Pre- and Postnatal Development of Topographic Transformations in the Brain

DISSERTATION
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich für Physik
der Johann Wolfang Goethe-Universität
in Frankfurt am Main

von

Urs Michael Bergmann

aus

Ludwigsburg

Frankfurt (2010)

vom Fachbereich für Physik der

Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Dirk-Hermann Rischke

Gutachter: Prof. Dr. Christoph von der Malsburg, Prof. Dr. Jochen Triesch

Datum der Disputation: 4. April 2011

Acknowledgements

First of all, special gratitude is dedicated to my supervisor, Christoph von der Malsburg, for the positive and creative support throughout the development of my thesis. I would like to thank Jochen Triesch, for being my second supervisor and for many interesting discussions. Further, I would like to voice my appreciation for my colleagues at FIAS, with its interdisciplinary background, that inspired and broadened my perspective even beyond the already interdisciplinary field of Neuroscience.

Special thanks goes to Gervasio Puertas, for fruitful collaboration and many interesting discussions on the analytical parts of chapter 2. I am very grateful to Junmei Zhu for the collaboration on turning chapter 4 into a paper. Further, I would like to thank Jenia Jitsev, Gervasio Puertas and Junmei Zhu for proof-reading parts of the thesis. Additional thanks goes to Claudius Gros, Andreas Braun, Dominik Heide, Andreea Lazar, Jan Scholz, Christian Wolff and Philip Wolfrum for many useful and inspiring discussions.

Finally, this thesis would not have been possible without the backing of my parents, and especially my wife Gwendolyn's enduring support, who gave me confidence and love to overcome the more difficult days during my thesis.

Contents

List of Figures

1 Introduction

A long standing mystery surrounding the functioning of the brain is how it manages to handle the enormous variety of different neuronal activity patterns and organize these so that they are related in a meaningful way. These patterns can be sensory, motor, but also internally generated patterns. A major difficulty is the identification of a particular pattern's semantics, because the relation of patterns with the same meaning can be extremely complicated, for example the firing patterns of retinal cells generated by objects seen from a different angle. Patterns with identical meaning can even be structurally unrelated, like many synonyms in language. It is obvious that the brain needs to *transform* patterns from all modalities to a representation that is closer to their semantics than it is to their appearance, so that it can build interrelations with the ability to generalize. In other words: an abstract way of thinking is more powerful than only relating instances of the abstract problem. Moreover, the brain seems to build abstract representations of the world, in the extreme case even in single cells. A prominent example is a cell in the medial temporal lobe that fires exclusively for stimuli containing Jennifer Aniston (Quiroga et al., 2005), independent of the visual appearance, that is, the specific pattern elicited on the retina. In this thesis we present a framework that shows how transformations, which solve the task of generating more abstract representations in the visual domain, can be organized.

A big part of the anatomical diversity and physiological complexity of the brain is already developed at birth. The study of these enormously complex patterns, and how biology is able to organize them, is therefore an interesting research question. Most models for postnatal learning in theoretical neuroscience either presume an ad hoc structure to be already in place or assume a very simple unstructured initial structure. The latter seems theoretically preferable due to Occam's Razor because it is simple. However, this assumption does not correspond to the brain at the time of eye-opening: the brain at this stage has been structured by self-organization, guided by the genetic program that developed over millions of years, to implement specific functions. These functions can be very specific, like primitive reflexes, for example the "rooting reflex" of human babies that assists breastfeeding, or they can be more fundamental and general, like the learning programs that form the basis for reinforcement learning (Sutton and Barto, 1998; Bergmann et al., 2009), a broadly applicable learning paradigm for behaving agents. This genetically driven self-organization process seems to be so precise and powerful, that it is possible that a person who never experienced a sensory phenomenon from a given modality can still perceive this modality. For example, a person who is born without limbs can perceive phantom limbs (Brugger et al., 2000), and a person who was born

blind might experience vision during sleep (Bértolo et al., 2003). In conclusion, the prenatal organization of the brain is essential to its postnatal function and its subsequent development and the initial structure has therefore to be taken into account for the understanding of brain function and postnatal development.

However, there is a problem in the direct applicability of the argument for the design of models for the brain: we only have a rough knowledge and do not know the precise organization of cortical networks at the time of eye-opening. A major goal of this thesis therefore is to describe structures, in particular the before-mentioned transformations, that can be self-organized prenatally from simple local interactions and that are consistent with neuroscientific findings. To this end, we follow the hypothesis of innate learning (e.g. Albert et al., 2008), which assumes that the brain starts to organize and learn already before eye-opening in its postnatal mode. Following this hypothesis, we generally apply the methodology and key ideas of models that have been designed for, and shown to be, postnatally successful to the prenatal stage.

1.1 Invariance in Vision

We constrain our investigations to the visual domain, because the visual parts of the cortex are the best investigated structures in the brain and it is known that much of the primate neocortex is involved in vision; approximately 60% of the macaque cortex. However, there is a long-lasting, still not completely solved open computational question (Pitts and McCulloch, 1947): How is the brain able to recognize stimuli of an object, while the retinal firing patterns it produces vary considerably due to translation, rotation or even more complicated three dimensional transformation? In fact, it is possible that two different objects at the same position are more similar, in the sense of an Euclidean metric on the retinal neural firing patterns, than the same object under different transformations(see e.g. Duda et al., 2001, p.189). As the number of possible appearances times the number of different identities of objects is enormous, it is impossible to store all of them. Further, in this case the system would be unable to generalize to new views of a known object, which our visual system is capable of (e.g. Biederman and Cooper, 1991; Wang et al., 2005, see however Cox and DiCarlo, 2008). This problem is termed the *invariance problem* and many different models for its solution have been proposed in the literature. These can roughly be grouped into two groups (Wiskott, 2006): normalization- and invariant feature-based.

1.1.1 Feature-based Invariance

The general idea of feature-based approaches is to make the response of output neurons invariant with respect to changes in the inputs that correspond to object transformations, for example translation. A possible, and widespread way to achieve this response is to use alternating layers of feature detectors ("simple cells") and pooling cells ("complex cells") (Rosenblatt, 1961; Fukushima, 1980; LeCun et al., 1989; Riesenhuber and Poggio, 1999; Deco and Rolls, 2004). Simple cells differing in the desired invariant parameter (for example position, scale etc.), but with otherwise identical receptive fields, project

to the pooling cells, which hence become invariant through a non-linear response on its inputs, for example a maximum operation (Riesenhuber and Poggio, 1999). Alternatively, the non-linear response functions of output cells can be learned by slowness (Wiskott and Sejnowski, 2002; Einhäuser et al., 2002). The exclusive focus on building invariant responses has the disadvantage that the information of how the invariant representation came about is lost, i.e. in particular *where* the information was pooled from. Resolving ambiguities, however, needs the interaction of higher level representations with lower, less invariant representations (see e.g. Lee and Mumford, 2003). This can be seen for example for segmentation or figure ground segmentation. It seems therefore necessary to explicitly model the processes that lead to the invariant responses.

1.1.2 Normalization

The normalization approach transforms the appearance of an object to a standardized form in its output. For example, if an object is translated or scaled in the input it is actively transformed to a standardized position or scale. The standardized representation then is invariant with respect to the set of transformations and we therefore shall call them invariance transformations. Selection of the necessary transformation can be based on a segmented version of the object in the input, which also allows for learning (Loos and Malsburg, 2002). Alternatively, and more common, is the correspondence-based selection of the transformation: this approach seeks a transformation of the input to the output that minimizes the distance (usually measured with least squares) of the output to a pattern stored in memory (e.g. Lades et al., 1993; Olshausen et al., 1993; Wiskott and von der Malsburg, 1996; Arathorn, 2002; Lücke et al., 2008; Wolfrum et al., 2008). The active seeking of a transformation in these networks has the advantage that information loss is minimal and that important *where-information* is accessible. It has been shown recently that psychophysical data is well in line with the paradigm of coordinate transformations (Graf, 2006).

A particular focus in this thesis is on the development of the necessary invariance transformations that are needed for the normalization approach and chapters 3 and 4 show that simple transformations can already be self-organized before birth. The next subsection introduces the main framework we build upon throughout the thesis, which potentially allows for an integration of both normalization- and feature-based approaches, and hence can benefit from their respective advantages.

1.1.3 Bilinear Models and Modulatory Synapses

The traditional model of a rate-coding neuron is to assume a linearly weighted sum of its inputs, that then gets passed through a non-linear transfer function to yield the neuron's output rate. From a computational perspective it has been argued, that networks with multiplicatory interactions are more efficient in performing complex computation (e.g. Durbin and Rumelhart, 1989; Koch, 1999). We therefore here shortly motivate and introduce the modulatory networks we used in this thesis.

The transfer function of the traditional model implements several biophysical constraints

of a neuron, for example its firing threshold and the saturation frequency of a neuron, which it cannot exceed. In general, however, there are many effects that seem to make this model much too simplistic. For example, different states of neural firing (regular firing or bursting), complex interactions and computations in the dendrite (Häusser et al., 2000), like for example dendritic spikes, and short-term synaptic plasticity, such as depression (Markram and Tsodyks, 1996) or facilitation (Markram et al., 1998) of synaptic transmission, significantly complicate the input-output relation. To abstract from this enormous complexity we therefore assume the neuron to calculate an unknown function of its input:

$$y = f(\boldsymbol{x}), \tag{1.1}$$

where y denotes the neuron's output activity and $\boldsymbol{x}$ is its input activity vector. For simplicity we assume this function to be stationary throughout this thesis. Assuming f to be a twice differentiable real function, its Maclaurin approximation of second order is:

$$\begin{aligned} y &\approx f(\mathbf{0}) + \sum_i x_i \frac{\partial}{\partial x_i} f(\mathbf{0}) + \frac{1}{2!} \sum_i \sum_j x_i x_j \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{0}) + \mathcal{O}(x^3), \\ &= c + \sum_i w_i x_i + \sum_i \sum_j w_{ij} x_i x_j, \end{aligned} \tag{1.2}$$

where in the second line we replaced the gradient of $f(\mathbf{0})$ and its Hessian with the parameters w_i and w_{ij}. From this formula we see that the traditional model corresponds to an approximation up to linear order plus compensating the errors to a real neuron with a nonlinear transfer function. Instead, throughout this thesis we will take the second-order term into account, but for mathematical simplicity will neglect a nonlinear transfer function[1]. Note that the multiplicative second-order term is nonlinear and hence allows for complex computations (Hertz et al., 1991). Each summand in the this term is linear in x_i, if x_j is assumed constant (and vice versa), this is why it is called *bilinear* (Tenenbaum and Freeman, 2000). Similar to the Maclaurin expansion presented here, a Volterra or Wiener expansion can be used to analyze non-linear systems (Dayan and Abbott, 2001). As an example, this technique has been applied to a neuronal chain in the catfish retina and was shown to be superior to linear analysis methods and can be truncated with small error after the second order term (Marmarelis and Naka, 1972).

There is plenty of physiological and theoretical evidence for multiplicative effects. For example, it has been discovered recently that glia cells, which have so far been considered to play a passive role in information processing, can modulate synaptic transmission (Haydon, 2001; Möller et al., 2007). However, this interaction is most likely too slow to account for the fast modulatory effects of equation 1.2. The nonlinear effective transfer function in a leaky integrate and fire (LIF) neuron model has been shown to yield a logarithmic transformation of its synaptic inputs to firing rates (Tal and Schwartz, 1997). The sum of the outputs of several neurons therefore corresponds to a product of their inputs. Similarly, there is direct evidence that specific neurons in locusts perform a logarithm of their inputs, sum these up and exponentiate afterwards, and

[1] An exception is chapter 2 where we use only the linear approximation.

therefore implement a multiplication of synaptic inputs (Gabbiani et al., 2004). Short-term synaptic depression has been shown to have effective multiplicatory effects (Rothman et al., 2009) and shunting inhibition at the same synapse (Volgushev et al., 1996; Kubota et al., 2007) and supra-linear interaction of apical and distal inputs (Larkum et al., 1999; Schaefer et al., 2003) implements multiplications as well. Finally, it has been shown that multiplications can also be implemented at the network level (Salinas and Abbott, 1996).

1.2 Topography

According to von Waldeyer-Hartz's neuron doctrine the brain consists of discrete units, the neurons, and adding the standard assumption that (chemical) synapses are the sole connections between neurons[2], information processing in the brain is based exclusively on neurons and their interconnections. It follows that the locations of neurons do not play a functional role, only their connectivity.

However, and luckily so for modern neuroscientific methods (for example fMRI), it has been found that neurons are arranged systematically. For example, visual responsive cells in the retina project topographically to lateral geniculate nucleus (LGN), that is, neighboring retinal cells innervate neighboring LGN cells. The same holds further downstream for simple cells of primary visual cortex (V1) (Hubel and Wiesel, 1968). This particular ordering of projections leads to a topographic representation that has been termed "*retinotopic* map" and is a special case of topographic maps. Retinotopy can also be found to a lesser degree in extrastriate areas and topographic maps have been found in all sensory systems as well as in many motor systems. Further, there are also functional topographic maps, for example the orientation map in V1, tonotopy maps in auditory cortex (Talavage et al., 2000) or odor maps in olfactory cortex (see Imai et al., 2009), to name but a few.

Why are neurons systematically ordered, despite any obvious functional relevance? Common answers to this question are of anatomic nature: the topographic organization minimizes wiring length, and therefore volume and energy consumption of the brain (Durbin and Mitchison, 1990). This argument holds if the assumption that functionally similar neurons need more connections is valid. Further, it is known that electrical synapses are numerous in the brain (Connors and Long, 2004) leading to local interactions of neurons. Also for synaptic plasticity, it has very recently been found in hippocampus that astrocytes control thousands of excitatory synapses in their vicinity (Henneberger et al., 2010). Both effects lead to statistical dependencies of nearby neurons and together with statistical learning might lead to topographic representations.

There is also a functional theory that explains topography (Hyvärinen, 2002): the general idea is based on the assumption that simple cell activities are used in further computations, for example pooling into complex cells (Hyvärinen et al., 2001), and on the assumption that functionally similar cells need to be considered significantly more often in these computations. Application of any statistical learning, like Hebbian learning or generalizations thereof, on the basis of topographically ordered simple cells, then leads

[2]This is also sometimes considered to be part of the neuron doctrine

to a massively reduced search space, by focusing connections only to a local area. In Topographic Independent Component Analysis (Hyvärinen et al., 2001), simple cells are ordered according to their (residual) statistical dependencies, which naturally leads to topography, simple and complex cell responses.

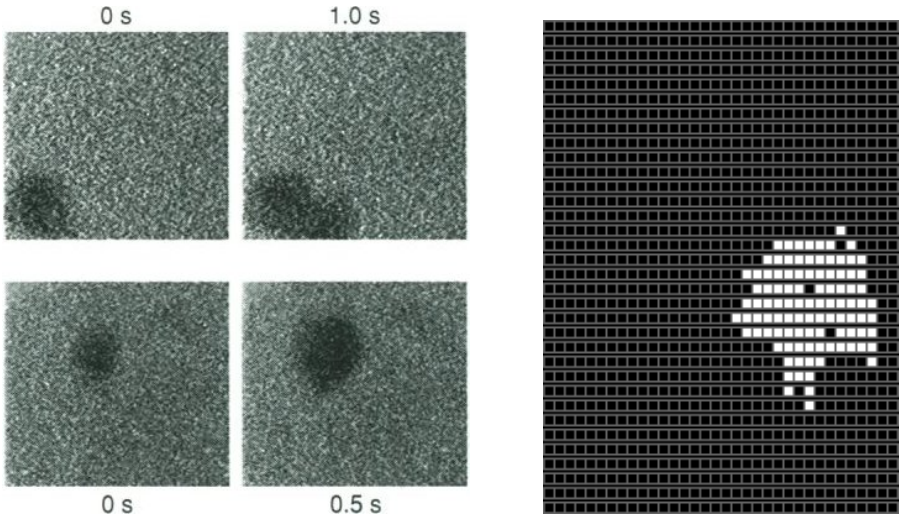
Finally, it has also been shown that topography is a useful representation for fast dynamically changing networks, such as the Dynamic Link Architecture (Lades et al., 1993), which is especially applicable to the subtle differences in face recognition. These networks implicitly encode relations between different input dimensions by preserving relative positions of features, an uncommon idea in statistical neuroscience. This is an idea which we will systematically exploit in chapter 5. For chapters 3 and 4 topography is of crucial importance, because it sets up a coordinate system that can be used to organize real-world transformations prenatally.

1.2.1 Ontogeny of Topography

It has been found that retinotopy in superior colliculus (SC) (Chalupa and Snider, 1998), lateral geniculate nucleus (LGN) (Jeffery, 1985) and primary visual cortex (V1) (Cang et al., 2008) develops before photoreceptors are able to transduce light. Likewise, eye-segregation of inputs in LGN, ocular dominance columns in V1 and even orientation-selectivity in V1 start to develop before vision begins (Huberman et al., 2008).

Historically, there have been two competing theories explaining the prenatal emergence of retinotopy. The chemoaffinity hypothesis was proposed by Roger Sperry (Sperry, 1963) and states that each neuron possesses unique receptors on its projecting fibers and the target neurons posses corresponding unique cytochemical markers (ligands), so that each fiber gets connected to its correct target neuron to establish a topographic map. As the number of chemical signals would be prohibitive large in this model, he suggested gradients of ligands and receptors to code for positions in the input and the output neuronal fields. Indeed, in the recent years corresponding ligands, ephrin family molecules (Drescher et al., 1995) and Wnts (Schmitt et al., 2006) have been identified together with their complementary receptors, Eph and Ryk/Frizzled(s). Both ligands can have attractive or repulsive effects, depending on the receptors on a fiber. Disruptions in map formation have been observed for EphA, EphB and Wnt/ryk signaling knockout mice (Huberman et al., 2008). Interestingly, it has been found that solitary cells find their correct location (Gosse et al., 2008), thus excluding competition effects of cells to be a necessity for rough retinotopy formation.

An alternative activity-based theory has been suggested by Willshaw and von der Malsburg (Willshaw and von der Malsburg, 1976) and is based on the idea of correlated activity of neighboring cells in the retina to code for neighborhood relationships. These input correlations could be interpreted as a prenatal model of the neighborhood correlations in postnatal image statistics (see e.g. Hyvärinen, 2002). Assuming lateral correlations and Hebbian learning of cells in the target can then be used to exploit the input correlations to build a topographic map. An analytically solvable model inspired by this idea is described in more detail in chapter 2 and many similar models have been proposed (see for a review Goodhill, 2007). Indeed, retinal waves impose lateral



(a) (b)

Figure 1.1: Subfigure (a) shows recorded retinal waves from ferrets before eye-opening, taken from (Feller et al., 1996), and subfigure (b) plots simulated retinal waves, that were simulated as described in (Godfrey and Swindale, 2007).

correlations on RGCs (retinal ganglion cells) (see Figure 2.1).

Contemporary interpretations of the ontogeny of retinotopy include both suggested models (Goodhill, 2007; Huberman et al., 2008), although each by itself could describe the emergence independently. It has been shown that the prevention of retinal waves in ephrin knockout ($ephrin - A2/5^{-/-}$) mice leads to the abolishment of topography in the naso-temporal dimension (Pfeiffenberger et al., 2006). Chemoaffinity-based mechanisms precede activity-based refinement of retinotopy and therefore set the stage for the latter mechanism, which otherwise would either need time-dependent parameter tuning (Zhu, 2008) or expensive long-range lateral interactions (see section 2.4).

1.3 Prenatal Waves

For all prenatal models we describe in this thesis, the candidate input patterns are prenatal waves. Therefore, we here give a short review of the physiological findings.

Retinal waves consist of active retinal ganglion cells that often form a simply-connected area of activity but can also consist of several simply-connected areas. They are ignited by the firing of solitary or few spontaneously firing cells that then activate neighboring cells (Wong, 1999). Activated cells burst for a given period of time and subsequently are in a refractory period, when they cannot get activated, even by neighboring activities. This leads to an active area of varying size on the retina, 200μm up to 1mm in an in vitro macaque retina (Warland et al., 2006) that migrates, which is hence called a "wave". Figure 1.1a shows recorded retinal waves from ferrets and Figure 1.1b shows simulations of retinal waves, which we produced using the method described in (Godfrey

and Swindale, 2007). Retinal waves are phylogenetically stable and can be found in both low and high vertebrate species (Wong, 1999). Suppression of retinal waves from P11 to P15 in mice by tetrotoxin (TTX) shows that they are critical for retinotopy refinement in this late ontogenetic phase (Hooks and Chen, 2006) and hence are crucial for correct visual system development. For a more extensive review on retinal waves see (Wong, 1999; Huberman et al., 2008).

Cortical Waves have been investigated less, in comparison to retinal waves. However, cortical calcium waves have been found in slices of rat cortex, that lead to long-range correlations of 8mm and more (Garaschuk et al., 2000), and therefore presumably influence long range connections. Further, in vivo recordings in ferrets showed that synchronous activity bursts with patchy correlations of a mean distance of 1mm (Chiu and Weliky, 2001), suggesting that cortical correlations are a stable phenomenon across different mammalian species. Importantly, the correlations persisted even under blockage of LGN and are therefore generated by cortical circuits.

1.4 Outline of Thesis

In chapter 2, we describe a model for the establishment of retinotopic mappings, that can be solved analytically, due to Gaussian assumptions. Chapter 3 shows that correlation-based learning on prenatal waves can account for the self-organization of transformations necessary for normalization-based object recognition. In chapter 4 we show that for translations, it is possible to organize the connectivity structure also for multilayer circuits. Finally, chapter 5 derives a probabilistic model that allows for the simultaneous learning of transformations *and* features in a bilinear setting.

1.5 Notational Conventions and List of Symbols

x_i	activity of input unit i
N_i	size of input neuronal field
y_o	activity of output unit o
N_o	size of output neuronal field
c_k	activity of control unit k
z_l	activity of memory or second output layer unit l
I_{ni}	active region resembling a prenatal wave (i is the input position, n indexes different regions)
$\hat{x}_i, \hat{y}_o, \hat{c}_k$	inputs to units i, o and k
w_{oi}	forward weight from input unit i to output unit o
g_{io}	generative weight from output unit o to input unit i
w_{koi}	forward bilinear weights for control unit k
$\tilde{w}_{ki}$	effective input weight footprint of control unit k
g_{iok}	generative bilinear weights for control unit k
$\boldsymbol{\xi}$	vector of uncorrelated noise
$\mathcal{N}(\boldsymbol{x}, \boldsymbol{x}_0, \Gamma)$	Gaussian function centered around $\boldsymbol{x}_0$ with covariance matrix Γ
α	unspecific weight growth rate
$C_{ij}^{I/O}$	effective lateral interaction of units i and j in the input (I) or output (O) layer
$C_{oo'ii'}$	coupling matrix for weights w_{oi} and $w_{o'i'}$
F_{oi}	weight cooperation term for the weight w_{oi}
B_{oi}	competition term for the weight w_{oi}
ω	relative weighting of input and output based competition
κ_k	gain modulating variable of control unit k
$\zeta(t)$	stimulus specificity at time t
$s(k)$	average synaptic spread of control unit k
$S(k)$	scale factor of the transformation implemented by control unit k
$r(k, o)$	synaptic center-of-mass for control unit k and output unit o
ν	frequency or wave number

2 A Gaussian Generative Model for the Ontogeny of Retinotopy

In this chapter, we present a novel model for the emergence of topographic mappings and establish a first link to information processing. In section 2.2 we analytically show all possible solutions to the system. Further, we derive equations for the estimation of the hidden variables in section 2.3, as well as a learning rule for the weights. Using these equations we show that an implementation of the model reliably yields the expected results.

2.1 Introduction

Modeling of topographic mappings by now has a long tradition in computational neuroscience and dates back more than 30 years (Willshaw and von der Malsburg, 1976), see section 1.2.1. Like many models for topography formation (Goodhill, 2007), we formulate an activity-based model that makes use of lateral correlations in the input and the output neural fields.

In contrast to most modeling studies of topography, we shall however take a different approach. Our goal here is not to directly model the emergence of topography, but instead we build a simple model that attempts to build a good code for the input statistics it receives. Topography then is a result, if reasonable neurophysiological constraints, i.e. long-range excitatory connections or positive firing rates, are taken into account.

Methodologically we therefore build on top of the both technically and neuroscientifically successful class of generative models, which we introduce taking vision as an example: Why is computer graphics so advanced these days that it can be hard to tell apart real and rendered pictures and the reverse problem, computer vision, is still far from competing with the visual systems of animals? One reason why the vision problem is difficult is because we only receive a two-dimensional input from the three-dimensional world, and hence inputs are necessarily ambiguous. Therefore, any visual system needs prior information to interpret the inputs. Visual illusions illustrate that prior information can lead to wrong interpretations, even in humans. The generation of inputs from an internal model on the other hand, like in computer graphics, is fully determined and hence simpler. In probabilistic terms, this generation process is formalized with a generative distribution:

$$p(\boldsymbol{x}|\boldsymbol{y};\mathcal{G}), \tag{2.1}$$

where $\boldsymbol{x}$ here denotes the input and the y_i are called latent variables or causes. $\mathcal{G}$ is a set of parameters of the model. Further, prior information on $\boldsymbol{y}$ is usually incorporated in the model: $p(\boldsymbol{y}|\mathcal{G})$, a term, that is unconditional with respect to $\boldsymbol{x}$. **Recognition** in this type of model then amounts to the estimation of the latent variables $\boldsymbol{y}$ given an input. Luckily, this reversal of probabilities is mathematically given by Bayes' theorem:

$$p(\boldsymbol{y}|\boldsymbol{x};\mathcal{G}) = \frac{p(\boldsymbol{x}|\boldsymbol{y};\mathcal{G})p(\boldsymbol{y}|\mathcal{G})}{p(\boldsymbol{x}|\mathcal{G})}. \tag{2.2}$$

Thus, we have a way to relate a comparatively simpler generative model with a recognition model[1].

The marginal distribution of a generative model,

$$p(\boldsymbol{x}|\mathcal{G}) = \int p(\boldsymbol{x}|\boldsymbol{y};\mathcal{G})p(\boldsymbol{y}|\mathcal{G})d^N y, \tag{2.3}$$

gives the probability distribution of the generated inputs, if sampled according to the prior $p(\boldsymbol{y}|\mathcal{G})$, and hence is the unconditioned firing distribution of the upstream units. A faithful model therefore should approximate the real input distribution $p(\boldsymbol{x})$ as close as possible by adapting its parameters $\mathcal{G}$. This is the key idea of **representational learning**:

$$p(\boldsymbol{x}|\mathcal{G}) \stackrel{!}{=} p(\boldsymbol{x}). \tag{2.4}$$

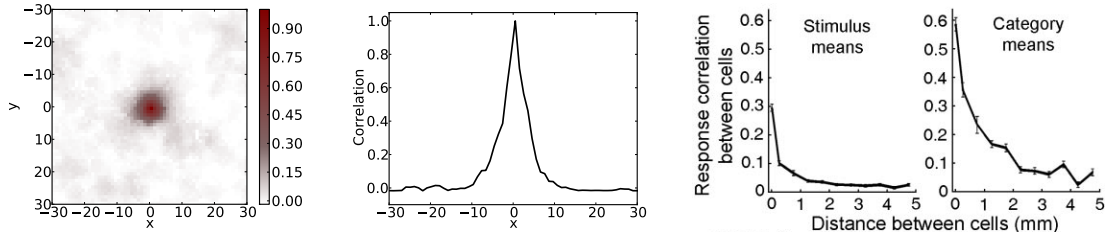
Using this idea, we show that the parameters $\mathcal{G}$ of a simple Gaussian model (one that takes only statistics up to second order into account) become similar to topographic maps in the brain.

2.2 An Analytically Solvable Model of Retinotopy

Neighborhood correlations are a widespread phenomenon in the brain and can be found subcortically, e.g. in the immature ferret lateral geniculate nucleus (LGN) (Ohshiro and Weliky, 2006), and in all cortical regions. Figure 2.1 shows two examples in the visual domain: the two-dimensional Pearson correlation coefficient[2] has been calculated from neuronal activities of simulated retinal waves, as described in section 1.3, see Figure 2.1a, and a one-dimensional cut through its maximum is shown in Figure 2.1b. Qualitatively, the shape of the correlation function can be seen to be similar to the correlations shown for postnatally recorded activities in inferotemporal cortex (IT), see Figure 2.1c (this plot is taken from Kiani et al., 2007). As even these very different neuronal fields yield similar second-order statistics, we therefore in the following make the reasonable assumption, that the lateral correlations are the same, or very similar, on all levels. We formalize the

[1] Unfortunately, a formula for the probabilities of the hidden variables, $p(\boldsymbol{y}|\boldsymbol{x};\mathcal{G})$, can usually not be given in closed form even for moderately complex systems, due to unsolvable integrals.

[2] The Pearson correlation coefficient is defined as the covariance of two variables divided by their respective standard deviations.



(a) (b) (c)

Figure 2.1: The Pearson correlation coefficient calculated from activities of prenatal retinal waves is shown in 2D in (a) and a one-dimensional cut in (b), whereas (c) shows the correlation from postnatally recorded data in inferotemporal cortex (taken from Kiani et al., 2007). Both share a monotonic decrease from the center, which we modeled with an exponential function.

input distribution with a multivariate Gaussian probability density function (pdf):

$$\begin{aligned} p(\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{x}, \boldsymbol{x}_0, \Gamma), \\ &= \frac{1}{\sqrt{\det 2\pi\Gamma}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^T \Gamma^{-1} \left(\boldsymbol{x}-\boldsymbol{x}_0\right)\right), \end{aligned} \tag{2.5}$$

where $\boldsymbol{x}_0$ denotes the center of the Gaussian and Γ is a real symmetric positive-semidefinite matrix that parameterizes $\boldsymbol{x}$'s covariances (i.e. the lateral correlations of the input neuronal layer). We assume the input to be homogeneous, i.e. all components of $\boldsymbol{x}_0$ are equal to a constant. This corresponds to the same average firing activity of all neurons in the input layer.

Standard second-order statistical methods, like probabilistic principal component analysis (pPCA) or factor analysis (FA) assume uncorrelated causes to build an efficient and non-redundant code of their inputs (see e.g. Dayan and Abbott, 2001). This is inconsistent with the finding of neighborhood correlations in the activities of neurons on all levels in the cortex, which implies redundant coding. We therefore explicitly model correlations in the prior probability distribution of the hidden variables. For simplicity we assume that the dimensionality of the latent variables is equal to the input dimensionality ($\dim \boldsymbol{y} = \dim \boldsymbol{x}$), and as discussed above, we assume the same distribution with the same covariances for the latent variables $\boldsymbol{y}$ as for the inputs $\boldsymbol{x}$, equation 2.5:

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}, \boldsymbol{y}_0, \Gamma), \tag{2.6}$$

where like for the input $\boldsymbol{x}_0$, we here also assume homogeneity of $\boldsymbol{y}_0$ with the same average firing activity as in the input. Note that for $\Gamma = I$ the prior of this model is identical to the priors of both pPCA and FA.

To complete the generative model, we need to define a generative distribution, equation 2.1. In this chapter, we focus on the simple and common approach (see e.g. Olshausen and Field, 1996; Bell and Sejnowski, 1997) that uses a linear generative model for the

magnitudes of upstream units:

$$\boldsymbol{x} = G\boldsymbol{y}, \tag{2.7}$$

where $G = (g_{io})$ is a matrix[3] containing the generative weights of the model. Hence, given the $\boldsymbol{y}$ activities, the input is determined by equation 2.7 and we have a deterministic generative model. The corresponding generative distribution is given by:

$$p(\boldsymbol{x}|\boldsymbol{y}; G) = \delta(\boldsymbol{x} - G\boldsymbol{y}). \tag{2.8}$$

In the following, we will identify the generative matrix G with the parameters $\mathcal{G}$, as the matrix G contains all parameters of the model.

Using equations 2.6 and 2.8 the integral appearing in the calculation of the marginal distribution, equation 2.3, can be solved in closed form:

$$\begin{aligned} p(\boldsymbol{x}|G) &= \int_{-\infty}^{\infty} \delta(\boldsymbol{x} - G\boldsymbol{y})\, \mathcal{N}(\boldsymbol{y}, \boldsymbol{y}_0, \Gamma)\, d^N y \\ &= \frac{1}{\sqrt{\det 2\pi\Gamma}} \exp\left(-\frac{1}{2}\left(\tilde{G}\boldsymbol{x} - \boldsymbol{y}_0\right)^T \Gamma^{-1} \left(\tilde{G}\boldsymbol{x} - \boldsymbol{y}_0\right)\right), \end{aligned} \tag{2.9}$$

where $\tilde{G}$ is the inverse of G. Rearranging terms the marginal becomes:

$$\begin{aligned} p(\boldsymbol{x}|G) &= \frac{1}{\sqrt{\det 2\pi\Gamma}} \exp\left(-\frac{1}{2}\left(\boldsymbol{x} - G\boldsymbol{y}_0\right)^T \tilde{G}^T \Gamma^{-1} \tilde{G}\left(\boldsymbol{x} - G\boldsymbol{y}_0\right)\right) \\ &= \sqrt{\det GG^T}\, \mathcal{N}(\boldsymbol{x}, G\boldsymbol{y}_0, G\Gamma G^T), \end{aligned} \tag{2.10}$$

and is therefore proportional to a Gaussian distribution.

To estimate the parameters of the generative model, we invoke the representational learning idea, equation 2.4:

$$\begin{aligned} p(\boldsymbol{x}) &\stackrel{!}{=} p(\boldsymbol{x}|G), \\ \mathcal{N}(\boldsymbol{x}, \boldsymbol{x}_0, \Gamma) &\stackrel{!}{=} \sqrt{\det GG^T}\, \mathcal{N}(\boldsymbol{x}, G\boldsymbol{y}_0, G\Gamma G^T). \end{aligned} \tag{2.11}$$

The probability distributions of equation 2.11 can only be equal if $\det GG^T = 1$, if the means are equal and if the covariance matrices are equal. For non-vanishing input and latent means $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$, which is necessarily the case for the non-negative firing rates of real neurons, it follows from $G\boldsymbol{y}_0 \stackrel{!}{=} \boldsymbol{x}_0$, and with the homogeneity assumptions for the input and the output, that the generative matrix is normalized:

$$\sum_o g_{io} \stackrel{!}{=} 1, \quad \forall i. \tag{2.12}$$

The equality of the covariance matrices turns the determination of the parameters G into an algebraic problem:

$$G\Gamma G^T \stackrel{!}{=} \Gamma, \tag{2.13}$$

[3]This matrix is also called mixing matrix in independent component analysis.

which means that the covariance matrix Γ is congruent to itself with the matrix G. In the following, we shall assume, that G is non-negative, because for neuronal systems long-range projections are commonly from pyramidal cells, and are hence excitatory[4]. Further, non-negativity constraints have turned out as a powerful ingredient to the learning of part-based representations (Lee and Seung, 1999).

