Unsupervised Learning of Place Cells, Head-Direction Cells, and Spatial-View Cells with Slow Feature Analysis on Quasi-Natural Videos

Mathias Franzius, Henning Sprekeler, Laurenz Wiskott

April 13, 2007

Institute for Theoretical Biology, Humboldt-Universität zu Berlin.
{m.franzius, h.sprekeler, l.wiskott}@biologie.hu-berlin.de

We present a model for the self-organized formation of place cells, head-direction cells, and spatial view cells in the hippocampal formation based on unsupervised learning on quasi-natural visual stimuli. The model comprises a hierarchy of Slow Feature Analysis (SFA) nodes, which were recently shown to be a good model for complex cells in the early visual system (Berkes and Wiskott, 2005). The system extracts a distributed grid-like representation of position and orientation, which is transcoded into a localized place field, head direction, or view representation, respectively, by sparse coding. The type of cells that develops depends solely on the relevant input statistics, i.e. the movement pattern of the simulated animal. The numerical simulations are complemented by a mathematical analysis that allows us to accurately predict the output of the top SFA layer.

1 Introduction

The brain needs to extract behaviorally relevant information from sensory inputs in order to successfully interact with the environment. Position and head orientation of an animal in the space surrounding it is part of this relevant information. Neural representations of a rodent's spatial position - termed *place cells* - have been found more than 35 years ago in hippocampal areas CA1 and CA3 (O'Keefe and Dostrovsky, 1971), correlates of head orientation - termed *head-direction cells* - twenty years later (Taube et al., 1990), and recently non-localized representations termed *grid cells* were found in entorhinal cortex (EC) of rats (Hafting et al., 2005). While primates also have head-direction cells, no place cells were found in primates yet. Instead, they have *spatial view cells,* which do not encode the animal's own (idiothetic) position but fire whenever the animal views a certain part of the environment (Rolls, 1999, 2006).

All of these cells selectively encode some aspects of position and/or orientation of the animal, while being invariant to others. Head-direction cells are strongly selective for the direction of the animal's head and largely invariant to its position (Sharp et al., 2001). They typically have a single peak of activity with a Gaussian or triangular shape and a tuning width of roughly 60° to 150° (Taube and Bassett, 2003) depending on brain area. In contrast, most place cells recorded in open fields are invariant to head direction while being selective for the animal's position. Interestingly, the degree of orientation-invariance depends on the behavioral task of of the animal and possibly on the structure

of the environment. In linear track environments and for repeated linear paths in open environments most place cells are orientation-specific (Markus et al., 1995). Grid cells in entorhinal cortex also exhibit conjunctive representations of position and orientation (Sargolini et al., 2006). Spatial view cells in primates show very different firing properties. These cells are neither position-invariant nor orientation-invariant but fire when a certain part of the environment is in the animal's field of view, resembling head-direction cells for the case of an infinitely distant view. Figure 1 illustrates the difference between grid cells, place cells, head-direction cells and spatial view cells.

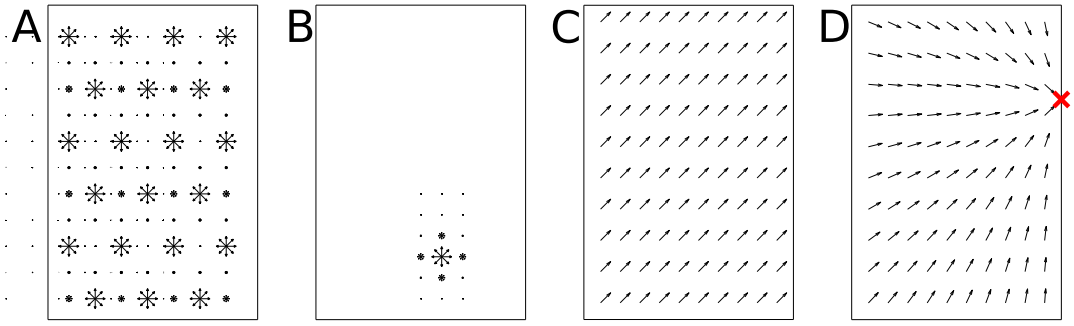

Figure 1: **Spatial and orientation tuning of an idealized grid cell (A), place cell (B), head-direction cell (C) and a spatial view cell (D).** The activity of a grid cell is mostly orientation-invariant and not spatially localized but repeats in a hexagonal grid, whereas a place cell is also orientation-invariant but spatially localized. The activity of a head-direction cell shows a global direction preference but is spatially invariant, and the spatial view cell is maximally active when a specific view is fixated (indicated by 'x') with an amplitude that is independent of spatial position.

