the effect of coordinate transformations is illustrated by a simple example.

Table 1 shows two objective functions and the corresponding dynamics terms they induce under different coordinate transformations. The first objective function, L, is linear in the weights and induces constant weight growth (or decay) under coordinate transformation $\mathcal{C}^1$. The growth of one weight does not depend on other weights. This term can be useful for dynamic link matching to introduce a bias for each weight depending on the similarity of the connected neurons. The second objective function, Q, is a quadratic form. The induced growth rule for one weight includes other weights and is usually based on correlations between input and output neurons, $\langle a_\tau a_\rho \rangle = \sum_j D_{ij} w_j$, and possibly also the mean activities of out-

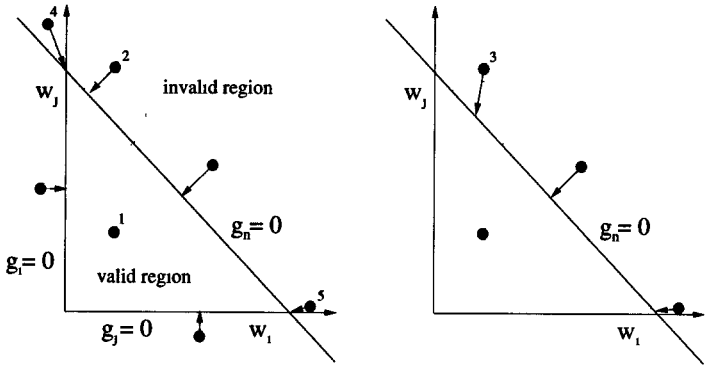


Figure 4: Different constraints and different ways in which constraints can be violated and enforced. The constraints along the axes are given by $g_i = w_i \geq 0$ and $g_j = w_j \geq 0$, which keep the weights w_i and w_j nonnegative. The constraint $g_n = 1 - (w_i + w_j) \geq 0$ keeps the sum of the two weights smaller or equal to 1. Black dots indicate points in state-space that may have been reached by the growth rule. Dot 1: None of the constraints is violated, and no normalization rule is applied. Dot 2: $g_n \geq 0$ is violated, and an orthogonal subtractive normalization rule is applied. Dot 3: $g_n \geq 0$ is violated, and a nonorthogonal multiplicative normalization rule is applied. Notice that the normalization does not follow the gradient of g_n; it is not perpendicular to the line $g_n = 0$. Dot 4: Two constraints are violated, and the respective normalization rules must be applied simultaneously. Dot 5: $g_n \geq 0$ is violated, but the respective normalization rule violates $g_j \geq 0$. Again both rules must be applied simultaneously. The dotted circles indicate regions considered in greater detail in Figure 5.

Whether a normalization rule is orthogonal depends on the coordinate system in which it is applied. This is illustrated in Figure 6 and discussed in greater detail below. The same rule can be nonorthogonal in one coordinate system but orthogonal in another. It is important to find the coordinate system in which an objective function can be derived and the normalization rules are orthogonal. This then is the coordinate system in which the model can be most conveniently analyzed. Not all nonorthogonal normalization rules can be transformed into orthogonal ones. In Wiskott and von der Malsburg (1996), for example, a normalization rule is used that affects a group of weights if single weights grow beyond their limits. Since the constraint surface depends on only one weight, only that weight can be affected by an orthogonal normalization rule. Thus, this normalization rule cannot be made orthogonal.

5.2 Constraints Can Be Enforced in Different Ways. For a given constraint, orthogonal normalization rules can be derived using various meth-

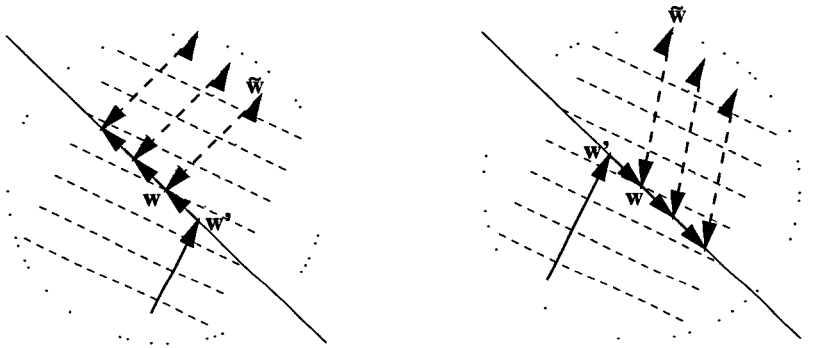


Figure 5: The effect of orthogonal versus nonorthogonal normalization rules. The two circled regions are taken from Figure 4. The effect of the orthogonal subtractive rule is shown on the left, and the nonorthogonal multiplicative rule is shown on the right. The growth dynamics is assumed to be induced by an objective function, the equipotential curves of which are shown as dashed lines. The objective function increases to the upper right. The growth rule (dotted arrows) and normalization rule (dashed arrows) are applied iteratively. The net effect is different in the two cases. For the orthogonal normalization rule, the dynamics increases the value of the objective function, while for the nonorthogonal normalization, the value decreases and the objective function that generates the growth rule is not even a Lyapunov function for the combined system.

ods. These include the method of Lagrangian multipliers, the inclusion of penalty terms, and normalization rules that are integrated into the weight dynamics without necessarily having any objective function. The former two methods are common in optimization theory. The latter is more specific to a model of neural map formation. It is also possible to substitute a constraint by a coordinate transformation.

5.2.1 Method of Lagrangian Multipliers. Lagrangian multipliers can be used to derive explicit normalization rules, such as equations 2.4–2.6. If the constraint $g_n(\mathbf{w}) \geq 0$ is violated for $\tilde{\mathbf{w}}$ as obtained after one integration step of the learning rule, $\tilde{w}_i(t + \Delta t) = w_i(t) + \Delta t\, \dot{w}_i(t)$, the weight vector has to be corrected along the gradient of the constraint function g_n, which is orthogonal to the constraint surface $g_n(\mathbf{w}) = 0$,

$$\text{if } g_n(\tilde{\mathbf{w}}) < 0: \qquad w_i = \tilde{w}_i + \lambda_n \frac{\partial g_n}{\partial \tilde{w}_i} \qquad \text{for all } i, \tag{5.1}$$

where $(\partial g_n/\partial \tilde{w}_i) = (\partial g_n/\partial w_i)$ at $\mathbf{w} = \tilde{\mathbf{w}}$ and $\lambda_n = \lambda_n(\tilde{\mathbf{w}})$ is a Lagrangian multiplier and determined such that $g_n(\mathbf{w}) = 0$ is obtained. If no constraint

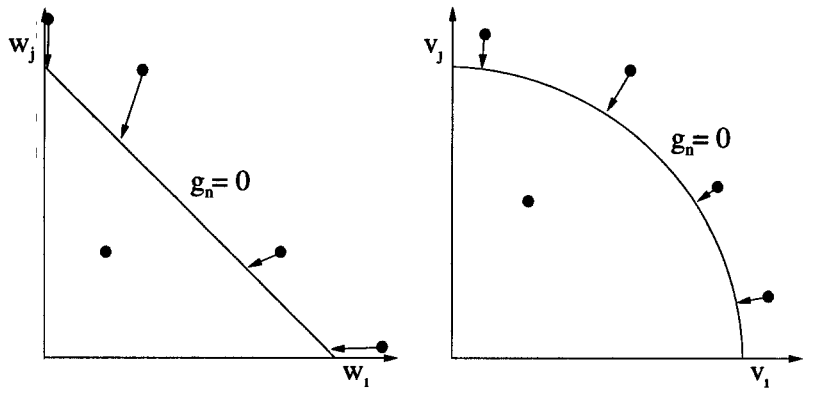


Figure 6: The effect of a coordinate transformation on a normalization rule. The constraint function is $g_n = 1 - (w_i + w_j) \geq 0$, and the coordinate transformation is $w_i = \frac{1}{4}v_i^2, w_j = \frac{1}{4}v_j^2$. In the new coordinate system $\mathcal{V}^w$ (right), the constraint becomes $g_n = 1 - \frac{1}{4}(v_i^2 + v_j^2) \geq 0$ and leads there to an orthogonal multiplicative normalization rule. Transforming back into $\mathcal{W}$ (left) then yields a nonorthogonal multiplicative normalization rule.

is violated, the weights are simply taken to be $w_i = \tilde{w}_i$. The constraints that must be taken into account, either because they are violated or because they become violated if a violated one is enforced, are called *operative*. All others are called *inoperative* and do not need to be considered for that integration step. If there is more than one operative constraint, the normalization rule becomes

$$\text{if } g_n(\tilde{\mathbf{w}}) < 0: \qquad w_i = \tilde{w}_i + \sum_{n \in N_O} \lambda_n \frac{\partial g_n}{\partial \tilde{w}_i} \qquad \text{for all } i, \tag{5.2}$$

where N_O denotes the set of operative constraints. The Lagrangian multipliers λ_n are determined such that $g_{n'}(\mathbf{w}) = 0$ for all $n' \in N_O$ (cf. Figure 4). Computational models of neural map formation usually take another strategy and simply iterate the normalization rules (see equation 5.1) for the operative constraints individually, which is in general not accurate but may be sufficient for most practical purposes. It should also be mentioned that in the standard method of Lagrangian multipliers as usually applied in physics or optimization theory, the two steps, weight growth and normalization, are combined in one dynamical equation such that $\mathbf{w}$ remains on the constraint surface. The steps were split here to obtain explicit normalization rules independent of growth rules.

Consider now the effect of coordinate transformations on the normalization rules derived by the method of Lagrangian multipliers. The constraint in equation 2.17 can be written as $g_n(\mathbf{w}) = \theta_n - \sum_{i\in I_n} w_i \geq 0$ and leads to a subtractive normalization rule as in the example above (see equation 2.5). Under the coordinate transformation $\mathcal{C}^w$ ($w_i = \frac{1}{4}v_i^2$), the constraint becomes $g_n(\mathbf{v}) = \theta_n - \sum_{i\in I_n} \frac{1}{4}v_i^2 \geq 0$, and in the coordinate system $\mathcal{V}^w$, the normalization rule is:

$$\text{if } g_n(\tilde{\mathbf{v}}) < 0: \qquad v_i = \tilde{v}_i - 2\left(\frac{\sqrt{\theta_n}}{\sqrt{\sum_{j\in I_n}\frac{1}{4}\tilde{v}_j^2}} - 1\right)\left(-\frac{1}{2}\tilde{v}_i\right) \tag{5.3}$$

$$= \frac{\sqrt{\theta_n}\,\tilde{v}_i}{\sqrt{\sum_{j\in I_n}\frac{1}{4}\tilde{v}_j^2}} \qquad \text{for all } i \in I_n. \tag{5.4}$$

Taking the square on both sides and applying the backtransformation from $\mathcal{V}^w$ to $\mathcal{W}$ leads to

$$\text{if } g_n(\tilde{\mathbf{w}}) < 0: \qquad w_i = \frac{\theta_n \tilde{w}_i}{\sum_{j\in I_n}\tilde{w}_j} \qquad \text{for all } i \in I_n. \tag{5.5}$$

This is a multiplicative normalization rule in contrast to the subtractive one obtained in the coordinate system $\mathcal{W}$ (see also Figure 6). It is listed as normalization rule $N^w_{\geq}$ in Table 1 (or $N^w_{=}$ for constraint $g(\mathbf{w}) = 0$). This multiplicative rule is commonly found in the literature (cf. Table 2), but it is not orthogonal in $\mathcal{W}$, though it is in $\mathcal{V}^w$.

For a more general coordinate transformation $w_i = w_i(v_i)$ and a constraint function $g(\mathbf{w})$, an orthogonal normalization rule can be derived in $\mathcal{V}$ with the method of Lagrangian multipliers and transformed back into $\mathcal{W}$, which results in general in a nonorthogonal normalization rule:

$$\text{if constraint is violated:} \qquad w_i = \tilde{w}_i + \lambda\left(\frac{\mathrm{d}w_i}{\mathrm{d}\tilde{v}_i}\right)^2 \frac{\partial g}{\partial \tilde{w}_i} + O(\lambda^2). \tag{5.6}$$

The λ actually would have to be calculated in $\mathcal{V}$, but since $\lambda \propto \Delta t$, second- and higher-order terms can be neglected for small Δt and λ calculated such that $g(\mathbf{w}) = 0$. Notice the similar effect of the coordinate transformation on the growth rules (see equation 4.8), as well as on the normalization rules (see equation 5.6). In both cases, a factor $(\mathrm{d}w_i/\mathrm{d}v_i)^2$ is added to the modification rate. As for gradient flows derived from objective functions, for a more general coordinate transformation $w_i = w_i(\mathbf{v})$, equation 5.6 would have to be modified accordingly.

We indicate these normalization rules by a subscript = (for an equality) and $\geq$ (for an inequality), because the constraints are enforced immediately and exactly.

5.2.2 *Integrated Normalization Without Objective Function.* Growth rule and explicit normalization rule as derived by the method of Lagrangian multipliers can be combined in one dynamical equation. As an example, consider the growth rule $\dot{w}_i = f_i$, that is, $\tilde{w}_i(t + \Delta t) = w_i(t) + \Delta t f_i(t)$, where f_i is an arbitrary function in $\mathbf{w}$ and can be interpreted as a fitness of a synapse. Together with the normalization rule $N^w_{\simeq}$ (see equation 5.5) and assuming $\sum_{j\in I} w_j(t) = \theta$, it follows that (von der Malsburg & Willshaw, 1981):

$$w_i(t + \Delta t) = \frac{\theta\,[w_i(t) + \Delta t f_i(t)]}{\sum_{j\in I}[w_j(t) + \Delta t f_j(t)]} \tag{5.7}$$

$$= w_i(t) + \Delta t f_i(t) - \Delta t \frac{w_i(t)}{\theta} \sum_{j\in I} f_j(t) + O(\Delta t^2) \tag{5.8}$$

$$\Longrightarrow \qquad \dot{w}_i(t) = f_i(t) - \frac{w_i(t)}{\theta} \sum_{j\in I} f_j(t), \tag{5.9}$$

and with $W(t) = \sum_{i\in I} w_i(t)$

$$\dot{W}(t) = \left(1 - \frac{W(t)}{\theta}\right) \sum_{j\in I} f_j(t), \tag{5.10}$$

which shows that $W = \theta$ is indeed a stable fixed point under the dynamics of equation 5.9. However, this is not always the case. The same growth rule combined with the subtractive normalization rule $N^1_{\simeq}$ (see equation 2.5) would yield a dynamics that provides only a neutrally stable fixed point for $W = \theta$. An additional term $(\theta - \sum_{j\in I} w_j(t))$ would have to be added to make the fixed point stable. This is the reason that this type of normalization rule is listed in Table 1 only for $\mathcal{C}^w$. We indicate these kinds of normalization rules by the subscript $\simeq$ because the dynamics smoothly approaches the constraint surface and will stay there exactly.

Notice that this method differs from the standard method of Lagrangian multipliers, which also yields a dynamics such that $\mathbf{w}$ remains on the constraint surface. The latter applies only to the dynamics at $g(\mathbf{w}) = 0$ and always produces neutrally stable fixed points because $\sum_i \dot{w}_i(t) \frac{\partial g}{\partial w_i} = 0$ is required by definition. If applied to a weight vector outside the constraint surface, the standard method of Lagrangian multipliers yields $g(\mathbf{w}) = \text{const} \neq 0$.

An advantage of this method is that it provides one dynamics for the growth rule as well as the normalization rule and that the constraint is enforced exactly. However, difficulties arise when interfering constraints are combined; that is, different constraints affect the same weights. This type of formulation is required for certain types of analyses (e.g., Häussler & von der Malsburg, 1983). A disadvantage is that in general there no longer exists an

objective function for the dynamics, though the growth term itself without the normalization term still has an objective function that is a Lyapunov function for the combined dynamics.