For **the special case** $\Gamma \propto I$ it follows from equation 2.13 that G is orthogonal, that means in particular:

$$\sum_o g_{io}^2 = 1, \tag{2.14}$$

and hence $g_{io}^2 \leq 1$. In the following we use the term stochastic matrix: a matrix is stochastic, if it consists exclusively of real non-negative values and if its rows are normed to 1. G therefore is an orthogonal stochastic matrix (see equation 2.12).

Theorem 1. *An orthogonal stochastic matrix G needs to be a permutation matrix.*

Proof. We proove the theorem by contradiction. Let us concentrate on an arbitrary row i of G and assume that $g_{io}^2 < 1, \forall o$, inconsistent with the assumption of permutation matrices. Due to non-negative g_{io}, we have $g_{io}^2 \leq |g_{io}|, \forall o$. As the rows are normalized, equation 2.12, at least one entry o in row i must be positive and hence this implies $g_{io}^2 < |g_{io}|$ for this particular o. From equations 2.12 and 2.14 follows the contradiction:

$$1 = \sum_o g_{io}^2 < \sum_o |g_{io}| = \sum_o g_{io} = 1. \tag{2.15}$$

Therefore, there is a k for which $g_{ik}^2 = 1$ and $g_{io} = 0, \forall o \neq k$. As we chose i to be an arbitrary row, this must hold for all rows and since G is invertible, all rows of G have to have a 1 at a different position, and hence G is a permutation matrix. □

Theorem 2. *For **the general case of an arbitrary covariance matrix** Γ, G is orthogonal.*

Proof. Let us first derive a characterization of all solutions of equation 2.13. The matrix Γ is an invertible covariance matrix and hence symmetric and positive-definite. It can therefore be eigen-decomposed:

$$\Gamma = Q \Lambda Q^T, \tag{2.16}$$

where $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_N)$ is a diagonal matrix with the real positive eigenvalues of Γ on its diagonal and Q is an orthogonal matrix. We substitute Γ for its eigen-decomposition, equation 2.16, in equation 2.13:

$$\begin{aligned} GQ\Lambda Q^T G^T &= Q \Lambda Q^T, \\ Q^T G Q \Lambda Q^T G^T Q &= \Lambda, \end{aligned} \tag{2.17}$$

[4]Note that there are exceptions to the rule of excitatory projection neurons, e.g. the GABAergic cerebellar Purkinje cell (Purves et al., 2004, p.143). However, this cell is in the cerebellum and the current theory focuses on connections between retina, LGN and cortex.

where for the second line we multiplied with Q^T from the left and Q from the right and made use of the orthogonality of Q. As Γ is positive-definit, Λ contains only positive eigenvalues. We therefore substitute $\Lambda = \sqrt{\Lambda}\sqrt{\Lambda}$ and multiply with $\sqrt{\Lambda}^{-1}$ from the left and right side:

$$\begin{aligned} \sqrt{\Lambda}^{-1} Q^T G^T Q \sqrt{\Lambda}\sqrt{\Lambda} Q^T G Q \sqrt{\Lambda}^{-1} &= I, \\ H^T H &= I, \end{aligned} \tag{2.18}$$

where we defined $H = \sqrt{\Lambda} Q^T G Q \sqrt{\Lambda}^{-1}$, which is orthogonal and hence has a full set of eigenvalues, all of which have an absolute value of 1. As G is similar to H (with the similarity transformation $Q\sqrt{\Lambda}^{-1}$) it has the same eigenvalues.

Theorem 1 showed that in case of orthogonality, a stochastic G must be a permutation matrix. A permutation matrix is orthogonal. Therefore, the statement that G is orthogonal is equivalent to the statement of G being a permutation matrix. We will now show that if G is not a permutation matrix, it violates the constraint that all eigenvalues need to have an absolute value of 1 and hence G must be orthogonal.

A matrix G that differs from a permutation matrix must differ in at least one row-vector $\boldsymbol{v}$, with $0 \leq v_i < 1, \forall i$. Further, the normalization property of G means that $\sum_i v_i = 1$. Let's now construct the matrix M that does not contain any entries of 1 anymore, by cutting out each column and row of G which contains a 1. This is illustrated in the following example:

$$G = \begin{pmatrix} & & \boldsymbol{v^1} & & & \\ & & \boldsymbol{v^2} & & & \\ 0 & \cdots & 1 & \cdots & 0 & 0 \\ & & \cdots & & & \\ 0 & \cdots & 0 & \cdots & 1 & 0 \\ & & \vdots & & & \\ & & \boldsymbol{v^k} & & & \end{pmatrix} = \begin{pmatrix} V_1^1 & x & V_2^1 & x & V_3^1 \\ 0 & 1 & 0 & 0 & 0 \\ V_1^2 & x & V_2^2 & x & V_3^2 \\ 0 & 0 & 0 & 1 & 0 \\ V_1^3 & x & V_2^3 & x & V_3^3 \end{pmatrix}, \tag{2.19}$$

with $\boldsymbol{v}^i$ being an arbitrary vector with the same constraints as $\boldsymbol{v}$ and x denotes an arbitrary value. For this exemplary case the matrix M is:

$$M = \begin{pmatrix} V_1^1 & V_2^1 & V_3^1 \\ V_1^2 & V_2^2 & V_3^2 \\ V_1^3 & V_2^3 & V_3^3 \end{pmatrix}. \tag{2.20}$$

Note that for an eigenvector $\boldsymbol{w} = (\boldsymbol{w}^1, \boldsymbol{w}^2, \boldsymbol{w}^3)$ of M (with the size of the vectors $\boldsymbol{w}^i$ being equal to the size of the blocks V) the vector $(\boldsymbol{w}^1, 0, \boldsymbol{w}^2, 0, \boldsymbol{w}^3)$ is an eigenvector of G with the same eigenvalue. Hence, if M violates the eigenvalue constraint, so does G.

If M is reducible, that is, if there exists a permutation matrix P such that M can be transformed to an upper triangular form:

$$\tilde{M} = PMP^T = \begin{pmatrix} M_1 & M_2 \\ 0 & M_3 \end{pmatrix}, \tag{2.21}$$

where M_1, M_2 and M_3 are non-trivial (each is of size bigger than or equal to 1) square matrices. Then we only need to investigate the matrix M_1, as $\tilde{M}$ and M share their eigenvalues (they are similar) and hence the eigenvector $\boldsymbol{w} = (\boldsymbol{w}^1, 0)$, where $\boldsymbol{w}^1$ is an eigenvector of M_1, is an eigenvector of M with the same eigenvalue. If the square matrix M_1 is a single value, $(1, \cdots, 0)$ is an eigenvector with eigenvalue $M_1 < 1$! If M_1 is not a single value, but reducible, we recursively perform the reduction exemplified with M. If M_1 is not reducible and not a single value, we can employ the Perron-Frobenius theorem for irreducible matrices:

1. the matrix M_1 has a so called Perron-Frobenius eigenvalue r, which is positive and real

2. the number h of eigenvalues with absolute value of r is called the period of M_1. These eigenvalues are given as $\lambda_l = re^{\frac{2\pi il}{h}}$ $(l = 0, \cdots, h-1)$ and are simple roots of the characteristic polynomial of M_1.

For the case we consider, $r = 1$ as all eigenvalues of G need to have an absolute value of 1. For the same reason M_1 needs to have full periodicity, that is $h = K$ with K being the size of matrix M_1. The characteristic polynomial of the matrix M_1 therefore is:

$$\lambda^K = 1. \tag{2.22}$$

The Cayley-Hamilton theorem states that every square matrix (over a commutative ring) satisfies its own characteristic polynomial. Application of the theorem yields:

$$M_1^K = I. \tag{2.23}$$

We now use that, by construction, the sum over the rows of M_1 is smaller or equal to 1 and all entries are non-negative and *strictly* below one to show that the equality of equation 2.23 is impossible. Consider first the quadratic case (we write M instead of M_1 for simplicity of notation from now on):

$$M_{ij}^2 = (MM)_{ij} = \sum_l M_{il} M_{lj} < \sum_l M_{il} \leq 1. \tag{2.24}$$

As $M_{ij}^2 < 1$, we can apply the same argument to get $(MM^2)_{ij} < 1$. We see via induction that therefore $M^K = I$ is impossible.

Therefore, if G is not orthogonal, at least one eigenvalue has an absolute value different from 1, in contradiction with the similarity of G to the orthogonal matrix H. □

Theorem 3. *If Γ is a non-degenerate matrix all solutions G are symmetric permutation matrices.*

Proof. Theorem 2 showed that G is orthogonal. Multiplying equation 2.13 with G from the right hand side therefore yields:

$$G\Gamma = \Gamma G, \tag{2.25}$$

hence the commutator of G and Γ vanishes, $[G, \Gamma] = 0$. We substitute the eigendecomposition for Γ:

$$\begin{aligned} GQ\Lambda Q^T &= Q\Lambda Q^T G, \\ \tilde{G}\Lambda &= \Lambda\tilde{G}, \end{aligned} \tag{2.26}$$

where we defined the matrix $\tilde{G} = Q^T G Q$. For non-degenerate Γ, Λ is a diagonal matrix with mutually different entries on its diagonal. Hence $\tilde{G}$ is an arbitrary diagonal matrix. The orthogonality constraint of $\tilde{G}$, $\tilde{G}\tilde{G}^T = 1$, implies for the square of its diagonal elements to equal unity. $\tilde{G}$ is therefore a signature matrix, a diagonal matrix with ± 1 on its diagonal. It is easy to see that signature matrices are involutory matrices:

$$\tilde{G}\tilde{G} = I, \tag{2.27}$$

which means that G needs to be involutory as well, as

$$Q^T G Q Q^T G Q = I \;\Leftrightarrow\; GG = I. \tag{2.28}$$

An involutory orthogonal matrix is symmetric. □

Taking together the results of theorem 1 and theorem 2, G in general must be a permutation matrix. Substituting a permutation P for G in equation 2.13, we get

$$P\Gamma P^T \stackrel{!}{=} \Gamma, \tag{2.29}$$

and see that all P which leave Γ invariant under exchange of rows and columns by applying P, are possible solutions. Note that the topographic identity solution $\Gamma = I$ is always a solution and it is a unique solution, precisely if no other permutation fulfills the congruency invariance of Γ.

The analytic result presented leaves open if the method works for differently sized input and output fields. In the next subsection we therefore show with simulations that the model indeed converges to topographic solutions in a probabilistic model, even without the constraints on G from this section.

2.3 The Probabilistic Topographic Model

In this section we provide simulations of the proposed generative model for retinotopy formation. To this end, the generative distribution needs to be probabilistic, as will be seen in the following. We therefore generalize equation 2.7, so that the magnitudes of upstream neurons are given by:

$$\boldsymbol{x} = G\boldsymbol{y} + \boldsymbol{\xi}, \tag{2.30}$$

where we assume ξ_i to be uncorrelated Gaussian noise with variance σ^2, i.e. $\sigma_i = \sigma,\ \forall i$. The corresponding generative distribution is a multivariate Gaussian:

$$p(\boldsymbol{x}|\boldsymbol{y}; G) = \mathcal{N}(\boldsymbol{x}, G\boldsymbol{y}, \sigma^2 I) \tag{2.31}$$

$$= \frac{1}{(2\pi\sigma^2)^{N_x/2}}\ \exp\left(-\frac{(\boldsymbol{x}-G\boldsymbol{y})^T(\boldsymbol{x}-G\boldsymbol{y})}{2\sigma^2}\right). \tag{2.32}$$

To complete the model, we take the same prior on the latent variables $\boldsymbol{y}$ as in the previous section, equation 2.6. We also introduce an optional Laplacian prior on the parameters:

$$p(G) = \prod_{io} \frac{\alpha}{2} \exp\left(-\alpha|g_{io}|\right), \tag{2.33}$$

which is shown in the results section 2.3.1 to have a similar effect as constraining the weight matrix to be non-negative, like in the previous section. This prior forces the weights to be sparse, similar to forcing sparse activities in cortex (Olshausen and Field, 1996)[5]. In qualitative accordance with the suggested prior $p(G)$, the probability distribution of synaptic weights in visual cortex has been shown to be heavy-tailed (Song et al., 2005).

For the simulations, the representational learning idea (equation 2.4) has to be formulated to work online, i.e. the parameters $\mathcal{G}$ of the model have to be modified at each presentation of an input $\boldsymbol{x}$ independently, to finally converge to a solution of equation 2.4. We derive an online learning rule by minimizing the Kullback-Leibler divergence, which measures the difference of two probability distributions:

$$D_{KL}(p(\boldsymbol{x}), p(\boldsymbol{x}; G)) = \int p(\boldsymbol{x})\ \ln\frac{p(\boldsymbol{x})}{p(\boldsymbol{x}; G)}\ d^N x \tag{2.34}$$

$$= -\int p(\boldsymbol{x})\ \ln p(\boldsymbol{x}; G)\ d^N x + \text{const.}, \tag{2.35}$$

where the second term is independent of the parameters and hence considered constant. As inputs in an online simulation are sampled from the distribution $p(\boldsymbol{x})$, we can neglect the integral and directly *maximize the log likelihood*:

$$\mathcal{L} = \ln p(\boldsymbol{x}; G) \tag{2.36}$$

$$= \ln \int p(\boldsymbol{x}|\boldsymbol{y}; G)\, p(\boldsymbol{y})\, p(G)\, d^N y. \tag{2.37}$$

In the following we want to identify the estimates of $\boldsymbol{y}$ with actual neural responses, and therefore we assume that actual latent variables $\hat{\boldsymbol{y}}$ are at the *maximum a posteriori* probability[6]. Hence we approximate the integral in equation 2.37 by evaluating it at its maximum, which is a standard approach if the integral is intractable (e.g. Olshausen

[5] For an explanation see chapter 5.

[6] Alternative theories of neural coding could however be implemented. For example, particle filtering or alike (Lee and Mumford, 2003).

and Field, 1996; Karklin and Lewicki, 2009). The maximum a posteriori estimation $\hat{\boldsymbol{y}} = \text{argmax}_{\boldsymbol{y}} \tilde{\mathcal{L}} = \text{argmax}_{\boldsymbol{y}} \ln p(\boldsymbol{x}|\boldsymbol{y}; G)\, p(\boldsymbol{y})\, p(G)$ can be written in closed form:

$$\nabla_{\boldsymbol{y}} \tilde{\mathcal{L}} = \frac{G^T \boldsymbol{x} - G^T G \boldsymbol{y}}{\sigma^2} + \Gamma^{-1} (\boldsymbol{y}_0 - \boldsymbol{y}) \stackrel{!}{=} 0, \tag{2.38}$$

$$\hat{\boldsymbol{y}} = \left[G^T G + \sigma^2 \Gamma^{-1}\right]^{-1} \left(G^T \boldsymbol{x} + \sigma^2 \Gamma^{-1} \boldsymbol{y}_0\right). \tag{2.39}$$

Note that this estimation is only influenced by Γ for non-vanishing noise $\sigma \neq 0$.

For each presentation of an input, the generative weights are then adapted by moving them along the gradient of the log likelihood at the estimate $\hat{\boldsymbol{y}}$:

$$G(t+1) = G(t) + \eta \frac{d\tilde{\mathcal{L}}}{dG} \tag{2.40}$$

$$\frac{d\tilde{\mathcal{L}}}{dG} = \sigma^{-2} (\boldsymbol{x} - G\hat{\boldsymbol{y}})\, \hat{\boldsymbol{y}}^T - \alpha \,\text{sgn}\,(G), \tag{2.41}$$

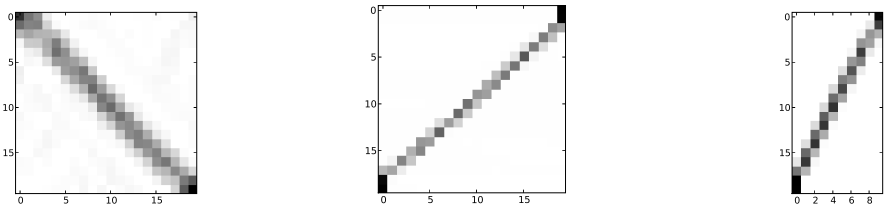
where $\text{sgn}(X)$ is the signum function and returns a matrix with the sign of its elements. η is a learning rate and $\alpha\sigma^2$ parameterizes the relative strength of the generative power of the model (equation 2.31) and the sparseness of the generative weights (equation 2.33). Interestingly, for $\alpha = 0$, the learning rule in equation 2.41 is equivalent to the Oja learning rule (Oja, 1989), which is known to lead to weights that span the subspace (in case of lower output than input dimensionality) that the weights with highest variance in PCA span[7]. Note that the topography prior on the latent variables $\boldsymbol{y}$ (equation 2.6), which correlates neighboring latent variables as given by the hyperparameter Γ, does not enter the gradient in equation 2.41 (the first term results from the generative distribution, equation 2.31, and the second from the Laplacian prior on the parameters, equation 2.33). However, we see from equation 2.39 that the hyperparameter Γ enters proportionally to the variance of the probabilistic generative model. A deterministic model, like the one from the previous section, corresponds to vanishing variance, and hence vanishing influence of the topography prior, if the simulations are performed with the derived equations. Hence, for the simulations, a non-vanishing variance is required for the model to develop topography.

2.3.1 Simulations

For the simulations, we sampled an input $\boldsymbol{x}_t$ from the multivariate Gaussian distribution, equation 2.5, for each iteration t. The generative weights were initialized to uniform random values in $[0, 0.5]$. Maximum a posteriori estimates for $\boldsymbol{y}_t$ were then calculated for each input using equation 2.39 and the weights subsequently updated according to equation 2.41.

Parameters. For all simulations shown, we used an input noise of $\sigma = 1$, a constant learning rate $\eta = 0.02$ and an input field of 20 units, while the size of the output field

[7] In contrast to the Oja network, G in our system is a generative weight matrix and not a feed-forward linear transformation.



(a) (b) (c)

Figure 2.2: Resulting mappings after $t = 80000$ iterations. The y-axis in each subfigure indexes the input domain and the x-axis the output domain. Subfigure (a) shows the result for equally sized input and output neuronal fields, mean-free Gaussians in the input and the prior of the output ($\boldsymbol{x}_0 = \boldsymbol{y}_0 = 0$) and a non-negativity constraint on the generative matrix. In subfigure (b) faster convergence is observed when not enforcing non-negativity but weight sparseness $\alpha = 0.01$. Subfigure (c) shows that the system generalizes to differently sized input and output fields.

was varied. The hyperparameter matrix Γ was set to:

$$\Gamma_{i,j} = \exp(-\gamma|i - j|), \tag{2.42}$$

to qualitatively fit the shape of the measured correlations in Figure 2.1. The precise shape turns out to be not important, as long as a monotonic decrease is present. For example, simulations with a Gaussian in equation 2.42 yielded similar results. The lateral extend of the correlations was set to $\gamma = \sqrt{5}/N$, with N being the size of the input or output field. Note that for bigger γ values (corresponding to a smaller range of the lateral correlations) the system is not guaranteed to yield a consistent topographic mapping, i.e. the weight matrix might converge to local minima with only piecewise topographic solutions. Smaller γ values, on the other hand, yield consistent topographic mappings, but the system's convergence to the final diagonal is slower. Results on varying $\boldsymbol{y}_0$, α and restricting the generative weights to be positive are described in the next paragraph.

Results. Figure 2.2 shows the resulting generative weight matrix at $t = 80000$ iterations. For vanishing mean ($\boldsymbol{x}_0 = \boldsymbol{y}_0 = 0$) of the Gaussians and clamping all weights to non-negative values after each learning step, the weight matrix is topographic, see Figure 2.2a, yet not fully converged to a one-to-one mapping. Setting the means $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$ to positive values significantly speeds up convergence, as then the linear superposition in equation 2.30 is (mostly) additive and no negative effects can help in the reconstruction of the inputs. A similar faster convergence can be seen in Figure 2.2b, where we did not restrict the weights to be non-negative, but alternatively set the sparseness constraint on the weights to $\alpha = 0.01$. Without a sparseness constraint, the weights converge to arbitrary orthogonal vectors spanning the PCA subspace. The most sparse vectors spanning this subspace consist of only one non-zero entry, thus constraining the estimation enormously.

Comparing Figure 2.2a and 2.2b, we see that two different solutions are possible. As

mentioned above (equation 2.29), multiple solutions are related to permutation invariances of Γ and in this case reflect the inversion symmetry $\Gamma_{i,j} = \Gamma_{-i,-j} = \Gamma_{N-i+1,N-j+1}$ (see equation 2.42; N is the size of the covariance matrix).

Finally, Figure 2.2c shows the result for a smaller output size, $N_o = 10$, relative to the input, N_i. The simulation shown used positive means and non-negative weights, but sparseness constraints yield similar results. This demonstrates that the system can handle different scales, which is of particular importance in chapter 3.

2.4 Discussion

In this chapter we derived a model for the formation of topographic mappings by employing probabilistic methods. Assuming all probability distributions to be Gaussians, the model can be solved analytically. To our knowledge, this is the first model that offers direct analytical insight in the solution of a topographic model. However, Häussler (Häussler and von der Malsburg, 1983) offered a detailed mathematical analysis of retinotopy formation for periodic boundary conditions. In contrast, the presented account does not need this biologically unrealistic assumption and is mathematically simpler and more compact. Further, it establishes a link to the necessary requirements (positive weight matrix or sparse weight matrix) and to permutation invariances in the covariance structure of the input and the output fields.

The performed simulations revealed that long-ranging lateral correlations need to be imposed to both the input and the output neuronal fields for a consistent topographic mapping to emerge. This finding is consistent with the Häussler theory, see chapter 3, and seems therefore to be necessary for activity-based topographic mappings. However, long-range lateral connections are expensive. Nature therefore saves resources by preorganizing a rough topographic map using chemical gradients (Sperry, 1963; Huberman et al., 2008) and only in a later stage refines these mappings with activity-based mechanisms similar to the proposed method. An alternative approach to long-range lateral correlations has been suggested in (Zhu, 2008), where a fine-tuning of unspecific synaptic growth has been used to organize consistent mappings. However, compared to the biological solution, still more resources are needed, as a full all-to-all connectivity needs to be established initially, while in biology single synapses have been shown to find their targets (Cang et al., 2008) and even get sorted during migration (Imai et al., 2009).

The current model was motivated from both information theoretical ideas (building an efficient code of the input data) and neurophysiological evidence, for example the lateral correlations in the output. From the purely information theoretic perspective it remains elusive why the model should build a topographic code. A potential advantage might be redundant coding of the inputs to compensate for the high noise in the firing of cortical neurons (London et al., 2010). Indeed, a reduction of information content due to second-order correlations has been observed in the retina (Puchalla et al., 2005), primary visual cortex (Reich et al., 2001; Gawne et al., 1996) and IT (Gawne and Richmond, 1993). In chapter 5 we show that higher order statistical relationships can be used to build a topographic code and reduce redundancy simultaneously.

3 Self-Organization of Topographic Bilinear Networks for Invariant Recognition

In this chapter, we present a model for the emergence of ordered fiber projections that serve as a basis for invariant recognition. After invariance transformations are self-organized, so-called control units competitively activate fiber projections for different transformation parameters. The model presented in this chapter builds upon an abstraction of the activity-based mechanism for the development of retinotopy, as described in chapter 2. In contrast to organizing a single identity mapping from input to output, activity regions of varying position and size are employed to install different transformations. We provide a detailed analysis for the case of 1D input and output fields for schematic input patterns that shows how the model is able to develop specific mappings. We further discuss results that show that the proposed learning scheme is stable for more complex, biologically more realistic, input patterns and that the model generalizes to 2D neuronal fields. Some parts of this chapter have been published in (Bergmann and von der Malsburg, 2011).

3.1 Introduction

We discussed in section 1.1 that a central problem for visual perception is the generation of representations of environmental input patterns that are invariant to transformation (e.g. translation and scale). These representations, which may reside in inferotemporal cortex (Tanaka, 1996), enable the brain to recognize and examine objects in spite of rapidly changing retinal images. A possible mechanism for the construction of invariant representations is based on variable fiber projections (Hinton, 1981), called dynamic links (von der Malsburg, 1981) or shifter circuits (Anderson and van Essen, 1987; Wolfrum and von der Malsburg, 2007b). As described in section 1.1.2, the underlying idea of these approaches is to normalize input patterns by applying a transformation to yield a standardized representation: active fiber projections route the input information to their corresponding invariant output position, while fiber projections that are inconsistent with the current invariance transformation get switched off. Hence, the approach factorizes the representation of input data into the invariant code representing the input independent of the transformation **and** a code for representing which transformation was needed to achieve this invariance (coded in the activity of the fibers themselves). Invariant object recognition has been modeled on this basis (Olshausen et al., 1993; Lades et al., 1993; Wiskott and von der Malsburg, 1996; Arathorn, 2002; Lücke et al., 2008; Wolfrum et al., 2008). Dynamic links may be controlled by temporal correlations of neuronal activities

in a rapid reversible version of Hebbian learning, and face recognition has been modeled in this way (Wiskott and von der Malsburg, 1996). Unfortunately, this mode of control is too slow to handle the required numbers of specific bindings in realistic time, taking more than a hundred times longer than adult object recognition. However, the switching of connections can be accelerated with the help of control units, dedicated units able to modulate the links they contact. Control units are uncommon in that the signal flow on their neurites is bidirectional, incoming during unit activation, outgoing during synaptic control (Hinton, 1981): on the one hand, these units collect information on the similarity of activity patterns at the pre- and postsynaptic side of the links under their control, and on the other hand, they modulate the strength of these links and therefore implement transformations. Consequently we call the set of their neurites receptive-and-projective fields (RPFs). Control units could be a special type of cells which learn to contact with their processes synapses that are experiencing strong pre- and postsynaptic signal correlations. This is akin to Hebbian learning, with the difference that not the synapse is strengthened but the modulatory contact of the control unit. It is not clear which cortical cell type or types are to be identified with control units (see, however, the hypothesis that astrocytes play this role, Möller et al., 2007). On the other hand, the control unit hypothesis could be interpreted as being a mathematical abstraction of nonlinear neuronal network effects (see section 1.1.3), that in detail would inevitably be more complicated. Although there are plenty of theoretical and experimental investigations of possible multiplicatory neuronal interactions (see the discussion in section 1.1.3), it is up to future research how the bidirectional multiplications required for normalization-based approaches can be implemented in detail.

In many dynamic link models, like dynamic link matching (Lades et al., 1993; Wiskott and von der Malsburg, 1996), the development of the strength of each link is independent. This needs a huge number of control units, causes a big search space and leads to the inability of the system to make use of the statistics of transformations it is confronted with. These issues can be tackled by letting each control unit modulate several links, see Figure 3.1. As proposed in (Olshausen et al., 1993; Zhu and von der Malsburg, 2004), the RPFs of control units should be intermediate between controlling single links, as in (Lücke et al., 2008; Wolfrum et al., 2008), and the total set of connections involved in a transformation from the input to the whole invariant output window. The latter, though most efficient (in that a single active control unit could project a whole figure into an invariant output window in inferotemporal cortex) is unrealistic for several reasons: First, due to the limited spatial range of neurites. Second, because a whole projection has to traverse several cortical areas (e.g. V1 - V2 - V4 - IT), as modeled in (Anderson and van Essen, 1987; Wolfrum and von der Malsburg, 2007b). And finally, since the number of required control units would be too large to cover the space of all possible projection patterns. Nevertheless, for the sake of simplicity, the concrete model of this chapter lets each global projection pattern be controlled by a single control unit. However, if we considered the output field of our model as being only a patch of the whole output window, the RPFs of the model would indeed be of intermediate size.

Mathematically, we implement the control unit idea by employing bilinear models (Tenenbaum and Freeman, 2000; Grimes and Rao, 2005; Olshausen et al., 2007;

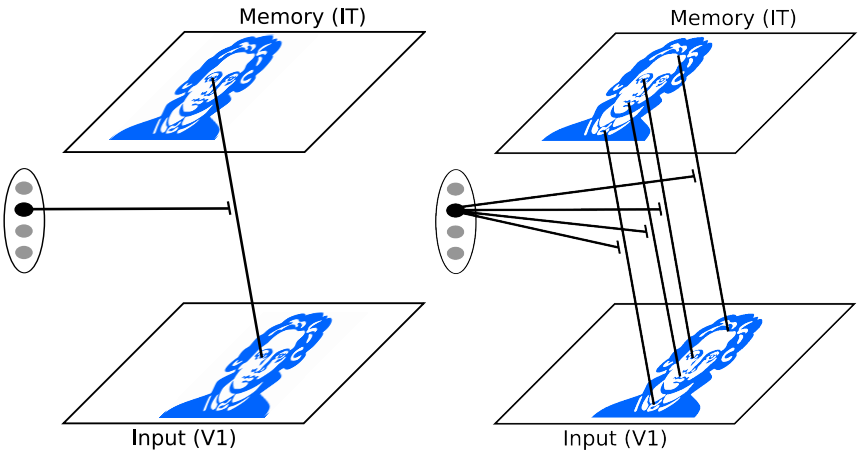


Figure 3.1: Two different connectivity patterns for control units. Left: A single control unit controls a single link, and hence maps a single input point to a single output point. Right: Control units that control many links allow for a whole mapping from a set of input points to a set of output points and hence implement a transformation from the input field to the output field. The effective search space to find correspondences of the input and the output can therefore be considerably reduced and the transformations can be adapted to the statistics of the inputs.

Berkes et al., 2009; Memisevic and Hinton, 2010; Bergmann and von der Malsburg, 2010), for which output activities are proportional to the product of input and control unit activity (see section 1.1.3). In comparison with the general bilinear responses as given in equation 1.2, where the output activity is given by a bilinear sum of input neurons, we in the following assume that multiplicative effect is restricted to the interaction of control units with "feature units", i.e. no multiplicative effects within these populations are assumed. If control units responsible for alternate transformations compete in a winner-take-all fashion, the stage is set for very rapid transformation detection and subsequent transformation-specific signal routing. This leads to the compensation of variations of input patterns and generates an invariant representation that can be exploited for object recognition (Olshausen et al., 1993; Arathorn, 2002; Lücke et al., 2008; Wolfrum et al., 2008).

Note that the invariance problem (for example in vision) is a phylogenetically very old problem. Even the very first animals with a visual receptor field have had an evolutionary advantage in being able to recognize invariantly, for example in recognizing a predator or food. In particular, recognizing a predator independent of its position in the visual field and appropriately activating corresponding motor patterns significantly increased probability of progeny. If we take into account that animals might also be confronted with predators in early postnatal life, we conclude that it would be advantageous to have the invariance problem solved already before birth. The age of the problem and the evolutionary pressure involved as well as the conservatism of evolutionary designs

suggests a common prenatal solution in many species.

The result of this prenatal process is best observed in precocial animals, i.e. animals that "are (relatively) mobile at the moment of birth or hatching" (Starck and Ricklefs, 1998). For moving, a reliable (visual) sensory system is extremely beneficial. Taking into account the high dimensionality (approximately 10^8 rod cells in the human retina) of the highly nonlinear sensory inputs, and given the very short time until recognition functions, postnatal learning of the real-world transformations is unlikely in these animals. Therefore, the sensory system must already be structured to be able to solve the invariance problem. For altricial animals the necessity of a prenatal invariance network is not so obvious, as they in principle have more time to learn the statistics of their environments. Nevertheless, it is known for human newborns to track schematic human faces (Goren et al., 1975; Morton and Johnson, 1991) as well as veridical faces (Cassia et al., 2004), which implies at least a partial solution of the invariance problem at birth[1]. It has been hypothesized that the neonate face likeness results from prenatal PGO waves (Bednar and Miikkulainen, 2004). After eye-opening and under the influence of visual input, further learning can then refine mappings and can deform them appropriately to take into account varying magnification factors due to retinal inhomogeneities (fovea, visual streak etc.).

In the following we are concerned with the problem of the ontogenetic development of appropriate connectivity patterns for control units which correspond to meaningful real-world transformations (e.g. translations), which hence can be used postnatally to generate invariant representations for recognition. No developmental mechanism for the organization of the transformations has been proposed so far. To compensate for object transformations in a recognition system (see e.g. Arathorn, 2002; Wolfrum et al., 2008), the transformations need to be mutually consistent in the output representations they produce. Take for example the special case of translations, where this means that a pattern at one position on the retina needs to yield an identical output as the same pattern at a different position. This implies that the RPFs of the two control units which implement the two transformations need to have the same connectivity, just translated. In particular, the ordering of the connections needs to be identical. We solve this consistency problem by demanding the same ordering in the output representation as in the input representation, thus we demand **topography** to be preserved. Note that this assumption is well in line with neuroscientific evidence, where topography has been found in most brain areas and most species.

In a nutshell, our model is based on the idea that synaptic plasticity, driven by temporal signal correlations, generates a tendency towards neighborhood-preserving projection patterns, see chapter 2. In contrast to the process of the establishment of retinotopy, which leads to a single mapping, we assume that under the influence of slow activity waves a whole set of different topographic projection patterns are gradually installed. Each mapping corresponds to a different transformation parameter, induced by the position and size of an activity wave. The waves necessary for our model might be

[1]Note that newborn vision is two orders of magnitude less accurate than adult vision (Hendrickson, 1994), simplifying the finding of correspondence with a face scheme.

projected retinal waves (Meister et al., 1991; Warland et al., 2006; Huberman et al., 2008) and/or spontaneous cortical waves (Chiu and Weliky, 2001). Eventually, each control unit has a receptive-and-projective field (RPF) in the form of a topographic mapping. We propose a related model for the ontogenesis of multiple topographic mappings for the simpler case of one dimensional input and output neuronal fields with periodic boundary conditions and restricted to translations, but potentially over several layers, in chapter 4 (see also (Bergmann and von der Malsburg, 2008; Zhu et al., 2010)).

What transformations are most likely to be already implemented at eye-opening? At least some of the real-world transformations are likely to be rather complicated. It is therefore unlikely that the whole set of transformations which the organism is confronted with postnatally can be instantiated prenatally. Yet, under the premise of topographic representations, all transformations should be more or less distorted mappings from the input to the output. A good approximation to all of them are affine transformations (i.e. the set of first order Taylor expansions of all possible topographic mappings). Hence, a system incorporating translations, scales and rotations is able to solve all mapping problems at least approximately. Note that newborn acuity of vision is two orders of magnitude worse than in adults (Hendrickson, 1994) and therefore this initial approximation may not be a big drawback, if mappings get refined later. We show in this chapter, that it is the set of affine transformations that a prenatally plausible model can self-organize.