Throughout this paper, *oriospatial cells* will be used as a superordinate term for place cells, grid cells, head-direction cells, and spatial view cells. While the precise role of these oriospatial cells is still discussed, they probably form the neural basis for the ability of an animal to self-localize and navigate (Knierim et al., 1995).

Stimuli available to oriospatial cells can be classified as either *idiothetic*, including motor feedback, proprioception, and vestibular input, or as *allothetic*, which includes all information from sensors about the external environment, e.g. vision or olfaction. While place cells are influenced by several modalities they seem to be driven primarily by visual input (e.g. Jeffery and O'Keefe, 1999), but since their firing properties remain stable in the absence of external sensory cues for several minutes, proprioceptive stimuli must play a major role for place cell firing as well (Save et al., 2000). Even in complete absence of allothetic sensory information an animal can integrate idiothetic self-motion cues to estimate its position and orientation in space. This process, called *path integration* (or dead reckoning), inherently accumulates errors over longer time scales, which can only be corrected by allothetic information. For the head-direction cells it is commonly assumed that idiothetic input from the vestibular system is dominant (e.g. Sharp et al., 2001), but like place cells they need external sensory stimuli to correct for drift.

We introduce here a model for the self-organized formation of hippocampal place cells, head-direction cells, and spatial view cells based on unsupervised learning on quasi natural visual stimuli. Our model has no form of memory and receives raw high-dimensional visual input. The former means that our model cannot perform path integration, the latter means that positional information has to be extracted from complex images. While such a model can certainly not be a complete model of

oriospatial cells, it can show how far a memoryless purely sensory-driven system can model oriospatial cells already. The learning rule of the model is based on the concept of slowness or temporal stability, which is motivated by the observation that raw sensory signals (like a camera's individual pixel values) typically vary much more quickly than some behaviorally relevant features of the animal or its environment, like the animal's position in space. By extracting slowly varying features from the sensory input one can hope to obtain a useful representation of the environment. This *slowness principle* forms the basis for a variety of learning rules (e.g. Földiak, 1991; Mitchison, 1991; Stone and Bray, 1995). The implementation used here is *Slow Feature Analysis* (SFA) as introduced by Wiskott (Wiskott, 1998; Wiskott and Sejnowski, 2002). For a given set of time-dependent training data, in our case video sequences, we are looking for a nonlinear scalar function from a given function space that generates the slowest possible output signal $y(t)$ when applied to the training data. The slowness of the signal is measured in terms of its Δ-value, which is given by the mean square of the signal's temporal derivative (see section 2). As small Δ-values correspond to slowly varying signals, the objective is to find the function that minimizes the Δ-value. To avoid the trivial constant solution, the signal is required to have unit variance and zero mean. Furthermore, we can find a second function that optimizes the objective under the additional constraint that its output signal is uncorrelated to the first, a third function, whose output is uncorrelated to the first two signals and so on. In this manner we generate a sequence of functions with increasing Δ-value that extract slowly varying features from the training data. More details on the approach as well as its mathematical formalization can be found in section 2.

SFA has been successfully applied as a model for the self-organized formation of complex cell receptive fields in primary visual cortex (Berkes and Wiskott, 2005). Here, we embed this approach in a biologically inspired hierarchical network of visual processing of a simulated rat where each layer learns the slowest features from the previous layer by SFA (see experimental methods in section 3). We find that the output of the highest layer performing SFA forms a distributed oriospatial representation. In a subsequent linear step the model applies a mechanism for sparse coding resulting in localized oriospatial codes. The same model in the same environment can reproduce the firing characteristics of place cells, head-direction cells, and spatial view cells, depending solely on the movement statistics of the simulated rat. For roughly uncorrelated head direction and body movement, the system learns head-direction cells or place cells, depending on the relative speed of head rotation and body movement. If the movement statistics is altered such that spots in the room are fixated for a while during simulated locomotion, the model learns spatial view cell characteristics.

We introduce a mathematical framework in section 4 that analytically explains the results of the SFA output. The mathematically less inclined reader may consider skipping this section. Both analytical and computer simulation results are presented in section 5.