5.2.3 Penalty Terms. Another method of enforcing the constraints is to add penalty terms to the objective function (e.g., Bienenstock & von der Malsburg). For instance, if the constraint is formulated as an equality $g(\mathbf{w}) = 0$, then add $-\frac{1}{2}g^2(\mathbf{w})$; if the constraint is formulated as an inequality $g(\mathbf{w}) \leq 0$ or $g(\mathbf{w}) \geq 0$, then add $\ln|g(\mathbf{w})|$. Other penalty functions, such as g^4 and $1/g$, are possible as well, but those used here induce the required terms as used in the literature.

The effect of coordinate transformations is the same as in the case of objective functions. Consider, for example, the simple constraint $g_i(\mathbf{w}) = w_i \geq 0$ ($\mathrm{I}_{\geq}$ in Table 1), which keeps weights w_i nonnegative. The respective penalty term is $\ln|w_i|$ ($\mathrm{I}_{>}$) and the induced dynamics under the four different transformations considered in Table 1 are $\frac{1}{w_i}$, $\frac{\alpha_i}{w_i}$, 1, and α_i.

An advantage of this approach is that a coherent objective function, as well as a weight dynamics, is available, including growth rules and normalization rules. A disadvantage may be that the constraints are only approximate and not enforced strictly, so that $g(\mathbf{w}) \approx 0$ and $g(\mathbf{w}) < 0$ or $g(\mathbf{w}) > 0$. We therefore indicate these kinds of normalization rules by subscripts $\approx$ and $>$. However, the approximation can be made arbitrarily precise by weighting the penalty terms accordingly.

5.2.4 Constraints Introduced by Coordinate Transformations. An entire
different way by which constraints can be enforced is by means of a coo
nate transformation. Consider, for example, the coordinate transforma
C^w ($w_i = \frac{1}{4}v_i^2$). Negative weights are not reachable under this coor
transformation because the factor $(\mathrm{d}w_i/\mathrm{d}v_i)^2 = w_i$ added to the grow
(see equation 4.8) as well as to the normalization rules (see equatio
lows the weight dynamics of weight w_i to slow down as it approa
so that positive weights always stay positive (This can be generali
itive and negative weights by the coordinate transformation v
Thus the coordinate transformation C^w (and also $C^{\alpha w}$) implici
limitation constraint $\mathrm{I}_{>}$. This is interesting because it shows
nate transformation can substitute for a constraint, which
optimization theory.

The choice of whether to enforce the constraints by
tion, an integrated dynamics without an objective fun
or even implicitly a coordinate transformation depe
well as the methods applied to analyze it. Table 1 sh
functions and their corresponding normalization r
ent coordinate systems and by the three different
Not shown is normalization implicit in a coord

Table 1: Objective Functions, Constraint Functions, and the Dynamics Terms Induced in Different Coordinate Systems.

		Coordinate Transformations			
		$\mathcal{C}^1$	$\mathcal{C}^\alpha$	$\mathcal{C}^w$	$\mathcal{C}^{\alpha w}$
		$w_i = v_i$	$w_i = \sqrt{\alpha_i} v_i$	$w_i = \frac{1}{4} v_i^2$	$w_i = \frac{1}{4} \alpha_i v_i^2$
		$\left(\frac{dw_i}{dv_i}\right)^2 = 1$	$\left(\frac{dw_i}{dv_i}\right)^2 = \alpha_i$	$\left(\frac{dw_i}{dv_i}\right)^2 = w_i$	$\left(\frac{dw_i}{dv_i}\right)^2 = \alpha_i w_i$
	Objective Functions $H(\mathbf{w})$	Growth Terms: $\dot{w}_i = \cdots + \cdots$ or $\tilde{w}_i = w_i + \Delta t(\cdots + \cdots)$			
L	$\sum_i \beta_i w_i$	β_i	$\alpha_i \beta_i$	$\beta_i w_i$	$\alpha_i \beta_i w_i$
Q	$\frac{1}{2} \sum_{ij} w_i D_{ij} w_j$	$\sum_j D_{ij} w_j$	$\alpha_i \sum_j D_{ij} w_j$	$w_i \sum_j D_{ij} w_j$	$\alpha_i w_i \sum_j D_{ij} w_j$
	Constraint Functions $g(\mathbf{w})$	Normalization Rules (if constraint is violated): $w_i = \cdots \quad \forall i \in I_n$			
$I_=$, $I_\geq$	$\theta_i - w_i$	θ_i	θ_i	θ_i	θ_i
$N_=$, $N_\geq$	$\theta_n - \sum_{j \in I_n} \beta_j w_j$	$\tilde{w}_i + \lambda_n \beta_i$	$\tilde{w}_i + \lambda_n \alpha_i \beta_i$	$\tilde{w}_i + \lambda_n \beta_i \tilde{w}_i$	$\tilde{w}_i + \lambda_n \alpha_i \beta_i \tilde{w}_i$
$Z_=$, $Z_\geq$	$\theta_n - \sum_{j \in I_n} \beta_j w_j^2$	$\tilde{w}_i + \lambda_n \beta_i \tilde{w}_i$	$\tilde{w}_i + \lambda_n \alpha_i \beta_i \tilde{w}_i$	$\tilde{w}_i + \lambda_n \beta_i \tilde{w}_i^2$	$\tilde{w}_i + \lambda_n \alpha_i \beta_i \tilde{w}_i^2$
	Constraint Functions $g(\mathbf{w})$	Normalization Terms: $w_i = \cdots$ or $\tilde{w}_i = w_i + \Delta t(\cdots)$			
$N_\simeq$	$\theta_n - \sum_{j \in I_n} w_j$			$f_i - \frac{w_i}{\theta_n} \sum_j f_j$	
	Penalty Functions $H(\mathbf{w})$	Normalization Terms: $\dot{w}_i = \cdots + \cdots$ or $\tilde{w}_i = w_i + \Delta t(\cdots + \cdots)$			
$I_\approx$	$-\frac{1}{2} \gamma_i (\theta_i - w_i)^2$	$\gamma_i (\theta_i - w_i)$	$\alpha_i \gamma_i (\theta_i - w_i)$	$\gamma_i w_i (\theta_i - w_i)$	$\alpha_i \gamma_i w_i (\theta_i - w_i)$
$I_>$	$\gamma_i \ln \lvert \theta_i - w_i \rvert$	$-\frac{\gamma_i}{\theta_i - w_i}$	$-\frac{\alpha_i \gamma_i}{\theta_i - w_i}$	$-\frac{\gamma_i w_i}{\theta_i - w_i}$	$-\frac{\alpha_i \gamma_i w_i}{\theta_i - w_i}$
$N_\approx$	$-\frac{1}{2} \gamma_n (\theta_n - \sum_{j \in I_n} \beta_j w_j)^2$	$\beta_i \gamma_n \times$ $(\theta_n - \sum_j \beta_j w_j)$	$\alpha_i \beta_i \gamma_n \times$ $(\theta_n - \sum_j \beta_j w_j)$	$\beta_i \gamma_n w_i \times$ $(\theta_n - \sum_j \beta_j w_j)$	$\alpha_i \beta_i \gamma_n w_i \times$ $(\theta_n - \sum_j \beta_j w_j)$

Note $\mathcal{C}$ indicates a coordinate transformation that is specified by a superscript. L indicates a linear term. Q indicates a quadratic term that is usually induced by correlations $\langle a_\tau, a_\rho \rangle = \sum_{ij} D_{ij} w_j$. But it can also account for mean activities $\langle a_\tau \rangle = \sum_j A_{ij} w_j$. I indicates a limitation constraint that limits the range for individual weights (I may stand for "interval"). N indicates a normalization constraint that limits the sum over a set of weights. Z is a rarely used variation of N (the symbol Z can be thought of as a rotated N). Subscript signs distinguish between the different ways in which constraints can be enforced. $I^w_{\approx}$, for instance, indicates the normalization term $\gamma_i w_i (\theta_i - w_i)$ induced by the penalty function $-\frac{1}{2} \gamma_i (\theta_i - w_i)^2$ under the coordinate transformation $\mathcal{C}^w$. Subscripts n and i for θ, λ, and γ denote different constraints of the same type, for example, the same constraint applied to different output neurons. Normalization terms are integrated into the dynamics directly, while normalization rules are applied iteratively to the dynamics of the growth rule. f_i denotes a fitness by which a weight would grow without any normalization (cf. section 5.2.2).

interesting that there are only two types of constraints. All variations arise from using different coordinate systems and different methods by which the normalization rules are implemented. The first type is a limitation constraint I, which limits the range of individual weights. The second type is a normalization constraint N, which affects a group of weights, usually the sum, very rarely the sum of squares as indicated by Z. In the next section we show how to use Table 1 for analyzing models of neural map formation and give some examples from the literature.

6 Examples and Applications

6.1 How to Use Table 1. The aim of Table 1 is to provide an overview of the different objective functions and derived growth terms as well as the constraint functions and derived normalization rules and terms discussed in this article. The terms and rules are ordered in columns belonging to a particular coordinate transformation $\mathcal{C}$. Only entries in the same column may be combined to obtain a consistent, constrained optimization formulation for a system. However, some terms can be derived under different coordinate transformations. For instance, the normalization rule $\mathrm{I}_{=}$ is the same for all coordinate transformations, and term $\mathrm{L}^{\alpha w}$ with $\beta_i = 1/\alpha_i$ is the same as term L^{w} with $\beta_i = 1$.

To analyze a model of neural map formation, first identify possible candidates in Table 1 representing the different terms of the desired dynamics. Notice that the average activity of output neurons is represented by $\langle a_\tau \rangle = \sum_j A_{\tau j} w_j$ and that the correlation between input and output neurons is represented by $\langle a_\tau, a_\rho \rangle = \sum_j D_{\tau j} w_j$. Usually both terms will be only an approximation of the actual mean activities and correlations of the system under consideration (cf. section 2.1). Notice also that normalization rules $\mathrm{N}^{w}_{=}$, $\mathrm{N}^{\alpha w}_{=}$, $\mathrm{Z}^{1}_{=}$, and $\mathrm{Z}^{\alpha}_{=}$ are actually multiplicative normalization rules and not subtractive ones, as might be suggested by the special form in which they are written in Table 1.

Next identify the column in which all terms of the weight dynamics can be represented. This gives the coordinate transformation under which the model can be analyzed through the objective functions and constraint or penalty functions listed on the left side of the table. Equivalent models (cf. section 6.4) can be derived by moving from one column to another and by using normalization rules derived by a different method. Thus, Table 1 provides a convenient tool for checking whether a system can be analyzed within the constrained optimization framework presented here and for identifying the equivalent models. The function of each term can be coherently interpreted with respect to the objective, constraint, and penalty functions on the left side. The table can be extended with respect to additional objective, constraint, and penalty functions, as well as additional coordinate transformations. Although the table is compact, it suffices to

explain a wide range of representative examples from the literature, as discussed in the next section.

6.2 Examples from the Literature. Table 2 shows representative models from the literature. The original equations are listed, as well as the classification in terms of growth rules and normalization rules listed in Table 1. Detailed comments for these models and the model in Amari (1980) follow below. The latter is not listed in Table 2 because it cannot be interpreted within our constrained optimization framework. The dynamics of the introductory example of section 2 can be classified as Q^1 (see equation 2.3), $I^1_{\geq}$ (see equation 2.4), and $N^1_{\geq}$ (see equations 2.5 and 2.6).

The models are discussed here mainly with respect to whether they can be consistently described within the constrained optimization framework, that is, whether growth rules and normalization rules can be derived from objective functions and constraint functions under one coordinate transformation (that does not imply anything about the quality of a model). Another important issue is whether the linear correlation model introduced in section 3 is an appropriate description for the activity dynamics of these models. It is an accurate description for some of them, but others are based on nonlinear models, and the approximations discussed in section 2.1 and appendix A have to be made.

Models typically contain three components: the quadratic term Q to induce neighborhood-preserving maps, a limitation constraint I to keep synaptic weights positive, and a normalization constraint N (or Z) to induce competition between weights and to keep weights limited. The limitation constraint can be waived for systems with positive weights and multiplicative normalization rules (Konen & von der Malsburg, 1993; Obermayer et al., 1990; von der Malsburg, 1973) (cf. section 5.2.4). A presynaptic normalization rule can be introduced implicitly by the activity dynamics (cf. section A.2 in the appendix). In that case, it may be necessary to use an explicit presynaptic normalization constraint in the constrained optimization formulation. Otherwise the system may have a tendency to collapse on the input layer (see section 6.3), a tendency it does not have in the original formulation as a dynamical system. Only few systems contain the linear term L, which can be used for dynamic link matching. In Häussler and von der Malsburg (1983) the linear term was introduced for analytical convenience and does not differentiate between different links. The two models of dynamic link matching (Bienenstock & von der Malsburg, 1987; Konen & von der Malsburg, 1993) introduce similarity values implicitly and not through the linear term. The models are now discussed individually in chronological order.

von der Malsburg (1973): The activity dynamics of this model is nonlinear and based on hexagon patterns in the output layer. Thus, the applicability of the linear correlation model is not certain (cf. section 2.1). The weight

Table 2: Examples of Weight Dynamics from Previous Studies.

Reference	Weight Dynamics	Equation	Classification
von der Malsburg (1973)	$\tilde{w}_{\tau\rho} = w_{\tau\rho} + h a_\rho a_\tau$ $w_{\tau\rho} = \tilde{w}_{\tau\rho} \cdot 19 \cdot \frac{w}{2} / \tilde{w}_\tau, \quad \tilde{w}_\tau = \sum_{\rho=1}^{19} \tilde{w}_{\tau\rho}$		Q^1 $N^w_=$
Whitelaw and Cowan (1981)	$\dot{w}_{\tau\rho} = \alpha_{\tau\rho} a_\rho a_\tau - \alpha a_\tau + \Omega$ (Ω: small noise term) $\sum_{\rho'} w_{\tau\rho'} = 1, \sum_{\tau'} w_{\tau'\rho} = 1$	(2) (5)	$Q^\alpha - Q^1 + ?$ $N^?_=$
Häussler and von der Malsburg (1983)	$\dot{w}_{\tau\rho} = f_{\tau\rho} - \frac{1}{2N} w_{\tau\rho} \left(\sum_{\tau'} f_{\tau'\rho} + \sum_{\rho'} f_{\tau\rho'} \right)$ $f_{\tau\rho} = \alpha + \beta w_{\tau\rho} C_{\tau\rho}$ $C_{\tau\rho} = \sum_{\tau'\rho'} D_{\tau\tau'} D_{\rho\rho'} w_{\tau'\rho'}$	(2.1) (2.2) (2.3)	$(I^w_> + Q^w) - (L^w + N^w_\simeq)$
Linsker (1986)	$\dot{w}_{\tau\rho} = k_1 + \frac{1}{N_G} \sum_{\rho'} \left(Q^F_{\rho\rho'} + k_2 \right) w_{\tau\rho'}$ $+ R_b \sum_{\tau'} f_{\tau\tau'} \left[k_{1a} + \frac{1}{N_G} \sum_{\rho'} \left(Q^F_{\rho\rho'} + k_2 \right) w_{\tau'\rho'} \right]$ $= k'_1 - \frac{A_\rho - k_2}{N_G} \sum_{\tau'\rho'} D_{\tau\tau'} A_{\rho'} w_{\tau'\rho'} + \frac{1}{N_G} \sum_{\tau'\rho'} D_{\tau\tau'} D_{\rho\rho'} w_{\tau'\rho'}$ ($k'_1 = k_1 + R_b k_{1a} \sum_{\tau'} f_{\tau\tau'}$, $D_{\tau\tau'} = R_b f_{\tau\tau'} + \delta_{\tau\tau'}$ ($\delta_{\tau\tau'}$ Kronecker), $D_{\rho\rho'} = \langle a_\rho a_{\rho'} \rangle$, $A_\rho = \langle a_\rho \rangle$, $k_2 < 0$) some $w_{\tau\rho} \in [0, 1]$ and some $w_{\tau\rho} \in [-1, 0]$ or all $w_{\tau\rho} \in [-0.5, 0.5]$	(5)	$L^1 + Q^1$ $I^1_\geq$
Bienenstock and von der Malsburg (1987)	$H = -\sum_{\tau\tau'\rho\rho'} D_{\tau\tau'} w_{\tau'\rho'} w_{\tau\rho} D_{\rho\rho'}$ $+ \gamma' \sum_\tau \left(\sum_\rho w_{\tau\rho} - p' \right)^2 + \gamma' \sum_\rho \left(\sum_\tau w_{\tau\rho} - p' \right)^2$ $w_{\tau\rho} \in [0, T_{\tau\rho}]$	(2)	Q^1 $+ N^1_\approx$ $I^1_\geq$

Table 2: Continued.