The mechanism proposed in this chapter is an extension of a model for the process of topographic self-organization (Häussler and von der Malsburg, 1983), described in section 3.2, applied to a bilinear model of neuronal information processing (Tenenbaum and Freeman, 2000; Grimes and Rao, 2005; Olshausen et al., 2007), and is described in section 3.3. The results of our model in chapter 3.4 show that visual stimuli are not necessary to organize a network with real-world invariance transformations. Indeed they might even make it more difficult, as it is non-trivial to disentangle pattern information and transformation information contained in visual stimuli.

3.2 The Häussler System

Instead of explicitly modeling the emergence of topographic maps by the model described in chapter 2, we here describe an abstract formulation, that is both simpler to understand and faster to simulate, as activity variables have been removed by adiabatic elimination (von der Malsburg, 1995) and therefore the dynamics is autonomously formulated in the weight variables.

On an abstract level all self-organizing map-formation mechanisms (for a review see Swindale, 1996; Goodhill, 2007) share two essential ingredients (see Figure 3.2): **cooperation** of neighboring connections and divergent and convergent **competition**. The first biases for a topographic structure, because a strong connection supports the growth of a neighboring connection with a close input coordinate and a close output coordinate. The second ensures convergence to a one-to-one mapping: at an output neuron o all incoming connections compete, and similarly, at an input neuron i all

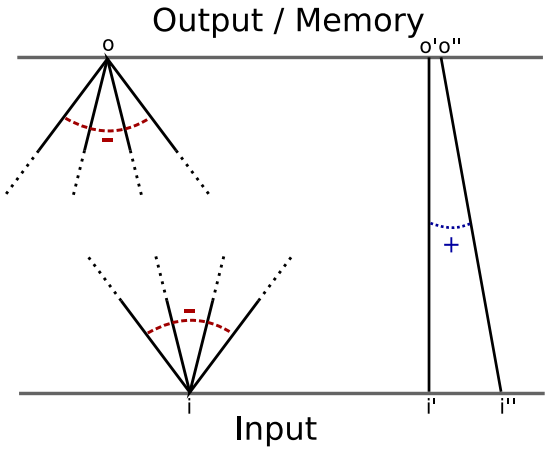


Figure 3.2: Interactions leading to the self-organization of topographic mappings. At an output neuron o all incoming connections compete, as depicted by the minus sign. Similarly, at an input neuron i all outgoing connections compete. On the other hand, there is cooperation (denoted by the plus sign) between two connections that run from neighboring input units (i' and i'') to neighboring output units (o' and o''). The strength of this cooperation falls off with the distance of neurons within the two fields.

outgoing connections compete. In the following, we formalize map formation in terms of an abstract model that has been proposed by (Häussler and von der Malsburg, 1983) and is termed the Häussler system in the following. It has the advantage that it is simple and analyzable, can be simulated efficiently, and is compatible with a full range of models (e.g. Swindale, 1996; Goodhill, 2007; Hyvärinen et al., 2001). All detailed models for the topography mechanism, see for example the model of chapter 2, involve statistical dependencies in neural activity patterns to encode neighborhood relationships in the connected neural fields and to drive synaptic plasticity.

The mapping from the input to the output area is represented by a set of links (o, i), with o and i representing output and input coordinates, respectively. The weight value w_{oi} of link (o, i) indicates the strength with which output unit o and input unit i are connected. In the Häussler system, weight values are positive and a value of zero represents the absence of a connection. The set of all links forms a mapping $W = (w_{oi})$.

The Häussler system, which is formulated as a set of differential equations, was inspired by an equation proposed by (Eigen, 1971) for the evolution of species in theoretical biology:

$$\dot{w}_{oi} = \alpha + w_{oi}F_{oi}(W) - w_{oi}B_{oi}(\alpha + WF(W)), \tag{3.1}$$

where α is a non-negative unspecific growth term. Qualitatively, this equation leads to a strong competition (in case of $\alpha = 0$ a hard winner-take-all competition) of weights that compete in the B-term. $F_{oi}(W)$ mediates weight **cooperation** of neighboring weights. The explicit form of the cooperation coefficient F_{oi} is derived in (von der Malsburg, 1995)

using Hebbian learning and yields a strictly positive weighted sum of neighboring weights with the coupling matrix C:

$$F_{oi}(W) = \sum_{o',i'} C_{oo'ii'} w_{o'i'}. \tag{3.2}$$

The coupling matrix $C_{oo'ii'}$ is a monotonically falling function of both $|o - o'|$ and $|i - i'|$ and describes the mutual cooperative support that link (o, i) receives from its neighbors (o', i').

The **competition** term B_{oi} contains as argument besides α the matrix WF (the component-wise Hadamard product $(WF(W))_{oi} = w_{oi}F_{oi}(W)$), and is the average of growth rates of all competitors to w_{oi}:

$$B_{oi}(M) = \left(\sum_{o'} m_{o'i}/N_o + \sum_{i'} m_{oi'}/N_i \right) /2, \tag{3.3}$$

where $M = (m_{oi})$ is a matrix. The two different sums in this term implement the divergent and convergent competition, shown in Figure 3.2, that is necessary for map-formation. Biophysically this term might be implemented by a normalization of weights on the output side and a competition for growth factors mediated by the input side.

For the case of periodic boundary conditions, the system can be treated analytically (Häussler and von der Malsburg, 1983). It has been shown that starting from a deviation of the homogeneous solution to the system, $W = \mathbf{1}$ (the matrix in which all entries equal 1), the system (equation 3.1) converges to a diagonal matrix (in case of the same size of the input and the output fields, $N_i = N_o$), hence a topographic one-to-one mapping of the input to the output. Numerical simulations show that in the case of non-periodic boundary conditions the system reliably converges to the identity mapping, similar to the system proposed in chapter 2. Due to biological plausibility, we restrict our studies in the current chapter to the investigation of systems with non-periodic boundary conditions.

3.3 The Model

The model (see Figure 3.3) consists of two layers of neurons, called input field and output field. There are short-range connections within the fields and excitatory all-to-all connections between them. In addition, there is a set of control units that are able to modulate the connections between the layers.

The purpose of the model is to demonstrate that different transformations can be organized on the basis of spontaneous activity in the input field. Activity in the input field is restricted to active regions, described in subsection 3.3.1, each anticipating the later appearance of object images with different position and size. Using the learning rule described in subsection 3.3.2, a single control unit active for a region slightly biases its RPF to get restricted to that region and to form a topographic map from that region to the whole output field. Subsection 3.3.3 describes the unsupervised winner-take-all

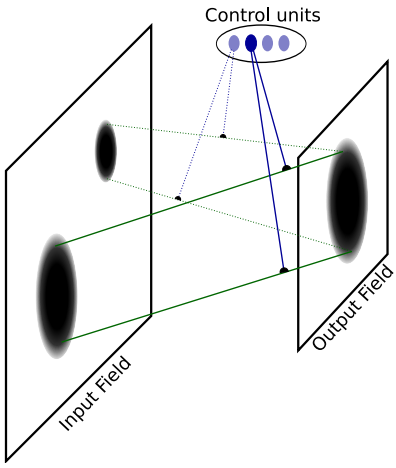


Figure 3.3: Training of control units. Different regions are active at different times in the input field (presumably primary visual cortex, left plane). Control units (full circles) learn to be activated by, and to activate connections. After self-organization, a control unit activates connections that form a topographic projection from an active region to the output field, which thus forms an invariant window (in an upstream cortical area, presumably IT), right plane.

(WTA) mechanism that was used to determine which control unit gets activated for a given input region. The increased bias of a control unit for a region, introduced during learning, increases its chances to win for that region on its next appearance. Iteration of the process leads to the emergence of topographic maps specialized to input regions. In a nutshell, restricted active regions of input activity yield the selection of a dedicated control unit, which then refines its RPF to a transformation from this region to the output region. Note that although not motivated probabilistically, the proposed method bears some similarities to the Expectation-maximization (EM) algorithm: control unit activities are unobserved latent variables and are set in the WTA mechanism, this corresponds to the E-step. Their RPFs are then updated using these activities, corresponding to the M-step. This process is iterated until convergence.

3.3.1 Input and Output Activities

Unit activities are denoted by x_i ($i \in \{1, .., N_i\}$), y_o ($o \in \{1, .., N_o\}$) or c_k ($k \in \{1, .., K\}$) referring to input field, output field or control units, respectively.

Inputs. In order for the model to work, there are some constraints on the input signal statistics. The main goal of the model is to demonstrate that control units can develop different transformations. In particular, for different translations and scales, this means that the connectivity of control units needs to get specialized to different input regions. We therefore assume the input signal to be constrained to regions of activity and to be zero elsewhere. Note that this assumption is well in line with prenatally observed retinal

waves (see section 1.3).

To motivate the topography-generating cooperation between connections described in equation 3.2, activity-based Hebbian models invoke short-range correlations in the input activities (von der Malsburg, 1973; Goodhill, 2007). We shall follow this approach and demand the correlations to fall off monotonously with the relative distance of two cells.

We distinguish two possibilities to generate signals with the desired constraints: First, the input signal could be generated by coupling independent spontaneous noise sources restricted to the active region of shape $\boldsymbol{I}$:

$$x_i = I_i \sum_j^{N_i} C^I_{ij} \xi_j, \tag{3.4}$$

where ξ_i is taken to be independent, identically distributed (i.i.d.) noise with mean value $\langle \xi_i \rangle = \mu_1$ and second moment $\langle \xi_i^2 \rangle = \mu_2$. Hence, $\langle \xi_i \xi_j \rangle = \mu_1^2 + \mu_d \delta_{ij}$ with $\mu_d = \mu_2 - \mu_1^2$. See section 3.3.4 for the dynamical diffusion model we used to derive the explicit form of the coupling matrix C^I_{ij}, which is strictly positive and monotonously decreasing with $|i - j|$. Alternatively, neurons in the input could be allowed to produce noise if they are in the area of the current active region and coupling of neighboring cells happens only afterwards:

$$x_i = \sum_j^{N_i} C^I_{ij} \xi_j I_j, \tag{3.5}$$

hence in this case activity in the input field is not constrained to the area $\boldsymbol{I}$, due to lateral correlations. In the rest of the chapter we focus on the latter possibility, equation 3.5, because it is harder in the sense that the input region constraint is weakened by the lateral correlations. However, we also performed simulations for the simpler case of equation 3.4 and found the results to be qualitatively similar.

Outputs. The output units receive input in the form of a bilinear term (Tenenbaum and Freeman, 2000) (see equation 1.2):

$$\hat{y}_o = \sum_k^K \sum_i^{N_i} w_{koi} x_i c_k. \tag{3.6}$$

According to this, the input $\hat{y}_o$ to output unit o depends linearly on the input activities x_i if the control unit activities c_k are constant, and vice versa. The coefficient w_{koi} is the connection-strength of control unit k to the link from input unit i to output unit o. Similar to the input field, the output field also needs to code neighborhood relationships for a Hebbian-based mechanism to develop topographic mappings. Therefore we assume that the final output activities result from the input $\hat{y}_o$ by multiplication with a coupling matrix C^O_{op}:

$$y_o = \sum_p^{N_o} C^O_{op} \hat{y}_o = \sum_p^{N_o} C^O_{op} \sum_k^K \sum_i^{N_i} w_{kpi} x_i c_k, \tag{3.7}$$

where the explicit shape of the coupling matrix C^O_{op} is given in section 3.3.4.

3.3.2 Synaptic weight dynamics

We now concentrate on the development of the synaptic weights w_{koi} of the control units, that is, of their RPFs. These are to be shaped by two tendencies. On the one hand, they have to concentrate on one active input region. On the other hand, they have to develop a topographic structure, connecting the input region to the output field with one-to-one connections linking neighbors to neighbors. Note that the specific map formation mechanism we chose here, the Häussler system described in section 3.2, is not critical for the mechanism to work and could be substituted by other map formation mechanisms. As for the Häussler system in section 3.2, the learning rule is formulated as a simple third-order differential equation:

$$\dot{w}_{koi} = \alpha + w_{koi} \operatorname{cov}(y_o, x_i) - w_{koi} B_{oi}(\alpha + W_k \operatorname{cov}(\boldsymbol{y} \otimes \boldsymbol{x})), \tag{3.8}$$

$$B_{oi}(M) = \left(\sum_{o'} m_{o'i}/N_o + \sum_{i'} m_{oi'}/(\omega N_i) \right) /(1 + 1/\omega), \tag{3.9}$$

where $[W_k \operatorname{cov}(\boldsymbol{y} \otimes \boldsymbol{x})]_{oi} = w_{koi} \operatorname{cov}(y_o, x_i)$ and the standard definition of the covariance $\operatorname{cov}(x, y) = \langle (x - \langle x \rangle) \cdot (y - \langle y \rangle) \rangle$ (the brackets $\langle\rangle$ denote expectation values). Weights are modified by this equation only for a single control unit k, which is determined by a winner-take-all mechanism from the current active input region, see section 3.3.3. We therefore set $c_k = 1$ for only one control unit and $c_{k'} = 0$ for all other units. Like in the Häussler system, α is a non-negative unspecific growth term.

As for the Häussler system (section 3.2), the Hebbian-like covariance term $\operatorname{cov}(y_o, x_i)$ mediates weight **cooperation** and by substituting equations 3.5 and 3.7 we see that it is a weighted sum of connections with neighboring o and i:

$$\begin{aligned} F_{oi} &:= \operatorname{cov}(y_o, x_i), \\ &= \sum_p \sum_{jlm} C^O_{op} w_{kpj} C^I_{jl} C^I_{im} I_l I_m \left(\langle \xi_l \xi_m \rangle - \mu_1^2 \right) \\ &= \mu_d \sum_p \sum_{jl} C^O_{op} w_{kpj} I_l^2 C^I_{jl} C^I_{il}, \end{aligned} \tag{3.10}$$

or in matrix notation

$$F = \operatorname{cov}(\boldsymbol{y} \otimes \boldsymbol{x}) = \mu_d C^O W_k \tilde{C} \tilde{C}^T, \tag{3.11}$$

where we defined

$$\tilde{C}_{ij} := C^I_{ij} I_j. \tag{3.12}$$

The form of the derived cooperation term, equation 3.11, therefore is a special case of the cooperation term in the Häussler system, equation 3.2, because the weight interaction is separable. In addition to the Häussler system, however, the current active region $\boldsymbol{I}$ modulates this cooperation term so that only weights connected to an active input unit are allowed to cooperate. This active region-driven modulation results in the specialization of the mapping of control unit k to specialize to an input region.

The **competition** term B essentially has the same form as for the original Häussler system, with the extension of a parameter ω to rescale the influence of the input competition. The preferred resulting scale of a mapping is influenced by the relative strength of the input and output sums in the competition term, equation 3.9. For $\omega = 1$ the preferred scale of the resulting mapping, like in the Häussler system (Häussler and von der Malsburg, 1983), corresponds to the connection of the whole input field to the whole output field. For $\omega < 1$ the influence of the competition in the input gets boosted, so that the preferred resulting mapping scale becomes the relation of size ωN_i to the whole output N_o. A value of $\omega < 1$ proved necessary so that it roughly corresponds to the size of the active region $\boldsymbol{I}$ and to counter-act the tendency of the final mappings to spread to larger input regions than the actual active region, due to the spread-out effect in equation 3.5 (see Figure 3.4). However, as will be seen in section 3.4, a single value of ω is able to organize a whole range of differently scaled mappings. If not mentioned otherwise, we set $\omega = 0.4$ and Euler-iterate equation 3.8 with iteration constant $\Delta t = 0.1$.

3.3.3 WTA Control Unit Selection

In the last sections we described the nature of the input signals we assume and the learning rule for a control unit k. In order for different transformations to emerge, different control units have to compete for the different active input regions and then specialize their RPFs for a region. The current section describes the process of competition between control units, which we implemented with a winner-take-all (WTA) mechanism.

Correspondence-finding networks for recognition need to evaluate the similarity of an input pattern with a stored memory pattern (see e.g. Wolfrum et al., 2008), which we denote z_l in the following. In the bilinear framework we use, each control unit needs to evaluate this match with respect to its RPF. We assume that during activation of an active region $\boldsymbol{I}$, the activity patterns x_i and z_l fluctuate randomly and independently. The control units then low-pass filter this noise on their input side:

$$\hat{c}_k = \sum_{li} w_{kli} \langle x_i z_l \rangle \,. \tag{3.13}$$

For the prenatal case we consider, we assume the memory patterns to be unstructured, without loss of generality $\langle z_l \rangle := 1$. Then the control unit is only a function of the inputs x_i:

$$\hat{c}_k = \mu_1 \sum_{lij} w_{kli} C_{ij} I_j, \tag{3.14}$$

where we have made use of equation 3.5 and of the property $\langle \xi_i \rangle = \mu_1$.

The control unit activities are then determined by a winner-take-all mechanism (for a neurally plausible WTA mechanism see e.g. Fukai and Tanaka, 1997; Lücke, 2004a):

$$c_k = \begin{cases} 1: & k = \arg\max_{k'} \{\hat{c}_{k'}\}, \\ 0: & \text{otherwise.} \end{cases} \tag{3.15}$$

Accordingly, when a region $\boldsymbol{I}$ is active, the control unit is selected, whose RPF $\{w_{koi}\}$ has the greatest overlap with the current input activity distribution x_i, as calculated in equation 3.14, and all others are switched off.

The model described so far couples a WTA mechanism (equation 3.15) to learning (equation 3.8). A unit winning for a given input therefore increases its probability to win for the same input again in the future. Note that the introduced learning rule does not include an explicit weight normalization (for example, the sum of the weights could have been forced to equal unity, the so-called L1 normalization). Therefore, a unit's increase in winning probability for a given input does not necessarily decrease the probability for all (or most) of the other inputs. On the contrary, for mutually overlapping input regions (see e.g. Figure 3.4a) a unit winning for any such input region increases its winning probability for all other overlapping input regions as well. As a result, only a small subset of control units usually win and organize their RPFs, while the other units would remain undifferentiated. We prevent this instability by introducing an additional gain modulation, which was inspired by the concept of intrinsic plasticity (IP) (DeSieno, 1988; Desai et al., 1999; Zhang and Linden, 2003; Triesch, 2005). The effect of this homeostatic mechanism is to let control units fire with equal probability. To balance the winning probabilities we introduce the gain modulating variable κ_k for each control unit, that is down-regulated while a unit is active and is up-regulated otherwise:

$$\dot{\kappa}_k = \eta \left(p_k^{goal} - c_k(t) \right), \tag{3.16}$$

and replace the WTA mechanism (equation 3.15) by:

$$c_k = \begin{cases} 1: & k = \arg\max_{k'} \{\kappa_{k'} \hat{c}_{k'}\}, \\ 0: & \text{otherwise.} \end{cases} \tag{3.17}$$

We set $p_k^{goal} = 1/K$ for all k, so that these parameters can be interpreted as probabilities, and the κ_k will stabilize at values such that the $\langle c_k \rangle_T = p_k^{goal}$ (with an averaging period T spanning many stimulations), that is, the control units all fire with the same probability.

How can the parameter η be chosen to get good performance? The rest of this section gives a rough estimate, in a linear approximation. Note that the dynamics of equation 3.16 serves a double role: 1) it controls the winning probabilities in equation 3.17 and 2) it estimates the winning probabilities from the current activities $c_k(t)$. The first point suggests a dynamic that is as fast as possible, $\eta \to \infty$. But concerning the second point, the stochasticity of random inputs and hence the random winning of units on a small time scale, puts an upper bound on η, that we shall approximate in the following. The expected gain-modulation of equation 3.16 becomes:

$$\langle \dot{\kappa}_k \rangle = \eta \left(p_k^{goal} - \left\langle p_k^{win} \right\rangle \right). \tag{3.18}$$

The question now becomes how accurate we can estimate $\left\langle p_k^{win} \right\rangle$. Assuming that the real approximate winning probability is equal to $p_k^{goal} = 1/K$ for all k (that is approximately the case if the mechanism of this section works) and noticing that the inputs to the

system are chosen randomly, we can model the probability of winning n times in an interval T as a Poisson distribution:

$$p_T(n) = \frac{(T/K)^n}{n!} \exp\left(-T/K\right), \tag{3.19}$$

hence the expected number of winnings in a time interval of size T is $\langle n \rangle_T = T/K$. For Poisson processes the associated variance is equal to the mean: $\sigma^2 = \langle n^2 \rangle_T - \langle n \rangle_T^2 = T/K$. The relative error of our estimate of $\langle p_k^{win} \rangle$ therefore becomes:

$$\frac{\sqrt{\sigma^2}}{\langle n \rangle_T} = \sqrt{\frac{K}{T}} \stackrel{!}{=} \hat{\sigma}, \tag{3.20}$$

which we want to be of a predetermined size $\hat{\sigma}$. Thus, the time we have to integrate to achieve this relative error is $T = K/\hat{\sigma}^2$. But how does this time relate to the parameter η? From equation 3.17 we see that the winning probability rises with κ_k. The simplest Ansatz for the winning probability therefore is $< p^{win} >= b\kappa$ (where we dropped the index k for convenience), with b being a constant. Inserting this in equation 3.18 we get

$$\langle \dot{\kappa} \rangle = \eta \left(p_k^{goal} - b\kappa \right) \tag{3.21}$$

which has the solution

$$\kappa(t) = \frac{p_k^{goal}}{b} + e^{-\eta b t} \kappa(0). \tag{3.22}$$

Thus for $t \to \infty$ the gain-control $\kappa \to p_k^{goal}/b$ and by our Ansatz $\langle p^{win} \rangle \to p^{goal}$. The second term in equation 3.22 shows that initial values decrease exponentially with $e^{-\eta b t}$, indicating the timescale of integration. We arbitrarily choose a decay to $1/e$ to equal the timescale for our desired relative error, $e^{-\eta b T} = 1/e$, hence:

$$\eta = \frac{\hat{\sigma}^2}{bK} \tag{3.23}$$

Normalizing $\sum_k \kappa_k = \sum_k \langle p_k^{win} \rangle = 1$ yields $b = 1$ and we have obtained a direct relationship between the desired relative error of our estimate and η. Unless mentioned explicitly we used $\hat{\sigma} = 0.05$ for all simulations.

3.3.4 Equilibrium Solution of the Neural Fields

Neighborhood relations in neural fields are encoded in terms of short-range signal correlations. We here introduce a simple dynamic model of the activity $\tilde{x}(i,t)$ for the input layer, where i indicates the position in the input field and t the time. The dynamics of the output layer is assumed to be of the same type. Consider the function

$$E = \int_i \left(\tilde{x}\left(i,t\right) - \tilde{x}\left(i + \Delta i, t\right) \right)^2. \tag{3.24}$$

Minimizing this function means maximizing the similarity - and therefore the correlation - of cells Δi apart. Therefore the correlation imposing term in the differential equation we seek is of the form

$$-\frac{\partial E}{\partial \tilde{x}(i,t)} = \tilde{x}(i-\Delta i,t) - 2\tilde{x}(i,t) + \tilde{x}(i+\Delta i,t). \tag{3.25}$$

If we now perform the limit $\Delta i \to 0$ this becomes the definition of the second derivative in space, that is an activity diffusion term. A simple linear PDE that contains this correlation imposing term is of the diffusion type:

$$\frac{\partial \tilde{x}(i,t)}{\partial t} = -\tau \tilde{x}(i,t) + D\frac{\partial^2 \tilde{x}(i,t)}{\partial i^2} + \hat{x}(i,t). \tag{3.26}$$

Here, $\tau > 0$ denotes the time constant for activity decay, $D > 0$ is the analogon to a diffusion constant and determines the spatial range of correlations, and $\hat{x}(i,t)$ is the input at a given position i.

To solve equation 3.26 we first perform a spatial Fourier transform[2]:

$$\frac{\partial \tilde{x}^{ft}(\nu,t)}{\partial t} = -\tau \tilde{x}^{ft}(\nu,t) - 4\pi^2 D\nu^2 \tilde{x}^{ft}(\nu,t) + \hat{x}^{ft}(\nu,t), \tag{3.27}$$

As the basis functions of the Fourier transforms are plane waves, which are the eigenfunctions of the Laplace operator, the set of ODEs is now decoupled and the solution can be obtained by the method of variation of parameters as

$$\tilde{x}^{ft}(\nu,t) = e^{-(4\pi^2 D\nu^2+\tau)(t-t_0)}\left(\tilde{x}^{ft}(\nu,t_0) + \int_{t_0}^{t} e^{(4\pi^2 D\nu^2+\tau)(t'-t_0)}\hat{x}^{ft}(\nu,t')dt'\right). \tag{3.28}$$

Under the adiabatic assumption that $\hat{x}^{ft}(\nu,t)$ changes much more slowly compared to the exponential decay, that is, we can find a t_0 so that $\hat{x}^{ft}(\nu,t')$ is approximately constant in the integral and $e^{-(4\pi^2 D\nu^2+\tau)(t-t_0)}$ is close to zero, we get the approximate solution

$$\tilde{x}^{ft}(\nu,t) \approx \frac{1}{4\pi^2 D\nu^2+\tau}\hat{x}^{ft}(\nu,t). \tag{3.29}$$

As the prefactor is a falling function of the wavenumber ν, this is a low-pass filtered version of the input signal $\hat{x}^{ft}(\nu,t)$. Taking the inverse Fourier transform and making use of the convolution theorem, this yields

$$x(i,t) \approx \left(\frac{1}{2\sqrt{\tau D}}e^{-\sqrt{\frac{\tau}{D}}|i|}\right) * \hat{x}(i,t), \tag{3.30}$$

where $*$ denotes a convolution. The parameter τ/D determines the size of the cooperative weight interactions in equation 3.10. In our simulations we set $\tau/D = 30/N^2$, where N is the number of units in the input or the output.

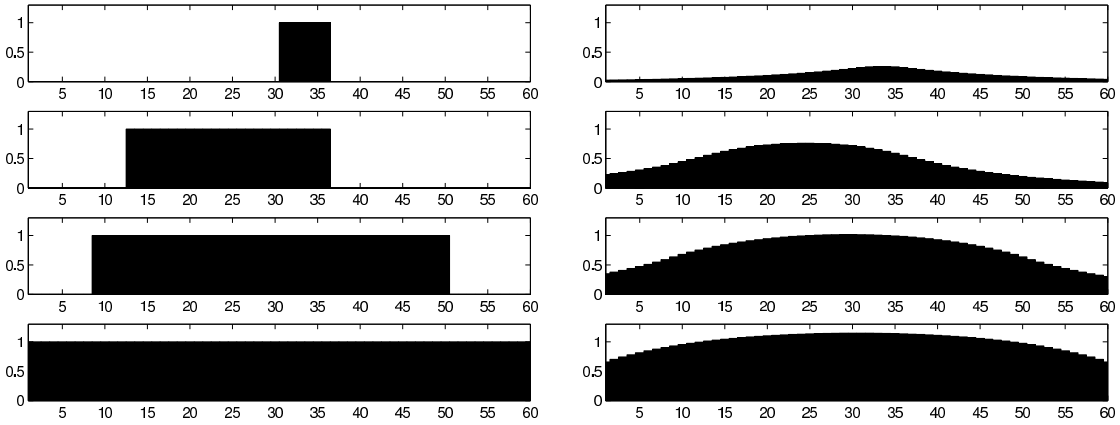


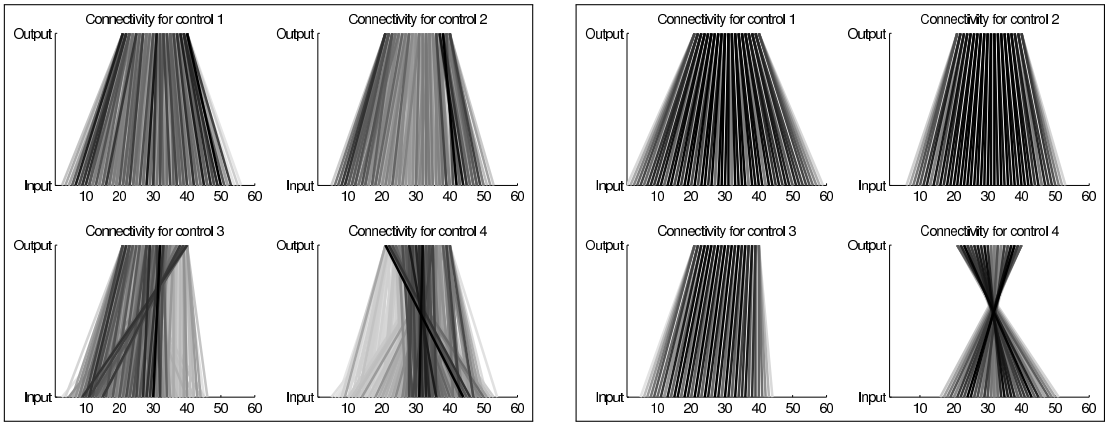
Figure 3.4: Left: Four active regions of different size, $\boldsymbol{I}_n$, where $n = 1, .., 4$ specifies the different regions used in an experiment with four control units. In each time step, one of the $\boldsymbol{I}_n$ is selected at random. These regions are subjected to smoothing and noise (see equation 3.5). Right: The resulting expected values $\langle x_i \rangle$ for the four active regions $\boldsymbol{I}_n$. Through a WTA mechanism (equation 3.17) the active regions determine which control unit is activated and permitted to learn.

3.4 Results

The system considered in this section consists of $N_i = 60$ input units, $N_o = 20$ output units, $K = 4$ control units and simulations were run with parameters $\alpha = 0$ and $\omega = 0.4$. Although we have run many simulations with different dimensions of the input and output and larger numbers of possible active regions and control units, we limit ourselves to the description of this system here for clarity. Figure 3.4a shows four active regions in the input field. The active regions are binary $I_{ni} \in \{0, 1\}$ (n specifies the different regions) and are subjected to smoothing by lateral signal exchange (see Figure 3.4b) and noise, equation 3.5. Given that the signal correlations are important for the establishment of topography, we included this smoothing for consistency although it does not serve any function in the simulations (see section 3.3.1). We intentionally use active regions of different size and with full overlap in order to show that the system is able to discriminate these regions.

At each time step, a random region n is chosen and the winning control unit is determined using equation 3.17. This winning unit is then permitted to iterate one step of the RPF learning rule of equation 3.8 to update its weights. This repeated selection of control units and subsequent weight modification leads to the reorganization and refinement of the RPF mappings. Figure 3.5a shows the RPFs of the four control

[2]Note that for the Fourier transformation we use the definition $\nu = 1/\lambda$ that leads to no additional 2π factors for the integrals.



(a) (b)

Figure 3.5: Control unit RPFs (receptive-projective fields) are shown for a simulation with four control units and $N_i = 60$, $N_o = 20$, $\alpha = 0$ and $\omega = 0.4$. Left panels: Learning at an intermediate stage, $t = 2000$. Black denotes strong links, while white is weakest. Right: The (nearly) converged weights are drawn ($t = 80000$ inputs).

units at an intermediate stage of learning ($t = 2000$ inputs), with the input field on the bottom and the output field on the top. As described in section 3.3.2, the parameter value $\omega = 0.4$ (see equation 3.8) pushes RPFs towards smaller scale factors (the ratio of input size to output size), which is necessary to counter-act the tendency of RPFs to spread in the input field, which happens due to the neighborhood correlations, see equation 3.5. The shade of a link renders the strength of the connection to and from control units – black is strongest and white is weakest. RPF entries with less than 10% of the maximum are visually clipped. From inspection of the RPFs it is hard to tell how well they specialized to the inputs while it is easy to see that topography has already emerged for units 1 and 2. Evidently, the symmetry of the initial RPFs was broken, resulting in one of the two possible map orientations possible in the one-dimensional case, the "ascending" and "descending" orientation. In our simulation, control units 1 and 2 preserve the orientation of the inputs. The RPF of control unit 3 and 4 still contain both map orientations, although the links corresponding to the ascending orientation seem more pronounced in unit 3. The final map orientation of an RPF is mainly determined by initial random weight values, but may be influenced by the random selection of active regions in the input as well.

In the final state, after application of $t = 80000$ inputs (when the mean relative weight changes have fallen below 0.25), the RPFs have converged to high specificity for one input pattern as well as to good topography, see Figure 3.5b. Comparing the final state (Figure 3.5b) to the intermediate stage of learning (Figure 3.5a), we see that the map orientations prominent in the intermediate stage were already stable and did not change in the consecutive RPF refinement.

We have performed simulations with a range of different parameter values and the

system proves robust against many changes. To consistently obtain topographic mappings, a critical lower bound on the size of the lateral cooperation in the fields, $^{D}/_{\tau} \gtrapprox {}^{N^2}/_{50}$ in equation 3.30, is to be observed, however. For smaller values the emergent transformations tend to be only piecewise topographic. If, like in Figure 3.4, the center points of active regions fall on varying positions, a non-vanishing positive α allows for a reorganization and a migration of the RPFs, as demonstrated in section 3.5 for the two-dimensional case. For the relatively small simulation presented, however, α can be set to zero. Non-vanishing α values also allow for significantly smaller cooperation in the fields, $^{D}/_{\tau}$, while still preserving final consistent topographic transformations (Zhu, 2008).

3.4.1 Quantitative Characterization of RPF Development

In order to be able to gain better insight into progress and parameter dependence of the system we introduce three quantities:

Input Specificity $\zeta(t)$ is a measure to analyze the assignment of the input regions to the control units. Let $\nu\left(k|n,t\right)$ be the conditional normalized winning frequency of control unit k given region n. We estimate this quantity with the help of a leaky integrator (with time constant $5 \cdot 10^{-3}$) and the constraint $\sum_k \nu\left(k|n,t\right) = 1$. We define the input specificity of the system as

$$\zeta(t) = \left\langle \max_k \nu\left(k|n,t\right) \right\rangle_{\{n\}}, \tag{3.31}$$

which is the highest winning frequency for a given active region, averaged over all possible active regions. Due to the normalization of ν, $\zeta \in [0,1]$. The maximum $\zeta = 1$ is reached when all active regions considered are assigned reliably to a specific control unit. Note that due to the intrinsic plasticity modulation of the WTA mechanism (section 3.3.3) the (unconditional) winning frequencies of all control units are close to the equidistributed goal probability $1/K$, ensuring that different control units are assigned to different regions.

Scale factor $S(k)$**.** A control unit k has the footprint

$$\tilde{w}_{ki} = \sum_o w_{koi}. \tag{3.32}$$

After learning, all control units have a footprint in the form of a smoothed version of one of the active input regions, similar to those in Figure 3.4 but with less smoothing. For the calculation of the scale factor, we include all entries of the footprint, which are bigger than 10% of the maximum entry and define the scale of a transformation $S(k)$ as the ratio of this footprint size and the (fixed) size of the output field.

Synaptic spread $s(k)$**.** The goal of control unit self-organization is the establishment of RPFs in the form of one-to-one mappings. Progress towards this goal can be assessed with the help of the synaptic spread, which is a measure for the size of the input patch to which an output unit is significantly connected. We define it as the synaptic standard deviation $s(k)$:

$$s(k) = \left\langle \left\langle w_{koi} \left(r_i - r\left(k,o\right)\right)^2 \right\rangle_{\{i\}}^{1/2} \right\rangle_{\{o\}}, \tag{3.33}$$

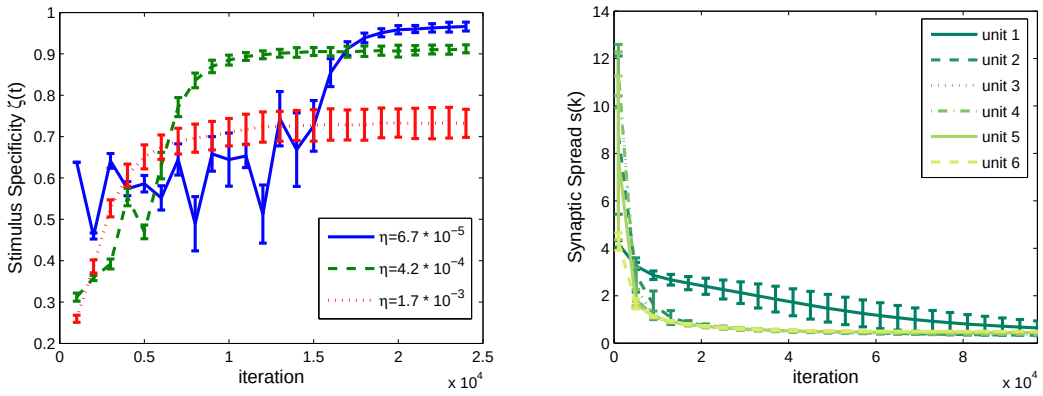


Figure 3.6: Left: Time course of the input specificity $\zeta(t)$ for 6 units, averaged over 50 trials. Higher η values lead to faster convergence, but at the cost of a smaller final value. The value $\eta = 4.2 \cdot 10^{-4}$ follows from the theoretical analysis of section 3.3.3, see equation 3.23. Right: The synaptic spread averaged over 50 trials is shown for 6 simulated units for the case of $\eta = 4.2 \cdot 10^{-4}$. Units have been ordered according to the active regions they specialized in after each trial.

in which we make use of the center of mass of the receptive-projective field of output unit o under control unit k:

$$r(k, o) = \sum_i w_{koi} r_i, \tag{3.34}$$

where r_i denotes the position of unit i in the input field.

The exact time course of our system depends on the details of its formulation, which cannot be fixed on the basis of current biological information. Moreover, topography formation can be seen as a constraint optimization problem, and many dynamic formulations have been shown to be consistent with a single optimization problem (Wiskott and Sejnowski, 1998). We nevertheless find it useful to discuss the parameter dependence of the developmental time course of input specificity and topography in our system.

Figure 3.6 (left) shows the average progress of input specificity, with error bars indicating standard deviations over 50 trials. Specificity converges within approximately 10^4 iterations to its maximal value. A higher rate coefficient η of the gain-control dynamics (equation 3.16) leads to faster initial growth but reduces the final specificity. This has to do with the necessity to obtain a sufficiently large sample of the randomly selected active regions.

Figure 3.7 (left) shows a bar plot of the final scales $S(k)$ ($T = 60000$ iterations) of a system with 6 control units and active regions $\boldsymbol{I}$ ranging from a third of the output size to twice the output size. The right side of the Figure shows their temporal development. For both plots, units are sorted at the end of each trial by the active region they specialized in. Initially, several of the scales decrease. The reason is strong effective cooperation of weights in the center of the weight matrix of each control unit relative to the boundary,

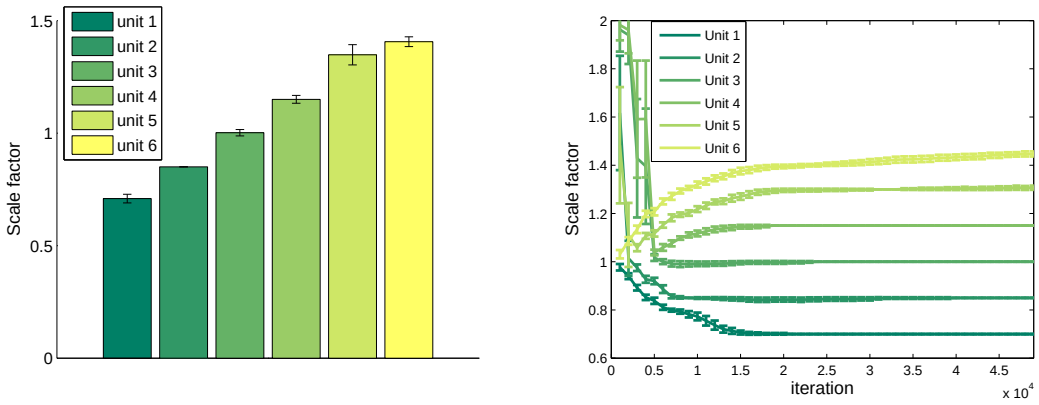


Figure 3.7: Left: Final transformation scale S for 6 units, $\omega = 7$ and $\eta = 4.2 \cdot 10^{-4}$. Error bars are standard deviations for 50 trials. Right: Time course of the scale S for the same set of simulations.

where there are fewer cooperating partners. Comparing this Figure with Figure 3.6 (left panel for $\eta = 4.2 \cdot 10^{-4}$) we see that with increasing specificity also the scale increases again for these units. Finally, the scales converge to a stable value. Note that the final scaling factors do not only depend on the active input regions, Figure 3.4a. For very small regions, the scale is increased significantly due to activity leakage in the input (the effective lateral interaction C in equation 3.5 that is illustrated in Figure 3.4b). This is the main motivation for an ω value smaller than 1, which counteracts this effect (see section 3.3.2). For large regions on the other hand, the effective lateral interaction C is relatively small and therefore the effect is negligible.

In Figure 3.6 (right) the synaptic spread, averaged over 50 trials, is shown for all simulated units, which are again sorted by the active regions they specialized in. Interestingly, those units develop faster that specialize to larger active input regions and scaling factors (compare Figure 3.7). The smallest unit stagnates for a long time, which is due to the coexistence of the two map orders that are possible in 1D (compare Figure 3.5a).

3.4.2 Signal Processing Analysis of Final RPFs

In this section we look at the emergent mappings resulting from the self-organization process described from the viewpoint of signal processing. In particular, we look at potential aliasing effects, that could generate artifacts in the transformations and hence would generate wrong wave numbers in the invariant output representations of the inputs.

We have seen that mappings can grow without unspecific weight growth, i.e. $\alpha = 0$ in equation 3.8. We here also analyze the effect of the parameter α on the converged RPFs and show that a non-vanishing parameter α leads to a suppression of high wave numbers. This low-pass behavior has the advantage of dampening aliasing effects.

The final synaptic spread of RPFs is controlled by the unspecific weight growth

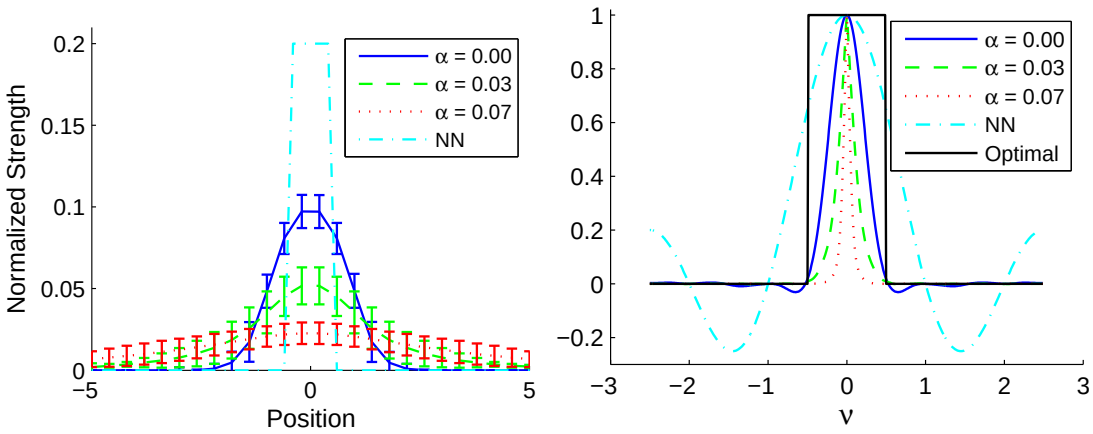


Figure 3.8: Left: The average RPF of all output units of control unit 2 from Figure 3.5 (standard deviations are evaluated over all output units and 20 trials), shifted to common center-of-mass position $r(k, o)$ for various values of α after $t = 5 \cdot 10^5$ steps ($\Delta t = 0.5$). NN denotes nearest neighbor. Right: Comparison of the optimal transfer function to self-organized transfer functions.

parameter α, see equation 3.8. In Figure 3.8 we plot the RPF of one control unit, averaged over all output units.

All RPFs have been translated (with sub-pixel accuracy) using linear interpolation, so that their center of mass position $r(k, o)$ (equation 3.34) is shifted to the origin[3]. As can be seen, the synaptic spread grows with α. Furthermore, the weights have a strictly monotonic decrease with distance to $r(k, o)$. Note that the arborization of both dendrites and axons seems to follow a simple rule for most neurons (Snider et al., 2010): they can be characterized by a Gaussians with different standard deviations in the three dimensions (or two for the map case). The Gaussian-like RPFs we get from our simulations therefore are well in line with these recent experimental results. Hard winner-take-all competition of the weights, on the other hand, would imply that only one weight in each column and row of the weight matrix could survive. This would correspond to the survival of only the nearest neighbor, which we have drawn for comparison. In practice however, several weights per column and/or row can survive also for zero unspecific growth, $\alpha = 0$, even for very long iteration times[4]. Hence, the transformations show a low-pass behavior and are able to suppress aliasing effects that would be produced in the nearest-neighbor case.

For the example case drawn, the output is half as big as the input, while at the same time it is supposed to represent the same "image". A perfect factor two zoom means that all $r(k, o)$ are at integer positions in the input and hence, due to the interpolation

[3]The precise type of interpolation did not significantly change the result of this section – we also tested nearest neighbor, cubic spline and piecewise cubic Hermite interpolation.

[4]This can also be observed in Figure 3.6, right panel, where the synaptic spread does not converge to 0 but stabilizes at a small, but non-finite value.

condition (i.e. an imagined reconstructed continuous image I(x) should be identical to I_i at the sample point x_i and therefore independent of all $I_j, j \neq i$ for $I(x_i)$) no interpolation is necessary and the mapping can just copy every second pixel and neglect the others. The emergent mappings are in general not this very special case, i.e. the $r(k, o)$ usually are not integer numbers. Therefore, the output values in this case depend on an interpolated value of the input.

The (approximate) independence of the drawn weights from their actual output positions (which can be seen from the small standard deviations in Figure 3.8), except a translation in the input, allows for an interpretation as convolution kernels. We therefore can compare the emerged kernels with the optimal image processing kernel for interpolation (see e.g. (Jähne, 2005)). To do this we performed a discrete Fourier transform of the kernels to obtain the corresponding transfer functions in Figure 3.8(right)[5]:

$$f(\hat{\nu}) = \sum_j f(j) e^{-2\pi i j \hat{\nu}/N},\ \hat{\nu} = 0..N-1. \tag{3.35}$$

Note that we had to use an oversampling factor $\beta = 5$ (i.e. we represented the kernels on a grid with more sampling points than the input) of the input, to get information about wave numbers higher than two times the Nyquist wavenumber[6]. This procedure is justified, as by sub-pixel wise translations of the individual RPFs and consecutive averaging over many trials, we get rid of the limitations of the original sampling of the RPFs. We introduce the new variable $\nu = \hat{\nu}\beta/N$ to make the visualization independent of the oversampling. Theoretically one can think of an optimal transformation constructed in three steps: first, we reconstruct from the sampled image I_i a continuous image $I(x)$. Second, on this we can now apply an arbitrary transformation. Finally, we have to re-sample with our output points. Using this construction and Shannon's sampling theorem we see that the optimal transfer function preserves wave numbers up to the Nyquist wavenumber $\nu_{Ny} = 1/2\Delta x$ and dampens all exceeding ones to 0. From the Figure we see that the nearest neighbor transfer function shows a slight low-pass behavior and it creates many artificial (aliasing) wave numbers. This transfer function can be analytically calculated to be $\mathrm{sinc}(\nu) = \sin(2\pi a\nu)/\pi\nu$ ($a = 0.5$ here corresponds to half of the width of the step function). Note that the same function (in position space) would be the optimal kernel, as it corresponds to the inverse Fourier transform of the optimal transfer function. Implementation in our model is impossible, as it involves negative values, which would correspond to inhibitory weights. The resulting transfer function for $\alpha = 0$ still creates artificial wave numbers, although much less compared to the nearest neighbor case. Finally, for $\alpha \gtrsim 0.07$ aliasing effects practically vanished but a strong low-pass behavior persists.

3.4.3 Specificity Problem

In this section we shall analyze how the system is able to achieve high specificity values (see equation 3.31 and Figure 3.6). In particular, we investigate how the model described

[5] We interpret negative wave numbers as $e^{-2\pi i(j-1)(-\hat{\nu})/N} = e^{-2\pi i(j-1)(N-\hat{\nu})/N}$.

[6] DFT only gives results up to two times the Nyquist wavenumber.

in section 3.3 normalizes the RPF of the control units, i.e. how the weights of a control unit are constrained (e.g. the sum of all entries could equal a constant – we will refer to this normalization as $L1$ norm), because these constraints influence the input drive to the control units (equation 3.14), given the active input region we use. We shall first show that some nomalizations would not be able to achieve arbitrarily high specificity and then present evidence that the system uses a superposition of these norms to solve this problem, which we ascribe to the intermediate scope of the B-term in equation 3.8, which is neither a global sum over all weights, nor only dependent on a single weight. Note, that the system can achieve high specificity as well if the intrinsic plasticity values $\boldsymbol{\kappa}$ are chosen appropriately – we neglect this possibility here, as we shall see that already the weight learning itself solves the problem.

We start by discussing "locally saturated" weights. In this case the constraints are only local for each entry of the weight matrix. The weight values are then assumed to lie in an interval between zero and the maximal possible value, which is assumed to be the same value for all weights. The learning rule we consider (equation 3.8) does not have a fully global normalization and hence might suggest a local saturation – in this case the RPF footprint (see equation 3.32) would be $\tilde{\boldsymbol{w}}_k = \gamma_1 \boldsymbol{I}_n$ for unit k being specialized to input region n. We assume $I_{ni} \in \{0, 1\}$ and for the sake of simplicity of the argument we ignore lateral interactions in the input field ($C_{ij} = \delta_{ij}$ in equation 3.14). If this normalization would be the case, the output of a unit was just the sum of the active input units with which its RPF footprint had overlap times the saturation value γ_1. Hence, a unit specialized for a certain input region $\boldsymbol{I}_n$ and a unit specialized to an input region $\boldsymbol{I}_{n'}$, which has the same active units and additional active entries compared to $\boldsymbol{I}_n$, would result in the same input drive from input region $\boldsymbol{I}_n$.

On the other hand, if the weights would be globally $L1$ normalized, i.e. $\tilde{\boldsymbol{w}}_k = \gamma_2 \boldsymbol{I}_n / ||\boldsymbol{I}_n||_1$, the discussed problem would be avoided. But in this case a similar problem occurs: consider a unit specialized for a big input region and a unit specialized for a smaller input region fully included in the big input region. The drive to the two units from the big input region then would be the same and they therefore could not be differentiated.

A possible solution to the discussed specificity problem is a superposition of local saturation and $L1$ norm: $\tilde{\boldsymbol{w}}_k = \gamma_1 \boldsymbol{I}_n + \gamma_2 \boldsymbol{I}_n / ||\boldsymbol{I}_n||_1$. For linearly increasing input regions $\boldsymbol{I}_n$ the input drive $\tilde{\boldsymbol{w}}_k \cdot \boldsymbol{I}_{n'}$ for an input region then is $\gamma_1 \min(k, n') + \gamma_2 \min(k,n')/||k||$. In Figure 3.9 (left panel) the response of units specialized to different sizes is shown for an input region of size 20 for the three norms discussed. It can be seen that only the superimposed normalization has a unique maximum and therefore allows for high specificity. Figure 3.9 (right panel) shows the sum over the weights $\sum_i \tilde{w}_{ki}$ for 8 control unit RPFs from the final state of 20 simulation trials plotted over their developed input region size preference. For comparison also theoretical plots for $L1$ and local saturation are plotted. The superposition norm has been fitted to the simulation results and turned out to be a good match with $\gamma_1 = 0.47$ and $\gamma_2 = 0.51$.

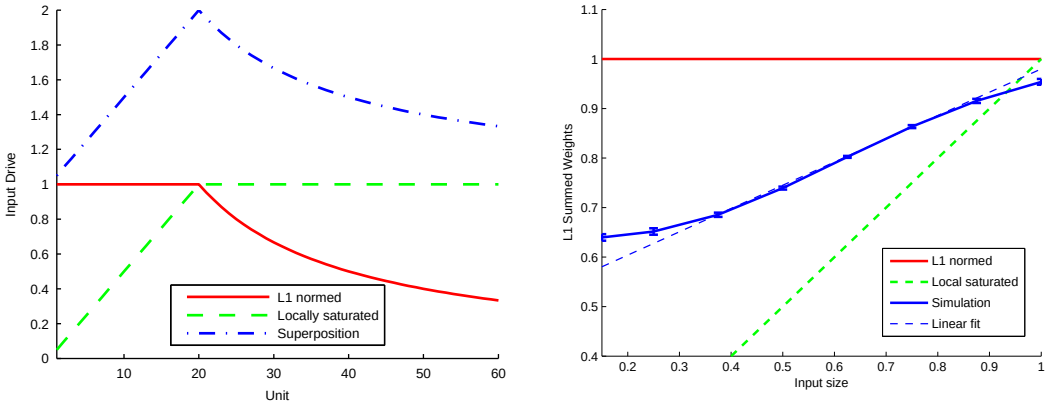


Figure 3.9: Left: The input drive to control units, which are specialized to different region sizes (x-axis), to an active region $\boldsymbol{I}$ of size 20 for three different norms. For better visibility, the plot of the superposition norm has an offset of +1. Right: The sum of the RPFs of the control units plotted over their preferred input region size for the three different norms. The superposition norm has been fitted to the simulation results: $\gamma_1 = 0.47$ and $\gamma_2 = 0.51$.

3.4.4 Complex Inputs

In this section we shortly describe the stability of the proposed learning scheme to more complex active input regions. The number of different active input regions so far exactly matched the number of possible control units. This is an unrealistic assumption that was only chosen for simplicity and analysis purposes. It is known that e.g. retinal waves (Meister et al., 1991; Warland et al., 2006; Huberman et al., 2008) start at a random position and then migrate on the retina. Therefore the number of active biological regions can be assumed to outnumber by far the number of control units. We simulated this by picking a random position (or size) at each iteration and found topographically consistent mappings, varying in the desired parameters, to emerge. Hence, the competitive learning scheme used is able to develop mappings for representative input regions and the WTA mechanism categorizes non-representative input regions to belong to a certain class. From this perspective the system can be seen as a generalization of vector quantization or K-means clustering, but instead of selecting prototype vectors the system develops prototype transformations.

Finally, biological retinal waves are not necessarily simply-connected. However, the desired mappings are supposed to map a simply-connected area from the input to the output. We therefore performed experiments with non-linear superpositions of two input regions, each simply-connected (as in Figure 3.4) at random positions, with a cut-off of the superposition at 1. Surprisingly, the resulting mappings were topographically consistent and map a simply-connected area from the input to the output. A reason is that the combination of cooperation of neighboring weights with the competition of

distant weights (see equation 3.8) leads to a benefit of neighboring links in the input.

3.5 2D Results

Generalization to two-dimensional input and output domains is straightforward on the basis of equation 3.8, if only indices are replaced by two-dimensional integer vectors $\boldsymbol{i} = (i_1, i_2)$ and $\boldsymbol{o} = (o_1, o_2)$ for input and output units:

$$\dot{w}_{k\boldsymbol{oi}} = \alpha + F_{\boldsymbol{oi}} w_{k\boldsymbol{oi}} - w_{k\boldsymbol{oi}} B_{\boldsymbol{oi}}(\alpha + FW_k), \tag{3.36}$$

$$B_{\boldsymbol{oi}}(X) = \left(\sum_{\boldsymbol{o}'} x_{\boldsymbol{o}'\boldsymbol{i}}/N_{\boldsymbol{o}} + \sum_{\boldsymbol{i}'} x_{\boldsymbol{oi}'}/(\omega N_{\boldsymbol{i}}) \right) /(1 + 1/\omega).$$

Again, $N_{\boldsymbol{i}}$ and $N_{\boldsymbol{o}}$ are the numbers of units in the input and output field. For the 2D simulations we chose $\omega = 0.2$.

To speed-up computing performance we chose separable input regions,

$$I_{\boldsymbol{n}} = \boldsymbol{I}_{n_1} \otimes \boldsymbol{I}_{n_2}^T, \tag{3.37}$$

which for the present simulation are 6x6 regions of the type

$$= \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}. \tag{3.38}$$

As in this example, input regions for our simulations were constructed by choosing two **1**-vectors with either the first two or the last two entries set to zero[7]. This procedure gives four different active input regions, each occupying a simply-connected active region in the input. We also chose a separable form $C_{\boldsymbol{ii}'} = C_{i_1 i_1'} C_{i_2 i_2'}$ for the interaction kernel, which is used to convolve the active input regions in analogy to equation 3.5. In particular we chose it to be of a Gaussian form. This leads to the smoothing matrix $\tilde{C}_{\boldsymbol{ii}'} = e^{-(i_1 - i_1')^2/2\sigma_i^2} I_{n_1 i_1'} e^{-(i_2 - i_2')^2/2\sigma_i^2} I_{n_2 i_2'}$ ($\boldsymbol{n} = (n_1, n_2)$ specifies the input region) that generalizes equation 3.12. In our simulations, the standard deviations of the Gaussian were set to be a fourth of the length of the input or output field, $\sigma_x = N_{\boldsymbol{x}}/4$ (x being either i or o). Using the separability assumption, the calculation of the cooperation tensor $F(W)$ factors to four generalized matrix multiplications (i.e. a sum along one dimension at a time) instead of two summations along two dimensions for each new value (the input and the output field kernels are considered independent and hence are trivially separable).

[7] Note that the active regions in two dimensions do not need to be rectangles. For example, we also tried separable Gaussians and got similar results.

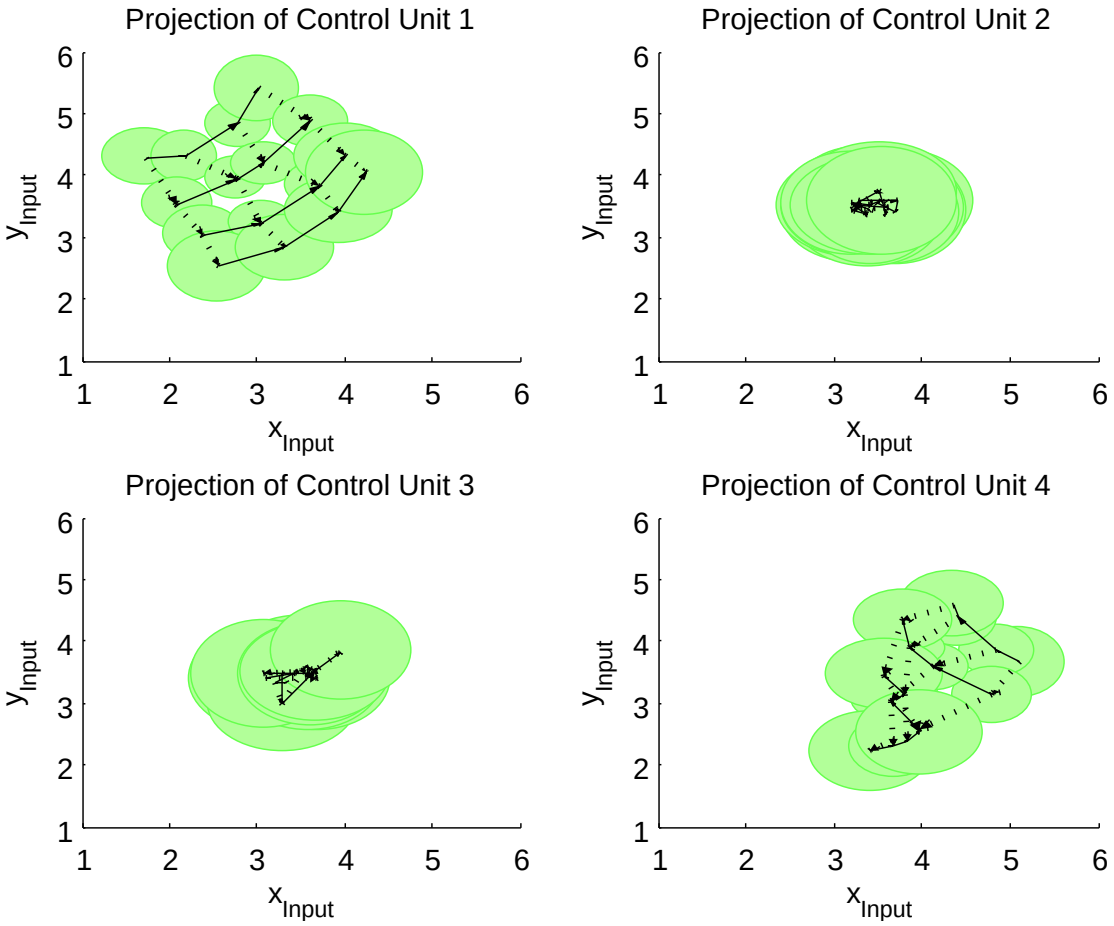


Figure 3.10: 2D Projections of 4x4 output units to the 6x6 input space for $K = 4$ control units at an intermediate stage, at $t = 700$ iterations. The horizontal and vertical axes in each plot denote the coordinates in the input layer. Control units 1 and 4 have specialized their RPFs to the left and right parts of the input field, respectively. Unit 2 just started winning and organizing its RPF while unit 3 is still in its initial state.

Figure 3.10 shows the projection of the weight vectors for all units of a 4x4 output field to the 6x6 input field for $K = 4$ control units at an intermediate stage at $t = 700$ iterations. Nodes of the visualized graphs are the centers of mass $r(k, o)$ of the output units, calculated as in equation 3.34, while the ellipses indicate 2D synaptic standard deviations $s(k, o)$, calculated as in equation 3.33 without averaging. Arrows connect neighboring output units in increasing order (solid arrows indicates the first dimension in the output and the dashed ones the second dimension). For the simulation shown we used a value of $\Delta t = 0.6$ and $\eta = 2.5 \cdot 10^{-3}$ to Euler-iterate equation 3.36. This relatively high iteration constant Δt, in comparison to the η of the IP mechanism (see section 3.3.3), leads to fast emergence of topography before the winning probabilities of the control units are balanced. Hence, it can be seen that control unit 1 specialized to the two left active input regions and control unit 4 to the two right ones. Control unit 2 just started winning and only deviates slightly from its initial configuration while the weights of control unit 3 are still in their initial state. We chose a non-vanishing unspecific growth parameter $\alpha = 0.05$ to allow for more flexible reorganization of the mappings. As this leads to bigger standard deviations in the weights (see Figure 3.8), we set $\alpha = 0$ after $t = 8000$ iterations. After $t = 10000$ iterations the mappings converged to their final configuration, see Figure 3.11, in which each control unit has specialized to a single input region $I_{\boldsymbol{n}}$. In comparison to the intermediate stage at $t = 700$, the RPF of control unit 1 had to migrate to the lower left corner due to competition with control unit 2, which specialized for the upper left corner. Similarly, competition of control unit 4 with unit 3 resulted in the migration of its RPF to the lower right corner in the input. Note that topography is stable during migration.

The 3D cross-product of two arrows in the plots can be used to determine if the corresponding mappings preserve mirror-symmetry: if the cross product of a solid times a dashed arrow points out of the sheet, the corresponding mapping preserves mirror-symmetry, otherwise it violates it. All mappings except number 1 preserve this symmetry. As in the one-dimensional case, the initial values of the weights mainly determine the mirror-symmetry. Comparison of the converged weights with the intermediate stage of learning shows that analogous to the observation in Figure 3.5 the initially established mirror-symmetry stays stable during the refinement of the mappings.

3.6 Probability-based Scale Organization

In the presented simulations, the organization of transformations with different scales needed differently sized input regions, which seems to be consistent with in-vitro findings of retinal waves in fetal macaques (Warland et al., 2006). However, in this section, we demonstrate that differently sized active input regions are not a theoretical necessity to organize transformations that differ in their scale parameters. Assume the active input regions to be at random positions in the input and of approximately the same size throughout the learning procedure. Transformations with big scale factors cover a large area in the input and therefore should be responsible for many inputs, while transformations with small scale factors should develop selectivity for only a small set of

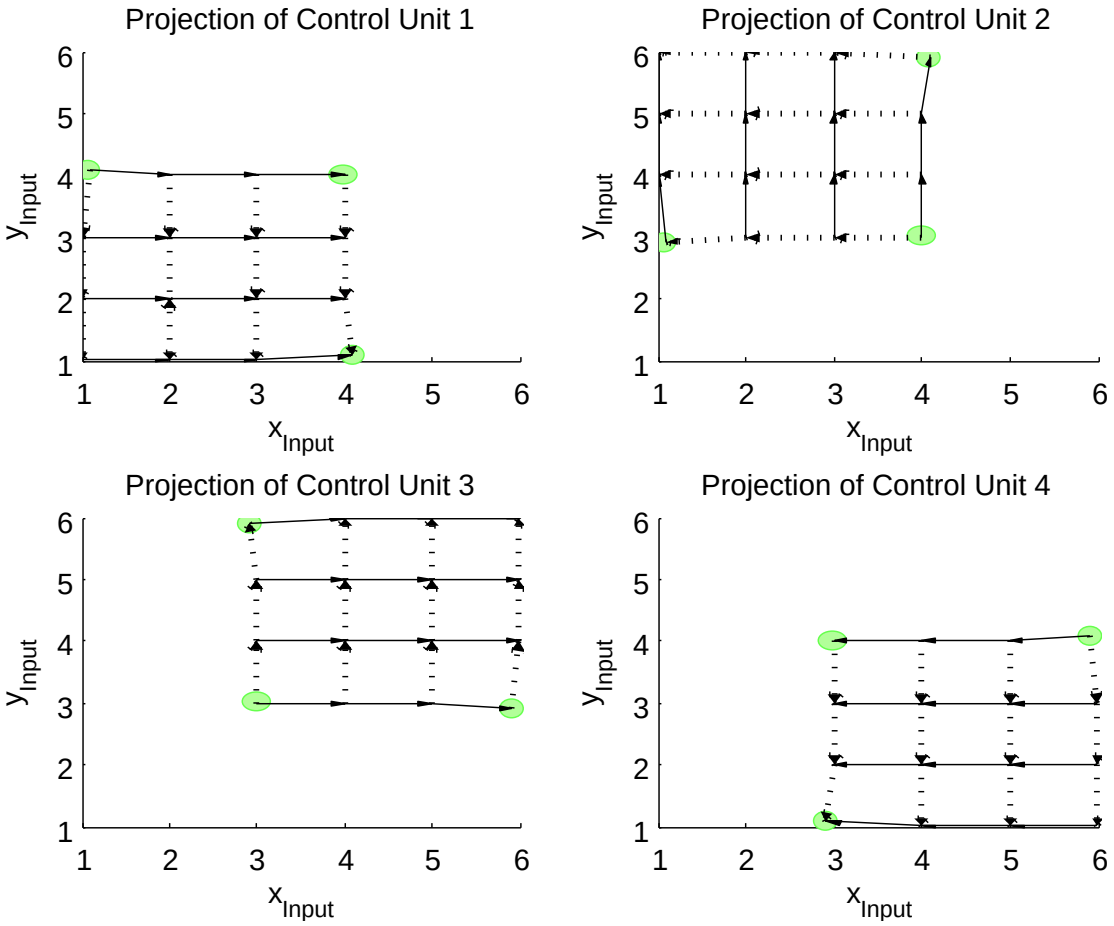


Figure 3.11: 2D Projections of 4x4 output units to the 6x6 input space for $K = 4$ control units. Shown are the weights in their (nearly) converged state after $t = 10000$ iterations. The horizontal and vertical axes in each plot denote the coordinates in the input layer. Compared to the intermediate stage of learning (Figure 3.10) the RPFs are more balanced in size and input specificity. Note that although the RPFs changed between the two Figures, their mirror-symmetries are conserved.

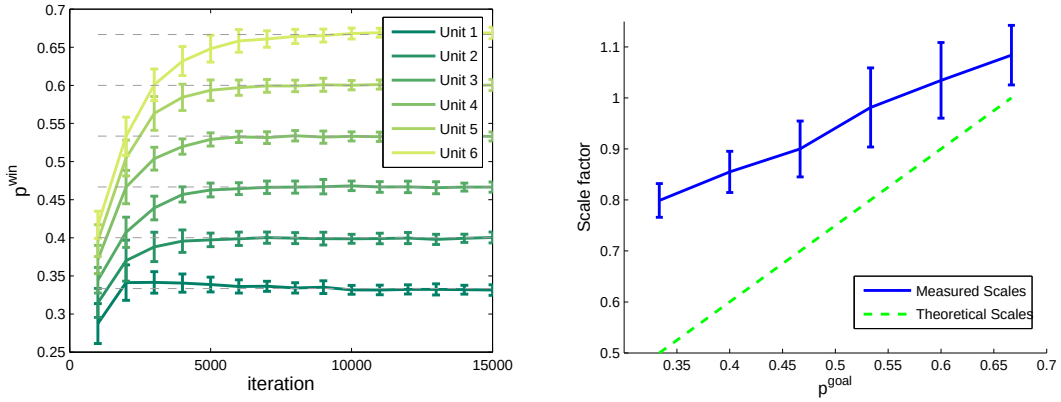


Figure 3.12: Development of transformations varying in scale based on different a priori goal probabilities p^{goal} of the control units. Left: The measured winning probabilities of the 6 control units in the simulations reliable converge to their respective goal winning probabilities (gray dashed lines). Right: Higher winning probabilities translate to bigger scale factors of the transformations. The green dashed line shows the theoretically calculated scales. Error bars are estimated over 50 trials.

active regions. Hence, control units implementing big scales should get activated more often than controls for smaller scales. We implemented this idea by imposing a different a priori winning probability for different control units, i.e. we set the values of p_k^{goal} to different values depending on k in equations 3.16 and 3.17. In addition, as different transformations map from overlapping input areas, different control units need to win for the same input region.