Reference	Weight Dynamics	Equation	Classification
Miller, Keller, and Stryker, 1989	$\dot{w}^L_{\tau\rho} = \lambda\alpha_{\tau\rho}\sum_{\tau'\rho'} D_{\tau\tau'}\left[D^{LL}_{\rho\rho'}w^L_{\tau'\rho'} + D^{LR}_{\rho\rho'}w^R_{\tau'\rho'}\right] - \left[\gamma w^L_{\tau\rho} + \epsilon\alpha_{\tau\rho}\right]$ a) $\sum_{\rho'}(w^L_{\tau\rho'} + w^R_{\tau\rho'}) = 2\sum_{\rho'}\alpha_{\tau\rho'}, \quad w^L_{\tau\rho} = \tilde{w}^L_{\tau\rho} + \lambda_\tau\alpha_{\tau\rho}$ b) $\sum_{\tau'} w^L_{\tau'\rho} = \text{const}, \quad w^L_{\tau\rho} = \tilde{w}^L_{\tau\rho} + \lambda_\tau\alpha_{\tau\rho}$ $w^L_{\tau\rho} \in [0, 8\,\alpha_{\tau\rho}]$ (If weights were cut due to $I^\alpha_{\geq}$: $w^L_{\tau\rho} = \tilde{w}^L_{\tau\rho} + \lambda_\tau\tilde{w}^L_{\tau\rho}$) Interchanging L (left eye) and R (right eye) yields equations for $w^R_{\tau\rho}$.	(1) (Note 23)	Q^α - $I^\alpha_{\approx}$ $N^\alpha_=$ $N^\alpha_=$ $I^\alpha_{\geq}$ ($N^w_=$)
Obermayer et al. (1990)	$w_{\tau\rho}(t+1) = \frac{w_{\tau\rho}(t)+\epsilon(t)a_\tau(t)a_\rho}{\sqrt{\sum_{\rho'}\left(w_{\tau\rho'}(t)+\epsilon(t)a_\tau(t)a_{\rho'}\right)^2}}$	(4)	$\frac{Q^1}{Z^1_=}$
Tanaka (1990)	$\dot{w}_{\tau\rho} = w_{\tau\rho}\left[\kappa_0 - \kappa_1\sum_{\rho'}\beta_{\rho'}w_{\tau\rho'}\right] + gm_\tau w_{\tau\rho}a_\rho + \gamma_{\tau\rho}$ (later in the article $\beta_{\rho'} = 1$)	(2.1)	$N^{\alpha w}_{\approx} + Q^w + I^w_{>}$ ($N^{\alpha w}_{\approx} = N^w_{\approx}$)
Goodhill (1993)	$w_{\tau\rho} = w_{\tau\rho} + \alpha a_\rho a_\tau$ a) $w_{\tau\rho} = \begin{cases} w_{\tau\rho}-t & \text{if } w_{\tau\rho}-t>0 \\ 0 & \text{otherwise} \end{cases}$, $t = \frac{\sum_{\rho'} w_{\tau\rho'} - N_\tau}{n_\tau}$, $n_\tau = \sum_{\{\rho' \mid 0<w_{\tau\rho'}\}} 1$ (if some weights have become zero due to $I^1_{\geq}$: $w_{\tau\rho} = \frac{N_\tau w_{\tau\rho}}{\sum_{\rho'} w_{\tau\rho'}}$) b) $w_{\tau\rho} = \frac{N_\rho w_{\tau\rho}}{\sum_{\tau'} w_{\tau'\rho}}$		Q^1 $\begin{cases} N^1_= \\ I^1_{\geq} \end{cases}$ ($N^w_=$) $N^w_=$
Konen and von der Malsburg (1993)	$w_{\tau\rho} \to w_{\tau\rho} + \epsilon w_{\tau\rho}\alpha_{\tau\rho}a_\tau a_\rho$ $\to w_{\tau\rho}/\sum_{\rho'}\frac{w_{\tau\rho'}}{\alpha_{\tau\rho'}}$ $\to w_{\tau\rho}/\sum_{\tau'}\frac{w_{\tau'\rho}}{\alpha_{\tau'\rho}}$ ($w_{\tau\rho}$ are the "effective couplings" $J_{\tau\rho}T_{\tau\rho}$)	(3.5)	$Q^{\alpha w}$ $N^{\alpha w}_=$ $N^{\alpha w}_=$

Note: The original equations are written in a form that uses the notation of this article. The classification of the original equations by means of the terms and coordinate transformations listed in Table 1 are shown in the right column (the coordinate transformations are indicated by superscripts). See section 6.2 for further comments on these models.

dynamics is inconsistent in its original formulation. However, Miller and MacKay (1994) have shown that constraints $N^w_=$ and $Z^1_=$ have a very similar effect on the dynamics, so that the weight dynamics could be made consistent by using $Z^1_=$ instead of $N^w_=$. No limitation constraint is necessary because neither the growth rule nor the multiplicative normalization rule can lead to negative weights, and the normalization rule limits the growth of positive weights.

Amari (1980): This is a particularly interesting model not listed in Table 2. It is based on a blob dynamics, but no explicit normalization rules are applied, so that the derivation of correlations and mean activities as discussed in section 3 cannot be used. Weights are prevented from growing infinitely by a simple decay term, which is possible because correlations induced by the blob model are finite and do not grow with the total strength of the synapses. Additional inhibitory inputs received by the output neurons from a constantly active neuron ensure that the average activity is evenly distributed in the output layer, which also leads to expanding maps. In this respect, the architecture deviates from Figure 2. Thus, this model cannot be formulated within our framework.

Whitelaw and Cowan (1981): The activity dynamics is nonlinear and based on blobs. Thus, the linear correlation model is only an approximation (cf. section 2.1). The weight dynamics is difficult to interpret in the constrained optimization framework. The normalization rule is not specified precisely, but it is probably multiplicative because a subtractive one would lead to negative weights and possibly infinite weight growth. The quadratic term $-Q^1$ is based on mean activities and would lead by itself to zero weights. The Ω term was introduced only to test the stability of the system.

Häussler and von der Malsburg (1983): This model is directly formulated in terms of weight dynamics; thus, the linear correlation model is accurate. The weight dynamics is consistent; however, as argued in section 5.2.2, there is usually no objective function for the normalization rule $N^w_\simeq$, but by replacing $N^w_\simeq$ by $N^w_=$ or $N^w_\approx$, the system can be expressed as a constrained optimization problem without qualitatively changing the model behavior. The limitation term $I^w_>$ and the linear term L^w are induced by the constant α and were introduced for analytical reasons. The former is meant to allow weights to grow from zero strength, and the latter limits this growth. α needs to be small for neural map formation, and for a stable one-to-one mapping, α strictly should be zero. Thus, these two terms could be discarded if all weights would be initially larger than zero. Notice that the linear term does not differentiate between different links and thus does not have a function as suggested for dynamic link matching (cf. sections 4 and 6.5).

Linsker (1986): This model is also directly formulated in terms of weight dynamics; thus, the linear correlation model is accurate. The weight dynamics is consistent. Since the model uses negative and positive weights

and weights have a lower and an upper bound, no normalization rule is necessary. The weights converge to their upper or lower limit.

Bienenstock and von der Malsburg (1987): This is a model of dynamic link matching and was originally formulated in terms of an energy function. Thus the classification is accurate. The energy function does not include the linear term. The features are binary, black versus white, and the similarity values are therefore 0 and 1 and do not enter the dynamics as continuous similarity values. The $T_{\tau\rho}$ in the constraint $\mathrm{I}^{1}_{\geq}$ represent the stored patterns in the associative memory, not similarity values.

Miller et al. (1989): This model is directly formulated in terms of weight dynamics; thus, the linear correlation model is accurate. One inconsistent part in the weight dynamics is the multiplicative normalization rule $\mathrm{N}^{w}_{=}$, which is applied when subtractive normalization leads to negative weights. But it is only an algorithmic shortcut to solve the problem of interfering constraints (limitation and subtractive normalization). A more systematic treatment of the normalization rules could replace this inconsistent rule (cf. section 5.2.1). Another inconsistency is that weights that reach their upper or lower limit become frozen, or fixed at the limit value. With some exception, this seems to have little effect on the resulting maps (Miller et al., 1989, n. 23). Thus, this model has only two minor inconsistencies, which could be modified to make the system consistent. Limitation constraints enter the weight dynamics in two forms, $\mathrm{I}^{\alpha}_{\approx}$ and $\mathrm{I}^{\alpha}_{\geq}$. The former tends to keep $w^{L}_{\tau\rho} = -\frac{\epsilon}{\gamma}\alpha_{\tau\rho}$ while the latter keeps $w^{L}_{\tau\rho} \in [0, 8\alpha_{\tau\rho}]$, which can unnecessarily introduce conflicts. However, $\gamma = \epsilon = 0$, so that only the latter constraint applies and the $\mathrm{I}^{\alpha}_{\approx}$ term is discarded in later publications. In principle, the system can be simplified by using coordinate transformation $\mathcal{C}^{1}$ instead of $\mathcal{C}^{\alpha}$, thereby eliminating $\alpha_{\tau\rho}$ in the growth rule Q^{α} as well as in the normalization rule $\mathrm{N}^{\alpha}_{=}$, but not in the normalization rule $\mathrm{I}^{\alpha}_{\geq}$. This is different from setting $\alpha_{\tau\rho}$ to a constant in a certain region. Using coordinate transformation $\mathcal{C}^{1}$ would result in the same set of stable solutions, though the trajectories would differ. Changing $\alpha_{\tau\rho}$ generates a different set of solutions. However, the original formulation using $\mathcal{C}^{\alpha}$ is more intuitive and generates the "correct" trajectories—those that correspond to the intuitive interpretation of the model.

Obermayer et al. (1990): This model is based on an algorithmic blob model and the linear correlation model is only an approximation (cf. the appendix). The weight dynamics is consistent. It employs the rarely used normalization constraint Z, which induces a multiplicative normalization rule under the coordinate transformation $\mathcal{C}^{1}$. No limitation constraint is necessary because neither the growth rule nor the multiplicative normalization rule can lead to negative weights, and positive weights are limited by the normalization rule.

Tanaka (1990): This model uses a nonlinear input-output function for the neurons, which makes a clear distinction between membrane potential and

firing rate. However, this nonlinearity does not seem to play a specific functional role and is partially eliminated by linear approximations. Thus, the linear correlation model seems to be justified. The weight dynamics includes parameters $\beta_{\rho'}$ (f_{SP} in the original notation), which make it inconsistent. The penalty term $N_{\approx}^{\alpha w}$, which induces the first terms of the weight dynamics, is $-\frac{1}{2\kappa_1}\sum_{\tau'}(\kappa_0 - \kappa_1 \sum_{\rho'} \beta_{\rho'} w_{\tau'\rho'})^2$, which has to be evaluated under the coordinate transformation $\mathcal{C}^{\alpha w}$ with $\alpha_{\tau\rho} = 1/\beta_\rho$. Later in the article, the parameters $\beta_{\rho'}$ are set to 1, so that the system becomes consistent. Tanaka gives an objective function for the dynamics, employing a coordinate transformation for this purpose. The objective function is not listed here because it is derived under a different set of assumptions, including the nonlinear input-output function of the output neurons and a mean field approximation.

Goodhill (1993): This model is based on an algorithmic blob model and the linear correlation model is only an approximation (cf. the appendix). Like the model in Miller et al. (1989), this model uses an inconsistent normalization rule as a backup, and it freezes weights that reach their upper or lower limit. In addition, it uses an inconsistent normalization rule for the input neurons. But since this inconsistent multiplicative normalization for the input neurons is applied after a consistent subtractive normalization for the output neurons, its effect is relatively weak, and substituting it by a subtractive one would make little difference (G. J. Goodhill, personal communication). To avoid dead units (neurons in the output layer that never become active), Goodhill (1993) divides each output activity by the number of times each output neuron has won the competition for the blob in the output layer. This guarantees a roughly equal average activity of the output neurons. With the probabilistic blob model (cf. the appendix), dead units do not occur as long as output neurons have any input connections. The specific parameter setting of the model even guarantees a roughly equal average activity of the output neurons under the probabilistic blob model because the sum over the weights converging on an output neuron is roughly the same for all neurons in the output layer. Thus, despite some inconsistencies, this model can probably be well approximated within the constrained optimization framework.

Konen and von der Malsburg (1993): The activity dynamics is nonlinear and based on blobs. Thus the linear correlation model is only an approximation (cf. section 2.1). The weight dynamics is consistent. Although this is a model of dynamic link matching, it does not contain the linear term to bias the links. It introduces the similarity values in the constraints and through the coordinate transformation $\mathcal{C}^{\alpha w}$ (see section 6.4). No limitation constraint is necessary because neither the growth rule nor the multiplicative normalization rule can lead to negative weights, and positive weights are limited by the normalization rule.

6.3 Some Functional Aspects of Term Q. So far the focus of the considerations has been only on formal aspects of models of neural map formation.

In this section some remarks on functional aspects of the quadratic term Q are made.

Assume the effective lateral connectivities in the output layer, and in the input layer are sums of positive and/or negative contributions. Each contribution can be either a constant, C, or a centered gaussian-like function, G, which depends on only the distance of the neurons, for example, $D_{\rho\rho'} = D_{|\rho-\rho'|}$ if ρ is a spatial coordinate. The contributions can be indicated by subscripts to the objective function Q. First index indicates the lateral connectivity of the input layer, the second index the one of the output layer. A negative gaussian (constant) would have to be indicated by $-G$ ($-C$). $Q_{(-C)G}$, for instance, would indicate a negative constant $D_{\rho\rho'}$ and a positive gaussian $D_{\tau\tau'}$. $Q_{G(G-G')}$ would indicate a positive gaussian $D_{\rho\rho'}$ and a $D_{\tau\tau'}$ that is a difference of gaussians. Notice that negative signs can cancel each other, for example $Q_{(G-C)G} = -Q_{(C-G)G} = -Q_{(G-C)(-G)}$. We thus discuss the terms only in their simplest form: $-Q_{CG}$ instead of $Q_{(-C)G}$. All feedforward weights are assumed to be positive. Assuming all weights to be negative would lead to equivalent results because Q does not change if all weights change their sign. The situation becomes more complex if some weights were positive and others negative. A term Q is called positive if it can be written in a form where it has a positive sign and only positive contributions; for example, $-Q_{(-C)G} = Q_{CG}$ is positive, while $Q_{(G-C)G}$ is not. Since Q is symmetrical with respect to $D_{\rho\rho'}$ and $D_{\tau\tau'}$, a term such as $Q_{(G-C)G}$ has the same effect as $Q_{G(G-C)}$ with the role of input layer and output layer exchanged. A complicated term can be analyzed most easily by splitting it into its elementary components. For instance, the term $Q_{G(G-C)}$ can be split into $Q_{GG}-Q_{GC}$ and analyzed as a combination of these two simpler terms.

Some elementary terms are now discussed in greater detail. The effect of the terms is considered under two types of constraints. In constraint A, the total sum of weights is constrained, $\sum_{\rho'\tau'} w_{\rho'\tau'} = 1$. In constraint B, the sums of weights originating from an input neuron, $\sum_{\tau'} w_{\rho\tau'} = 1/R$, or terminating on an output neuron, $\sum_{\rho'} w_{\rho'\tau} = 1/T$, are constrained, where R and T denote the number of input and output neurons, respectively. Without further constraints, a positive term always leads to infinite weight growth and a negative term to weight decay.

Terms $\pm Q_{CC}$ simplify to $\pm Q_{CC} = \pm D_{\rho\rho} D_{\tau\tau} (\sum_{\rho'\tau'} w_{\rho'\tau'})^2$ and depend on only the sum of weights. Thus, neither term has any effect under constraints A or B.