For the simulations, we used Gaussians with a standard deviation of $1/12 N_i$, with N_i=60 being the number of input units. The a priori goal probabilities of the $K = 6$ control units were set linearly from $p_1^{goal} = 0.33$ to $p_6^{goal} = 0.66$, thus summing to $N_{win} = 3$ winners for each input region. The winners were selected by a k-WTA mechanism: the three units with the biggest overlap with the current input region (see equation 3.14) were allowed to organize their RPFs. All other parameters were as in the other 1D simulations.

Figure 3.12 (left panel), shows the development of a leaky integrator-based estimation (with timescale $8 \cdot 10^{-4}$) of the winning probabilities p_k^{win} for 50 different trials of the simulation. The 6 gray dashed lines indicate the 6 goal win probabilities p_k^{goal}. We can see from the Figure that the control unit winning probabilities converge reliably to their respective goal winning probabilities within approximately $3000 - 6000$ iterations. The measured scale factors on the right panel of Figure 3.12 show that increased winning probability indeed translates to bigger scale factors of the transformations. The green dashed line shows the theoretical scale factors as calculated by the input area that is expected to be covered by a transformation from its goal winning probability. The measured transformations are bigger for mainly two reasons: first, as already discussed in subsection 3.4.1, due to activity leakage in the input field and second, due to the finite

size of the Gaussian input regions we used.

3.7 Discussion

The model we propose in the current chapter is based on a generalization of classic retinotopy mechanisms (Willshaw and von der Malsburg, 1976; Häussler and von der Malsburg, 1983). The essential difference is that shifter circuits in the form of a whole set of topographic mappings are installed, each of which can be switched on and off under the command of a control unit. We implemented this idea employing bilinear networks (Tenenbaum and Freeman, 2000; Grimes and Rao, 2005; Olshausen et al., 2007). Our simulations show that different types of transformation — translation, scaling, rotation and reflection — can be reliably self-organized. Our main focus is on the investigation of the one-dimensional case but we also demonstrate the extension to two dimensions. After eye opening (corresponding to the state when the proposed mechanism has converged) the network is immediately able to build invariant representations of patterns, relying on the prenatally self-organized transformations. Further, it is potentially able to separate contents and transformations of input patterns, hence both "what" and "where" information are explicitly represented. The emergent network therefore facilitates further development of invariant recognition and can, in its subsequent life, adapt to the statistics of real-world transformations as well as build up a memory for content (e.g. knowledge of faces and objects).

The necessary active input regions for our model could arise in random locations as small and gradually growing and migrating activity regions. They may correspond to the retinal waves as observed in prenatal mammals (Meister et al., 1991; Warland et al., 2006; Huberman et al., 2008), projected up to visual cortex and/or they might emerge spontaneously in cortex (Chiu and Weliky, 2001). In our model, these patterns, varying in position and size, serve as teaching signals to control units and lead to transformations that project topographically to an invariant window in an upstream cortical area. Interestingly, although at first sight simply-connected regions of activity in the neural input field seem crucial for our model, simulations on non-simply connected inputs indicate that this is not necessary.

If indeed the active input regions necessary for the organization of alternate projection patterns are generated spontaneously in retina or cortex, in a way simulating the postnatal appearance of segmented figures separated from a background, the self-organization of projection patterns can take place prenatally or before eye-opening. This is not only biologically desirable but is even likely to alleviate the organization process decisively. In line with the idea that retinal waves mimic the postnatal appearance of segmented figures, is the observation that waves propagate more often along the nasal-temporal than the dorsal-ventral axis of the retina Anishchenko and Feller (2009).

Our analysis of the temporal development of connectivity showed that control unit specificity to input region location and size develops rather quickly, Figure 3.6 (left). It reaches a maximum value that depends on the time constant of the homeostatic regulation which evens out the firing probability of the control unit. If this time constant is made too

short, the accidental over-representation or under-representation of some input regions acts to lower the specificity that is finally reached. Note that high specificity values are reached in spite of the absence of inhibitory weights of the control units (apart from those that may be required to implement the WTA mechanism, equation 3.17), consistent with the finding that GABA plays an excitatory role prenatally (Ben-Ari et al., 2007). The refinement of topography, as measured by $s(k,t)$ in Figure 3.6 (right), progresses somewhat more slowly than unit specificity, although regions of small size may take longer by more than an order of magnitude. Note that this relative order of specificity vs. topography holds for the case of $\alpha = 0$. For the more complex case of a non-vanishing α, which allows for the migration of whole mappings, the learning rate can be increased. In this case, it is possible for topography to develop before control unit specificity and before subsequent re-organization/migration, as demonstrated by the 2D results.

We would like to draw attention to the effect of the unspecific weight growth parameter α in equation 3.8 on the point spread (or sampling) function of the resulting mappings, see Figure 3.8. With increasing α, the final sampling function shows more and more low-pass filtering. Note that the optimal sampling function (the inverse Fourier transformation of the step function, a $\mathrm{sinc}(x)$ function) would need the inclusion of inhibitory connections between input and output.

3.8 Future Perspectives and Conclusion

There are several issues that will need further consideration. Receptive field size increases within the cortical hierarchy and each level therefore can be expected to have a different feature basis system to optimally represent the statistics of its inputs. For example, the receptive fields of primary cortical neurons are specific for stimulus orientation and size. When shifter circuits are to be used to normalize the image of a given object under change not only in position but also size and orientation, then also feature types are to be transformed (as modeled in (Sato et al., 2006) for the simplified assumption of higher cortical areas sharing the same representation as V1) to establish correspondence to a stored model. What has been treated here as a single link between the input and output fields therefore could be interpreted as a whole trunk of connections between all feature units in the image and model points connected by the link. In the functional adult state, the system must be able to activate links that establish the correct correspondences not only between points but also between feature units. Appropriate control structure for the latter will have to be modeled in future work.

The correspondence mapping problem requires a control space of very high dimensionality. It is quantitatively unrealistic to assume the existence of a separate control unit (as our model suggests) for each combination of retinal location, size and orientation. An obvious solution would be to factorize this high-dimensional space, so that one set of control units is responsible for position, another for scale, a third for orientation. A given link would then be activated under the influence of several control units. To cope with deformation, the control units should not encompass the whole projection from an input segment to the output field but should, as proposed in (Olshausen et al.,

1993) or (Zhu and von der Malsburg, 2004), control only the projections between smaller patches in the input and output fields. Further, the number of parameters, which is cubic for the three-way weights used, could be reduced significantly by factorizing them into outer products (Memisevic and Hinton, 2010). This is not only a more efficient way of representing the transformations, but might also speed up the prenatal learning process, due to the smaller parameter space.

An interesting future extension of the system would include the additional organization and incorporation of feature-preferences like prenatal orientation-specificity as has been observed in V1 (Wiesel and Hubel, 1974; White et al., 2001). Emergence of orientation-specificity has already been modeled by Linsker for linear networks (Linsker, 1986) and more recently also exploiting waves (Grabska-Barwinska and von der Malsburg, 2008). A first step in incorporating feature organization is done in chapter 5, yet we test this model only on postnatal natural inputs.

Another simplifying assumption of our model is the assumption of direct links between input field and output field, which would require an unrealistic number of fibers to converge on a single target unit. As proposed under the names of dynamic connections (Feldman, 1982) or shifter circuits (Olshausen et al., 1993) this problem can be solved, in analogy to telephone exchange systems, by making several consecutive line selections. This is also in line with anatomical and physiological evidence of intermediate cortical areas between V1 and IT. As shown in an optimization study (Wolfrum and von der Malsburg, 2007b), quite modest and realistic numbers of intermediate layers and convergence/divergence factors (and correspondingly modest numbers of control units) are sufficient to connect a million points in V1 to an area in IT. We therefore investigate a model, which can develop on a shifter circuit in the next chapter.

4 Multi-layer Organization of Translations

This chapter shows that a bilinear model with local competition can account for the emergence of translations. In contrast chapter 3, where translations, scalings and rotations were organized, the model proposed in this chapter has no specific assumptions on its input stimuli, except that neighboring cells are correlated (which retinal wave inputs are, see Figure 2.1). Therefore, the model builds exclusively on self-organization and can be analyzed formally by linear approximations (Zhu et al., 2010). Further, we show that the model can be applied to shifter circuits (Anderson and van Essen, 1987; Wolfrum and von der Malsburg, 2007b), a multi-layered routing structure that is necessary to save resources (e.g. links), and that is still able to connect any point in the input to any point in the output.

Parts of this work have been published in (Bergmann and von der Malsburg, 2008; Zhu et al., 2010).

4.1 Model Description

As in chapter 3, the outputs of the system are given by a bilinear model, see equation 3.6, with the 3-way connections w_{koi}, by which an active control unit k activates a topographic mapping between two neuronal chains of units, which are indexed by o and i. The purpose of the system in this chapter is to self-organize the control unit mappings w_{koi} to form different translations for all k. In a postnatal stage, a control unit k can then be activated to compensate for translations of input stimuli. We restrict the discussion of the proposed system to the one dimensional case.

The system is a generalization of the original Häussler system, see section 3.2, which is the special case for $K = 1$ control units:

$$\dot{w}_{koi} = \alpha + w_{koi}F_{koi}(W) - w_{koi}B_{koi}(\alpha + WF), \tag{4.1}$$

where we define $(WF)_{koi} = w_{koi}F_{koi}$. The system incorporates the necessary interactions for topography emergence: competition of incoming and outgoing links and neighborhood cooperation. For more than one control unit, the fundamental ingredient to guarantee different translations is competition of the emergent mappings. In contrast to the model proposed in chapter 3, where the competition was implemented by a competition of whole control units for input stimuli, we here propose that the competition is implemented at the level of single connections: each control unit competes with all other units for the control of each single link in a hard Winner-Take-All (hWTA) fashion, see Figure 4.1.

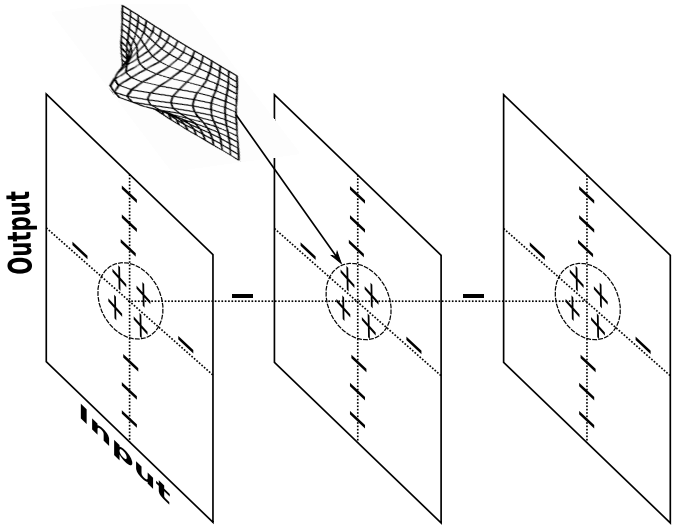


Figure 4.1: Interaction of three different mappings: In addition to the original Häussler interactions – competition of incoming and outgoing links and neighborhood cooperation (denoted by a Gaussian in the Figure) – the model is extended with hWTA competition of different control units over control for the same link. This leads to the emergence of mutually non-overlapping translations in this link space, if the map orientations of all control units are the same. To guarantee the same orientations, sequential organization, control unit cooperation and shifter circuit constraints can be imposed.

In the following, as in the previous chapter, the term "map orientation" refers to the two possible solutions of a translation in 1D, an ascending and a descending solution, as discussed in section 2.3.1, see Figure 2.2. As different translations of the same map orientation are topographic and mutually non-overlapping in link space (the outer product of the input and output, as links connect all-to-all), they are solutions to the system. The competitive B-term therefore gets extended to the k dimension:

$$B_{koi}(M) = \frac{1}{3}\left(\frac{1}{N}\sum_{o'} m_{ko'i} + \frac{1}{N}\sum_{i'} m_{koi'} + \frac{1}{K-1}\sum_{k'\neq k} m_{k'oi}\right), \tag{4.2}$$

where the sum over the control units excludes self-competition to comply with the original Häussler system for the case of $K = 1$, and where $M = (m_{koi})$ is a parameter to B. The additional last competition term can be implemented locally, and therefore in a biologically plausible way. For example, by competition of growing neurites of control units for a neurotrophic factor released at a link.

Unfortunately, translations of different 1D map orientation are not necessarily mutually non-overlapping in link space. In particular, the implementation of all possible translations of each map orientation needs all possible input-output connections and therefore covers the whole link space. Any translation of the respective mirror orientation therefore competes at all its links. Hence, different map orientations interfere with the organization

process. Therefore, the system has to organize all mappings with the same orientation. A possible solution would be to impose initial conditions, which are (strongly) biased to a single map orientation. As this, however, would itself be part of the solution to the problem, we show several other solutions to deal with this problem.

Sequential organization. In a developing animal, it is very unlikely that all control units develop at the same time. A single non-interacting control unit, however, is able to break the orientation symmetry spontaneously (or depending on it's random initial conditions), see chapter 2. Suppose we add a single additional control unit: all mappings of the opposite orientation to the first one lead to competition, while translations of the same orientation are non-overlapping and therefore do not compete. The latter ones therefore have a significant growth benefit. This leads to the sequential development of translations of the same orientation as the initial one.

Link Cooperation. An alternative way to consistently break the orientation symmetry is to let the control unit links cooperate with their neighbors. A control unit that already developed a slight orientation preference therefore imposes this preference on its neighbors as well. Further, as neighboring links are supported, the final translations can be expected to be ordered topographically with respect to the index of the translation, i.e. neighboring control units can be expected to implement similar translations. The weight cooperation term F_{koi} is derived from local bilinearly generalized Hebbian learning in section 4.1.1 and is of the form:

$$F_{koi}(W) = \sum_{k',o',i'} C_{kk'oo'ii'} w_{k'o'i'}, \tag{4.3}$$

where C is a cooperative coupling matrix. The derivation shows that the C matrix is separable, $C_{kk'oo'ii'} = C^K_{kk'} \hat{C}^O_{oo'} \hat{C}^I_{ii'}$. For the simulations we constructed C as an outer product of Gaussians.

Shifter Circuit Constraints. All-to-all connections between two sheets of neural tissue are very expensive as they scale quadratically. For the number of neurons in visual cortex, the number of connections therefore becomes prohibitive. A possible solution is to introduce intermediate layers, which allow a significant reduction in the number of links needed (Anderson and van Essen, 1987; Wolfrum and von der Malsburg, 2007b). Figure 4.2 shows a simple one-dimensional example for the case of $N = 9$ input and output units and $l = 2$ layers[1], that has been organized using optimization principles as suggested in (Wolfrum and von der Malsburg, 2007b). The number of links in these networks reduces to $lN^{\frac{l+1}{l}}$, thus is quadratic for the one layer case but less for more layers.

It has been shown that shifter circuits can be grown prenatally (Wolfrum and von der Malsburg, 2007a) and we therefore can assume a corresponding network to be present before we start the control unit organization for consistent topographic maps. The orientation symmetry in these networks is already broken, constraining control unit neurites to shifter circuit links therefore solves the problem of interfering orientations.

[1] For multi-layer architectures we refer to the number of "layers" to be the number of weight layers and not the number of unit layers.

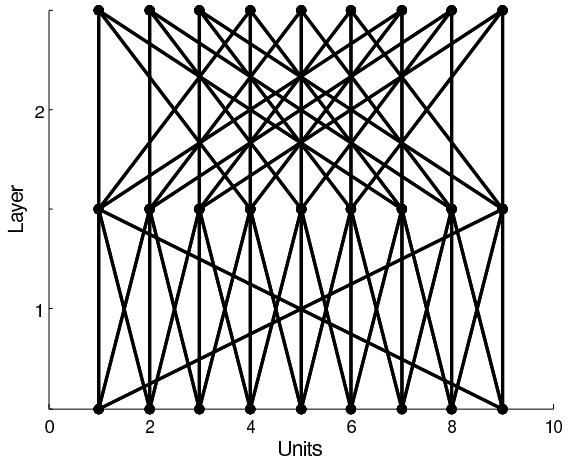


Figure 4.2: A shifter circuit with $N = 9$ input and output units and $l = 2$ weight layers. The shown shifter circuit has been optimized for minimal resources (links and units), as described in (Wolfrum and von der Malsburg, 2007b).

4.1.1 Derivation of the Weight Cooperation from Local Hebbian Rules

In this section we show that the non-local coupling matrix $C_{kk'oo'ii'}$ of equation 4.3 can be derived from local interactions. For this derivation, we assume three different populations of neurons: input units, output units and control units. Input units and control units play a similar role in that they multiplicatively feed the output units. We assume the input drive to the output units, $\hat{y}$, to be given by a simple bilinear model (see section 1.1.3):

$$\hat{y}_o = \sum_{k'} \sum_{i'} w_{k'oi'} c_{k'} x_{i'}, \tag{4.4}$$

where y_o, x_i and c_k denote the firing rates of output, input and control units at positions o, i and k, respectively.

Taking lateral interactions in each layer into account, the equilibrium state of the output activities as a function of the given input drives in each layer can generally be approximated by a linear equation (see chapter 3.3.4 or von der Malsburg (1995); Bergmann and von der Malsburg (2011) for a derivation):

$$x_i = \sum_{i'} C_{ii'} \hat{x}_{i'}, \tag{4.5}$$

We assume that the input layer as well as the control layer get i.i.d. (independent and identically distributed) noise, with $\langle \xi_i \xi_j \rangle_t = \delta_{ij}$, as inputs. The activity of the units in each layer is then fully determined by equations 4.4 and 4.5.

Given these assumptions we now derive the expected bilinear generalized Hebbian plasticity term:

$$\begin{aligned} F_{koi} &= \langle c_k y_o x_i \rangle_t \\ &= \sum_{k'o'i'} C^K_{kk'} \hat{C}^O_{oo'} \hat{C}^I_{ii'} w_{k'o'i'}, \end{aligned} \tag{4.6}$$

where the second line results from substituting equations 4.4 and 4.5 and using the i.i.d. property of the noise, where we defined:

$$\hat{C}^O_{oo'} = \sum_{o''} C^O_{oo''} C^O_{o'o''} \tag{4.7}$$

$$\hat{C}^I_{ii'} = \sum_{i''} C^I_{ii''} C^I_{i'i''}. \tag{4.8}$$

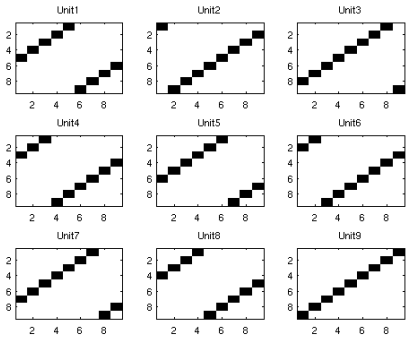
4.2 Simulations

We here present simulations of systems with $N = 9$ input and output units under the various modes for breaking the orientation symmetry, discussed in the model description in section 4.1. The differential equations 4.1 are integrated by Euler's method with a stepsize of $\Delta t = 0.05$. The 3-way weights w_{koi} are initialized to be close to the systems's homogeneous stationary state $W_0 = (1 - \epsilon)\mathbf{1} + \epsilon\Xi$, where Ξ is a random matrix with each entry is sampled uniformly from the interval $[0, 1]$. For all simulations $\epsilon = 0.1$. The unspecific weight growth rate is $\alpha = 0.1$. The cooperation function C in equation 4.3 is set to a separable Gaussian with standard deviation $\sigma = 1.125$ for input and output dimensions, while the standard deviation in the control unit dimension depends on the experiment performed. The Gaussians are normalized to one for each dimension $\sum_{x'} C^X_{xx'} = 1, \forall x$, where x denotes the input i, the output o or the control unit dimension k. If not mentioned otherwise, the simulations are performed using periodic (wrap-around) boundary conditions.

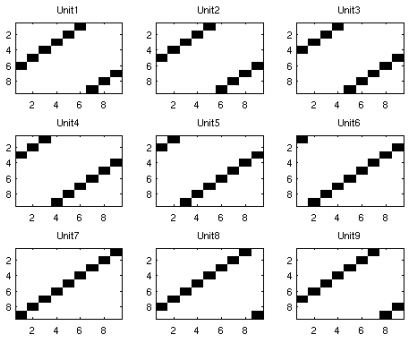
The final, converged weight configurations are shown in Figure 4.3. For the sequential organization process, a single control unit was organized first and every 10 timesteps (or every $10/\Delta t$ iterations) a new unorganized control unit was added. The effective lateral interaction for the control units was set to $C^K_{kk'} = \delta_{kk'}$, i.e. there was no lateral cooperation of control units. The final weight configurations in Figure 4.3a of the 9 organized units show that all units organized different translations of the same orientation.

In simultaneous organization mode, for subfigure 4.3b, the effective lateral interaction of the control units was set to a Gaussian with standard deviation $\sigma = 1.125$ and all 9 control units were organized simultaneously. Similar to the sequential case, all control units organized different translations of the same orientation. However, control units are topographically organized (with periodic boundaries) with respect to the transformation they implement, i.e. neighboring control units implement neighboring translations. For example, control unit 7 implements the mirrored identity mapping, while control unit 6 implements a translation of one pixel to the right and control unit 8 corresponds to a translation of one pixel to the left.

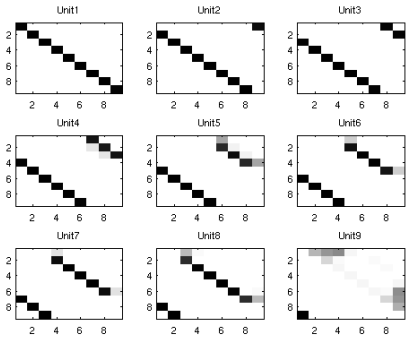
Combining cooperative coupling of control units with sequential organization allows for the organization of translations with non-periodic boundary conditions. This is not possible in simultaneous mode, as the identity and its mirror version have more support in this case than other translations and therefore both tend to get organized. As the control units compete in a hard WTA fashion for each link and the existence of the identity and



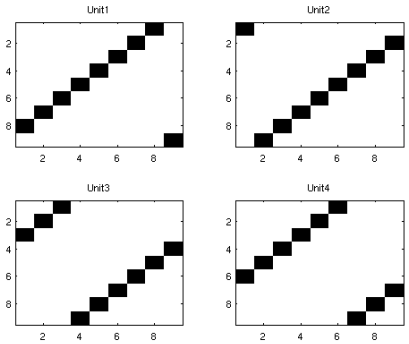
(a) Sequential



(b) Cooperative Controls



(c) Non-periodic boundaries



(d) 4 units

Figure 4.3: Converged weight configurations for $N = 9$ input and output units. The y-axis in each plot shows the output while the x-axis the input dimension. In the sequential organization mode in subfigre (a), all units organize different translations of the same orientation, while subfigure (b) shows that cooperative coupling of control units yields also a topography of neighboring control units, that is they implement close translations. Subfigure (c) combines sequential organization with cooperative coupling and allows for the organization of mappings with non-periodic boundary conditions. Finally, subfigure d) shows that the number of control units can differ (here $K = 4$) from the number of input and output units.

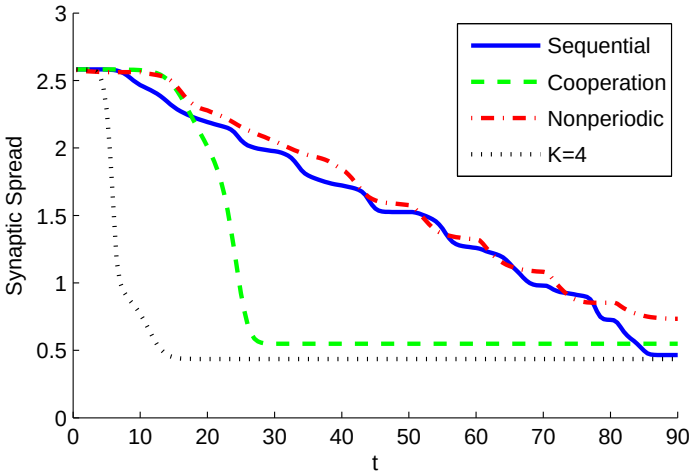


Figure 4.4: Synaptic spread $s(k)$ development for the different modes. Sequential organization ("Sequential" and "Nonperiodic" in the figure) is in general slower than the simultaneous organization ("Cooperation" and "$K = 4$") of topographic maps.

its mirror version (or translated versions thereof) would share single links, this disrupts the organization process. Figure 4.3c shows that in sequential mode the first control unit becomes the identity (or its mirror version, depending on the initial conditions), because it has the biggest support from neighboring links. The cooperative coupling of control units then most strongly supports links on both sides of this diagonal for the subsequent control unit. A symmetry breaking occurs and one side is chosen, here the lower diagonal. For the third control unit the support from the second unit is bigger compared to the first one, due to the monotonous decrease of $C_k(k - k')$. Therefore, the sum of the first two cooperative terms is bigger below the already organized mappings compared to the diagonal above the identity. Hence, the symmetry is already broken and the third mapping evolves below the second one for the case shown. This process continues until all translations are organized. Note that due to the nature of the dynamics, one link in each column and row survives. Hence, the mappings become "pseudo-periodic". Note however, that the side, which is not determined by the cooperative coupling of other previous control units, does not necessarily organize the same orientation and develops wrong link connections. This is why the last ($k = 9$) control unit is not able to organize a translation, due to the competitive influences from all other mappings that have wrong links. The first 8 control units were very stable in organizing translations over many trials.

Finally, Figure 4.3d shows that the number of control units does not need to be equal to the number of units in the input and output layers for simultaneous organization. The standard deviation of the Gaussian for cooperative coupling of control units was set to $\sigma = 0.5$ and its normalization was set to $\sum_{k'} C^K_{kk'} = 2, \forall k$.

We analyzed the development over time of the different modes of organizing translations

by plotting the average synaptic spread $s(k)$, a measure of how refined the mappings are (see equation 3.33), of the control units. Figure 4.4 shows this development for the four different simulations performed so far. In general it can be seen, that simultaneous organization is faster than sequential organization. For the sequential organization it can be seen that the convergence of the weights of each control unit to a translation becomes faster with each new mapping, as the search space gets smaller (the competition from the already organized control units restrict the dimensionality of the weight space): this is particularly apparent for the last 10 timesteps, when the last control unit gets added. For simultaneous organization, it can be seen that decreasing the number of control units significantly reduces the plateau of big synaptic spread in the beginning, and hence speeds up organization. This is because with fewer control units, each control unit has more possible final configurations compatible with the dynamics and therefore convergence to a solution is faster.

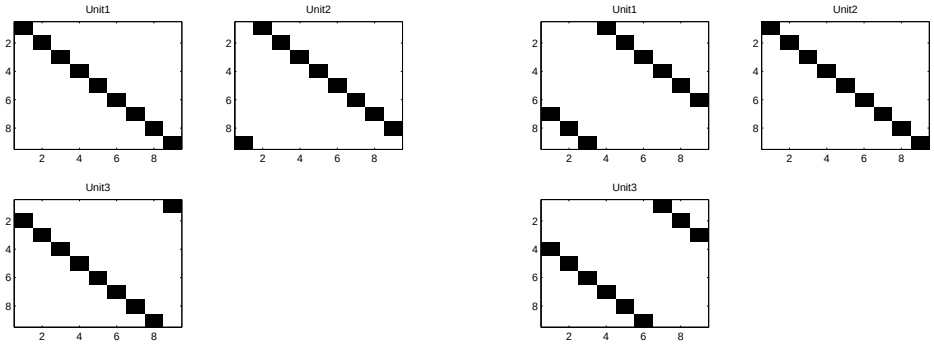
Shifter Circuit. As described in section 4.1, organization on shifter circuits saves resources and the orientation symmetry is broken. We performed simulations with the same parameters as the other simulations, but with no cooperative coupling of control units (i.e. $C^K_{kk'} = \delta_{kk'}$) and constraining the weights to the present links of a two-layered shifter circuit (with $N = 9$ units on all layers), see Figure 4.2. Sequential organization of the $K = 3$ control units on each layer, turns out to yield the correct topographic mappings, see Figure 4.5. For the cases considered, we found this also possible for simultaneous organization, in which case, however, diligent parameter tuning was necessary.

Each layer of the shifter circuit was simulated independently. A crucial ingredient to the derivation of the dynamics is lateral correlated noise, see section 4.1.1. Note that the projection of laterally correlated spontaneous random activity from the input to the intermediate layer, with subsequent relaxation to equilibrium (equation 4.5), again yields laterally correlated activity. Hence, a simultaneous simulation of all layers should be possible if the noise is simulated explicitly and not implicitly throught the dynamics of equation 4.1.

4.3 Conclusion and Discussion

In this chapter, we presented a dynamical system for the organization of translations in multiple control units. We analyzed two different modes of organization: sequential and simultaneous organization of all control units. It turned out that the sequential organization mode is slower than the simultaneous mode, but more stable in that it can also be used to organize mappings without periodic boundary conditions, due to less interference with other control units. Further, sequential organization needed less tuning and was also applicable to multi-layer networks, such as shifter circuits (Anderson and van Essen, 1987; Wolfrum and von der Malsburg, 2007b).

Comparing the model of this chapter to the one of chapter 3, there are several pluses on each side. The current model seems simpler and hence can be analyzed by linear approximation (Zhu et al., 2010). However, it is restricted to translations (at least in its current version). Further, the competition at local synapses, instead of the competition



(a) Layer 1 (b) Layer 2

Figure 4.5: Sequential organization without cooperative coupling of control units yields stable topographic mappings on a shifter circuit with $N = 9$ units on all 3 neuronal layers and $L = 2$ weight layers. The number of necessary translations on each layer is $K = 3$. (a) shows the final connections of the three control units if constrained to the first layer (the constraint matrix is equal to sum of the three converged weight matrices). (b) shows the result for 3 control units that organized their weights for the second layer.

of control units, as in chapter 3, makes it very sensitive to wrongly wired links, which through the competition term affect all control units. For example, all control units of the system need to have the same map orientation, as otherwise interfering competition occurs. This is not the case for the model of chapter 3, where different orientations pose no problem to the mechanism. A big advantage of the current model is the applicability of the model to shifter circuits, making it much less resource demanding.

It was shown that the non-local interactions of the proposed model can be derived from a purely local model with generalized Hebbian learning. However, for full biological or computer vision applicability, the model would have to be expanded and modified in various directions. A first step was to show its applicability to the case of non-periodic boundary conditions. Future work needs to assess the model's generalization to the two-dimensional case without boundary conditions. Furthermore, the incorporation of different feature types in the input and output, such as e.g. the orientation maps in V1, might help with the interference problem of different control units that organize mappings that go beyond translations.

5 Slowness yields consistent Features across Transformations in a Bilinear Model

We have seen in chapter 2 that a simple Gaussian generative model can account for the emergence of retinotopy, assuming neighborhood correlations in the retina. In this chapter, we extend this model to higher visual processing. In particular, we focus on the invariance problem (see section 1.1) by harvesting the power of multiplicative models, as introduced in section 1.1.3. We apply the model to real, natural image data to demonstrate that Gabor receptive fields, which are similar to cortical receptive fields, emerge. Parts of this chapter have been published in (Bergmann and von der Malsburg, 2010).

5.1 Introduction

It has long been known that models that take into account input statistics only up to second order, are not able to organize localized, bandpass-filter receptive fields. For a translation-invariant covariance matrix of the input data, it can be shown that PCA weights become sinusoids (Hyvärinen et al., 2009). For illustration, we applied Sanger's learning rule (Sanger, 1989) to natural image patches in Figure 5.1. This learning rule assigns the first unit to the weight vector with the highest variance in the input data. The second unit gets a weight vector which has highest variance in the orthogonal subspace to the first weight, and so forth. Hence, Sanger's rule is an online algorithm to arrive at weights identical to those of (offline) PCA. It can be seen from the Figure, that the DC component has highest variance (1st component), a horizontal low-frequency (2nd) and then a vertical low-frequency weight follow. This result is very consistent over many trials, and therefore reflects large enough steps in the variances associated with the weights. The weights of the fourth and higher units are not necessarily in the same order in each trial, as variance differences get smaller. But the general trend from low-frequeny to high frequency weight vectors is consistent over many trials and is due to the $f^{-\gamma}$ dependency in natural images, with $\gamma \approx 2$ (Field, 1987).

The models we built so far assume Gaussian responses for the firing rates of the neurons (this is made explicit in chapter 2). In contrast, cortical cells tend to fire with a non-Gaussian statistics. Figure 5.2, left, shows the firing rates of a cortical neuron in inferotemporal cortex (IT) of a macaque monkey freely watching a natural video (Baddeley

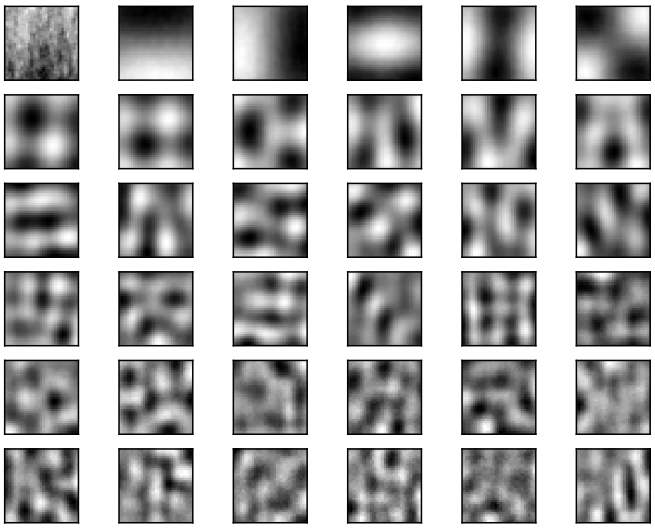


Figure 5.1: Sanger's learning rule (Sanger, 1989), which is an online version of principal component analysis (PCA), applied to natural image patches. The resulting weights are ordered by their variance from the upper left to the lower right. The weights converge to sinusoids with wave numbers that decrease with variance in the input.

et al., 1997), as measured in a given time window. The distribution is very close to an exponential distribution.

Furthermore, as the resulting weights of second-order models are not localized for natural inputs, it is not possible to define or organize weights that are topographic with respect to the input positions. Thus, we extend the model to also take into account higher-order statistical dependencies, which have been shown to yield localized weights (Olshausen and Field, 1996; Bell and Sejnowski, 1997). To construct an efficient code for the outputs $\{y_o\}$, note that the joint entropies[1] are less or equal to the sum of the single entropies:

$$\mathrm{H}(\{y_o\}) \begin{cases} = & \sum_o \mathrm{H}(y_o) \text{ iff } p\left(\{y_o\}\right) = \prod_o p\left(y_o\right), \\ < & \sum_o \mathrm{H}(y_o) \text{ else,} \end{cases} \tag{5.1}$$

where $p\left(\{y_o\}\right)$ is the joint probability distribution over the outputs and $p\left(y_o\right)$ is the probability distribution for a single output o. Hence, if we guarantee information preservation, lowering individual entropies can reduce higher-order statistical dependencies. This has been done using *sparse priors* for the latent variables, which has been found to be the underlying principle to organize localized, band-pass receptive fields (Olshausen and Field, 1996; Bell and Sejnowski, 1997; Weber and Triesch, 2008; Savin et al., 2010). Further, sparse coding systems have been shown to solve the bars problem (Lücke, 2004b; Gros and Kaczor, 2010), in which the network is presented with statistically independent bars that are non-linearly superimposed and it is supposed to detect the

[1] We here use the standard definition of Shannon entropy: $H(X) = -\int p(X) \log\left(p(X)\right)$.

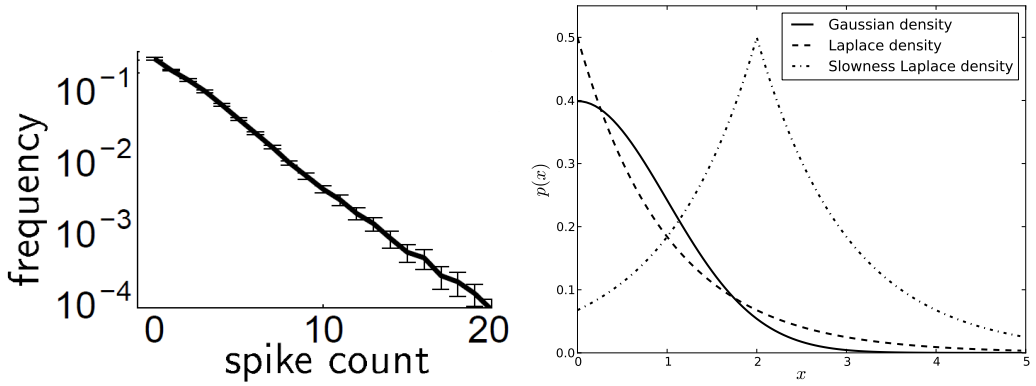


Figure 5.2: Probability density functions. Left: Response of a macaque IT cell to a natural video in a logarithmic plot (from Baddeley et al., 1997). Right: A Laplacian prior probability density function (pdf), i.e. $p(x) \propto \exp(-|x|)$ (dashed line), deviates from a Gaussian pdf (solid line), i.e. $p(x) \propto \exp(-x^2/2)$, in having higher probabilities for small and large values of x. To compensate, probabilities are lower in the intermediate range. All shown pdfs are normalized to $\langle p(x) \rangle = 1$. Moving the mean μ of the Laplacian, $p(x) \propto \exp(-|x - \mu|)$, can be used to implement slowness or top-down information (matching) in the model.

underlying generating patterns, that is, single bars. A sparse, or alternatively leptokurtotic or supergaussian (Barlow, 1972; Field, 1994) probability distribution deviates from a Gaussian distribution in having higher probabilities around and far away from the mean (also called a heavy-tail distribution), while it compensates this in having lower probabilities in the intermediate range (see Figure 5.2). Thus, compared with Gaussian priors, populations of units with sparse priors tend to code with fewer active units.

In this chapter, we develop a bilinear model with slowness in section 5.2 and show in section 5.3 that the proposed framework is able to organize topographic representations of the input, i.e. maps of different kind that are invariant to the simple transformations the network has been trained with. The model therefore builds "semantically ordered" invariant representations, in the sense of grouping statistically dependent outputs, while at the same time explicitly representing the underlying transformations that are necessary to yield invariance.

5.2 The Bilinear Topographic Model

In analogy to chapter 3, where we used a bilinear model in equation 3.6, we now define a probabilistic bilinear generative model. Here, the input data x_i is assumed to be generated by noise and by a bilinear transformation of the latent variables y_o and c_k:

$$x_i = \sum_{o}^{N_o} \sum_{k}^{K} g_{iok} y_o c_k + \eta_i, \tag{5.2}$$

where η_i is assumed to be Gaussian white noise with variance σ_i^2. Note that the generative weights g_{iok} are not identical to the "bottom-up" weights w_{koi} of chapter 3, but correspond to a generalization of the mixing matrix in ICA (Bell and Sejnowski, 1997) or the linear generative weights of sparse coding (Olshausen and Field, 1996). In the linear case, the generative weights can be shown to be closely related to the classical receptive fields of neurons. The latent variables y_o and c_k can be identified with the firing rates of cortical cells.

In the following we assume i.i.d. η_i, $\sigma_i = \sigma$, $\forall i$. Then the generative distribution of the generative model of equation 5.2 is:

$$p\left(\boldsymbol{x}|G,\boldsymbol{y},\boldsymbol{c}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N_i}{2}}}\exp\left(-\frac{\sum_i^{N_i}\left(x_i - \sum_o^{N_o}\sum_k^K g_{iok}y_o c_k\right)^2}{2\sigma^2}\right), \tag{5.3}$$

where N_i indicates the dimensionality of the input vector $\boldsymbol{x}$, N_o and K the dimensionality of the output and the control units, respectively.

5.2.1 Topography

We model the emergence of topography by following an idea discovered by Hyvärinen and Hoyer (2001): although the prior distributions over the causes in standard linear models, like sparse coding or ICA, bias the responses of the latent variables to be mutually independent, the responses to natural images are not fully independent. The reason is a too complex statistical structure of the natural images, which cannot be transformed to fully independent causes by linear models. If we now assume the cells coding for the latents y_o to project to another layer of cells z_l, we can however make use of the Central Limit Theorem (CLT) to order the y_o according to their residual statistical dependencies, simply by forcing the next-layer cells z_l to be non-Gaussian. This works, because if the cells y_o that project to cell z_l were independent, the response of z_l would tend to be more Gaussian. The resulting bilinear model, including this additional layer of units, is illustrated in Figure 5.3. The responses of the highest cells in the model are modeled similar to the standard energy model of complex cells: $z_l = \sum_o \Gamma(l,o)y_o^2$. The response of the cell z_l is therefore given by pooling over the squared outputs of nearby cells y_o. $\Gamma(l,o)$ defines the neighborhood function and is set to 1 if y_o is in the vicinity of z_l and otherwise to 0.

We do not model the units z_l explicitly, but instead employ an according prior for the output units y_o to implement the idea. As the resulting prior cannot be given in closed form under this missing-variables model, a lower bound approximation of the prior has been derived in (Hyvärinen et al., 2001):

$$\tilde{p}(\boldsymbol{y}) = \prod_l \exp\left(s\left(\sum_o \Gamma(l,o)y_o^2\right)\right). \tag{5.4}$$

It is well known, that the precise form of the sparsity forcing function s is not important for the results, as long as the overall shape of the function is correct. We use the sparsity

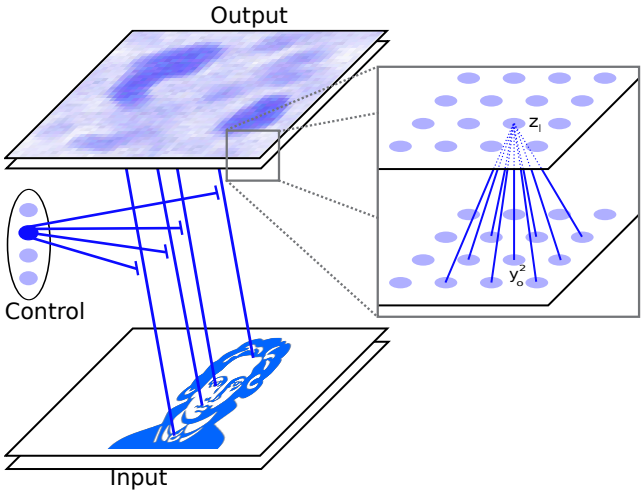


Figure 5.3: The bilinear generative model consists of three different groups of units: Feature units y_o, local pooling units z_l and control units c_k.

forcing function s as suggested in (Hyvärinen et al., 2001):

$$s(\xi) = -\alpha\sqrt{\xi + \epsilon}, \tag{5.5}$$

where ϵ guarantees numeric stability and is set to $\epsilon = 0.001$ for all our simulations. $\alpha = 2.0$ is an unimportant parameter, as it can be absorbed in the learning rate. Note that for $\Gamma(l, o) = \delta_{lo}$ and $\epsilon = 0$ the resulting prior becomes the standard Laplacian distribution (see Figure 5.2):

$$p(\boldsymbol{c}) = \prod_k \exp\left(-\lambda_c |c_k|\right), \tag{5.6}$$

which we used as a prior for the latent variables $\boldsymbol{c}$ with sparseness parameter λ_c.

5.2.2 Slowness

A central goal of the proposed model is to build topographic representations of the input data which are invariant to real-world transformations on the inputs (for example translation). Slowness has been proposed to play a central role in unsupervised learning of invariant responses (Földiák, 1991; Wiskott and Sejnowski, 2002; Wyss et al., 2006) and its rationale is that although the input data changes fast with time, the actual causes (e.g. objects) tend to change slowly with time. Hence, if neurons try to code with slowly changing responses, they are likely to catch the underlying invariant causes. Slow Feature Analysis (SFA) (Wiskott and Sejnowski, 2002) has been shown to be equivalent to probabilistic learning in a linear Gaussian model with an independent Markovian prior (Turner and Sahani, 2007). Similarly, we formalize this idea in our probabilistic model, by shifting the topographic prior probability in the direction of the latest responses,

thus increasing the probability of the estimates to lie closer to the latest values (see for an illustration the shifted Laplacian distribution in Figure 5.2):

$$p(\boldsymbol{y}) = \prod_l \exp\left(s\left(\sum_o \Gamma(l,o)\left(\frac{y_o - \hat{y}_o/\beta}{\sigma_s}\right)^2\right)\right), \tag{5.7}$$

where $\hat{y}_o$ is the output of the cell in the output field in the last timestep and σ_s parameterizes how much the estimated latent variable y_o is allowed to deviate from the β fraction of the last output $\hat{y}_o$. For the simulations we set $\sigma_s = 1$ and found the simulations to be not very sensitive to this parameter.

An important observation can be made for the case of $\beta = 1$. The inputs to our model are assumed to be selected randomly. Hence, the estimates $\hat{y}_o$ will be distributed randomly as well and follow a random walk. Due to the Central Limit Theorem (CLT), the $\hat{y}_o$ will therefore be normally distributed and with them the output variables y_o. This is undesired, as we wanted a non-Gaussian, sparse distribution for the output variables y_o. For the simulations, we therefore set $\beta = 3$ to weaken this effect, while at the same time we still have the advantages of slowness.

Note that slowness for the latents $\boldsymbol{y}$ is the main ingredient that breaks the symmetry of $\boldsymbol{y}$ and $\boldsymbol{c}$ and yields invariant responses in $\boldsymbol{y}$, while $\boldsymbol{c}$ codes for the transformations.

5.2.3 Dynamics and Learning Rule

After having defined the probabilistic model, we shall now derive the dynamics for the latent variables as well as the learning rule. Both can be derived by noting that the input statistics should be as close to the sample statistics from the generative model as possible. It has been shown in chapter 2 that this density estimation procedure corresponds to maximizing the average log likelihood $\langle \ln p(\boldsymbol{x}|G)\rangle$, where the marginal distribution is given by marginalizing over both $\boldsymbol{y}$ and $\boldsymbol{c}$:

$$p(\boldsymbol{x}|G) = \int p\left(\boldsymbol{x}|G,\boldsymbol{y},\boldsymbol{c}\right)\; p(\boldsymbol{y})\, p(\boldsymbol{c})\, d\boldsymbol{y}\, d\boldsymbol{c} \tag{5.8}$$

Unfortunately, the integral appearing in equation 5.8 is intractable. However, a common approximation is to replace the integral by an evaluation at the maximum a posteriori value (e.g. Olshausen and Field, 1996; Karklin and Lewicki, 2009):

$$(\hat{\boldsymbol{y}}, \hat{\boldsymbol{c}}) = \arg\max_{\boldsymbol{y},\boldsymbol{c}}\; p(\boldsymbol{y},\boldsymbol{c}|\boldsymbol{x},G), \tag{5.9}$$

thus, ignoring the volume around the maximum.

In general, maximizing the average log likelihood is equivalent to minimizing the estimate of code length:

$$\mathcal{L} \;\; = -\ln\left(p\left(\boldsymbol{x}|G,\boldsymbol{y},\boldsymbol{c}\right)\; p(\boldsymbol{y})\, p(\boldsymbol{c})\right). \tag{5.10}$$

By substituting the generative distribution, equation 5.3, and the prior distributions over the causes, equations 5.6 and 5.7, we arrive at the objective function of the model:

$$\begin{aligned}\mathcal{L} \quad &= \tfrac{N_i}{2}\ln\left(2\pi\sigma^2\right) + \textstyle\sum_i \frac{\left(x_i - \sum_o \sum_k g_{iok} y_o c_k\right)^2}{2\sigma^2} \\ &-\tfrac{\lambda_y}{2}\textstyle\sum_l s\left(\sum_o \Gamma(l,o)\left(\frac{y_o - \hat{y}_o/\beta}{\sigma_s}\right)^2\right) + \lambda_c \textstyle\sum_k |c_k|. \end{aligned} \tag{5.11}$$

For finding the minimum of the objective function, we define the residual image:

$$r_i = x_i - \sum_o \sum_k g_{iok} y_j c_k. \tag{5.12}$$

Then the gradients for the estimation of the latent variables become:

$$\begin{aligned}\frac{d}{dy_o}\mathcal{L} = \quad & -\tfrac{1}{\sigma^2}\textstyle\sum_i r_i \sum_k g_{iok} c_k \\ & -\lambda_y\left(\tfrac{y_o - \hat{y}_o/\beta}{\sigma_s}\right)\textstyle\sum_l s'\left(\sum_{o'} \Gamma(l,o')\left(\frac{y_{o'} - \hat{y}_{o'}/\beta}{\sigma_s}\right)^2\right)\Gamma(l,o), \end{aligned} \tag{5.13a}$$

$$\frac{d}{dc_k}\mathcal{L} = -\frac{1}{\sigma^2}\sum_i r_i \sum_o g_{iok} y_o + \lambda_c \operatorname{sgn}(c_k), \tag{5.13b}$$

where $s'(\xi)$ denotes the derivative of the function defined in equation 5.5. For the simulations, we used nonlinear conjugate (Polak-Ribière) gradient descent, using the log likelihood as defined in equation 5.11 and its gradient (equations 5.13) to estimate the latent variables $\boldsymbol{y}$ and $\boldsymbol{c}$.

After the estimation of the latent variables, a simple generalized bilinear Hebbian learning step on the generative weights follows:

$$\begin{aligned}\tilde{g}_{iok}(t) &= g_{iok}(t) + \eta_g \frac{d}{dg_{iok}}\mathcal{L} \\ \frac{d}{dg_{iok}}\mathcal{L} &= 2 r_i y_o c_k, \end{aligned} \tag{5.14}$$

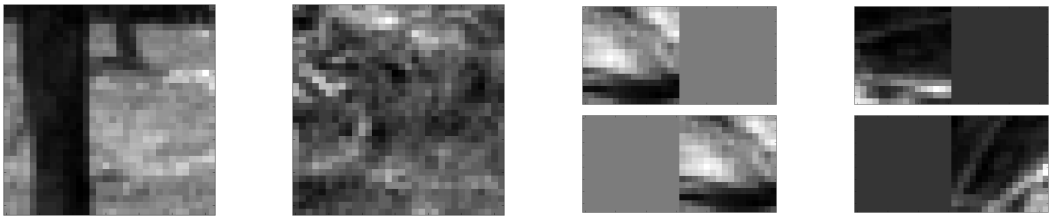
which is applied with a fixed learning rate $\eta_g = 0.005$. After each learning step the weights get L2 normalized:

$$g_{iok}(t+1) = \frac{\tilde{g}_{iok}}{\sqrt{\sum_i \tilde{g}_{iok}^2}}. \tag{5.15}$$

5.3 Simulations

5.3.1 Natural Inputs

We applied the proposed bilinear model to natural image data recorded by van Hateren and van der Schaaf (1998). Patches of size 32x32 were extracted at random positions



(a) 32x32 patch (b) 32x32 patch (c) Translated (d) Translated+45°

Figure 5.4: Subfigures (a) and (b) show randomly selected natural image patches of size 32x32 used for the simulations with a single control unit. For the case of two control units, 16x16 sized natural inputs were placed in succession in a 32x16 field (starting randomly on the left or the right side).

from the images for the single control unit simulations ($\dim(\boldsymbol{c}) = 1$), and patches of size 16x16 for the two control unit case. Figures 5.4a and 5.4b show two randomly selected patches. A common procedure in simplifying the estimation of higher-order models is to whiten the input data (for example in ICA). This means that second-order correlations are removed and variances of individual responses are normalized to one, to help the learning method to focus on higher-order dependencies. We did this algorithmically for all natural inputs using PCA. Note that whitening could be implemented by Oja's (or Sanger's) rule or via the model discussed in chapter 2 and is thought to be happening upstream of cortex in the LGN and the retina, where whitening can be related to the observed receptive fields (Dan et al., 1996; Graham et al., 2006). For the illustrations of the generative weights after learning, they are projected back to the input space.

In case of more control units, we placed the randomly extracted image patches at different positions in a bigger input field. To guarantee independent transformations underlying the input data, we placed the patches at disjoint positions. Figure 5.4c shows a pair of such inputs which are shown in succession, yet in random order, to the model and comply to a translation. In analogy to this procedure, different transformations have been used, for example a translation and rotation with 45° in Figure 5.4d. The presence of a single patch in a bigger field of vanishing activity can be interpreted as a schematic of early visual attention processing (Vanduffel et al., 2000; Hopf et al., 2006), which has been shown to supress activity peripheral to the representation of a stimulus.

5.3.2 Parameters

30000 randomly selected input patches were extracted as described in section 5.3.1 and each input was reduced to the 100 or 200 dimensions with highest variance using PCA, for the 16x16 or the 32x32 inputs, respectively. The assumed noise level in the input was kept constant at $\sigma = 1.0$ for all simulations. The sparseness parameters were set symmetrically to $\lambda_y = \lambda_c = 0.1$, i.e. $\boldsymbol{c}$ and $\boldsymbol{y}$ differ only in dimensionality and the estimation for $\boldsymbol{y}$ includes slowness. For their estimation, all components of $\boldsymbol{y}$ and $\boldsymbol{c}$ were initialized independently to a value drawn from a standard Gaussian distribution

with variance 0.1. Although we also tried bigger pooling sizes like 5x5 for $\Gamma(l, o)$, for all simulations shown the pooling size was a 3x3 neighborhood.

5.3.3 Results

In general, the learned generative fields are localized bandpass filters and are shown for the example of a single control unit in Figure 5.5. For a single control unit, the model essentially reduces to a linear one, with a single modulatory interaction with the control unit activity h, which can rescale and/or invert the activities of all feature coding units $\boldsymbol{y}$. The resulting basis functions are similar to the ones learned in ICA or sparse coding, with the additional property of being arranged topographically according to position, orientation and frequency, including clusters of low wave numbers and discontinuities in orientation, similar to pinwheels in cortex. As for TICA (Hyvärinen et al., 2001), the results resemble the cortical organization of receptive fields in V1.

To gain further insight into the resulting generative fields, we matched 2D Gabor functions to the results using least-square fitting as described in appendix A. Figure 5.6a shows a generative field of a single output unit from Figure 5.5 and Figure 5.6b shows the best Gabor fit. Both Figures look qualitatively similar and Figure 5.6c shows that the fit residual is comparatively small. Most other generative fields yielded very good Gabor fits, as well. A better impression of the generative field, the fitted Gabor function and the color coding of the Figures, can be gained by Figure 5.7, which shows a 3D surface plot of the result and the fit.

For the bilinear case of two control units, we shall first analyze the responses of the control units to the inputs. Figure 5.8 shows the development of the response of $\boldsymbol{c}$ to the inputs. For a given time, the responses of the last 150 estimations of c_2 are plotted against the responses of c_1. Red circles denote $\boldsymbol{c}$ estimates inferred from input patches from the left side in the input field, while green upside down triangles denote estimates from the right side. It comes at no surprise, that the initial responses of the control units are circular symmetric for both inputs, as can be seen from Figure 5.8a. This is clear, as all generative fields are random and the inputs are normalized. Learning does cluster the population code for the different inputs, i.e. specific combinations of control unit activities code for either the left or the right input, see Figure 5.8b. As learning proceeds, see Figure 5.8c, these clusters lie on one dimensional subspaces that are mutually orthogonal, thus reflecting a good code for the indepence in the inputs. Finally, sparseness forces the one dimensional subspaces to lie on the coordinate axes themselves, because this is the most sparse code possible, with only one unit coding for each input. Thus, the final code after generative field organization effectively resembles a Winner-take-All response, if the input is chosen accordingly, yet can be a population response, if necessary.

Figure 5.9 shows activity histograms of the output units y for the final 1000 iterations of a simulation. In figure 5.9a it can be seen that the output statistics fits well the prior Laplacian density. As has been discussed in section 5.2.2, for the case of $\beta = 1$ the output units in the case of two control units and active slowness follow a random walk (RW) that yields a Gaussian distribution, as can be observed in Figure 5.9b. Single inferences in this

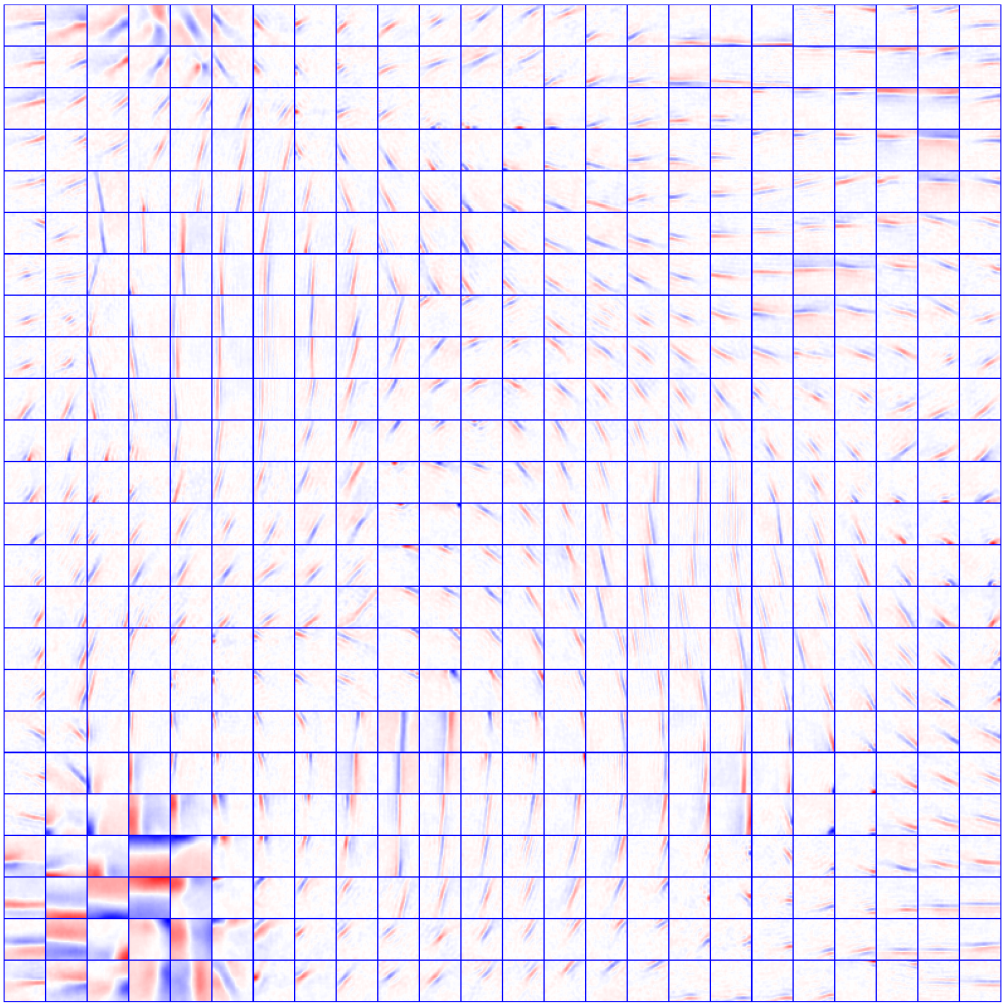
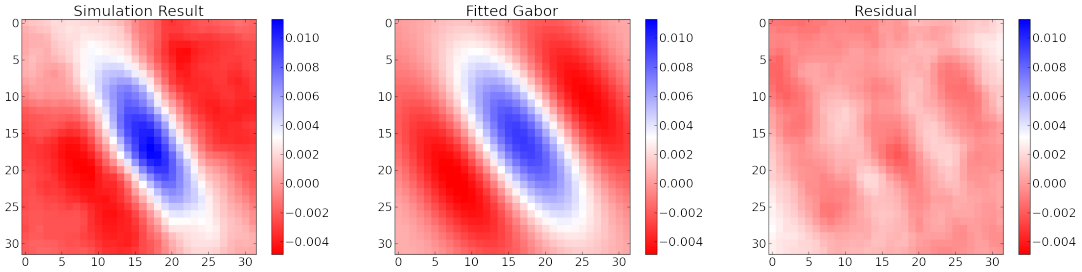
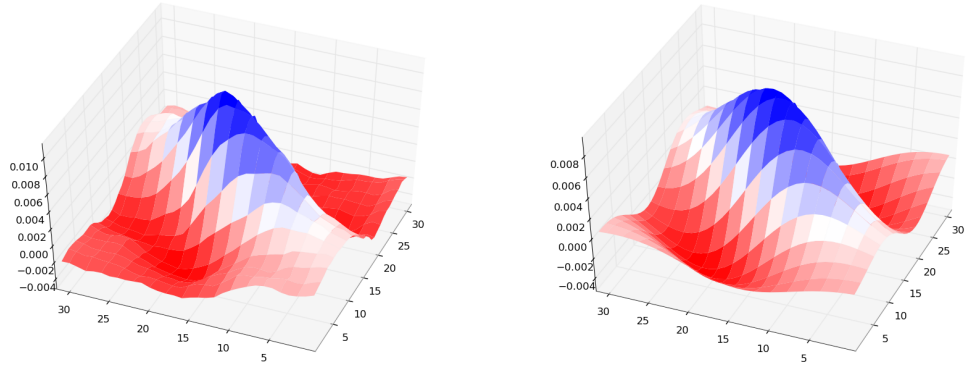


Figure 5.5: Generative bilinear weights without slowness for a single control unit. Input patch dimensions were 32x32 and the output units were arranged on a periodic 24x24 grid with 3x3 neighborhood pooling and periodic boundary conditions.



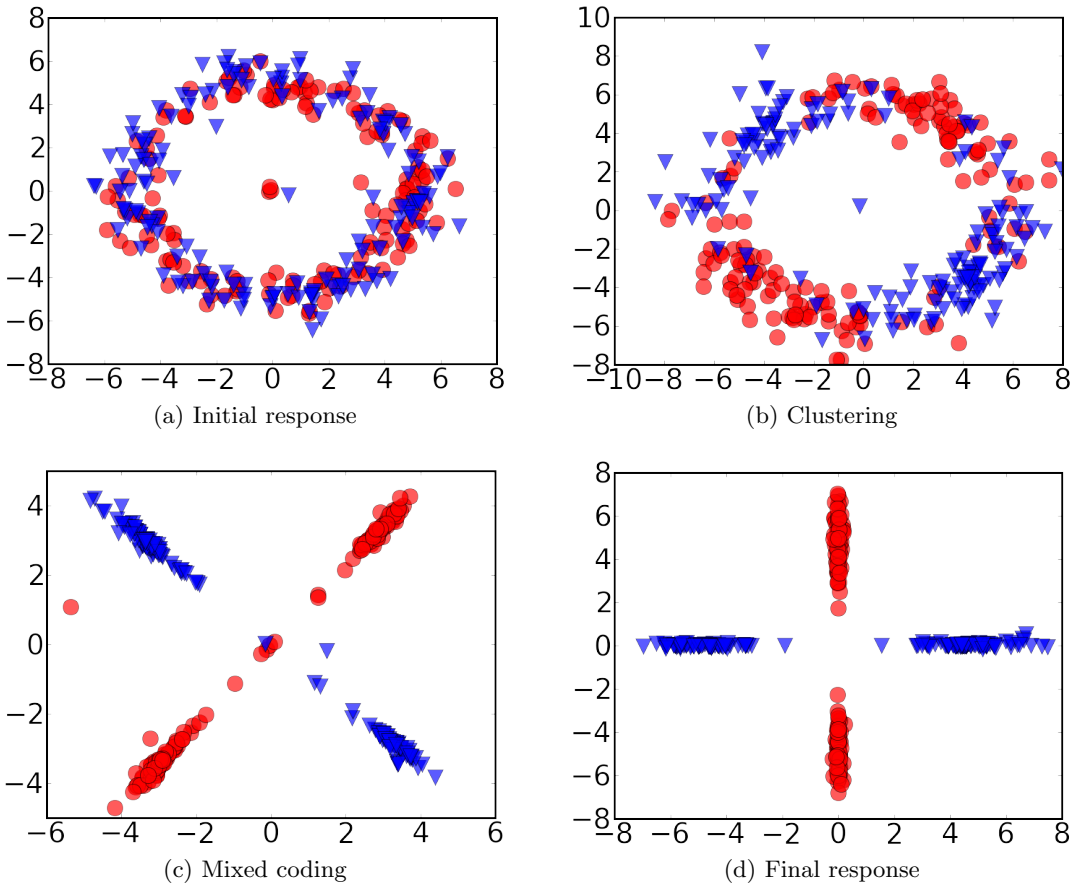
(a) Simulation Result (b) Fitted Gabor (c) Residual

Figure 5.6: Gabor fit analysis of a generative field from the simulation in Figure 5.5: (a) shows the resulting data, (b) the best fitted Gabor function (in the least squares sense) and (c) the residual, i.e. the result from a) minus b).



(a) Simulation Result (b) Fitted Gabor

Figure 5.7: An example generative field from the simulations to illustrate the color coding. Left: The resulting generative field from the simulations. Right: The best Gabor fit (in the mean least squared difference sense) to the generative field on the left.



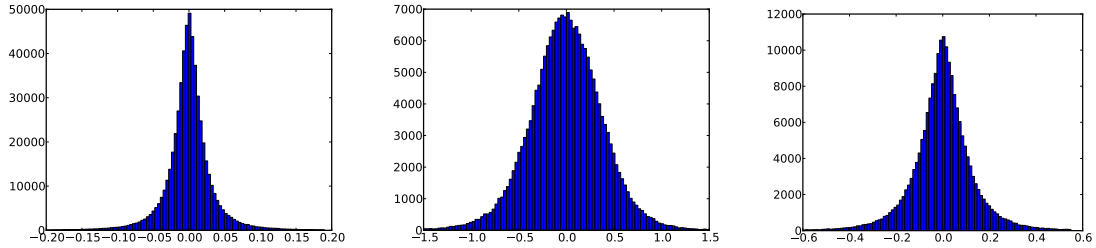
(a) Initial response

(b) Clustering

(c) Mixed coding

(d) Final response

Figure 5.8: Responses of the two control units to the two classes of transformed input patterns. In each subplot the last 150 estimations of c_2 are plotted against the responses of c_1. Red circles denote $\boldsymbol{c}$ estimates inferred from input patches from the left in the input field, while green upside down triangles denote estimates from the right. Subfigure (a) shows the initial response of the control units, while in subfigure (b), at $t = 5208$, it is shown that relatively short learning leads to clustering. Subfigure (c), at $t = 11444$, shows shrinkage to one-dimensional subspaces and subfigure (d) shows the final result with the sparsification of the responses.