Term $+Q_{CG}$ takes its maximum value under constraint A if all links terminate on one output neuron. The map has the tendency to collapse. This is because the lateral connections in the output layer are higher for smaller distances and maximal for zero distance between connected neurons. Under the constraint $\sum_{\tau'} w_{\rho\tau'} \leq 1, \sum_{\rho'} w_{\rho'\tau} \leq 1$, for instance, the resulting map connects the input layer to a region in the output layer that is of the size of the input layer even if the output layer is much larger. No topography is taken

into account because $D_{\rho\rho'}$ is constant and does not differentiate between different input neurons. Thus, this term has no effect under constraint B.

Term $-Q_{CG}$ has the opposite effect of $+Q_{CG}$. Consider the induced growth term $\dot{w}_{\rho\tau} = -D_{\rho\rho} \sum_{\tau'} D_{\tau\tau'} \sum_{\rho'} w_{\tau'\rho'}$. This is a convolution of $D_{\tau\tau'}$ with $\sum_{\rho'} w_{\tau'\rho'}$ and induces the largest decay in regions where the weighted sum over terminating links is maximal. A stable solution would require equal decay for all weights because constraint A can compensate only for equal decay. Thus, the convolution of $D_{\tau\tau'}$ with $\sum_{\rho'} w_{\tau'\rho'}$ must be a constant. Since $D_{\tau\tau'}$ is a gaussian, this is possible only if $\sum_{\rho'} w_{\tau'\rho'}$ is a constant, as can be easily seen in Fourier space. Thus, the map expands over the output layer, and each output neuron receives the same sum of weights. Constraint A could be substituted by a constant growth term L, in which case the expansion effect could be obtained without any explicit constraint. As $+Q_{CG}$, this term has no effect under constraint B.

Term $+Q_{GG}$ takes its maximum value under constraint A if all but one weight are zero. The map collapses on the input and the output layer. Under constraint B, the map becomes topographic because links that originate from neighboring neurons (high $D_{\rho\rho'}$ value) favorably terminate on neighboring neurons (high $D_{\tau\tau'}$ value). A more rigorous argument would require a definition of topography, but as argued in section 6.7, the term $+Q_{GG}$ can be directly taken as a generalized measure for topography.

Term $-Q_{GG}$ has the opposite effect of $+Q_{GG}$. Thus, it leads under constraint A to a map that is expanded over input and output layer. In addition, the map becomes antitopographic. Further analytical or numerical investigations are required to show whether the expansion is as even as for the term $-Q_{CG}$ and how an antitopographic map may look. Constraint B also leads to an antitopographic map.

6.4 Equivalent Models. The effect of coordinate transformations has been considered so far only for single growth terms and normalization rules. Coordinate transformations can be used to generate different models that are equivalent in terms of their constrained optimization problem. Consider the system in Konen and von der Malsburg (1993). Its objective function and constraint function are Q and $N_{\geq}$,

$$H(\mathbf{w}) = \frac{1}{2} \sum_{ij} w_i D_{ij} w_j, \qquad g_n(\mathbf{w}) = 1 - \sum_{j \in I_n} \frac{w_j}{\alpha_j} = 0, \tag{6.1}$$

which must be evaluated under the coordinate transformation $C^{\alpha w}$ to induce the original weight dynamics $Q^{\alpha w}$ and $N_{\geq}^{\alpha w}$,

$$\dot{w}_i = \alpha_i w_i \sum_j D_{ij} w_j, \qquad w_i = \frac{\tilde{w}_i}{\sum_{j \in I_n} \frac{\tilde{w}_j}{\alpha_j}}. \tag{6.2}$$

If evaluated directly (i.e., under the coordinate transformation C^1), one would obtain

$$\dot{w}_i = \sum_j D_{ij} w_j, \qquad w_i = \tilde{w}_i + \frac{1}{\sum_{j \in I_n} \alpha_j^{-2}} \left(1 - \sum_{j \in I_n} \frac{\tilde{w}_j}{\alpha_j} \right) \frac{1}{\alpha_i}. \tag{6.3}$$

As argued in section 5.2.4, an additional limitation constraint $I^1_>$ (or $I^1_\geq$) has to be added to this system to account for the limitation constraint implicitly introduced by the coordinate transformation $C^{\alpha w}$ for the dynamics above (see equation 6.2).

It follows from equation 4.8 that the flow fields of the weight dynamics in equations 6.2 and 6.3 differ, but since $dw_i/dv_i \neq 0$ for positive weights, the fixed points are the same. That means that the resulting maps to which the two systems converge, possibly from different initial states, are the same. In this sense, these two dynamics are equivalent.

This also holds for other coordinate transformations within the defined region as long as dw_i/dv_i is finite ($dw_i/dv_i = 0$ may introduce additional fixed points). Thus, this method of generating equivalent models makes it possible to abstract the objective function from the dynamics. Different equivalent dynamics may have different convergence properties, their attractor basins may differ, and some regions in state space may not be reachable under a particular coordinate transformation. In any case, within the reachable state space, the fixed points are the same. Thus, coordinate transformations make it possible to optimize the dynamics without changing its objective function.

Normalization rules derived by different methods can substitute each other without changing the qualitative behavior of a system. For instance, $I_=$ can be replaced by $I_\approx$, or $N_\geq$ can be replaced by $N_>$ under any coordinate transformation. These replacements will also generate equivalent systems in a practical sense.

6.5 Dynamic Link Matching. In the previous section, the similarity values α_i entered the weight dynamics in two places. In equation 6.2, the differential effect of α_i enters only the growth rule, while in equation 6.3, it enters only the normalization rule. Growth and normalization rules can, to some extent, be interchangeably used to incorporate feature information in dynamic link matching. However, the objective function (see equation 6.1) shows that the similarity values are introduced through the constraints and that they are transferred to the growth rule only by the coordinate transformation $C^{\alpha w}$. Similarity values can enter the growth rule more directly through the linear term L. An alternative objective function for dynamic

link matching is

$$H(\mathbf{w}) = \sum_i \beta_i w_i + \frac{1}{2} \sum_{ij} w_i D_{ij} w_j, \qquad g_n(\mathbf{w}) = 1 - \sum_{j \in I_n} w_j = 0, \qquad (6.4)$$

with $\beta_i = \alpha_i$. The first term now directly favors links with high similarity values. This may be advantageous because it allows better control over the influence of the topography versus the feature similarity term. Furthermore, this objective function is more closely related to the similarity function of elastic graph matching in Lades et al. (1993), which has been developed as an algorithmic abstraction of dynamic link matching (see section 6.7).

6.6 Soft versus Hard Competitive Normalization. Miller and MacKay (1994) have analyzed the role of normalization rules for neural map formation. They consider a linear Hebbian growth rule Q^1 and investigate the dynamics under a subtractive normalization rule $N^1_=$ (S1 in their notation) and two types of multiplicative normalization rules, $N^w_=$ and $Z^1_=$ (M1 and M2 in their notation, respectively). They show that when considering an isolated output neuron with the multiplicative normalization rules, the weight vector tends to the principal eigenvector of the matrix D, which means that many weights can maintain some finite value. Under the subtractive normalization rule, a winner-take-all behavior occurs, and the weight vector tends to saturate with each single weight having either its minimal or maximal value producing a more compact receptive field. If no upper bound is imposed on individual weights, only one weight survives, corresponding to a point receptive field.

von der Malsburg and Willshaw (1981) have performed a similar, though less comprehensive, analysis using a different approach. Instead of modifying the normalization rule, they considered different growth rules with the same multiplicative normalization rule $N^w_\simeq$. They also found two qualitatively different behaviors: a highly competitive case in which only one link survives (or several if single weights are limited in growth by individual bounds) (case μ=1 or μ=2 in their notation) and a less competitive case in which each weight is eventually proportional to the correlation between pre- and postsynaptic neuron (case μ=0).

Hence, one can either change the normalization rule and keep the growth rule or, vice versa, modify the growth rule and keep the normalization rule the same. Either choice generates the two different behaviors. As shown above, by changing both the growth and normalization rules consistently by a coordinate transformation, it is possible to obtain two different weight dynamics with qualitatively the same behavior. More precisely, the system (Q^w, N^w) is equivalent to (Q^1, N^1, I^1) and has the same fixed points; the former one uses a multiplicative normalization rule, and the latter uses a subtractive one. This also explains why changing the growth rule or changing the normalization rule can be equivalent.

It may therefore be misleading to refer to the different cases by the specific normalization rules (subtractive versus multiplicative), because that is valid only for the linear Hebbian growth rule Q^1. We suggest using a more generally applicable nomenclature that refers to the different behaviors rather than the specific mathematical formulation. Following the terminology of Nowlan (1990) in a similar context, the term *hard competitive* normalization could be used to denote the case where only one link survives (or a set of saturated links, which are limited by upper bounds); the term *soft competitive* normalization could be used to denote the case where each link has some strength proportional to its fitness.

6.7 Related Objective Functions. Objective functions also provide means for comparing weight dynamics with other algorithms or dynamics of a different origin for which an objective function exists.

First, maximizing the objective functions L and Q under linear constraints I and N is the quadratic programming problem, and finding an optimal one-to-one mapping between two layers of same size for objective function Q is the quadratic assignment problem. These problems are known to be NP-complete. However, there is a large literature on algorithms that efficiently solve special cases or find good approximate solutions in polynomial time (e.g., Horst, Pandalos, & Thoai, 1995).

Many related objective functions are defined only for maps for which each input neuron terminates on exactly one output neuron with weight 1, which makes the index $\tau = \tau(\rho)$ a function of index ρ. An objective function of this kind may have the form

$$H = \sum_{\rho\rho'} G_{\tau\rho\tau'\rho'}, \tag{6.5}$$

where G encodes how well a pair of links from ρ to $\tau(\rho)$ and from ρ' to $\tau'(\rho')$ preserves topography. A pair of parallel links, for instance, would yield high G values, while others would yield lower values. Now define a particular family of weights $\mathbf{w}$ that realize one-to-one connectivities:

$$\bar{w}_{\tau\rho} = \begin{cases} 1 & \text{if } \tau = \tau(\rho) \\ 0 & \text{otherwise.} \end{cases} \tag{6.6}$$

$\bar{\mathbf{w}}$ is a subset of $\mathbf{w}$ with $\bar{w}_{\tau\rho} \in \{0, 1\}$ as opposed to $w_{\tau\rho} \in [0, 1]$. It indicates that an objective function was originally defined for a one-to-one map rather than the more general case of an all-to-all connectivity. Then objective functions of one-to-one maps can be written as

$$H(\bar{\mathbf{w}}) = \sum_{\tau\rho\tau'\rho'} \bar{w}_{\tau\rho} G_{\tau\rho\tau'\rho'} \bar{w}_{\tau'\rho'} = \sum_{ij} \bar{w}_i G_{ij} \bar{w}_j, \tag{6.7}$$

with $i = \{\rho, \tau\}, j = \{\rho', \tau'\}$ as defined above. Simply replacing $\bar{\mathbf{w}}$ by $\mathbf{w}$ then yields a generalization of the original objective function to all-to-all connectivities.

Goodhill, Finch, and Sejnowski (1996) have compared 10 different objective functions for topographic maps and have proposed another, the C measure. They show that for the case of an equal number of neurons in the input and the output layer, most other objective functions can be either reduced to the C measure, or they represent a closely related objective function. This suggests that the C measure is a good unifying measure for topography. The C measure is equivalent to our objective function Q with $\bar{\mathbf{w}}$ instead of $\mathbf{w}$. Adapted to the notation of this article the C measure has the form

$$C(\bar{\mathbf{w}}) = \sum_{ij} \bar{w}_i G_{ij} \bar{w}_j, \tag{6.8}$$

with a separable G_{ij}, that is, $G_{ij} = G_{\rho\tau\rho'\tau'} = G_{\tau\tau'} G_{\rho\rho'}$. Thus, the objective function Q is the typical term for topographic maps in other contexts as well.

Elastic graph matching is an algorithmic counterpart to dynamic link matching and has been used for applications such as object and face recognition (Lades et al., 1993). It is based on a similarity function that in its simplest version is

$$H(\bar{\mathbf{w}}) = \sum_i \beta_i \bar{w}_i + \frac{1}{2} \sum_{ij} \bar{w}_i G_{ij} \bar{w}_j, \tag{6.9}$$

where $G_{ij} = -[(\mathbf{p}_\rho - \mathbf{p}_{\rho'}) - (\mathbf{p}_\tau - \mathbf{p}_{\tau'})]^2$, and $\mathbf{p}_\rho$ and $\mathbf{p}_\tau$ are two-dimensional position vectors in the image plane. This similarity function corresponds formally to the objective function in equation 6.4. The main difference between these two functions is hidden in G and D. The latter ought to be separable into two factors $D_{\rho\tau\rho'\tau'} = D_{\rho\rho'} D_{\tau\tau'}$ while the former is clearly not. G actually favors a metric map, which tends to preserve not only neighborhood relations but also distances, whereas with D, the maps always tend to collapse.

6.8 Self-Organizing Map Algorithm. Models of the self-organizing map (SOM) algorithm can be high-dimensional or low-dimensional, and two different learning rules, which we have called weight dynamics, are commonly used. The validity of the probabilistic blob model for the high-dimensional models is discussed in the appendix. A classification of the high-dimensional model by Obermayer et al. (1990) is given in Table 2. The low-dimensional models do not fall into the class of one-to-one mappings considered in the previous section, because the input layer is represented as a continuous space and not as a discrete set of neurons.

One learning rule for the high-dimensional SOM algorithm is given by

$$\tilde{w}_{\tau\rho}(t) = w_{\tau\rho}(t-1) + \epsilon B_{\tau\tau_0} B_{\rho\rho_0} \tag{6.10}$$

$$w_{\tau\rho}(t) = \frac{\tilde{w}_{\tau\rho}(t)}{\sqrt{\sum_{\rho'} \tilde{w}^2_{\tau\rho'}(t)}}, \tag{6.11}$$

as used, for example, in Obermayer et al. (1990). $B_{\tau\tau_0}$ denotes the neighborhood function (commonly indicated by h) and $B_{\rho\rho_0}$ denotes the stimulus pattern (sometimes indicated by x) with index ρ_0. $B_{\rho\rho_0}$ does not need to have a blob shape, so that ρ_0 may be an arbitrary index. Output neuron τ_0 is the winner neuron in response to stimulus pattern ρ_0. This learning rule is a consistent combination of growth rule Q^1 and normalization rule $\mathrm{Z}^1_=$ and an objective function exists, which is a good approximation to the extent that the probabilistic blob model is valid.

The second type of learning rule is given by

$$w_{\tau\rho}(t+1) = w_{\tau\rho}(t) + \epsilon B_{\tau\tau_0}(B_{\rho\rho_0} - w_{\tau\rho}(t)), \tag{6.12}$$

as used, for example, in Bauer, Brockmann, and Geisel (1997). For this learning rule, the weights and the input stimuli are assumed to be sum normalized: $\sum_\rho w_{\tau\rho} = 1$ and $\sum_\rho B_{\rho\rho_0} = 1$. For small ϵ this learning rule is equivalent to

$$\tilde{w}_{\tau\rho}(t) = w_{\tau\rho}(t-1) + \epsilon B_{\tau\tau_0} B_{\rho\rho_0} \tag{6.13}$$

$$w_{\tau\rho}(t) = \frac{\tilde{w}_{\tau\rho}(t)}{\sum_{\rho'} \tilde{w}_{\tau\rho'}(t)}, \tag{6.14}$$

which shows that it is a combination of growth rule Q^1 and normalization rule $\mathrm{N}^w_=$. Thus, this system is inconsistent, and to formulate it within our constrained optimization framework $\mathrm{N}^w_=$ would have to be approximated by $\mathrm{Z}^1_=$, which leads back to the learning rule in equations 6.10 and 6.11.

There are two ways of going from these high-dimensional models to the low-dimensional models. The first is simply to use fewer input neurons (e.g., two). A low-dimensional input vector is then represented by the activities of these few neurons. However, since the low-dimensional input vectors are usually not normalized to homogeneous mean activity of the input neurons and since the receptive and projective fields of the neurons do not codevelop in a homogeneous way, the probabilistic blob model is usually not valid.

A second way of going from a high-dimensional model to a low-dimensional model is by considering the low-dimensional input vectors and weight vectors as abstract representatives of the high-dimensional ones (Ritter, Martinetz, & Schulten, 1991; Behrmann, 1993). Consider, for example, the weight dynamics in equation 6.12 and a two-dimensional input layer. Let $\mathbf{p}_\rho$ be a

position vector of input neuron ρ. The center of the receptive field of neuron τ can be defined as

$$\mathbf{m}_\tau(\mathbf{w}) = \sum_\rho \mathbf{p}_\rho w_{\tau\rho}, \tag{6.15}$$

and the center of the input blob can be defined similarly,

$$\mathbf{x}(\mathbf{B}_{\rho_0}) = \sum_\rho \mathbf{p}_\rho B_{\rho\rho_0}. \tag{6.16}$$

Notice that the input blobs as well as the weights are normalized, that is, $\sum_\rho B_{\rho\rho_0} = 1$ and $\sum_\rho w_{\tau\rho} = 1$. Using these definitions and given a pair of blobs at locations ρ_0 and τ_0, the high-dimensional learning rule (see equation 6.12) yields the low-dimensional learning rule

$$\begin{aligned}
\mathbf{m}_\tau(\mathbf{w}(t+1)) &= \sum_\rho \mathbf{p}_\rho \left(w_{\tau\rho}(t) + \epsilon B_{\tau\tau_0}(B_{\rho\rho_0} - w_{\tau\rho}(t))\right) && (6.17)\\
&= \mathbf{m}_\tau(\mathbf{w}(t)) + \epsilon B_{\tau\tau_0}\left(\mathbf{x}(\mathbf{B}_{\rho_0}) - \mathbf{m}_\tau(\mathbf{w}(t))\right) && (6.18)\\
\iff \quad \mathbf{m}_\tau(t+1) &= \mathbf{m}_\tau(t) + \epsilon B_{\tau\tau_0}\left(\mathbf{x}_{\rho_0} - \mathbf{m}_\tau(t)\right). && (6.19)
\end{aligned}$$

One can first calculate the centers of the receptive fields of the high-dimensional model and then apply the low-dimensional learning rule, or one can first apply the high-dimensional learning rule and then calculate the centers of the receptive fields; the result is the same. Notice that the low-dimensional learning rule is even formally equivalent to the high-dimensional one and that it is the rule commonly used in low-dimensional models (Kohonen, 1990). Even though the high- and the low-dimensional learning rules are equivalent for a given pair of blobs, the overall behavior of the models is not. This is because the positioning of the output blobs is different in the two models (Behrmann, 1993). It is clear that many different high-dimensional weight configurations having different output blob positioning can lead to the same low-dimensional weight configuration. However, for a high-dimensional model that self-organizes a topographic map with point receptive fields, the positioning may be similar for the high- and the low-dimensional models, so that the stable maps may be similar as well.

These considerations show that only the high-dimensional model in equations 6.10 and 6.11 can be consistently described within our constrained optimization framework. The high-dimensional model of equation 6.12 is inconsistent. The probabilistic blob model in general is not applicable to low-dimensional models, because some assumptions required for its derivation are not valid. The simple relation between the high- and the low-dimensional model sketched above holds only for the learning step but not for the blob positioning, though the positioning and thus the resulting maps may be very similar for topographic maps with point receptive fields.

7 Conclusions and Future Perspectives

The results presented here can be summarized:

- A probabilistic nonlinear blob model can behave like a linear correlation model under fairly general conditions (see section 2.1 and the appendix). This clarifies the relationship between deterministic nonlinear blob models and linear correlation models and provides an approximation of the former by the latter.
- Coordinate transformations can transform dynamics with curl into curl-free dynamics, allowing the otherwise impossible formulation of an objective function (see section 4). A similar effect exists for normalization rules. Coordinate transformations can transform nonorthogonal normalization rules into orthogonal ones, allowing the normalization rule to be formulated as a constraint (see section 5.1).
- Growth rules and normalization rules must have a special relationship in order to make a formulation of the system dynamics as a constrained optimization problem possible: the growth rule must be a gradient flow, and the normalization rules must be orthogonal under the same coordinate transformation (see section 5.1).
- Constraints can be enforced by various types of normalization rules (see section 5.2), and they can even be implicitly introduced by coordinate transformations (see section 5.2.4) or the activity dynamics (see section A.2).
- Many all-to-all connected models from the literature can be classified within our constrained optimization framework based on only four terms: L, Q, I, and N (Z) (see section 6.2). The linear term L has rarely been used, but it can have a specific function that may be useful in future models (see section 6.5).
- Models may differ considerably in their weight dynamics and still solve the same optimization problem. This can be revealed by coordinate transformations and by comparing the different but possibly equivalent types of normalization rules (see section 6.4). Coordinate transformations make it in particular possible to optimize the dynamics without changing the stable fixed points.
- The constrained optimization framework provides a convenient formalism to analyze functional aspects of the models (see sections 6.3, 6.5, and 6.6).
- The constrained optimization framework for all-to-all connected models presented here is closely related to approaches for finding optimal one-to-one maps (see section 6.7) but is not easily adapted to the self-organizing map algorithm (see section 6.8).

- Models of neural map formation formulated as constrained optimization problems provide a unifying framework. It abstracts from arbitrary differences in the design of models and leaves only those differences that are likely to be crucial for the different structures that emerge by self-organization.

It is important to note that our constrained optimization framework is unifying in the sense that it provides a canonical formulation independent of most arbitrary design decisions, for example, due to different coordinate transformations or different types of normalization rules. This does not mean that most models are actually equivalent. But with the canonical formulation of the models as constrained optimization problems, it should be possible to focus on the crucial differences and to understand better what the essentials of neural map formation are.

Based on the constrained optimization framework presented here, a next step would be to consider specific architectures with particular effective lateral connectivities and to investigate the structures that emerge. The role of parameters and effective lateral connectivities might be investigated analytically for a variety of models by means of objective functions, similar to the approach sketched in section 6.3 or the one taken in MacKay and Miller (1990).

We have considered here only three levels of abstraction: detailed neural dynamics, abstract weight dynamics, and constrained optimization. There are even higher levels of abstraction, and the relationship between our constrained optimization framework and these more abstract models should be explored. For example, in section 6.7 our objective functions were compared with other objective functions defined only for one-to-one connectivities. Another possible link is with Bienenstock and von der Malsburg (1987) and Tanaka (1990), who have proposed spin models for neural map formation. An interesting approach is that taken by Linsker (1986), who analyzed the receptive fields of the output neurons, which were oriented edge filters of arbitrary orientation. He derived an energy function to evaluate how the different orientations would be arranged in the output layer due to lateral interactions. The only variables of this energy function were the orientations of the receptive fields, an abstraction from the connectivity. Similar models were proposed earlier in Swindale (1980), though not derived from a receptive field model, and more recently in Tanaka (1991). These approaches and their relationships to our constrained optimization framework need to be investigated more systematically.

A neural map formation model of Amari (1980) could not be formulated within the constrained optimization framework presented here (cf. section 6.2). The weight growth in this model is limited by weight decay rather than explicit normalization rules, which is possible because the blob dynamics provides only limited correlation values even if the weights would grow large. This model is particularly elegant with respect to the

way it indirectly introduces constraints and should be investigated further. Our discussion in section 6.3 indicates that the system L+Q might also show map expansion and weight limitation without any explicit constraints, but further analysis is needed to confirm this.

The objective functions listed in Table 1 have a tendency to produce either collapsing or expanding maps. It is unlikely that the terms can be counterbalanced such that they have the tendency to preserve distances directly, independent of normalization rules and the size of the layers, as does the algorithmic objective function in equation 6.9. A solution to this problem might be found by examining propagating activity patterns in the input as well as the output layer, such as traveling waves (Triesch, 1995) or running blobs (Wiskott & von der Malsburg, 1996). Waves and blobs of activity have been observed in the developing retina (Meister, Wong, Baylor, & Shatz, 1991). If the waves or blobs have the same intrinsic velocity in the two layers, they would tend to generate metric maps, regardless of the scaling factor induced by the normalization rules. It would be interesting to investigate this idea further and derive correlations for this class of models.

Another limitation of the framework discussed here is that it is confined to second-order correlations. As von der Malsburg (1995) has pointed out, this is appropriate only for a subset of phenomena of neural map formation, such as retinotopy and ocular dominance. Although orientation tuning can arise by spontaneous symmetry breaking (e.g., Linsker, 1986), a full understanding of the self-organization of orientation selectivity and other phenomena may require taking higher-order correlations into account. It would be interesting as a next step to consider third-order terms in the objective function and the conditions under which they can be derived from detailed neural dynamics. There may also be an interesting relationship to recent advances in algorithms for independent component analysis (Bell & Sejnowski, 1995), which can be derived from a maximum entropy method and is dominated by higher-order correlations.

Finally, it may be interesting to investigate the extent to which the techniques used in the analysis presented here can be applied to other types of neural dynamics, such as learning rules. The existence of objective functions for dynamics with curl may make it possible to formulate more learning rules within the constrained optimization framework, which could lead to new insights. Optimizing the dynamics of a learning rule without changing the set of stable fixed points may be an interesting application for coordinate transformations.

Appendix: Probabilistic Blob Model

A.1 Noise Model. Consider the activity model of Obermayer et al. (1990) as an abstraction of the neural activity dynamics in section 2.1 (see equations 2.1 and 2.2). Obermayer et al. use a high-dimensional version of the self-organizing map algorithm (Kohonen, 1982). A blob $B_{\rho'\rho_0}$ is located at

a random position ρ_0 in the input layer, and the input $i_{\tau'}(\rho_0)$ received by the output neurons is calculated as in equation 2.7. A blob $\bar{B}_{\tau'\tau_0}$ in the output layer is located at the position τ_0 of highest input, that is, $i_{\tau_0}(\rho_0) = \max_{\tau'} i_{\tau'}(\rho_0)$. Only the latter step differs in its outcome from the dynamics in section 2, the maximal input instead of the maximal overlap determining the location of the output blob.

The transition to the probabilistic blob location can be done by assuming that the blob $\bar{B}_{\tau'\tau_0}$ in the output layer is located at τ_0 with probability

$$p(\tau_0|\rho_0) = i_{\tau_0}(\rho_0) = \sum_{\rho'} w_{\tau_0\rho'} B_{\rho'\rho_0}. \tag{A.1}$$

For the following considerations, the same normalization assumptions as in section 2.1 are made, which leads to $\sum_{\tau'} i_{\tau'}(\rho_0) = 1$ and $\sum_{\tau_0} p(\tau_0|\rho_0) = 1$ and justifies the interpretation of $p(\tau_0|\rho_0)$ as a probability. The effect of different normalization rules, like those used by Obermayer et al. (1990), is discussed in the next section. The probabilistic blob location can be achieved by multiplicative noise η_τ with the cumulative density function $f(\eta) = \exp(-1/\eta)$, which leads to a modified input $l_\tau = \eta_\tau i_\tau$ with a cumulative density function

$$f_\tau(l_\tau) = \exp\left(-\frac{i_\tau(\rho_0)}{l_\tau}\right), \tag{A.2}$$

and a probability density function

$$p_\tau(l_\tau) = \frac{\partial f_\tau}{\partial l_\tau} = \frac{i_\tau(\rho_0)}{l_\tau^2} \exp\left(-\frac{i_\tau(\rho_0)}{l_\tau}\right). \tag{A.3}$$

Notice that the noise is different for each output neuron but always from the same distribution. The probability of neuron τ_0 having larger input l_{τ_0} than all other neurons τ', that is, the probability of the output blob being located at τ_0, is

$$\begin{aligned}
p(\tau_0|\rho_0) &= p(l_{\tau_0} > l_{\tau'} \; \forall \tau' \neq \tau_0) && \text{(A.4)}\\
&= \int_0^\infty p_{\tau_0}(l_{\tau_0}) \prod_{\tau' \neq \tau_0} f_{\tau'}(l_{\tau_0}) \, \mathrm{d}l_{\tau_0} && \text{(A.5)}\\
&= \int_0^\infty \frac{i_{\tau_0}(\rho_0)}{l_{\tau_0}^2} \exp\left(-\frac{1}{l_{\tau_0}} \sum_{\tau'} i_{\tau'}(\rho_0)\right) \mathrm{d}l_{\tau_0} && \text{(A.6)}\\
&= \frac{i_{\tau_0}(\rho_0)}{\sum_{\tau'} i_{\tau'}(\rho_0)} && \text{(A.7)}\\
&= i_{\tau_0}(\rho_0) \qquad \left(\text{since } \sum_{\tau'} i_{\tau'}(\rho_0) = 1\right), && \text{(A.8)}
\end{aligned}$$

which is the desired result. Thus, the model by Obermayer et al. (1990) can be modified by multiplicative noise to yield the probabilistic blob location behavior. A problem is that the modified input l_τ has an infinite mean value, but this can be corrected by consistently transforming the cumulative density functions by the substitution $l_\tau = k_\tau^2$, yielding

$$f_\tau(k_\tau) = \exp\left(-\frac{i_\tau(\rho_0)}{k_\tau^2}\right) \tag{A.9}$$

for the new modified inputs k_τ, the means of which are finite. Due to the nonlinear transformation $l_\tau = k_\tau^2$, the modified inputs k_τ are no longer a product of the original input i_τ with noise, whose distribution is the same for all neurons, but each input i_τ generates a modified input k_τ with a nonlinearly distorted version of the cumulative density function in equation A.2.

The probability for a particular combination of blob locations is

$$p(\tau_0, \rho_0) = p(\tau_0|\rho_0)p(\rho_0) = \sum_{\rho'} w_{\tau_0\rho'} B_{\rho'\rho_0} \frac{1}{R}, \tag{A.10}$$

and the correlation between two neurons defined as the average product of their activities is

$$\langle a_\tau a_\rho \rangle = \sum_{\tau_0\rho_0} p(\tau_0, \rho_0) \bar{B}_{\tau\tau_0} B_{\rho\rho_0} \tag{A.11}$$

$$= \sum_{\tau_0\rho_0} \sum_{\rho'} w_{\tau_0\rho'} B_{\rho'\rho_0} \frac{1}{R} \bar{B}_{\tau\tau_0} B_{\rho\rho_0} \tag{A.12}$$

$$= \frac{1}{R} \sum_{\tau'\rho'} \bar{B}_{\tau\tau'} w_{\tau'\rho'} \left(\sum_{\rho_0} B_{\rho'\rho_0} B_{\rho\rho_0} \right) \tag{A.13}$$

$$= \frac{1}{R} \sum_{\tau'\rho'} \bar{B}_{\tau\tau'} w_{\tau'\rho'} \bar{B}_{\rho'\rho}, \qquad \text{with} \quad \bar{B}_{\rho'\rho} = \sum_{\rho_0} B_{\rho'\rho_0} B_{\rho\rho_0}, \tag{A.14}$$

where the brackets $\langle \cdot \rangle$ indicate the ensemble average over a large number of blob presentations. This is equivalent to equation 2.13 if $\bar{B}_{\tau'\tau} = \sum_{\tau_0} B_{\tau'\tau_0} B_{\tau\tau_0}$. Thus, the two probabilistic dynamics are equivalent, though the blobs in the output layer must be different.

A.2 Different Normalization Rules. The derivation of correlations in the probabilistic blob model given above assumes explicit presynaptic normalization of the form $\sum_{\tau'} w_{\tau'\rho'} = 1$. This assumption is not valid for some models that use only postsynaptic normalization (e.g., von der Malsburg, 1973). The model by Obermayer et al. (1990) postsynaptically normalizes the square sum, $\sum_{\rho'} w_{\tau'\rho'}^2 = 1$, instead of the sum, which may make the applicability of the probabilistic blob model even more questionable.