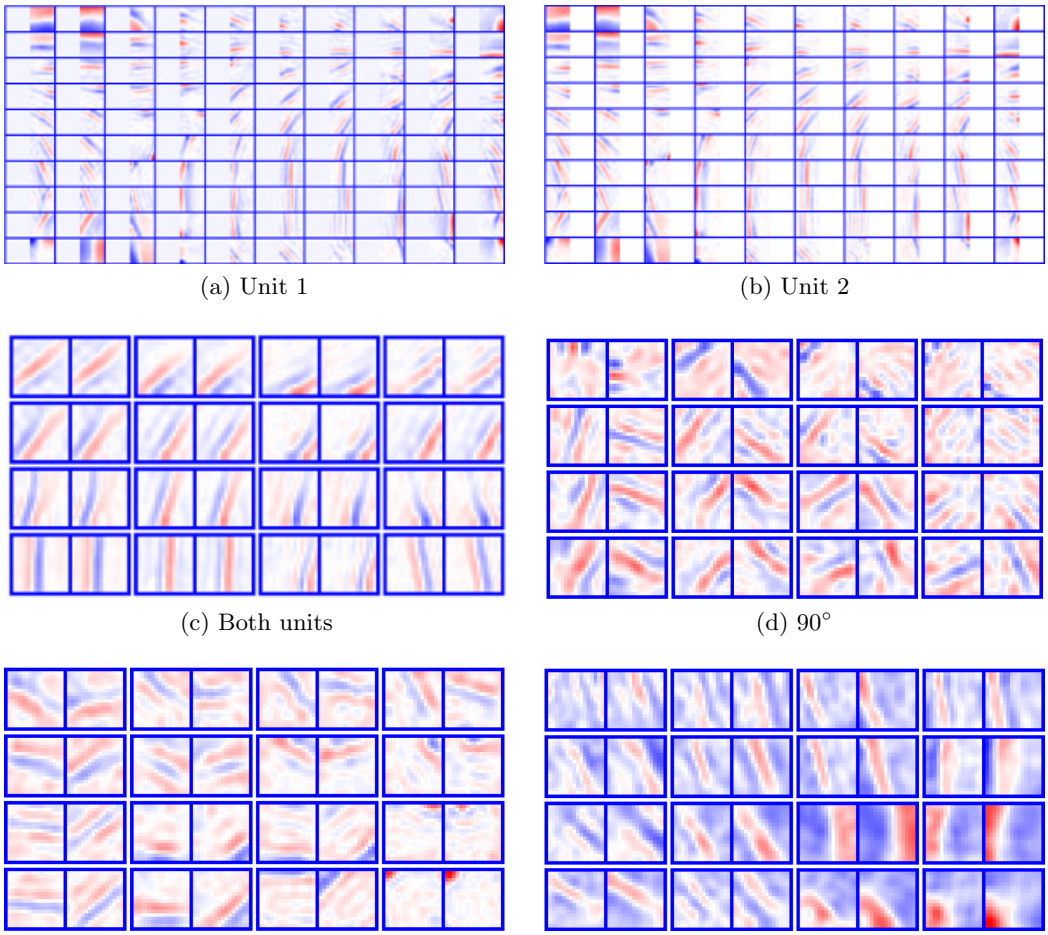
(a) Single Control (b) Two Controls RW (c) Two Controls no RW

Figure 5.9: The activity histograms of the output units are shown for different simulations for the last 1000 estimations of a simulation. Subplot (a) shows the activity histogram for a single control unit and inactive slowness. The distribution is Laplacian. In subplot (b) we see that for two control units and active slowness with $\beta = 1$, the distribution is Gaussian, as is predicted by the CLT and the random walk (RW) behavior of the estimates. (c) For $\beta = 3$ the outputs do not follow a random walk anymore, even in case of active slowness and the output distribution is sparse.

case still have a non-Gaussian distribution and therefore even in this case topography and oriented generative fields emerge. However, they are not as nicely localized as for the case of $\beta = 3$, for which the probability distribution deviates significantly from a Gaussian and is closer to a Laplace density (see Figure 5.9c). We therefore in the following show results for different β values.

The final generative fields can be visualized given an activity code for the control units. Figure 5.10a shows the resulting linear generative fields given $\boldsymbol{c} = (1, 0)$ and Figure 5.10b for $\boldsymbol{c} = (0, 1)$, respectively. Both simulations used a value $\beta = 3$. Due to the non-Gaussian statistics, the generative fields are nicely localized and oriented band-pass filters. Significant deviations from zero are only seen on one side of the fields, the side where the corresponding control unit got specialized in. We used this fact in the other subfigures of Figure 5.10, where only the non-zero parts of each given control unit field is plotted, side by side with the other control unit field, to simplify comparison of the generative fields. Figure 5.10c shows that slowness practically yielded identical generative fields for the two control units, despite the inputs being translated. Similarly, Figures 5.10d and 5.10e show that this result generalizes to orientations of the input stimuli around 90° and 45° degrees, respectively. The same holds if the right inputs are scaled 1.5 fold, see Figure 5.10f. For the simulations in Figure 5.10d to Figure 5.10f, the value of β was set to unity. Though not as nicely localized as for the simpler linear case without slowness or for $\beta = 3$, the generative fields are clearly oriented band-pass filters and some also show good localization.

Parameters extracted from the Gabor analyis, as described in appendix A, can be used to analyze the topographic continuity of the resulting generative fields with respect to the extracted parameters. For the case of a single control unit, maps extracted in this way are shown in Figures 5.11a to 5.11d. The first map, Figure 5.11a, shows that neighboring output units have similar input positions (the input position is coded with



(a) Unit 1 (b) Unit 2

(c) Both units (d) 90°

(e) 45° (f) 1.5 Upscaled

Figure 5.10: Generative weights for two control units. (a) Generative field given a control unit activity response of $\boldsymbol{c} = (1, 0)$ and in (b) for $\boldsymbol{c} = (0, 1)$. In this simulation $\beta = 3$. (c) shows that the non-zero parts of the generative fields are practically identical for both control unit activities, despite being translated. Similar, orientations around (d) 90°, (e) 45° and (f) 1.5x upscaling yields accordingly transformed generative fields. For (d)–(f) $\beta = 1$.

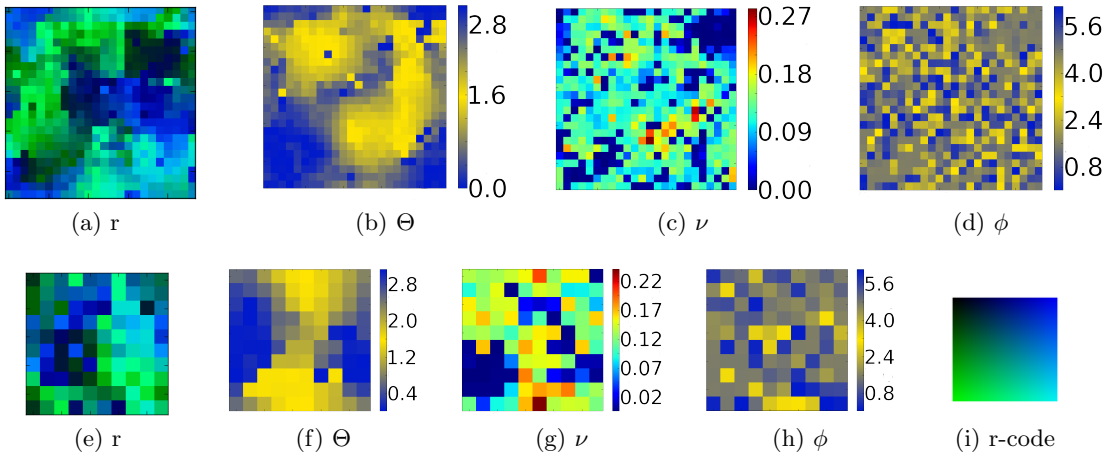


Figure 5.11: Visualization of the topographic maps with respect to position (r), orientation (Θ), frequency (ν) and phase (ϕ) of the generative fields extracted by fitting Gabor functions. Subplots (a)–(d) show results for a single control unit, while (e)–(h) show results for two control units. For the position plots, the color-coding in the input is given in subplot (i). Orientation is color-coded as given by the color-bar in $[0, \pi]$ and phases are given by their respective color-coding in $[0, 2\pi]$.

the colorcode shown in subfigure 5.11i), the map is continuous and mostly smooth with occasional discontinuities. This result corresponds to the finding of "retinotopy" in visual cortex. Figure 5.11b shows that the extracted orientations are even smoother, while for frequencies, Figure 5.11c, clusters can be seen, yet the map is noisy. Similar to simple cells in cortex, phases show no apparent topography, see Figure 5.11d.

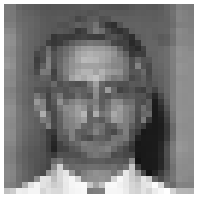
Figures 5.11e to 5.11h show the analogous results for the bilinear case and given $\boldsymbol{c} = (1, 0)$. The Gabor fits were restricted to the areas in the generative field that was significantly non-zero. Qualitatively the maps are very similar to the single control unit case. Orientation continuity and frequency topography is more pronounced, while position continuity is slightly more noisy.

5.3.4 Face experiments

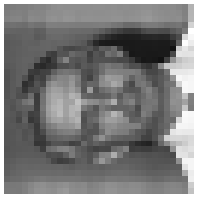
In addition to the experiments on natural inputs described in section 5.3.3, we applied the network to the FERET database (Phillips et al., 1998), a database containing facial images of 1196 individuals. We restricted ourselves to the fafb subset (containing images from 1195 individuals) of the FERET database, which contains frontal face photographs with varying facial expressions of the same subjects (see Figure 5.12a for 30 sample faces). The photographs have been taken on the same day. In contrast to the natural input data of section 5.3.3, face images are not rotation invariant, hence their statistics changes if the images are rotated. This should in principle allow for an inference of the angle of a presented face. Therefore, we tested if the current system is able to dissociate

(a) FERET faces

(b) Original



(c) Rescaled



(d) Rescaled+Rotated

Figure 5.12: Example images from the FERET database (Phillips et al., 1998). Subfigure (a) shows a sample of the subset fafb that we used for our simulations. The first face is shown in subfigure (b) in its original resolution, while subfigure (c) shows a rescaled version to 32x32 pixels - the size we used in the simulations. In temporal succession to the standard view of a face, an around 90° rotated version was shown, see subfigure (d).

rotated versions of faces. To this end, we presented a randomly selected image, which was rescaled to a 32x32 sized version (see Figure 5.12b for an example in its original size and Figure 5.12c for the corresponding rescaled version) and, in temporal succession, a 90° rotated version of the same image, to a network with two control units, 5x5 output units and slowness learning in the output layer $\boldsymbol{y}$. In contrast to the experiments with natural images, we here do not translate the inputs to disjoint positions, but the standard views of the faces and their rotated versions are presented at the same position. The response of the control units during learning was very similar to the responses shown in Figure 5.8 for natural inputs at different positions, the responses of the control units therefore essentially resembled a WTA responses, with each of the two control units firing for either the rotated or the standard view of a face. Similar to Figure 5.10, Figure 5.13 shows the generative weights given $\boldsymbol{c} = (1, 0)$ (Figure 5.13a) and Figure 5.13b for $\boldsymbol{c} = (0, 1)$, respectively. From the generative weights, we see that control unit 1 exclusively specialized to the 90° rotated version of the inputs, while control unit 2 specialized to the standard view of faces. The network therefore was able to detect the underlying rotations, that is, the hidden transformational causes, for the faces in this simple example. Importantly, the number of output units must be small (in comparison to the output dimension) for the case of faces, as otherwise control unit generative fields do not specialize to exclusive standard view or rotated versions. This is due to (nearly) rotation invariant components in faces that can be used for both versions.

(a) Control unit 1 (b) Control unit 2

Figure 5.13: Final generative weights for two control units and learning applied to the FERET database. Subfigure (a) shows the generative fields given a control unit activity response of $\boldsymbol{c} = (1, 0)$ and in subfigure (b) for $\boldsymbol{c} = (0, 1)$. Clearly, the two control units specialized on exclusively upright or rotated versions of faces.

5.4 Discussion

In this chapter, we introduced a topgraphic bilinear model for the purpose of building invariant representations. Like in (Hyvärinen et al., 2001), nearby output cell generative fields are organized to maximize the mutual information of their output activities. Using maximum likelihood estimation, we performed simulations on pre-whitened natural inputs. For the simplest case of a single control unit, most generative fields were well fitted with Gabor functions, which themselves are good fits to the linear receptive fields of the simple cells in V1. The Gabor analysis showed that topography emerges with respect to several parameters: location, orientation and frequency. This is akin to the corresponding maps observed in the cortex. The uppermost layer in the model represents complex cells and indeed complex cells have been shown to be constrained in their wiring length, a key assumption for topography in the model.

To investigate the invariance properties of the model, we used two control units and added slowness (Földiák, 1991; Wiskott and Sejnowski, 2002; Turner and Sahani, 2007) to the sparse prior of the output cells. For natural inputs presented at two disjoint positions the response of the control units becomes similar to a Winner-Take-All (WTA) response, due to the sparseness constraint on the control units. The model therefore is a consistent generalization of the models in chapter 3 and chapter 4, where a WTA mechanism is assumed to be at work. In case of only a translation of natural inputs, the generative fields of the two control units become similar, yet shifted versions, of a Gabor-like function. For additional input transformations, like rotations and scalings,

the generative fields transformed accordingly. Hence, it is reasonable to assume that the system learns representations of the input that are invariant to other transformations as well. Future work should attack the question how complex these transformations can be. Although the Gabor fits were not as good as for the simpler single control unit case, the resulting topographic maps extracted from the fits showed the same topographic features as in the simpler case.

Several models in the recent literature attempted to separate features and transformations using bilinear models (Tenenbaum and Freeman, 2000; Grimes and Rao, 2005; Olshausen et al., 2007; Memisevic and Hinton, 2010). All these models have in common that they do infer their latent variables on at least two successive images or a whole batch of images. In contrast, our model uses a single input image to infer both the feature and transformation unit activities. This is achieved by implementing slowness in the prior of the feature coding units, to break the symmetry with the control units, and simplified by the assumption of an attention signal in the input. Simple disjoint inputs were chosen for analysis purposes and future work will show how the model generalizes to more complex situations. A similar fully unsupervised bilinear approach, where each "identity unit" has its own linear subspace, was shown to reproduce V1 organization purely when learning from videos and it offered a new interpretation of simple and complex cells: simple cells are coding for the appearance of a feature while complex cells would code for their existence (Berkes et al., 2009).

Further, while all models for invariant recognition known to us either presume a topographic organization (Lades et al., 1993; Wolfrum et al., 2008, e.g.) or neglect topography altogether, our model learns topography from the higher-order dependencies in the input data. From this perspective the model can be seen as a bilinear extension of Topographic Independent Component Analysis with an additional relaxation of the completeness demand in ICA (Hyvärinen et al., 2001; Ma and Zhang, 2007).

Recent experiments indicate that, at least in some cases, learned shapes do not generalize across retinal position (Cox and DiCarlo, 2008) and even object identities can be confused depending on retinal position (Li and DiCarlo, 2008), hence there does not seem to be "full" invariance and invariance can be broken by changing the statistics of the environment (Li and DiCarlo, 2008; Wallis et al., 2009). This is in contrast to other studies, which showed full translation invariance (Biederman and Cooper, 1991). In the model of this chapter, the precise featural type as well as the position of a hidden cause (y_o) depends on the control unit activities, which are position specific. If input patterns that are presented in temporal succession and at varying positions correspond to different object identities, the slowness learning procedure therefore will lead to hidden causes, which vary accordingly with position. Thus, invariance is installed with respect to this statistics and therefore "broken". However, for a non-broken input statistics, the model develops the same hidden causes at different positions, as shown in the results, and therefore implements translation invariance. Therefore, if for the experiments that argue for non-existent translational invariance, new hidden causes must be developed, while for the experiments that claim the opposite, the set of hidden causes is sufficient, the model is consistent with both types of experiments - although they seem contradictory at first sight.

An interesting extension of the model would be to add a definition of sparseness also in time and not only in the population. It has been shown that an application of ICA to faces with independence in time leads to independent components that resemble localized face parts (for example the mouth) (Bartlett et al., 2002). Topographically ordering these should yield a parts-based representation that has, in contrast to holistic features, the flexibility for better representation of small topographic differences. Note that similar representations are used in object recognition systems (Lades et al., 1993; Wolfrum et al., 2008).

It is widely accepted in the neuroscientific community that the cortex is organized hierarchically (Felleman and Essen, 1991), and it would therefore be desirable to exploit this organization. However, standard linear methods in representational learning, like sparse coding (Olshausen and Field, 1996) or independent component analysis (Bell and Sejnowski, 1997), do not benefit from hierarchies, as representations would not change significantly in the hierarchy. The topographic bilinear model of this chapter, on the the other hand, is a well suited module in a hierarchy, due to its inherent non-linearity. Further, the shifting prior currently used for implementing slowness can be used for integration of top-down information, information from presumably higher cortical areas. In this way the bilinear model can be biased to a certain representation, for example a face, and hence would correspond to a generative version of matching.

6 Conclusion and Outlook

A central goal of this thesis was to demonstrate that invariance transformations, which are transformations that convert input patterns to a recognizable form, can be self-organized, both pre- and postnatally. Transformations as a basic principle to brain functioning, in contrast to the traditional view of mere feature detectors, have the significant advantage that they can be applied to unknown data. Feature detectors, on the other hand, are tuned to a specific set of input patterns. Therefore, the transformation paradigm, which inherently needs multiplicatory and therefore non-static connections, is vastly more powerful (von der Malsburg, 1981; Durbin and Rumelhart, 1989) in processing and generalization than traditional static neural networks. In particular, we believe that besides the advantages in the visual domain, the concept is also necessary for higher cortical processing or motor control. For example, grammatical rules in language, which are independent of specific words, could be implemented by transformations. Motor control can be seen as the inverse problem of the invariance problem, as here we try to get a specific pattern from a more abstract goal or representation. As the bilinear approach (Tenenbaum and Freeman, 2000) we have chosen is invertible, or even generative in chapter 5, these problems would be possible applications of the presented networks.

A crucial ingredient for the formation of the transformations was topography in the input and output domains. Therefore we investigated a model, which describes the establishment of retinotopic mappings in chapter 2. The model has the advantage that it is simple enough to offer analytical insights. Although the model's main motivation is to describe the formation of a topographic connectivity structure, it establishes a connection to information processing, because it is built on top of the idea of information preservation across cortical areas. Its key ingredients, like for all activity-based mechanisms for topography (Goodhill, 2007), were neighborhood correlations of cell activities in the input and the output layer and Hebbian learning. Chapter 3 extends an abstract model for topography formation (Häussler and von der Malsburg, 1983) to a bilinear framework, which allows for the organization of a set of invariance transformations. On the basis of schematic prenatal waves, it was demonstrated that the network can be used to organize translations, scales and orientations. The network was further analyzed in detail and its robustness to more complex stimuli and generalization to two dimensional input and output sheets was evidenced. In chapter 4 we presented a more compact model that was able to organize translations even without the assumption of prenatal waves. This simplification allows for easier analysis and was shown to be extensible to several layers. The extension to several layers is of particular importance, as it demands considerably less resources (Wolfrum and von der Malsburg, 2007b). Finally, we

derived a probabilistic model in chapter 5 that allowed for the simultaneous learning of transformations *and* features in a bilinear setting. By employing slowness and applying the model to transformed versions of natural inputs, the model was shown to organize efficient and consistent, or invariant, featural representations that compensate for transformations in the input data. The proposed model is in line with recent evidence (Li and DiCarlo, 2008; Wallis et al., 2009) which indicates that there is no full invariant recognition in monkeys and humans. Yet it also is compatible with older results (Biederman and Cooper, 1991), which claim there is full invariance, if the featural routing transformations are powerfull enough to estimate the invariant hidden causes of the input patterns.

An important note is that the central claim of chapters 3 and 4 is **not** that the organization of invariance transformations **stops** at birth or eye-opening and transformations are carved in stone for the rest of the animal's life. In contrast, the claim is that transformations can be organized, at least theoretically, before birth and therefore organization of systems for higher level vision **starts** before eye-opening. Support for this hypothesis comes also from theoretical experiments that show that lower level V1-like Gabor features can also be organized before birth by imposing sparseness (Albert et al., 2008) or slowness (Dähne et al., 2009) constraints on learning.

Future work should therefore address the integration of the prenatal and postnatal learning phases. The investigations we started in chapter 5 are a first step in this direction, and additionally incorporate feature learning, but so far without explicitly taking into account the prenatally organized initial conditions. To do this, more efficient representations for the transformations are necessary, like for example the approach recently proposed in (Memisevic and Hinton, 2010), so that the huge parameter space can be tackled efficiently. Using these networks it can then be investigated if it is possible to also organize the features in the topographic transformations before eye-opening or if this has to happen postnatally. Finally, the topographical organization makes these feature transformation networks an ideal candidate for a multilayer system like the brain (Felleman and Essen, 1991), as they are non-linear and have a neighborhood-sorted representation, that can be efficiently exploited by higher layers.

A Gabor Fitting of Generative Fields

We analyzed the spatial profile of the (bilinear) generative fields using a two-dimensional Gabor function (see Figure 5.7), which are commonly used in image processing for edge-detection and image analysis in the space and Fourier domain (Jähne, 2005). Gabor functions have been shown to very well match the receptive fields of simple cells in primary visual cortex (Jones and Palmer, 1987; Ringach, 2002). In the spatial domain, Gabor functions are sinusoidals modulated by a Gaussian envelope:

$$\Psi\left(x', y'\right) = A \exp\left(-\left(x'/\sqrt{2}\sigma_{x'}\right)^2 - \left(y'/\sqrt{2}\sigma_{y'}\right)^2\right) \cos\left(2\pi\nu x' + \phi\right), \tag{A.1}$$

where $\boldsymbol{x'} = (x', y')$ refers to a with $\boldsymbol{x}_0$ translated and with Θ rotated coordinate system:

$$\boldsymbol{x'} = \begin{pmatrix} \cos(\Theta) & -\sin(\Theta) \\ \sin(\Theta) & \cos(\Theta) \end{pmatrix} (\boldsymbol{x} - \boldsymbol{x}_0). \tag{A.2}$$

In the rotated coordinate system, the cosine only varies with frequency ν along the x' direction and like in (Ringach, 2002; Lücke, 2009), there was no need to vary the orientation of the Gaussian envelope with respect to the cosine plane wave. Thus, $\sigma_{x'}$ and $\sigma_{y'}$ give the width of the Gaussian envelope in the direction of the grating and orthogonal to it, respectively. Finally, A denotes the amplitude of the Gabor and ϕ the spatial phase of the cosine. The generative fields that resulted from the simulations, were matched to the Gabor functions using least squares error minimization. As this procedure can get stuck in local minima, each generative field was matched 30 times with different inital conditions of the Gabor function, and the best fit was selected.

Bibliography

Albert, Mark V, Adam Schnabel, and David J Field (2008), Innate visual learning through spontaneous activity patterns. *PLoS Comput Biol*, 4, e1000137.

Anderson, C. H. and D. C. van Essen (1987), Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proc Natl Acad Sci U S A*, 84, 6297–6301.

Anishchenko, Anastacia and Marla B Feller (2009), Go with the flow – but only in one direction. *Neuron*, 64, 152–154.

Arathorn, David W. (2002), *Map-Seeking Circuits in Visual Cognition*. Stanford University Press.

Baddeley, R., L. F. Abbott, M. C. Booth, F. Sengpiel, T. Freeman, E. A. Wakeman, and E. T. Rolls (1997), Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc Biol Sci*, 264, 1775–1783.

Barlow, H. B. (1972), Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*, 1, 371–394.

Bartlett, M. S., J. R. Movellan, and T. J. Sejnowski (2002), Face recognition by independent component analysis. *IEEE Trans Neural Netw*, 13, 1450–1464.

Bednar, James A. and Risto Miikkulainen (2004), Constructing visual function through prenatal and postnatal learning. *Neuroconstructivism*.

Bell, A. J. and T. J. Sejnowski (1997), The "independent components" of natural scenes are edge filters. *Vision Res*, 37, 3327–3338.

Ben-Ari, Y., J. Gaiarsa, R. Tyzio, and R. Khazipov (2007), Gaba: a pioneer transmitter that excites immature neurons and generates primitive oscillations. *Physiol Rev*, 87, 1215–1284.

Bergmann, Urs and Christoph von der Malsburg (2008), Ontogenesis of invariance transformations. In *Computational and Systems Neuroscience (Cosyne)*.

Bergmann, Urs and Christoph von der Malsburg (2010), A bilinear model for consistent topographic representations. In *ICANN, Part III, LNCS 6354*.

Bergmann, Urs and Christoph von der Malsburg (2011), Self-organization of topographic bilinear networks for invariant recognition. *Neural Computation (accepted)*.

Bergmann, Urs M., Reimer Kühn, and I.-O. Stamatescu (2009), Learning with incomplete information in the committee machine. *Biological Cybernetics*, 101, 401.

Berkes, Pietro, Richard E Turner, and Maneesh Sahani (2009), A structured model of video reproduces primary visual cortical organisation. *PLoS Comput Biol*, 5.

Bértolo, Helder, Teresa Paiva, Lara Pessoa, Tiago Mestre, Raquel Marques, and Rosa Santos (2003), Visual dream content, graphical representation and eeg alpha activity in congenitally blind subjects. *Brain Res Cogn Brain Res*, 15, 277–284.

Biederman, I. and E. E. Cooper (1991), Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, 20, 585–593.

Brugger, P., S. S. Kollias, R. M. Mri, G. Crelier, M. C. Hepp-Reymond, and M. Regard (2000), Beyond re-membering: phantom sensations of congenitally absent limbs. *Proc Natl Acad Sci U S A*, 97, 6167–6172.

Cang, Jianhua, Lupeng Wang, Michael P Stryker, and David A Feldheim (2008), Roles of ephrin-as and structured activity in the development of functional maps in the superior colliculus. *J Neurosci*, 28, 11015–11023.

Cassia, V. M., C. Turati, and F. Simion (2004), Can a nonspecific bias toward top-heavy patterns explain newborns' face preference? *Psychol Sci*, 15, 379–383.

Chalupa, L. M. and C. J. Snider (1998), Topographic specificity in the retinocollicular projection of the developing ferret: an anterograde tracing study. *J Comp Neurol*, 392, 35–47.

Chiu, C. and M. Weliky (2001), Spontaneous activity in developing ferret visual cortex in vivo. *J Neurosci*, 21, 8906–8914.

Connors, Barry W and Michael A Long (2004), Electrical synapses in the mammalian brain. *Annu Rev Neurosci*, 27, 393–418.

Cox, David D and James J DiCarlo (2008), Does learned shape selectivity in inferior temporal cortex automatically generalize across retinal position? *J Neurosci*, 28, 10045–10055.

Dähne, Sven, Niko Wilbert, and Laurenz Wiskott (2009), Learning complex cell units from simulated prenatal retinal waves with slow feature analysis. In *Proc. 18th Annual Computational Neuroscience Meeting, CNS 2009, Berlin, July 18–23.*

Dan, Y., J. J. Atick, and R. C. Reid (1996), Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J Neurosci*, 16, 3351–3362.

Dayan, Peter and L. F. Abbott (2001), *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press.

Deco, Gustavo and Edmund T Rolls (2004), A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res*, 44, 621–642.

Desai, N. S., L. C. Rutherford, and G. G. Turrigiano (1999), Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nat Neurosci*, 2, 515–520.

DeSieno, D. (1988), Adding a conscience to competitive learning. In *Neural Networks, 1988., IEEE International Conference on*, 117–124vol.1.

Drescher, U., C. Kremoser, C. Handwerker, J. Lschinger, M. Noda, and F. Bonhoeffer (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.

Duda, R., E. Hart, and D. Stork (2001), *Pattern Classification (2nd edition)*. John Wiley and Sons.

Durbin, R. and G. Mitchison (1990), A dimension reduction framework for understanding cortical maps. *Nature*, 343, 644–647.

Durbin, R. and D. E. Rumelhart (1989), Product units: A computationally powerful and biologically plausible extension to brackpropagation networks. *Neural Computation*, 1, 133–142.

Eigen, M. (1971), Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58, 465–523.

Einhäuser, Wolfgang, Christoph Kayser, Peter Knig, and Konrad P Körding (2002), Learning the invariance properties of complex cells from their responses to natural stimuli. *Eur J Neurosci*, 15, 475–486.

Feldman, J. A. (1982), Dynamic connections in neural networks. *Biol Cybern*, 46, 27–39.

Felleman, D. J. and D. C. Van Essen (1991), Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex*, 1, 1–47.

Feller, M. B., D. P. Wellis, D. Stellwagen, F. S. Werblin, and C. J. Shatz (1996), Requirement for cholinergic synaptic transmission in the propagation of spontaneous retinal waves. *Science*, 272, 1182–1187.

Field, D. J. (1987), Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc Am A*, 4, 2379–2394.

Field, David J. (1994), What is the goal of sensory coding? *Neural Comput*, 6, 559–601.

Földiák, P. (1991), Learning invariances from transformational sequences. *Neural Computation*, 3, 791–801.

Fukai, T. and S. Tanaka (1997), A simple neural network exhibiting selective activation of neuronal ensembles: from winner-take-all to winners-share-all. *Neural Computation*, 9, 77–97.

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.

Gabbiani, F., H. G. Krapp, N. Hatsopoulos, C. Mo, C. Koch, and G. Laurent (2004), Multiplication and stimulus invariance in a looming-sensitive neuron. *J Physiol Paris*, 98, 19–34.

Garaschuk, O., J. Linn, J. Eilers, and A. Konnerth (2000), Large-scale oscillatory calcium waves in the immature cortex. *Nat Neurosci*, 3, 452–459.

Gawne, T. J., T. W. Kjaer, J. A. Hertz, and B. J. Richmond (1996), Adjacent visual cortical complex cells share about 20% of their stimulus-related information. *Cereb Cortex*, 6, 482–489.

Gawne, T. J. and B. J. Richmond (1993), How independent are the messages carried by adjacent inferior temporal cortical neurons? *J Neurosci*, 13, 2758–2771.

Godfrey, Keith B and Nicholas V Swindale (2007), Retinal wave behavior through activity-dependent refractory periods. *PLoS Comput Biol*, 3, e245.

Goodhill, G. J. (2007), Contributions of theoretical modeling to the understanding of neural map development. *Neuron*, 56, 301–311.

Goren, C. C., M. Sarty, and P. Y. Wu (1975), Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, 56, 544–549.

Gosse, Nathan J, Linda M Nevin, and Herwig Baier (2008), Retinotopic order in the absence of axon competition. *Nature*, 452, 892–895.

Grabska-Barwinska, Agnieszka and Christoph von der Malsburg (2008), Establishment of a scaffold for orientation maps in primary visual cortex of higher mammals. *J Neurosci*, 28, 249–257.

Graf, Markus (2006), Coordinate transformations in object recognition. *Psychol Bull*, 132, 920–945.

Graham, Daniel J, Damon M Chandler, and David J Field (2006), Can the theory of "whitening" explain the center-surround properties of retinal ganglion cell receptive fields? *Vision Res*, 46, 2901–2913.

Grimes, D. B. and R. P. N. Rao (2005), Bilinear sparse coding for invariant vision. *Neural Computation*, 17, 47–73.

Gros, C. and G. Kaczor (2010), Semantic learning in autonomously active recurrent neural networks. *J. Algorithms in Cognition, Informatics and Logic*, 81.

Häusser, M., N. Spruston, and G. J. Stuart (2000), Diversity and dynamics of dendritic signaling. *Science*, 290, 739–744.

Häussler, A. F. and C. von der Malsburg (1983), Development of retinotopic projections: An analytic treatment. *J. Theor. Neurobiol.*, 2.

Haydon, P. G. (2001), Glia: listening and talking to the synapse. *Nat Rev Neurosci*, 2, 185–193.

Hendrickson, A. E. (1994), Primate foveal development: a microcosm of current questions in neurobiology. *Invest Ophthalmol Vis Sci*, 35, 3129–3133.

Henneberger, Christian, Thomas Papouin, Stphane H R Oliet, and Dmitri A Rusakov (2010), Long-term potentiation depends on release of d-serine from astrocytes. *Nature*, 463, 232–236.

Hertz, J., A. Krogh, and R. G. Palmer (1991), *Introduction to the theory of neural computation*. Addison Wesley.

Hinton, Geoffrey E. (1981), A parallel computation that assigns canonical object-based frames of reference. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*.

Hooks, Bryan M and Chinfei Chen (2006), Distinct roles for spontaneous and visual activity in remodeling of the retinogeniculate synapse. *Neuron*, 52, 281–291.

Hopf, J-M., C. N. Boehler, S. J. Luck, J. K. Tsotsos, H-J. Heinze, and M. A. Schoenfeld (2006), Direct neurophysiological evidence for spatial suppression surrounding the focus of attention in vision. *Proc Natl Acad Sci U S A*, 103, 1053–1058.

Hubel, D. H. and T. N. Wiesel (1968), Receptive fields and functional architecture of monkey striate cortex. *J Physiol*, 195, 215–243.

Huberman, A. D., M. B. Feller, and B. Chapman (2008), Mechanisms underlying development of visual maps and receptive fields. *Annu Rev Neurosci*, 31, 479–509.

Hyvärinen, A. and P. O. Hoyer (2001), A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Res*, 41, 2413–2423.

Hyvärinen, A., Patrick O. Hoyer, and Mika Inki (2001), Topographic independent component analysis. *Neural Comput*, 13, 1527–1558.

Hyvärinen, A., J. Hurri, , and P. O. Hoyer (2009), *Natural Image Statistics*. Springer-Verlag.

Hyvärinen, Aapo (2002), Topography as a property of the natural sensory world. *Natural Computing*, 1, 185–198.

Imai, Takeshi, Takahiro Yamazaki, Reiko Kobayakawa, Ko Kobayakawa, Takaya Abe, Misao Suzuki, and Hitoshi Sakano (2009), Pre-target axon sorting establishes the neural map topography. *Science*, 325, 585–590.

Jähne, Bernd (2005), *Digital Image Processing*. Springer.

Jeffery, G. (1985), Retinotopic order appears before ocular separation in developing visual pathways. *Nature*, 313, 575–576.

Jones, J. P. and L. A. Palmer (1987), An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol*, 58, 1233–1258.

Karklin, Yan and Michael S Lewicki (2009), Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457, 83–86.

Kiani, Roozbeh, Hossein Esteky, Koorosh Mirpour, and Keiji Tanaka (2007), Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J Neurophysiol*, 97, 4296–4309.

Koch, Christoph (1999), *Biophysics of computation: information processing in single neurons*. Oxford University Press.

Kubota, Yoshiyuki, Sayuri Hatada, Satoru Kondo, Fuyuki Karube, and Yasuo Kawaguchi (2007), Neocortical inhibitory terminals innervate dendritic spines targeted by thalamocortical afferents. *J Neurosci*, 27, 1139–1150.

Lades, M., J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen (1993), Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42, 300–311.

Larkum, M. E., J. J. Zhu, and B. Sakmann (1999), A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature*, 398, 338–341.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (1989), Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1, 541–551.

Lee, Daniel D. and Sebastian Seung (1999), Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788–791.

Lee, Tai Sing and David Mumford (2003), Hierarchical bayesian inference in the visual cortex. *J Opt Soc Am A Opt Image Sci Vis*, 20, 1434–1448.

Li, Nuo and James J. DiCarlo (2008), Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, 321, 1502–1507.

Linsker, R. (1986), From basic network principles to neural architecture: emergence of orientation-selective cells. *Proc Natl Acad Sci U S A*, 83, 8390–8394.

London, Michael, Arnd Roth, Lisa Beeren, Michael Husser, and Peter E Latham (2010), Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature*, 466, 123–127.

Loos, Hartmut S. and Christoph von der Malsburg (2002), 1-click learning of object models for recognition. In *Biologically Motivated Computer Vision 2002 (BMCV 2002)* (H.H. Bülthoff, S.-W. Lee, T.A. Poggio, and C. Wallraven, eds.), volume 2525, 377–386.

Lücke, J. (2004a), Hierarchical self-organization of minicolumnar receptive fields. *Neural Networks*, 17, 1377–1389.