To investigate the effect of these different normalization rules on the probabilistic blob model, assume that the projective (or receptive) fields of the input (or output) neurons codevelop in such a way that, at any given moment, all neurons in a layer have the same weight histogram. Neuron ρ, for instance, would have the weight histogram $w_{\tau'\rho}$ taken over τ', and it would be the same as those of the other neurons ρ'. Two neurons of same weight histogram have the same number of nonzero weights, and the square sums over their weights differ from the sums by the same factor c, for example, $\sum_{\tau'} w^2_{\tau'\rho'} = c \sum_{\tau'} w_{\tau'\rho'} = 1$ for all ρ' with $c \leq 1$. The weight histogram, and with it the factor c, may change over time. For instance, if point receptive fields develop from an initial all-to-all connectivity, the histogram has a single peak at $1/T$ in the beginning and has a peak at 0 and one entry at 1 at the end of the self-organization process, and $c(t)$ grows from $1/T$ up to 1, where T is the number of output neurons.

Consider first the effect of the square sum normalization under the assumption of homogeneous codevelopment of receptive and projective fields. The square sum normalization differs from the sum normalization by a factor $c(t)$ common to all neurons in the layer. Since the nonlinear blob model is insensitive to such a factor, the derived correlations and the learning rule are off by this factor c. Since this factor is common to all weights, the trajectories of the weight dynamics are identical, though the time scales differ by c between the two types of normalization.

Consider now the effect of pure postsynaptic normalization under the assumption of homogeneous codevelopment of receptive and projective fields. Assume a pair of blobs is located at ρ_0 and τ_0. With a linear growth rule, the sum over weights originating from an input neuron would change according to

$$\dot{W}_\rho = \sum_\tau \dot{w}_{\tau\rho} = \sum_\tau B_{\tau\tau_0} B_{\rho\rho_0} = B_{\rho\rho_0}, \tag{A.15}$$

since the blob $B_{\tau\tau_0}$ is normalized to one. Averaging over all input blob positions yields an average change of

$$\langle \dot{W}_\rho \rangle = \frac{1}{R} \sum_{\rho_0} B_{\rho\rho_0} = \frac{1}{R}, \tag{A.16}$$

since we assume a homogeneous average activity in the input layer, that is, $\sum_{\rho_0} B_{\rho\rho_0} = 1$. A similar expression follows for the postsynaptic sum:

$$\langle \dot{W}_\tau \rangle = \sum_{\rho_0 \tau_0} p(\tau_0, \rho_0) \sum_\rho B_{\tau\tau_0} B_{\rho\rho_0} \tag{A.17}$$

$$= \sum_{\rho_0 \tau_0} \left(\frac{1}{R} \sum_{\tau'\rho'} B_{\tau'\tau_0} w_{\tau'\rho'} B_{\rho'\rho_0} \right) \sum_\rho B_{\tau\tau_0} B_{\rho\rho_0} \tag{A.18}$$

$$= \frac{1}{R} \sum_{\tau_0} B_{\tau\tau_0} \sum_{\tau'} B_{\tau'\tau_0} \sum_{\rho'} w_{\tau'\rho'} \sum_{\rho_0} B_{\rho'\rho_0} \sum_{\rho} B_{\rho\rho_0} \tag{A.19}$$

$$= \frac{1}{T}, \tag{A.20}$$

where $\sum_{\rho'} w_{\tau'\rho'} = R/T$ is assumed due to the postsynaptic normalization rule and the blobs are normalized with respect to both of their indices. R and T are the number of neurons in the input and output layer, respectively. This equation shows that each output neuron has to normalize its sum of weights by the same amount, and it has to do that by a subtractive normalization rule if the system is consistent. The amount by which each single weight $w_{\tau\rho}$ is changed depends on the number of nonzero weights an output neuron receives. Since we assume the weight histograms are the same, each output neuron has the same number of nonzero weights, and each weight gets corrected by the same amount. Since we also assume same weight histograms for the projective fields, the sum over all weights originating from an input neuron is corrected by the same amount for each input neuron, namely, by $1/R$ per time unit. Thus, the postsynaptic normalization rule preserves presynaptic normalization.

It can even be argued that a postsynaptic normalization rule stabilizes presynaptic normalization. Assume that an input neuron has a larger (or smaller) sum over its weights than the other input neurons. Then this neuron is likely to have more (fewer) nonzero weights than the other input neurons. This results in a larger (smaller) negative compensation by the postsynaptic normalization rule, since each weight is corrected by the same amount. This then reduces the difference between the input neuron under consideration and the others. It is important to notice that this effect of stabilizing the presynaptic normalization is not preserved in the constrained optimization formulation. It may be necessary to use explicit presynaptic normalization in the constrained optimization formulation to account for the implicit presynaptic normalization in the blob model.

If the postsynaptic constraint is based on the square sum, then the normalization rule is multiplicative, and the projective fields of the input neurons need not have the same weight histograms. The system would still preserve the presynaptic normalization. Notice that the derivation given above does not hold for a nonlinear Hebbian rule, for example, $\dot{w}_{\tau\rho} = w_{\tau\rho} a_\tau a_\rho$.

These considerations show that the probabilistic blob model may be a good approximation even if the constraints are based on the square sum instead of the sum and if only the postsynaptic neurons are constrained and not the presynaptic neurons, as was required in the derivation of the probabilistic blob model above. The homogeneous codevelopment of receptive and projective fields is probably a reasonable assumption for high-dimensional models with a homogeneous architecture. For low-dimensional models, such as the low-dimensional self-organizing map algorithm (Kohonen, 1982), the assumption is less likely to be valid. However, numerical

simulations or more detailed analytical considerations are needed to verify the assumption for any given concrete model.

Acknowledgments

We are grateful to Geoffrey J. Goodhill, Thomas Maurer, Jozsef Fiser, and two anonymous referees for carefully reading the manuscript and offering useful comments. L. W. has been supported by a Feodor-Lynen fellowship by the Alexander von Humboldt-Foundation, Bonn, Germany.

References

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

Amari, S. (1980). Topographic organization of nerve fields. *Bulletin of Mathematical Biology*, *42*, 339–364.

Bauer, H.-U., Brockmann, D., & Geisel, T. (1997). Analysis of ocular dominance pattern formation in a high-dimensional self-organizing-map model. *Network: Computation in Neural Systems*, *8*(1), 17–33.

Behrmann, K. (1993). *Leistungsuntersuchungen des "Dynamischen Link-Matchings" und Vergleich mit dem Kohonen-Algorithmus* (Internal Rep. No. IR-INI 93–05). Bochum: Institut für Neuroinformatik, Ruhr-Universität Bochum.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*, 1129–1159.

Bienenstock, E., & von der Malsburg, C. (1987). A neural network for invariant pattern recognition. *Europhysics Letters*, *4*(1), 121–126.

Dirac, P. A. M. (1996). *General theory of relativity*. Princeton, NJ: Princeton University Press.

Ermentrout, G. B., & Cowan, J. D. (1979). A mathematical theory of visual hallucination patterns. *Biological Cybernetics*, *34*(3), 137–150.

Erwin, E., Obermayer, K., & Schulten, K. (1995). Models of orientation and ocular dominance columns in the visual cortex: A critical comparison. *Neural Computation*, *7*, 425–468.

Ginzburg, I., & Sompolinsky, H. (1994). Theory of correlations in stochastic neural networks. *Physical Review E*, *50*(4), 3171–3191.

Goodhill, G. J. (1993). Topography and ocular dominance: A model exploring positive correlations. *Biol. Cybern.*, *69*, 109–118.

Goodhill, G. J., Finch, S., & Sejnowski, T. J. (1996). Optimizing cortical mappings. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems* (Vol. 8, pp. 330–336). Cambridge, MA: MIT Press.

Häussler, A. F., & von der Malsburg, C. (1983). Development of retinotopic projections—An analytical treatment. *J. Theor. Neurobiol.*, *2*, 47–73.

Horst, R., Pardalos, P. M., & Thoai, N. V. (1995). *Introduction to global optimization*. Dordrecht: Kluwer.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern.*, *43*, 59–69.

Kohonen, T. (1990). The self-organizing map. *Proc. of the IEEE, 78*(9), 1464–1480.

Konen, W., Maurer, T., & von der Malsburg, C. (1994). A fast dynamic link matching algorithm for invariant pattern recognition. *Neural Networks, 7*(6/7), 1019–1030.

Konen, W., & von der Malsburg, C. (1993). Learning to generalize from single examples in the dynamic link architecture. *Neural Computation, 5*(5), 719–735.

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers, 42*(3), 300–311.

Linsker, R. (1986). From basic network principles to neural architecture: Emergence of orientation columns. *Ntl. Acad. Sci. USA, 83*, 8779–8783.

MacKay, D. J. C., & Miller, K. D. (1990). Analysis of Linsker's simulations of Hebbian rules. *Neural Computation, 2*, 173–187.

Meister, M., Wong, R. O. L., Baylor, D. A., & Shatz, C. J. (1991). Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science, 252*, 939–943.

Miller, K. D. (1990). Derivation of linear Hebbian equations from nonlinear Hebbian model of synaptic plasticity. *Neural Computation, 2*, 321–333.

Miller, K. D., Keller, J. B., & Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. *Science, 245*, 605–615.

Miller, K. D., & MacKay, D. J. C. (1994). The role of constraints in Hebbian learning. *Neural Computation, 6*, 100–126.

Nowlan, S. J. (1990). Maximum likelihood competitive learning. In D. S. Touretzky (Ed.), *Advances in neural information processing systems* (Vol. 2, pp. 574–582). San Mateo, CA: Morgan Kaufmann.

Obermayer, K., Ritter, H., & Schulten, K. (1990). Large-scale simulations of self-organizing neural networks on parallel computers: Application to biological modelling. *Parallel Computing, 14*, 381–404.

Ritter, H., Martinetz, T., & Schulten, K. (1991). *Neuronale Netze.* Reading, MA: Addison-Wesley.

Sejnowski, T. J. (1976). On the stochastic dynamics of neuronal interaction. *Biol. Cybern., 22*, 203–211.

Sejnowski, T. J. (1977). Storing covariance with nonlinearly interacting neurons. *J. Math. Biology, 4*, 303–321.

Swindale, N. V. (1980). A model for the formation of ocular domance stripes. *Proc. R. Soc. Lond. B, 208*, 243–264.

Swindale, N. V. (1996). The development of topography in the visual cortex: A review of models. *Network: Comput. in Neural Syst., 7*(2), 161–247.

Tanaka, S. (1990). Theory of self-organization of cortical maps: Mathematical framework. *Neural Networks, 3*, 625–640.

Tanaka, S. (1991). Theory of ocular dominance column formation. *Biol. Cybern., 64*, 263–272.

Triesch, J. (1995). *Metrik im visuellen System* (Internal Rep. No. IR-INI 95-05). Bochum: Institut für Neuroinformatik, Ruhr-Universität Bochum.

von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik, 14*, 85–100.

von der Malsburg, C. (1995). Network self-organization in the ontogenesis of

the mammalian visual system. In S. F. Zornetzer, J. Davis, and C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 447–463). San Diego: Academic Press.

von der Malsburg, C., & Willshaw, D. J. (1977). How to label nerve cells so that they can interconnect in an ordered fashion. *Proc. Natl. Acad. Sci. (USA), 74,* 5176–5178.

von der Malsburg, C., & Willshaw, D. J. (1981). Differential equations for the development of topological nerve fibre projections. *SIAM-AMS Proceedings, 13,* 39–47.

Whitelaw, D. J., & Cowan, J. D. (1981). Specificity and plasticity of retinotectal connections: A computational model. *J. Neuroscience, 1*(12), 1369–1387.

Willshaw, D. J., & von der Malsburg, C. (1976). How patterned neural connections can be set up by self-organization. *Proc. R. Soc. London, B194,* 431–445.

Wiskott, L., & von der Malsburg, C. (1996). Face recognition by dynamic link matching. In J. Sirosh, R. Miikkulainen, & Y. Choe (Eds.), *Lateral interactions in the cortex: structure and function* (Chap. 11) [Electronic book]. Austin, TX: UTCS Neural Networks Research Group. Available from http://www.cs.utexas.edu/users/nn/web-pubs/htmlbook96/.

Received April 25, 1997; accepted September 3, 1997.

LETTER Communicated by Helge Ritter

Breaking Rotational Symmetry in a Self-Organizing Map Model for Orientation Map Development

M. Riesenhuber
Department of Brain and Cognitive Sciences and Center for Biological and Computational Learning, Massachusetts Institute of Technology, E25-221, Cambridge, MA 02139, U.S.A.

H.-U. Bauer
D. Brockmann
T. Geisel
Max-Planck-Institut für Strömungsforschung, Postfach 28 53, 37018 Göttingen, Federal Republic of Germany

We analyze the pattern formation behavior of a high-dimensional self-organizing map (SOM) model for the competitive projection of ON-center-type and OFF-center-type inputs to a common map layer. We mathematically show, and numerically confirm, that even isotropic stimuli can drive the development of oriented receptive fields and an orientation map in this model. This result provides an important missing link in the spectrum of pattern formation behaviors observed in SOM models. Extending the model by including further layers for binocular inputs, we also investigate the combined development of orientation and ocular dominance maps. A parameter region for combined patterns exists; corresponding maps show a preference for perpendicular intersection angles between iso-orientation lines and ocularity domain boundaries, consistent with experimental observations.

1 Introduction

Topographic maps are a ubiquitous pattern of organization in the brain. Among the most intensely investigated such patterns are orientation maps and ocular dominance maps in the visual cortex. Various models have been formulated that generate topographic maps as a consequence of activity-driven self-organization processes (for two recent reviews, see Erwin, Obermayer, & Schulten, 1995; Swindale, 1996). The simulated maps coincide with observed maps in many aspects. Yet distinctive differences remain between simulated and observed maps, as well as between simulated maps from different modeling frameworks. The better we can relate such differences to specific underlying assumptions, the more differences we can eliminate, and the more experimental observations we can account for within uni-

Neural Computation **10, 717–730** (1998)

versal modeling frameworks, the more stable our understanding of map self-organization processes will become.

A few years ago, two map formation models were presented that generate oriented receptive fields from a competition of ON-center and OFF-center cell responses in the lateral geniculate nucleus (LGN) (Miller, 1992, 1994; Miyashita & Tanaka, 1992). These models elegantly explain how orientation-selective simple cell responses could be due to a self-organization process driven by nonpatterned input activity and could take place even before birth. Without discussing any of the details of these models, we note that their inputs are nonoriented, yet the resulting patterns break this symmetry.

A third, rather widely applied framework for map development is Kohonen's self-organizing map (SOM) algorithm. SOM-based models have sucessfully accounted for various aspects of visual (Obermayer, Ritter, & Schulten, 1990; Obermayer, Blasdel, & Schulten, 1992; Goodhill, 1993; Wolf, Bauer, & Geisel, 1994; Wolf, Bauer, Pawelzik, & Geisel, 1996; Bauer, Brockmann, & Geisel, 1997), auditory (Martinetz, Ritter, & Schulten, 1988) and somatosensory (Ritter & Schulten, 1986; Andres, Schlüter, Spengler, & Dinse, 1994) map formation. Yet in simulations of SOM-based models for ON-center- and OFF-center-cell competition, a break of rotational symmetry has not been observed so far, despite a lengthy search by several groups. This negative outcome could be the consequence of a suboptimal selection of parameters, or it could be the fingerprint of a fundamental difference between SOM-based models and the models by Miyashita and Tanaka, and by Miller. Clarification of this issue is an interesting problem, not only with regard to explanations of orientation map development but, in particular, with regard to theoretical consistency between modeling frameworks.

In this article, we report that SOMs can break rotational symmetry, albeit in a quite small regime of parameters. Using a recently described analysis technique (Riesenhuber, Bauer, & Geisel, 1996; Bauer et al., 1997), we first mathematically analyze the pattern formation behavior of the corresponding SOM model. After a brief introduction to the SOM in the second section, we describe analytical results in the third section. Guided by the mathematical analysis, we also performed simulations of the model; the results are given in the fourth section. Finally, we investigate the development of combined orientation and ocular dominance maps.

2 "High-Dimensional" SOM Model for the Development of Orientation Maps from Nonoriented Stimuli

Neurons in an SOM are characterized by positions $\mathbf{r}$ in a map lattice $\mathcal{A}$ and receptive fields $\mathbf{w_r}$ in a map input space $\mathcal{V}$. The input space is assumed to consist of one (or several) layer(s) of input channels. The typically large number of input channels give rise to the notion of a high-dimensional SOM, as opposed to a feature map description. As a consequence, stimuli $\mathbf{v} \in \mathcal{V}$ are activity distributions, and receptive fields $\mathbf{w_r}$ are synaptic weight distri-

butions. A stimulus $\mathbf{v}$ is mapped onto that neuron $\mathbf{s} \in \mathcal{A}$, whose receptive field $\mathbf{w}_\mathbf{s}$ matches $\mathbf{v}$ best,

$$\mathbf{s} = \arg\max_{\mathbf{r}\in\mathcal{A}}\{\mathbf{w}_\mathbf{r} \cdot \mathbf{v}\}. \tag{2.1}$$

Presenting a random sequence of stimuli and performing adaptation steps,

$$\Delta\mathbf{w}_\mathbf{r} = \epsilon h(\mathbf{r} - \mathbf{s})(\mathbf{v} - \mathbf{w}_\mathbf{r}), \tag{2.2}$$

the internal shape of individual receptive fields as well as the map layout self-organize simultaneously. The neighborhood function $h(\mathbf{r} - \mathbf{s})$,

$$h(\mathbf{r} - \mathbf{s}) = \exp\left(-\frac{\|\mathbf{r} - \mathbf{s}\|^2}{2\sigma^2}\right), \tag{2.3}$$

ensures that neighboring neurons align their receptive fields; that is, it imposes topography on the map. A comprehensive treatment of many theoretical and application-related aspects of SOMs can be found in Ritter, Martinetz, and Schulten (1992) and Kohonen (1995).

Within this general framework, we now consider a projection geometry analogous to that proposed by Miller (1992, 1994) and Miyashita and Tanaka (1992). Cells in ON-center and OFF-center input layers project to the map layer. As would result from a filtering of pointlike retinal stimuli by thalamic ON-center and OFF-center cells, we assume our stimuli to consist of an activity peak in one layer, plus an activity annulus in the other layer.

Mathematically, stimuli are represented as difference-of-gaussians (DOG) (stimulus center position: $\mathbf{x}_0$, widths of the two gaussians: $\sigma_{1,2}$, relative amplitude of the gaussians: k),

$$\mathbf{a}(\mathbf{x}; \mathbf{x}_0) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{x}_0\|^2}{2\sigma_1^2}\right) - k\exp\left(\frac{-\|\mathbf{x} - \mathbf{x}_0\|^2}{2\sigma_2^2}\right). \tag{2.4}$$

Furthermore, $\mathbf{a}_\bullet(\mathbf{x}; \mathbf{x}_0) = [\mathbf{a}(\mathbf{x}; \mathbf{x}_0)]_+$ denotes the activity distribution of the central peak of the DOG, and $\mathbf{a}_\circ(\mathbf{x}; \mathbf{x}_0) = [-\mathbf{a}(\mathbf{x}; \mathbf{x}_0)]_+$ the annulus-shaped activity distribution corresponding to the negative part of the DOG ($[\cdot]_+$ is the rectification operator). Naturally, ON-type and OFF-type stimuli are represented as two-component vectors,

$$\mathbf{v}_{\mathrm{ON}} = (\mathbf{a}_\bullet, \mathbf{a}_\circ), \quad \mathbf{v}_{\mathrm{OFF}} = (\mathbf{a}_\circ, \mathbf{a}_\bullet), \tag{2.5}$$

each component describing the (nonnegative) activity distribution in one of the input layers. In the simulations, the center positions $\mathbf{x}_0$ and polarity (that is, whether the stimulus is $\mathbf{v}_{\mathrm{ON}}$ or $\mathbf{v}_{\mathrm{OFF}}$) are chosen at random.

3 Mathematical Results on the Formation of Oriented Receptive Fields

Recently, Riesenhuber et al. (1996) and Bauer et al. (1997) described a new technique to calculate conditions on stimulus and map parameters for the emergence of nontrivial patterns in high-dimensional SOMs. This technique makes use of a distortion measure $E_{\mathbf{v}}$, which is evaluated for different potentially stable states of the map and is assumed to be minimized by the SOM algorithm. A crucial feature of the method is the way in which "potentially stable states" of a SOM are formalized. Although an explicit characterization of such states in terms of the weight vectors $\mathbf{w_r}$ seems impossible without actually simulating the SOM, the states can also be characterized by the way they distribute stimuli among map neurons (the tesselation). This is specific to the SOM, where the winner-take-all mapping rule (see equation 2.1) assigns a particular map neuron to each stimulus.

Denoting by $\Omega_{\mathbf{r}}$ all stimuli that are mapped to neuron $\mathbf{r}$ (the Voronoi cell of $\mathbf{r}$), we define a distortion measure,

$$E_{\mathbf{v}} = \sum_{\mathbf{r},\mathbf{r}'} h(\mathbf{r} - \mathbf{r}')\, w(\mathbf{r}, \mathbf{r}'), \tag{3.1}$$

$$w(\mathbf{r}, \mathbf{r}') = \sum_{\mathbf{v}\in\Omega_{\mathbf{r}}} \sum_{\mathbf{v}'\in\Omega_{\mathbf{r}'}} (\mathbf{v} - \mathbf{v}')^2. \tag{3.2}$$

Each term in $E_{\mathbf{v}}$ consists of the mean squared difference $w(\mathbf{r}, \mathbf{r}')$ between stimuli within the same, or between neighboring, Voronoi cells, weighted by the neighborhood function $h(\mathbf{r} - \mathbf{r}')$. The definition of $E_{\mathbf{v}}$ is motivated by the analogy of SOMs to vector quantizers (see Riesenhuber et al., 1996). Qualitatively different states of a map yield different values of $E_{\mathbf{v}}$ because they correspond to different tesselations $\{\Omega_{\mathbf{r}}\}$. What are the typical tesselations in the present context?

To simplify the analysis, we assume an equal number of ON-center input channels, OFF-center input channels, and map neurons, with a basically retinotopic projection. We further assume that the stimulus center positions $\mathbf{x}_0$ are constrained to the positions of the input channels, resulting in twice as many stimuli as map neurons. Three qualitatively different possible map states can be distinguished:

1. System $\mathcal{B}$: Each neuron responds to both an ON- and an OFF-stimulus, each located at the same retinal position. This tesselation yields neurons with orientation-insensitive receptive fields.
2. System $\mathcal{S}$: As in system $\mathcal{B}$, each neuron responds to stimuli of both polarities, but now displaced one step along one retinal coordinate. The displacement breaks isotropy. It causes the receptive fields to exhibit internal ON-center and OFF-center structure, with orientation specificity.

3. System $\mathcal{O}$: Each neuron responds to two retinally neighboring stimuli of identical polarity. Although this tesselation induces an orientation specificity, it also breaks the symmetry between ON-center and OFF-center inputs to each neuron. Neurons segregate into ON-center- and OFF-center-dominated populations, analogous to an ocular dominance map. While this state is an imaginable (and numerically observed, see below) state of SOMs, we do not consider this state as biologically interesting.

To evaluate $w(\mathbf{r}, \mathbf{r}')$ and $E_{\mathbf{v}}$ for these tesselations, we need to consider the difference δ^s between two stimuli of same polarity, separated by a distance of $\Delta\mathbf{x}_0 = \mathbf{x}_0 - \mathbf{x}_0'$,

$$\delta^s(\Delta\mathbf{x}_0) = \sum_{\mathbf{x}} \left(\mathbf{a}_\bullet(\mathbf{x}, \mathbf{x}_0) - \mathbf{a}_\bullet(\mathbf{x}, \mathbf{x}_0')\right)^2 + \left(\mathbf{a}_\circ(\mathbf{x}, \mathbf{x}_0) - \mathbf{a}_\circ(\mathbf{x}, \mathbf{x}_0')\right)^2, \quad (3.3)$$

and the difference δ^a between two oppositely polarized stimuli,

$$\delta^a(\Delta\mathbf{x}_0) = \sum_{\mathbf{x}} \left(\mathbf{a}_\bullet(\mathbf{x}, \mathbf{x}_0) - \mathbf{a}_\circ(\mathbf{x}, \mathbf{x}_0')\right)^2 + \left(\mathbf{a}_\circ(\mathbf{x}, \mathbf{x}_0) - \mathbf{a}_\bullet(\mathbf{x}, \mathbf{x}_0')\right)^2. \quad (3.4)$$

Using these distances, and exploiting the isotropy with respect to one retinal coordinate, we obtain for the interaction terms w (see equation 3.2):

$$w^{\mathcal{B}}(\Delta r) = \delta^s(\Delta r) + \delta^a(\Delta r), \quad (3.5)$$

$$w^{\mathcal{S}}(\Delta r) = \delta^s(\Delta r) + \frac{1}{2}\{\delta^a(\Delta r - 1) + \delta^a(\Delta r + 1)\}, \quad (3.6)$$

$$w^{\mathcal{O}}(\Delta r) = \sum_{i=s,a} \frac{1}{2}\delta^i(\Delta r) + \frac{1}{4}\{\delta^i(\Delta r + 1) + \delta^i(\Delta r - 1)\}. \quad (3.7)$$

Inserting equations 3.5–3.7 into equation 3.1 and performing a numerical summation, we obtain the distortion measures $E_{\mathbf{v}}^{\mathcal{B}}$, $E_{\mathbf{v}}^{\mathcal{S}}$, and $E_{\mathbf{v}}^{\mathcal{O}}$. This analysis predicts that, depending on the stimulus parameters $\sigma_{1,2}$, k, and the map neighborhood parameter σ, different final states of the map will be attained. Figure 1a shows a state diagram in the σ, k-plane, at widths $\sigma_1 = 1$, $\sigma_2 = 2$. At large values of the annulus amplitude k, receptive fields segregate into "monopolar" ON- and OFF-receptive fields. Large values of the neighborhood width σ prohibit internal structure of the receptive fields to occur. Only in a rather small regime of σ, k-values, the biologically interesting map state $\mathcal{S}$ is attained.

4 Numerical Results

To corroborate the mathematical analysis above and to obtain orientation maps, we also investigated the model numerically. In a first series we ran

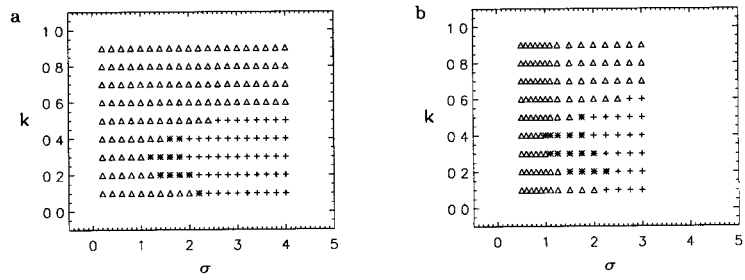


Figure 1: Analytical (a) and numerical (b) phase diagrams for the SOM orientation map model. The parameters σ and k denote the neighborhood width of the SOM algorithm and the annulus amplitude of the stimuli, respectively. $+$: the nonoriented state $\mathcal{B}$. $\star$: the oriented state $\mathcal{S}$. $\triangle$: the (nonbiological) state $\mathcal{O}$. In both diagrams, further parameters were $\sigma_1 = 1$, $\sigma_2 = 2$, in the numerical maps we applied 10^5 learning steps, learning step size was decreased from $\epsilon_{\text{init}} = 0.2$ to $\epsilon_{\text{final}} = 0.01$, and the neighborhood width σ was kept constant during the simulation.

simulations with 16×16 neuron maps, at various values of σ and k. Classifying the resulting receptive fields with regard to the states $\mathcal{B}, \mathcal{S}, \mathcal{O}$, we obtained the state diagram depicted in Figure 1b, which corresponds quite well to the mathematically obtained diagram.

To obtain receptive fields with a fine spatial resolution, we also simulated maps with 48×48-channel input layers, projecting to a 24×24-neuron map layer. Figure 2 shows exemplary receptive fields of neurons in a 4×8 segment of the map. The receptive fields show a multilobed structure and are clearly oriented. The variation of orientation over the whole map is shown in an angle map in Figure 3, where the preferred orientation of each cell is given by a circular color code. As in orientation maps obtained by optical imaging methods in the cat (Bonhoeffer & Grinvald, 1991) or monkey (Blasdel & Salama, 1986), we find a patchy arrangement of different preferred orientations and also pinwheel-like singularities.

In addition to the map of preferred orientation angles of the receptive fields, we also calculated the phases of the receptive field, that is, the shift angles that would occur in a Gabor function fit to the receptive field profile (see the caption for Figure 3). It has been hypothesized that the phase angle is also arranged in a topographic fashion in the primary visual cortex, with a topology of the combined orientation and phase stimulus space equivalent to that of a Klein bottle (Tanaka, 1995). In our simulated maps, we find the phase angle to vary indeed in a smooth way in many areas of the map (see the arrows in Figure 3). Phase and orientation values are not correlated.

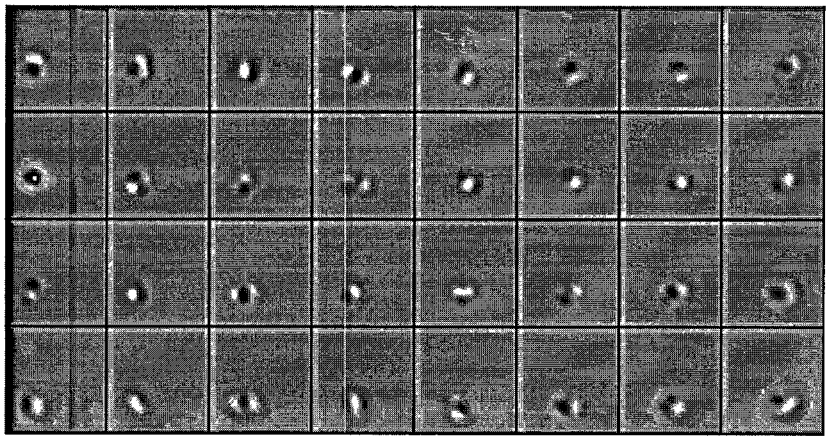


Figure 2: Sample receptive fields of an SOM orientation map (4 × 8 segment of a 24 × 24-neuron SOM, with two 48 × 48 input layers, periodic boundary conditions). Further parameters of the simulation were: $\sigma_1 = 3.4$, $\sigma_2 = 6.8$, $k = 0.3$, width of SOM neighborhood function $\sigma = 0.85$, 3×10^5 learning steps, $\epsilon = 0.1 \rightarrow 0.01$. For each neuron, the difference between ON-center and OFF-center cell layer connection strengths is shown as a gray-value image. The gray background means no connection strength; black and white regions indicate preferred connections to the ON- or OFF-center layers.

2π-singularities can be found in the phase, at locations other than those of orientation singularities. All the topological properties we could identify in our map are consistent with Tanaka's suggestion of the Klein bottle topology.

5 Development of Combined Orientation and Ocular Dominance Maps

Finally, we complemented the two ON-center and OFF-center input cell layers for one eye by two further ON-center and OFF-center cell layers for the other eye. The repertoire of possible patterns in this extended model should go beyond merely oriented receptive fields in an orientation map. It should also include monocular receptive fields and ocular dominance maps, and combinations of the two types of patterns.