Lücke, J. (2004b), *Information Processing and Learning in Networks of Cortical Columns*. Ph.D. thesis, Ruhr-Universität Bochum.

Lücke, J., C. Keck, and C. von der Malsburg (2008), Rapid convergence to feature layer correspondences. *Neural Computation*, 20, 2441–2463.

Lücke, Jörg (2009), Receptive field self-organization in a model of the fine structure in v1 cortical columns. *Neural Comput*, 21, 2805–2845.

Ma, Libo and Liqing Zhang (2007), A hierarchical generative model for overcomplete topographic representations in natural images. In *IJCNN*.

Markram, H. and M. Tsodyks (1996), Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature*, 382, 807–810.

Markram, H., Y. Wang, and M. Tsodyks (1998), Differential signaling via the same axon of neocortical pyramidal neurons. *Proc Natl Acad Sci U S A*, 95, 5323–5328.

Marmarelis, P. Z. and K. Naka (1972), White-noise analysis of a neuron chain: an application of the wiener theory. *Science*, 175, 1276–1278.

Meister, M., R. O. Wong, D. A. Baylor, and C. J. Shatz (1991), Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science*, 252, 939–943.

Memisevic, Roland and Geoffrey E Hinton (2010), Learning to represent spatial transformations with factored higher-order boltzmann machines. *Neural Comput*, 22, 1473–1492.

Möller, C., J. Lücke, J. Zhu, P. M. Faustmann, and C. von der Malsburg (2007), Glial cells for information routing? *Cognitive Systems Research*, 8, 28–35.

Morton, J. and M. H. Johnson (1991), Conspec and conlern: a two-process theory of infant face recognition. *Psychol Rev*, 98, 164–181.

Ohshiro, Tomokazu and Michael Weliky (2006), Simple fall-off pattern of correlated neural activity in the developing lateral geniculate nucleus. *Nat Neurosci*, 9, 1541–1548.

Oja, E. (1989), Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1, 61–68.

Olshausen, B. A., C. H. Anderson, and D. C. Van Essen (1993), A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J Neurosci*, 13, 4700–4719.

Olshausen, B. A., C. Cadieu, J. Culpepper, and D. K. Warland (2007), Bilinear models of natural images. In *SPIE Proceedings: Human Vision and Electronic Imaging XII* (B.E. Rogowitz, T.N. Pappas, and S.J. Daly, eds.), volume 6492, San Jose, California.

Olshausen, B. A. and D. J. Field (1996), Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381, 607–609.

Pfeiffenberger, Cory, Jena Yamada, and David A Feldheim (2006), Ephrin-as and patterned retinal activity act together in the development of topographic maps in the primary visual system. *J Neurosci*, 26, 12873–12884.

Phillips, P.J., H. Wechsler, J. Huang, and P.J. Rauss (1998), The feret database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 16, 295–306.

Pitts, W. and W. S. McCulloch (1947), How we know universals; the perception of auditory and visual forms. *Bull Math Biophys*, 9, 127–147.

Puchalla, Jason L, Elad Schneidman, Robert A Harris, and Michael J Berry (2005), Redundancy in the population code of the retina. *Neuron*, 46, 493–504.

Purves, Dale, George J. Augustine, David Fitzpatrick, William C. Hall, Anthony-Samuel LaMantia, James O. McNamara, and S. Mark Williams, eds. (2004), *Neuroscience (3rd edition)*. Sinauer Associates, Inc. Sunderland, Massachusetts U.S.A.

Quiroga, R. Quian, L. Reddy, G. Kreiman, C. Koch, and I. Fried (2005), Invariant visual representation by single neurons in the human brain. *Nature*, 435, 1102–1107.

Reich, D. S., F. Mechler, and J. D. Victor (2001), Independent and redundant information in nearby cortical neurons. *Science*, 294, 2566–2568.

Riesenhuber, M. and T. Poggio (1999), Hierarchical models of object recognition in cortex. *Nat Neurosci*, 2, 1019–1025.

Ringach, Dario L (2002), Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J Neurophysiol*, 88, 455–463.

Rosenblatt, F. (1961), *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, D.C.

Rothman, J. S., L. Cathala, V. Steuber, and R. A. Silver (2009), Synaptic depression enables neuronal gain control. *Nature*.

Salinas, E. and L. F. Abbott (1996), A model of multiplicative neural responses in parietal cortex. *Proc Natl Acad Sci U S A*, 93, 11956–11961.

Sanger, Terence D. (1989), An optimality principle for unsupervised learning. In *Advances in Neural Information Processing Systems 1* (D.S. Touretzky, ed.), 11–19, Morgan Kaufmann, San Mateo, CA.

Sato, Y.D., C. Wolff, P. Wolfrum, and C. von der Malsburg (2006), Dynamic link matching between feature columns of different scale and orientation. In *Proceedings ICONIP*.

Savin, Cristina, Prashant Joshi, and Jochen Triesch (2010), Independent component analysis in spiking neurons. *PLoS Comput Biol*, 6, e1000757.

Schaefer, Andreas T, Matthew E Larkum, Bert Sakmann, and Arnd Roth (2003), Coincidence detection in pyramidal neurons is tuned by their dendritic branching pattern. *J Neurophysiol*, 89, 3143–3154.

Schmitt, Adam M, Jun Shi, Alex M Wolf, Chin-Chun Lu, Leslie A King, and Yimin Zou (2006), Wnt-ryk signalling mediates medial-lateral retinotectal topographic mapping. *Nature*, 439, 31–37.

Snider, Joseph, Andrea Pillai, and Charles F. Stevens (2010), A universal property of axonal and dendritic arbors. *Neuron*, 66, 45–56.

Song, Sen, Per Jesper Sjstrm, Markus Reigl, Sacha Nelson, and Dmitri B Chklovskii (2005), Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol*, 3, e68.

Sperry, R. W. (1963), Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proc Natl Acad Sci U S A*, 50, 703–710.

Starck, J. M. and R. E. Ricklefs (1998), *Avian Growth and Development*, chapter 1: Patterns of Development: The Altricial-Precocial Spectrum. Oxford University Press, New York.

Sutton, R. S. and A. G. Barto (1998), *Reinforcement Learning: An Introduction*. The MIT Press.

Swindale, N. V. (1996), The development of topography in the visual cortex: a review of models. *Network*, 7, 161–247.

Tal, D. and E. L. Schwartz (1997), Computing with the leaky integrate-and-fire neuron: logarithmic computation and multiplication. *Neural Comput*, 9, 305–318.

Talavage, T. M., P. J. Ledden, R. R. Benson, B. R. Rosen, and J. R. Melcher (2000), Frequency-dependent responses exhibited by multiple regions in human auditory cortex. *Hear Res*, 150, 225–244.

Tanaka, K. (1996), Inferotemporal cortex and object vision. *Annu Rev Neurosci*, 19, 109–139.

Tenenbaum, J. B. and W. T. Freeman (2000), Separating style and content with bilinear models. *Neural Computation*, 12, 1247–1283.

Triesch, Jochen (2005), A gradient rule for the plasticity of a neuron's intrinsic excitability. In *Int. Conf. on Artificial Neural Networks (ICANN)*.

Turner, Richard and Maneesh Sahani (2007), A maximum-likelihood interpretation for slow feature analysis. *Neural Comput*, 19, 1022–1038.

van Hateren, J. H. and A. van der Schaaf (1998), Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc Biol Sci*, 265, 359–366.

Vanduffel, W., R. B. Tootell, and G. A. Orban (2000), Attention-dependent suppression of metabolic activity in the early stages of the macaque visual system. *Cereb Cortex*, 10, 109–126.

Volgushev, M., T. R. Vidyasagar, and X. Pei (1996), A linear model fails to predict orientation selectivity of cells in the cat visual cortex. *J Physiol*, 496 (Pt 3), 597–606.

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

von der Malsburg, C. (1981), The correlation theory of brain function. Technical report, MPI Biophysical Chemistry.

von der Malsburg, C. (1995), *An Introduction to Neural and Electronic Networks*, chapter 22: Network Self-Organization in the Ontogenesis of the Mammalian Visual System, 447–464.

Wallis, Guy, Benjamin T. Backus, Michael Langer, Gesche Huebner, and Heinrich Bülthoff (2009), Learning illumination- and orientation-invariant representations of objects through temporal association. *Journal of Vision*, 9, 1–8.

Wang, Gang, Shinji Obama, Wakayo Yamashita, Tadashi Sugihara, and Keiji Tanaka (2005), Prior experience of rotation is not required for recognizing objects seen from different angles. *Nat Neurosci*, 8, 1768–1775.

Warland, D. K., A. D. Huberman, and L. M. Chalupa (2006), Dynamics of spontaneous activity in the fetal macaque retina during development of retinogeniculate pathways. *J Neurosci*, 26, 5190–5197.

Weber, Cornelius and Jochen Triesch (2008), A sparse generative model of v1 simple cells with intrinsic plasticity. *Neural Comput*, 20, 1261–1284.

White, L. E., D. M. Coppola, and D. Fitzpatrick (2001), The contribution of sensory experience to the maturation of orientation selectivity in ferret visual cortex. *Nature*, 411, 1049–1052.

Wiesel, T. N. and D. H. Hubel (1974), Ordered arrangement of orientation columns in monkeys lacking visual experience. *J Comp Neurol*, 158, 307–318.

Willshaw, D. J. and C. von der Malsburg (1976), How patterned neural connections can be set up by self-organization. *Proc R Soc Lond B Biol Sci*, 194, 431–445.

Wiskott, L. and T. Sejnowski (1998), Constrained optimization for neural map formation: a unifying framework for weight growth and normalization. *Neural Computation*, 10, 671–716.

Wiskott, L. and C. von der Malsburg (1996), Recognizing faces by dynamic link matching. *Neuroimage*, 4, S14–S18.

Wiskott, Laurenz (2006), How does our visual system achieve shift and size invariance? In *23 Problems in Systems Neuroscience* (J. L. van Hemmen and T. J. Sejnowski, eds.), chapter 16, 322–340, Oxford University Press, New York.

Wiskott, Laurenz and Terrence J Sejnowski (2002), Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, 14, 715–770.

Wolfrum, P. and C. von der Malsburg (2007a), A marker-based model for the ontogenesis of routing circuits. In *Artificial Neural Networks – ICANN 2007*, volume 4669 of *LNCS*, 1–8, Springer.

Wolfrum, P. and C. von der Malsburg (2007b), What is the optimal architecture for visual information routing? *Neural Computation*, 19, 3293–3309.

Wolfrum, P., C. Wolff, J. Lücke, and C. von der Malsburg (2008), A recurrent dynamic model for correspondence-based face recognition. *J Vis*, 8, 34.1–3418.

Wong, R. O. (1999), Retinal waves and visual system development. *Annu Rev Neurosci*, 22, 29–47.

Wyss, Reto, Peter Knig, and Paul F M J Verschure (2006), A model of the ventral visual system based on temporal stability and local memory. *PLoS Biol*, 4, e120.

Zhang, W. and D. J. Linden (2003), The other side of the engram: experience-driven changes in neuronal intrinsic excitability. *Nat Rev Neurosci*, 4, 885–900.

Zhu, J. (2008), Synaptic formation rate as a control parameter in a model for the ontogenesis of retinotopy. In *Proc. ICANN*, volume 5164 II of *Lecture Notes in Computer Science*, 462–470, Springer.

Zhu, J. and C. von der Malsburg (2004), Maplets for correspondence-based object recognition. *Neural Networks*, 17, 1311–1326.

Zhu, Junmei, Urs Bergmann, and Christoph von der Malsburg (2010), Self-organization of steerable topographic mappings as basis for translation invariance. In *ICANN, Part II, LNCS 6353*.

Index

Zusammenfassung in deutscher Sprache

Dieses Kapitel enthält die deutsche Zusammenfassung der Dissertation. Die Einteilung der Kapitel entspricht der Einteilung der Arbeit. Unterkapitel wurden allerdings der besseren Lesbarkeit halber nicht übernommen. Fachbegriffe, die keine deutschen Übersetzungen haben, wurden in englischer Sprache beibehalten.

1 Einleitung

Ein großes Mysterium, dass die Funktionsweise des Gehirns betrifft, ist die Fähigkeit unterschiedlichste Muster miteinander in Verbindung zu setzen. Hierzu muss das Gehirn die zugrunde liegende Bedeutung, die Semantik, eines Musters erfassen. Dies ist kein triviales Problem, da die Muster hochdimensional sind und Muster mit gleicher oder ähnlicher Bedeutung oft keine einfachen mathematischen Zusammenhänge aufweisen – wie zum Beispiel Synonyme in der Sprache. In der vorliegenden Arbeit wird eine Theorie zur Selbstorganisation visueller Transformationen vorgestellt, die es ermöglicht abstraktere Repräsentationen zu erzeugen, die näher an der semantischen Bedeutung der Eingabemuster liegen.

Da das Säugetiergehirn eine lange Phase pränataler Ontogenese im Mutterleib durchläuft, haben sich zum Zeitpunkt der Geburt schon viele Verbindungsstrukturen organisiert. In der Mikrostruktur, das heißt auf der Ebene einzelner Synapsen, ist diese Verbindungsstruktur jedoch weitgehend unbekannt. Die Verbindungsstruktur zum Zeitpunkt der Geburt hat jedoch einen entscheidenden Einfluss auf das darauf folgende Lernen und die Organisation des Gehirns, weshalb es nötig ist mögliche pränatale Organisationsprinzipien aufzudecken, um Einblicke in die Mikrostruktur zu bekommen. Ein Fokus der Arbeit liegt deshalb auf der Demonstration, dass die oben erwähnten Transformationen bereits vorgeburtlich organisiert werden können.

Zur invarianten Objekterkennung, das heißt der Erkennung von Objekten unabhängig von ihrer mannigfaltigen Erscheinung auf der Netzhaut, gibt es im Wesentlichen zwei Ansätze: merkmalsbasierte und transformationsbasierte Erkennung. Für die erste Klasse von Theorien wird in der Regel eine Hierarchie von zunehmend invarianter werdenden Merkmalen angenommen, die Schritt für Schritt als unwichtig erachtete Information (in der Regel Information über den Ort eines Merkmals) vernachlässigt, um abstraktere Merkmale zu erzeugen. Die transformationsbasierte Erkennung hingegen basiert auf der Idee, Muster aktiv in eine normalisierte Repräsentierung zu transformieren. Wichtige Information zur Transformation (also auch über den Ort eines Objektes) bleibt deshalb bei letzterem Ansatz explizit erhalten. Um beide Ansätze zu vereinen, führen wir bilineare

Modelle ein, die auf gewichteten Multiplikationen der Aktivität jeweils zweier Einheiten basieren. Ein Satz von Einheiten, die wir im Folgenden Kontrolleinheiten nennen, kann somit aktiv den Informationsfluss in höhere Areale steuern, während die Gewichtungen die Merkmale definieren.

Ein grundlegendes und experimentell bekanntes Prinzip der Gehirnorganisation ist seine topographische Organisation. Zum Beispiel projizieren benachbarte retinale Zellen auf benachbarte Zellen im Thalamus. Diese *Retinotopie* kann auch in weiterführenden Arealen, wie dem primären visuellen kortikalen Areal V1, beobachtet werden. Viele dieser topographischen Karten sind bereits pränatal angelegt. Die klassische Interpretation der Nützlichkeit topographischer Organisation besagt, dass diese Volumen, Axonlänge und somit Energieverbrauch reduziert. In Ergänzung hierzu zeigen unsere Modelle, dass die topographische Organisation, die ein Koordinatensystem aufspannt, genutzt werden kann, um Transformationen bereits vorgeburtlich zu organisieren. Deshalb führt Kapitel 2 in ein neues Modell zur Organisation der Retinotopie ein und Kapitel 3 und 4 nutzen eine Erweiterung eines Retinotopiemechanismus, um Transformationen anzulegen. Kapitel 5 beschäftigt sich mit der postnatalen Organisation von topographischen Transformationen, die auch die Einbeziehung einer Merkmalsextrahierung ermöglicht.

2 Ein Generatives Modell zur Ontogenese der Retinotopie

Kapitel 2 führt zuerst in die theoretischen Grundlagen probabilistischer generativer Modelle ein, welche die Wahrscheinlichkeitsverteilung in der Eingabeschicht modellieren und somit den Informationsgehalt der Daten möglichst gut abdecken. Aufbauend auf dieser Methodik wird ein probabilistisches Modell zweiter Ordnung zur Entstehung retinotoper Abbildungen entwickelt. Standardmodelle zweiter Ordnung, wie zum Beispiel die probabilistische Hauptkomponentenanalyse (pPCA) oder die Faktoranalyse (FA), versuchen die Redundanz in den Ausgabeeinheiten zu reduzieren. Diese Annahme ist inkonsistent mit den allgegenwärtigen Nachbarschaftskorrelationen im Gehirn. Deshalb wird in dem vorgeschlagenen Modell sowohl die Nachbarschaftskorrelation in der Eingabeschicht, wie auch die a priori Wahrscheinlichkeitsverteilung der Ausgabeeinheiten, explizit mit Nachbarschaftskorrelationen modelliert. Da Verbindungen zwischen Arealen im Gehirn in der Regel exzitatorisch sind, nehmen wir außerdem nicht-negative Gewichte als generative Modellparameter an.

Die einfache Struktur des Modells führt zur Reduktion des probabilistischen Lernens auf eine algebraische Gleichung. Unter den Modellannahmen ist es möglich zu beweisen, dass die Gewichtsmatrix orthogonal sein muss und, dass eine solche orthogonale Matrix notwendig eine Permutationsmatrix ist. Zudem muss die Kovarianzmatrix, die die Nachbarschaften in der Eingabe- und Ausgabeschicht definiert, invariant unter dieser Permutationsmatrix sein. Für eine realistisch gewählte Kovarianzmatrix bleibt deshalb nur die Identität als Lösung, oder eine Spiegelung – beides retinotope Abbildungen.

Um die Analytik zu bestätigen wurden außerdem Simulationen des Modells durchgeführt. Hierzu wird bei Eingabe eines Musters die maximum a posteriori Lösung für die Ausgabeschicht berechnet und anschließend werden mit einer lokalen Hebbschen Lernregel,

die als Gradient des log likelihood definiert ist, die Gewichte angepasst. In Erweiterung zur Bestätigung der analytischen Resultate, zeigen die Simulationen weiterhin, dass die Annahme nicht-negativer Gewichte durch die Annahme dünn besiedelter Gewichte (*sparse weights*) getauscht werden kann, die außerdem zu schnellerer Konvergenz führt. Zudem zeigen die Resultate der Simulation, dass das Modell auch für verschieden große Eingabe- und Ausgabeschichten anwendbar ist, also verschiedene Skalen organisieren kann. Dies ist besonders wichtig als Grundlage für das Modell im dritten Kapitel, da hier Transformationen mit unterschiedlicher Skala organisiert werden.

3 Selbstorganisation Topographischer Bilinearer Netzwerke zur Invarianten Objekterkennung

Im Gegensatz zum vorherigen Kapitel, in dem ein detaillierter Retinotopiemechanismus eingeführt wurde um eine einzelne topographische Abbildung zu organisieren, wird im dritten Kapitel ein Modell zur pränatalen Organisation mehrerer topographischer Abbildungen vorgeschlagen. Postnatale Aktivierung einer dieser Abbildungen kann genutzt werden um Eingaben gezielt zu transformieren und somit etwaige Objekttransformationen auf der Retina (zum Beispiel Translation) zu normalisieren und invariante Objekterkennung zu ermöglichen.

Das dritte Kapitel baut auf dem Häussler System auf, einer abstrakten Formulierung des Retinotopiemechanismus, das durch adiabatische Elimination der Aktivitätsvariablen erreicht wird. Somit ist die Selbstorganisation als direkte Wechselwirkung der Gewichte formuliert. In Erweiterung zum klassischen Häussler Modell, das lediglich Nachbarschaftskorrelationen in der Eingabeschicht annimmt, werden die Gewichtsinteraktionen für Eingaben mit pränatalen (retinalen) Wellen in einem bilinearen Modell abgeleitet. Retinale Wellen starten in der Regel spontan an einer zufälligen Position der Retina und führen meist zu einfach zusammenhängenden Gebieten neuronaler Aktivität auf der Retina. Bei Anlegen eines Eingabemusters (einer Welle zu einem bestimmten Zeitpunkt), führt ein Winner-Take-All Wettbewerb unter den Kontrolleinheiten zu einer einzelnen aktiven Einheit. Selektiert wird die Kontrolleinheit, deren zugehörige Gewichte am Besten auf das aktuelle Muster passt. Durch Hebbsches Häussler Lernen spezialisiert sich diese Kontrolleinheit nun weiter auf dieses Muster, bei gleichzeitiger Organisation einer topographischen Abbildung aus dem aktiven Bereich aus der Eingabeschicht, der durch die pränatale Welle definiert ist, auf die ganze Ausgabeschicht.

Für den eindimensionalen Fall zeigen die Simulationen eine zuverlässige Spezialisierung der Kontrolleinheiten auf die einfach zusammenhängenden Eingabemuster. Da sich die verwendeten Muster in Position und Menge der aktiven Eingabeeinheiten unterscheiden, entwickeln sich topographische Transformationen mit unterschiedlichen Translations- und Skalenparametern. Zusätzlich, wie schon in Kapitel 2 beschrieben, können die Gewichte abhängig von den Anfangsbedingungen in eine spiegelsymmetrische topographische Lösung konvergieren. Weitere Simulationen zeigen, dass die Transformationen auch mit nicht einfach zusammenhängenden Eingabemustern organisiert werden können. Auch im biologisch interessanteren zweidimensionalen Fall zeigen die Simulationen eine sta-

bile Organisation mehrerer Transformationen, die abhängig von den zufällig gewählten Anfangsbedingungen in den Gewichten, eine Rotation der Eingabemuster in die Ausgabeschicht implementieren.

In einem zusätzlichen Unterkapitel wird demonstriert, dass die Organisation verschiedener Skalen nicht notwendig auf unterschiedlich großen Eingabemustern basieren muss. Hierzu werden gleich große Eingabemuster an verschiedenen Stellen in der Eingabeschicht angenommen. Durch eine unterschiedliche Gewinnwahrscheinlichkeit der Kontrolleinheiten entwickeln diese Präferenzen für eine unterschiedliche Anzahl an Eingabemustern: eine Kontrolleinheit mit hoher Gewinnwahrscheinlichkeit entwickelt eine Abbildung mit großer Skala, während eine seltener gewinnende Einheit eine Transformation mit kleinerer Skala organisiert.

4 Organisation von Translationen in Mehrschichtigen Netzwerken

Im dritten Kapitel wurde eine komplette Verbindungsstruktur von der Eingabeschicht zur Ausgabeschicht angenommen, das heißt jede Eingabeeinheit kann prinzipiell jede Ausgabeeinheit beeinflussen. Für die hochdimensionalen Eingabemuster im Gehirn ist diese Annahme jedoch unrealistisch, weshalb im vierten Kapitel ein bilineares Modell entwickelt wird, das auch auf ein mehrschichtiges Netzwerk, einen “shifter curcuit”, anwendbar ist: diese Netzwerkstruktur ermöglicht eine logarithmisch in der Anzahl der Eingabezellen wachsende Anzahl an Synapsen, statt der prohibitiv quadratischen Anzahl. Ausgenutzt wird in diesem Modell die Orthoginalität von Translationen im Raum der Verbindungsstrukturen um diese durch harten Wettbewerb an einzelnen Synapsen zu organisieren. Neurobiologisch ist dieser Mechanismus einfach durch Wettbewerb um einen wachstumsregulierenden Transmitter realisierbar. Im Gegensatz zu dem Modell des dritten Kapitels nimmt der Mechanismus, wie das Häussler System, nur nachbarschaftskorrelierte Eingaben an, um Translationen zu organisieren und benötigt keine pränatalen Wellen die ein einfach zusammenhängendes Gebiet definieren.

Da die entstehenden Gewichte der Kontrolleinheiten für das vorgeschlagene Modell im Raum der möglichen Verbindungen disjunkt sein müssen, Retinotopiemechanismen allerdings immer zwei mögliche Lösungen haben (eine, die die Ordnung der Eingabeeinheiten erhält und eine, die diese spiegelt – siehe Kapitel 2), muss das Modell diese Lösungen für alle Kontrolleinheiten konsistent einschränken. Hierzu untersuchen wir drei verschiedene Möglichkeiten. Erstens die sequentielle Organisation: die erste topographische Transformation entwickelt eine der beiden Möglichkeiten abhängig von den Gewichtsanfangsbedingungen und darauf folgende Transformationen fallen durch Wettbewerb mit der Ersten in die gleiche Richtung. Zweitens kann eine Nachbarschaftskooperation von Gewichten verschiedener Kontrolleinheiten genutzt werden um konsistente Lösungen zu erreichen. Drittens, ist die Lösung auf einem mehrschichtigen shifter circuit ist durch die bestehende Struktur schon im Voraus nur in eine Richtung möglich. Simulationen für alle drei Fälle zeigen eine stabile Organisation von topographischen Abbildungen mit verschiedenen Translationsparametern.

5 Slowness zur konsistenten Merkmalsorganisation in Topographischen Bilinearen Netzwerken

Kapitel 5 nutzt Methoden des probabilistischen Lernens, um ein generatives bilineares Modell auf das Lernen einer optimalen Repräsentation der Eingabestatistiken im postnatalen Fall zu optimieren. Da statistische Methoden zweiter Ordnung, wie zum Beispiel die erweiterte Faktoranalyse aus Kapitel 2, keine lokalisierten rezeptiven Felder ermöglichen und somit keine (örtliche) Topographie möglich ist, wird "sparseness" sowohl für die Ausgabeeinheiten wie auch die Kontrolleinheiten verwendet, um statistische Abhängigkeiten höherer Ordnung zu lernen und gleichzeitig Topographie zu implementieren.

Die Anwendung des Modells auf natürliche Bilder für den Fall einer einzigen Kontrolleinheit zeigt, dass lokalisierte, Bandpass filternde rezeptive Felder entstehen, die sehr ähnlich zu primären kortikalen rezeptiven Feldern sind. Die konvergierten Gewichte können gut mit einer Gaborfunktion gefittet werden. Mit dieser Methode können topographische Karten bezüglich der Parameter der Gaborfunktion, zum Beispiel Orientierung und Frequenz, analysiert werden. Die Ergebnisse zeigen, dass durch die erzwungene Topographie Orientierungs- und Frequenzkarten entstehen, das heißt benachbarte Ausgabeeinheiten ähneln sich in ihren Gaborparametern. Entsprechende Karten wurden auch experimentell im Kortex gefunden.

Im Falle von zwei Kontrolleinheiten wurden die natürlichen Bilder als Eingabemuster an unterschiedlichen Positionen und mit unterschiedlicher Transformation gezeigt (zum Beispiel einer Drehung um 45°). Obwohl prinzipiell ein Populationscode in den Kontrolleinheiten für das Modell dieses Kapitels möglich ist, führt die "sparseness" der Kontrolleinheiten nach dem Lernen der generativen Gewichte zur Aktivierung einer einzelnen Kontrolleinheit zu einem Zeitpunkt. Das Modell ist also als Erweiterung der Modelle der vorigen zwei Kapitel zu verstehen und ist für die benutzten Eingabemuster konsistent mit einem Winner-Take-All Mechanismus. Um konsistente Merkmalsrepräsentierungen in den Ausgabeeinheiten für verschiedene, nicht zeitgleich aktive Kontrolleinheiten zu erreichen, benutzt das Modell "slowness" der Ausgabezellen, das heißt es wird die zeitliche Kontinuität der Identität der Objekte in der Eingabeschicht ausgenutzt, obwohl ihre Repräsentierung sich ändern kann (zum Beispiel die erwähnten 45°). Simulationen zeigen, dass verschiedene Kontrolleinheiten konsistente und den Eingabetransformationen entsprechende rezeptive Felder entwickeln. Somit wird die Aktivierung verschiedener Kontrolleinheiten genutzt, um invariante Repräsentierungen bezüglich der gezeigten Eingaben in den Ausgaben zu erzeugen, während die Kontrollaktivität kodiert welche Transformation hierzu nötig ist und somit explizit den Ort des Objektes repräsentiert.

Wenn die Eingabemuster eine klare Orientierung aufweisen, zum Beispiel benutzen wir hierzu Bilder von Gesichtern aus der FERET Datenbank, so können rotierte Versionen (wir verwenden um 90° gedrehte Gesichter) dieser Muster an der gleichen Position in der Eingabeschicht gezeigt werden, da dann eine klare Signatur der Transformation in der Eingabe vorhanden ist. Simulationen für diesen Fall zeigen, dass die zwei Kontrolleinheiten ihre Gewichte auf jeweils eine Orientierung der Gesichter spezialisieren. Bei Präsentierung eines einzelnen Musters ist das Modell nun in der Lage die Identität des Gesichts in

einem Populationscode in der Ausgabeschicht zu repräsentieren, während die aktive Kontrolleinheit kodiert, ob es sich um ein aufrechtes oder gedrehtes Gesicht handelt – das Modell ist also völlig unüberwacht in der Lage die zugrunde liegenden Ursachen (Identität und Transformation) zu separieren und separat zu represäntieren.

6 Fazit und Ausblick

Ein zentraler Punkt dieser Arbeit ist, dass Invarianztransformationen, also Transformationen die es ermöglichen Eingabemuster in eine normalisierte Form zu konvertieren, bereits vorgeburtlich angelegt werden können. Im Gegensatz zur klassischen Perspektive einer immer abstrakter werdenden Merkmalsrepräsentierung der Eingabemuster, sind Transformationen als grundlegendes Prinzip der Gehirnfunktion auch auf unbekannte Eingabemuster anwendbar - und bieten somit ein grösseres Verallgemeinerungspotential. Merkmalsextrahierung hingegen ist immer auf eine bestimmte Eingabestatistik optimiert. Obwohl Transformationen als grundlegendes Prinzip des Gehirns in dieser Arbeit nur für die visuelle Domäne studiert wurde, sollte eine Anwendung in anderen Modalitäten ebenfalls von großem Vorteil sein. Zum Beispiel könnten grammatikalische Regeln, die unabhängig von spezifischen Wörtern sind, mit Transformationen implementiert werden. Außerdem kann das Problem der Kontrolle der Motorik als invers zum Invarianzproblem angesehen werden, da hier eine Instanz aus einer abstrakteren Repräsentierung abgeleitet werden muss. Da der bilineare Ansatz den wir verfolgen invertierbar ist und zudem im Falle von Kapitel 5 sogar generativ, sind motorische Kontrollprobleme zusätzliche mögliche Anwendwungsgebiete der eingeführten Netzwerke.

Es ist wichtig hervorzuheben, dass der zentrale These der Arbeit nicht lautet, dass die Organisation von Invarianztransformationen zur Geburt oder dem Augenöffnen enden. Im Gegenteil ist die Hauptthese, dass die Organisation der Invarianztransformationen bereits vor der Geburt anfängt.

Zukünftige Arbeiten sollten deshalb die Integration der pränatalen und postnatalen Lernphasen untersuchen. Ein erster Schritt in diese Richtung wurde in der postnatalen Studie aus Kapitel 5 unternommen, um auch eine Merkmalsextraktion zu erreichen, allerdings bisher ohne Einbeziehung der pränatalen Organisation. Desweiteren würde eine pränatale Studie des Modells aus dem fünften Kapitel zeigen, ob entsprechende Merkmale auch schon vorgeburtlich organisiert werden können. Letztlich ist die topographische Organisation der Ausgabeschicht der Modelle optimal geeignet um diese in Hierarchien zu implementieren und den Suchraum für Lernmechanismen sehr effizient einzuschränken. Eine Hintereinanderschaltung dieses kanonischen Netzes ist deshalb potentiell in der Lage mögliche Repräsentierungen in höheren Arealen des Gehirns aufzudecken und zu erklären.

Lebenslauf

Persönliche Daten

Adresse	Urs Michael Bergmann, Hufelandstrasse 31, 10407 Berlin E-Mail: ubergmann@fias.uni-frankfurt.de
Geburt	27. Februar 1980 in Ludwigsburg, Baden-Württemberg
Nationalität	Deutsch

Schulbesuch

8.1990-6.1999	Otto-Hahn Gymnasium Ludwigsburg

Zivildienst

7.1999-6.2000	Mobile Soziale Pflege, Johanniter-Unfall-Hilfe e.V. Ludwigsburg

Studium

10.2000-6.2006	Universität Heidelberg Physik Diplom und Informatik Zusatzzertifikat Diplomarbeit: "Unspecific reinforcement learning in one- and two-layered networks" Betreuer: Prof. Dr. I.-O. Stamatescu und Priv. Doz. Dr. Reimer Kühn Abschluss: Diplom-Physiker, Note: 1,1

Promotion

Seit 7.2006	Frankfurt Institute for Advanced Studies, Universität Frankfurt Arbeitsgruppe Prof. Dr. Christoph von der Malsburg

Publikationen (peer-reviewed)

Urs Bergmann, Reimer Kühn and I.-O. Stamatescu.
Learning with incomplete information in the Committee Machine,
Biological Cybernetics 101(5), 401–410, 2009.

Urs Bergmann and Christoph von der Malsburg.
A Bilinear Model for Consistent Topographic Representations,
In *Artificial Neural Networks – ICANN 2010*, Part III, LNCS 6354.

Junmei Zhu, Urs Bergmann and Christoph von der Malsburg.
Self-Organization of Steerable Topographic Mappings as Basis
for Translation Invariance,
In *Artificial Neural Networks – ICANN 2010*, Part II, LNCS 6353.

Urs Bergmann and Christoph von der Malsburg.
Self-Organization of Topographic Bilinear Networks for
Invariant Recognition,
Neural Computation, 2011 (accepted).

Urs Bergmann, Gervasio Puertas and Christoph von der Malsburg.
A Gaussian Generative Model for the Ontogeny of Topography.
2011 (in preparation).

Vorträge und Publikationen

Urs Bergmann.
Optimierung eines elastischen Netzes zur Ringfindung in
RICH Detektoren, *GSI Note* 2005.
Weblink: http://www.gsi.de/documents/DOC-2005-Nov-186_e.html

Learning with incomplete information on the Committee Machine.
Vortrag auf der *Delta Konferenz*, Heidelberg 2007.

Urs Bergmann and Christoph von der Malsburg.
Ontogenesis of invariance transformations.
In *Proceedings COSYNE*, 2008.

Ontogeny of Mappings for Invariant Recognition.
Vortrag auf der *Computational Developmental Neuroscience*
Konferenz, Edinburgh, 2008.