Stimuli in the extended model consist of activity distributions in all four input layers. Although the difference in the shape of the activity distributions between ON-center and OFF-center layers is the same as before, the partial stimuli are assumed to be of identical shape in the corresponding

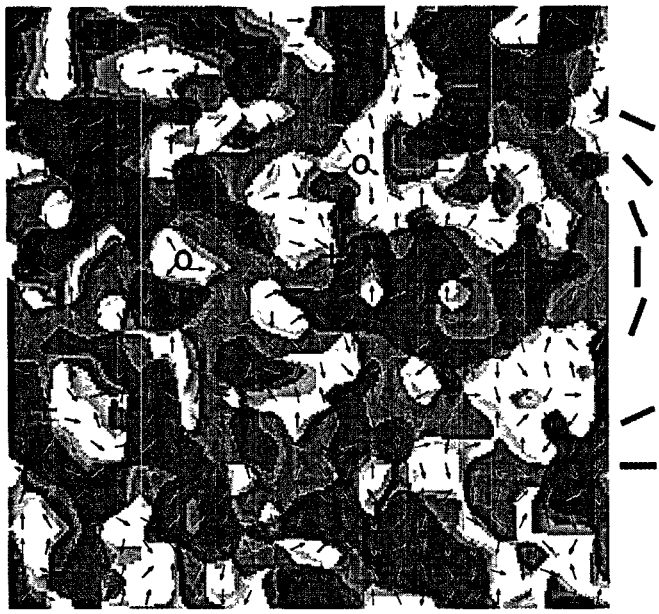


Figure 3: The complete map described in the caption of Figure 2, now depicted as an angle map using a circular color code for the preferred orientation angles of each cell. Superimposed on the color-coded orientation map, we show (as arrows) the phase angle of the receptive fields. The phase angle is calculated by rotating each receptive field by the inverse of its preferred orientation (so that the lobes of an ideal bilobed receptive would fall on different sides of the horizontal meridian after rotation) and then determining the phase shift as compared to a dampened sine wave (i.e., a Gabor filter) of the receptive field profile. A rightward arrow designates a phase of zero degrees, that is, the receptive field is bilobed with the positive lobe being in the upper hemisphere after rotation. An upward arrow represents a phase angle of 90 degrees, that is, a trilobed receptive field with the negative lobe in the middle. Crosses and circles show examplary locations of orientation and phase singularities, respectively.

layers for either eye, but attenuated by a factor of c, $0 \leq c \leq 1$, in one of the eyes (analogous to the assumptions underlying a recently analyzed SOM-based model for ocular dominance formation; Bauer et al., 1997; see also Goodhill, 1993). This yields, apart from the random variations of the

stimulus center, four different types of stimuli:

$$\mathbf{v}_{\mathrm{L,ON}} = \begin{pmatrix} \mathbf{a}_\bullet \\ \mathbf{a}_\circ \\ c \cdot \mathbf{a}_\bullet \\ c \cdot \mathbf{a}_\circ \end{pmatrix}, \quad \mathbf{v}_{\mathrm{L,OFF}} = \begin{pmatrix} \mathbf{a}_\circ \\ \mathbf{a}_\bullet \\ c \cdot \mathbf{a}_\circ \\ c \cdot \mathbf{a}_\bullet \end{pmatrix},$$
$$\mathbf{v}_{\mathrm{R,ON}} = \begin{pmatrix} c \cdot \mathbf{a}_\bullet \\ c \cdot \mathbf{a}_\circ \\ \mathbf{a}_\bullet \\ \mathbf{a}_\circ \end{pmatrix}, \quad \mathbf{v}_{\mathrm{R,OFF}} = \begin{pmatrix} c \cdot \mathbf{a}_\circ \\ c \cdot \mathbf{a}_\bullet \\ \mathbf{a}_\circ \\ \mathbf{a}_\bullet \end{pmatrix}. \tag{5.1}$$

The analysis technique introduced in section 3 can be applied to this more complicated case as well, considering the different tesselation possibilities for four stimuli per neuron. To save space, we omit the details of the rather lengthy calculations and proceed to a description of the results. Since we have five parameters in the model now (σ_1, σ_2, k for the DOG, c for the between-eye correlations, and σ as a map control parameter), the full state diagram cannot be depicted. Instead, we show in Figure 4a a section in the k-c-plane. Regions with orientation only and with ocular dominance only are found. Most important, there is a region with a combination of both orientation and ocular dominance at small values of k and c.

In computer simulations, we found maps with monocular receptive fields, oriented receptive fields, or combined monocular, oriented receptive fields, each in the parameter regimes predicted by the analysis (see Figure 4b). Figure 5 shows one combined map in a plot that displays the boundaries of the iso-ocularity domains superimposed on the color-coded orientation map.

Determining the transition lines between iso-ocularity regions in the simulated map, and computing the intersection angles with the iso-orientation lines at these locations, we compiled an angle histogram (see Figure 6). Iso-orientation lines intersect the boundaries between iso-ocularity regions preferably at larger angles, consistent with experimental observations by Bartfeld and Grinvald (1992) and Obermayer and Blasdel (1993).

6 Discussion

We showed mathematically and numerically how in a high-dimensional SOM model for the competitive projection of ON-center and OFF-center inputs to a common map layer, a rotation symmetry of stimuli can be broken to yield oriented receptive fields. This pattern formation behavior can be described only in a high-dimensional map formation framework, which also allows consideration of the internal structure of receptive fields. In low-

as two

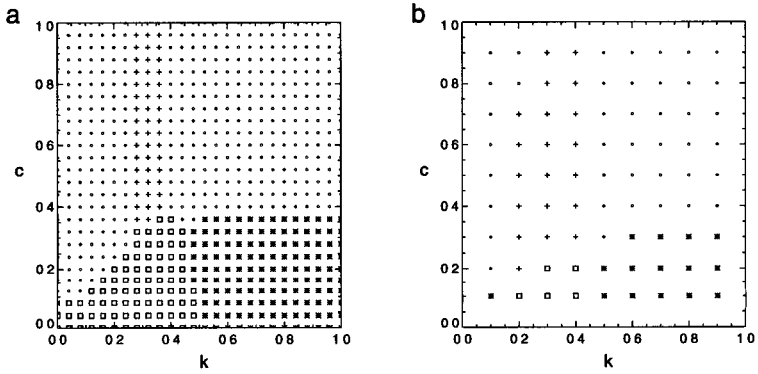


Figure 4: Phase diagram for combined orientation and ocular dominance maps, as a function of ON-OFF stimulus parameter k and between-eye-correlation parameter c, at $\sigma_1 = 0.84$, $\sigma_2 = 1.68$, $\sigma = 1.0$. ○: states with unoriented receptive fields. +: oriented, binocular receptive fields. ⋆: monocular receptive fields with type $\mathcal{O}$ ON-center OFF-center cell segregation. □: monocular, oriented receptive fields. (a) The analytically obtained diagram. (b) The diagram resulting from classification of the receptive fields of simulated maps.

dimensional feature map models, where each map dimension corresponds to a particular stimulus and receptive field parameter, a nontrivial structure along a particular dimension cannot be obtained if the stimuli have no extension along this dimension (as is the case for round stimuli with respect to an orientation dimension). A break of isotropy has already been observed in other frameworks for map development models (Miyashita & Tanaka, 1992; Miller, 1992, 1994). The results we describe here for the SOM framework close a somewhat puzzling gap in the qualitative behavior of these different frameworks, reducing the relative importance of the specific mathematical formalizations, and increasing the importance of common mechanisms.

Our results are based not only on numerical simulations but also on a mathematical analysis. The coincidence of the mathematically derived parameter regimes for particular map structures and the numerical observation of these structures underlines the value of our energy formalism to analyze pattern formation in high-dimensional SOM models and to guide simulations of these models. For the case of combined ocular dominance and orientation maps, the analysis turned out to involve a substantially larger number of map patterns, which need to be considered. The increase in effort necessary for the two-variable case suggests that this kind of analysis is not feasible for maps with three underlying symmetries. For the com-

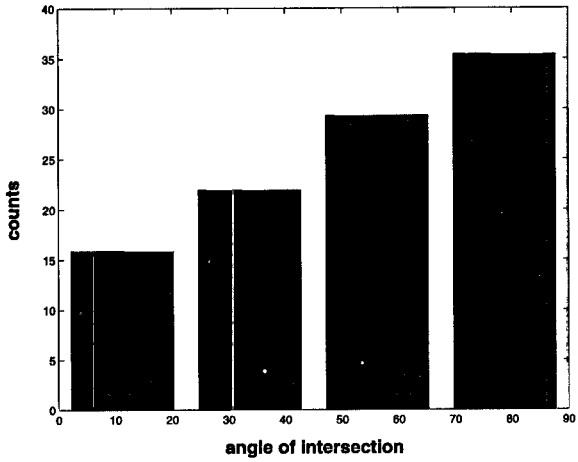


Figure 6: Histogram of angle of intersection of iso-orientation lines and iso-ocularity domain boundaries, computed for all cells along iso-ocularity domain boundaries of the map shown in Figure 5.

Acknowledgments

We gratefully acknowledge interesting discussions with Ken Miller and Fred Wolf. This work has been supported by the Deutsche Forschungsgemeinschaft through Sonderforschungsbereich 185 Nichtlineare Dynamik, TP E6.

References

Andres, M., Schlüter, O., Spengler, F., & Dinse, H. R. (1994). A model of fast and reversible representational plasticity using Kohonen mapping. In M. Marinao & P. G. Morasso (Eds.), *Proceedings of the ICANN94, Sorrento* (pp. 306–309). Berlin: Springer-Verlag.

Bartfeld, E., & Grinvald A. (1992). Relationship between orientation preference pinwheels, cytochrome oxidase blobs and ocular-dominance columns in primate striate cortex. *Proc. Nat. Acad. Sci. USA, 89*, 11905–11909.

Bauer, H.-U., Brockmann, D., & Geisel, T. (1997). Analysis of ocular dominance pattern formation in a high-dimensional self-organizing-map-model. *Network, 8*, 17–33.

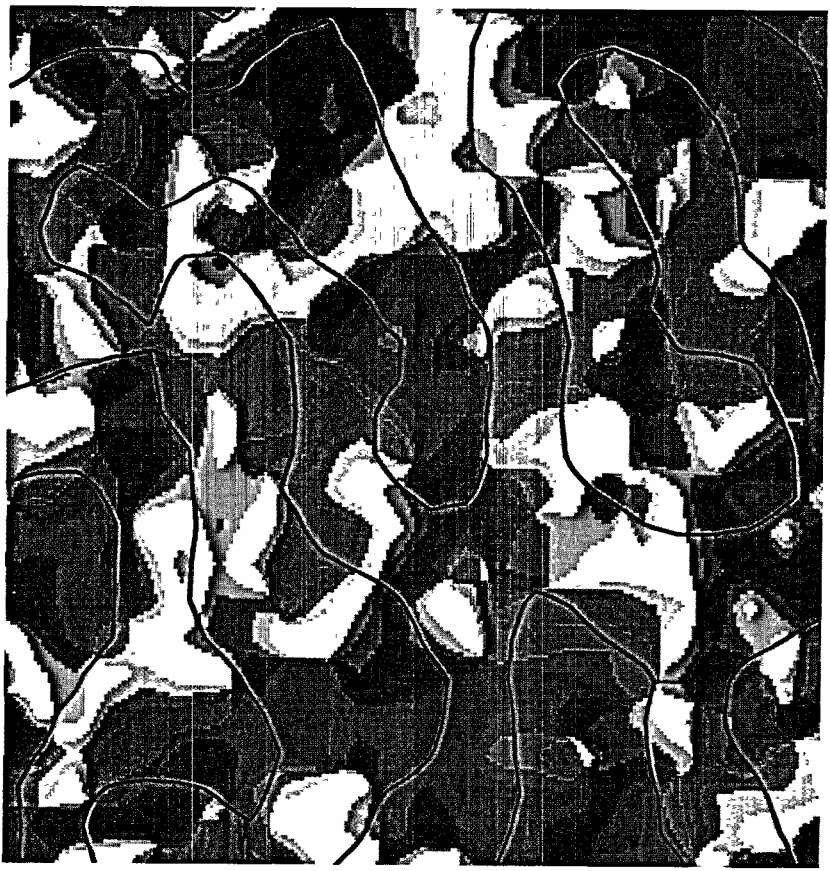


Figure 5: Combined ocular dominance and orientation map, with stimulus parameters as described in the caption of Figure 4. The orientation angle map is given in color code (as in Figure 3); the iso-ocularity domain boundaries are superimposed as black lines. Further parameters are: 24×24-neuron SOM, 34×34-channel input layers, $\sigma = 0.71$, 3×10^5 learning steps, $\epsilon = 0.1 \rightarrow 0.01$, periodic boundary conditions.

bined maps, we identified a parameter regime in which ocular dominance and orientation structure stably coexist. In the correlation-based framework, combined maps were numerically found to exist (Erwin & Miller, 1995), but the mathematical underpinnings were judged controversial (Piepenbrock, Ritter, & Obermayer, 1996, 1997).

Blasdel, G. G., & Salama, G. (1986). Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature, 321,* 579–585.

Bonhoeffer, T., & Grinvald, A. (1991). Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature, 353,* 429–430.

Erwin, E., & Miller, K. D. (1995). Modeling joint development of ocular dominance and orientation maps in primary visual cortex. In J. Bower (Ed.), *Computational neuroscience* (pp. 179–184). Boston: Kluwer.

Erwin, E., Obermayer, K., & Schulten, K. (1995). Models of orientation and ocular dominance columns in the visual cortex: A critical comparison. *Neur. Comp., 7,* 425–468.

Goodhill, G. J. (1993). Topography and ocular dominance: A model exploring positive correlations. *Biol. Cyb., 69* 109–118.

Kohonen, T., (1995). *The self-organizing map.* Berlin: Springer-Verlag.

Martinetz, T., Ritter, H., & Schulten, K. (1988). Kohonen's self-organizing map for modeling the formation of the auditory cortex of a bat. In SGAICO-Proceedings "Connectionism in Perspective," 403–412.

Miller, K. D. (1992). Development of orientation columns via competition between ON- and OFF-center inputs. *NeuroRep., 3,* 73–79.

Miller, K. D. (1994). A model for the development of simple-cell receptive fields and the ordered arrangement of orientation columns through activity dependent competition between On- and Off-center inputs. *J. Neurosci., 14,* 409–441.

Miyashita, M., & Tanaka, S. (1992). A mathematical model for the self-organization of orientation columns in visual cortex. *NeuroRep., 3,* 69–72.

Obermayer, K., & Blasdel, G. G. (1993). Geometry of orientation and ocular dominance columns in monkey striate cortex. *J. Neurosci., 13,* 4114–4129.

Obermayer, K., Blasdel, G. G., & Schulten, K. (1992). Statistical-mechanical analysis of self-organization and pattern formation during the development of visual maps. *Phys. Rev. A, 45,* 7568–7569.

Obermayer K., Ritter, H., & Schulten, K. (1990). A principle for the formation of the spatial structure of cortical feature maps. *Proc. Nat. Acad. Sci. USA, 87,* 8345–8349.

Piepenbrock, C., Ritter, H., & Obermayer, K. (1996). Linear correlation-based learning models require a two-stage process for the development of orientation and ocular dominance. *Neur. Proc. Lett., 3,* 1–7.

Piepenbrock, C., Ritter, H., & Obermayer, K. (1997). The joint development of orientation and ocular dominance: Role of constraints. *Neur. Comp., 9,* 959–970.

Riesenhuber, M., Bauer, H.-U., & Geisel, T. (1996). Analyzing phase transitions in high-dimensional self-organizing maps. *Biol. Cyb., 75,* 397–407.

Ritter, H., Martinetz, T., & Schulten, K., (1992). *Neural computation and self-organizing maps.* Reading, MA: Addison-Wesley.

Ritter, H., & Schulten, K. (1986). On the stationary state of Kohonen's self-organizing sensory mapping. *Biol. Cyb., 54,* 99–106.

Swindale, N. V. (1996). The development of topography in the visual cortex: A review of models. *Network, 7,* 161–247.

Tanaka, S. (1995). Topological analysis of point singularities in stimulus preference maps of the primate visual cortex. *Proc. R. Soc. London B, 261,* 81–88.

Wolf, F., Bauer, H.-U., & Geisel, T. (1994). Formation of field discontinuities and islands in visual cortical maps. *Biol. Cyb., 70,* 525–531.
Wolf, F., Bauer, H.-U., Pawelzik, K., & Geisel, T. (1996). Organization of the visual cortex. *Nature, 382,* 306–307.

Received May 6, 1997; accepted September 18, 1997.