We conclude that a purely sensory-driven model can capture the key properties of several major cell types associated with spatial coding, namely place cells, head-direction cells, spatial view cells, and to some extent grid-cells.

2 Slow Feature Analysis

Slow Feature Analysis solves the following learning task: Given a multidimensional input signal we want to find instantaneous scalar input-output functions that generate output signals that vary as slowly as possible but still carry significant information. To ensure the latter we require the output signals to be uncorrelated and have unit variance. In mathematical terms, this can be stated as follows:

Optimization problem: *Given a function space $\mathcal{F}$ and an I-dimensional input signal $\mathbf{x}(t)$ find a*

set of J real-valued input-output functions $g_j(\mathbf{x}) \in \mathcal{F}$ such that the output signals $y_j(t) := g_j(\mathbf{x}(t))$

$$\textit{minimize } \Delta(y_j) := \langle \dot{y}_j^2 \rangle_t \tag{1}$$

under the constraints

$$\begin{aligned} \langle y_j \rangle_t &= 0 \quad \textit{(zero mean)}, & (2)\\ \langle y_j^2 \rangle_t &= 1 \quad \textit{(unit variance)}, & (3)\\ \forall i < j : \langle y_i y_j \rangle_t &= 0 \quad \textit{(decorrelation and order)}, & (4) \end{aligned}$$

with $\langle \cdot \rangle_t$ and $\dot{y}$ indicating temporal averaging and the derivative of y, respectively.

Equation (1) introduces the Δ-value, which is a measure of the temporal slowness of the signal $y(t)$. It is given by the mean square of the signal's temporal derivative, so small Δ-values indicate slowly varying signals. The constraints (2) and (3) avoid the trivial constant solution and constraint (4) ensures that different functions g_j code for different aspects of the input.

It is important to note that although the objective is slowness, the functions g_j are instantaneous functions of the input, so that slowness cannot be enforced by low-pass filtering. Slow output signals can only be obtained if the input signal contains slowly varying features that can be extracted by the functions g_j, which are computed instantaneously for a given input.

In the computationally relevant case where $\mathcal{F}$ is finite-dimensional the solution to the optimization problem can be found by means of Slow Feature Analysis (Wiskott and Sejnowski, 2002; Berkes and Wiskott, 2005). This algorithm, which is based on an eigenvector approach, is guaranteed to find the global optimum. More biologically plausible learning rules for the optimization problem, both for graded response and spiking units exist (Hashimoto, 2003; Sprekeler et al., 2007).

If the function space is infinite-dimensional, the problem requires variational calculus and will in general be difficult to solve. In section 4 we demonstrate that the optimization problem for the high-dimensional visual input, as faced by the hierarchical model, can be reformulated for the low-dimensional configural input of position and orientation. In this case, the variational calculus approach becomes tractable and allows to make analytical predictions for the behavior of the full model.

3 Experimental methods

The outcome of an unsupervised learning rule, such as Slow Feature Analysis, is crucially determined by the statistics of the training data. As we want to show that oriospatial cells can be learnt from raw sensory stimuli, we approximate the retinal stimuli of a rat by video sequences generated in a virtual-reality environment. The input statistics of the training data are thus jointly determined by the structure of the virtual-reality environment and the movement pattern of the simulated rat. As this video data is very high-dimensional, nonlinear SFA in a single step is computationally infeasible. To overcome this problem, the model is organized as a hierarchy of SFA nodes in analogy to the hierarchy of the brain's visual system (see figure 2C).

Simulated environments

Many experimental place field data were recorded either in a linear track or in an open field apparatus. For our simulations we use a linear track of 10:1 side length, and a rectangular open field of 3:2 side length. We have also simulated radial mazes (e.g. plus or 8-arm mazes) as a third apparatus type but they can be considered as a combination of an open field in the center with linear tracks extending from it and simulation results for this type will not be presented here.

The input data consists of pixel images generated by a virtual-reality system based on OpenGL with textures from the Vision Texture Database (Picard et al., 2002). The virtual rat's horizontal field of view is 320° (see figure 2A for a top view of the environment, and figure 2B for a typical rat's view from this environment) and consistent with that of a biological rat (Hughes, 1978). The vertical field of view is reduced to 40° because outside this range usually only unstructured floor and ceiling are visible. An input picture has 40 by 320 color pixels (RGB, 1pixel/°). The input dimensionality for the system is thus 38400, while the dimensionality of the interesting oriospatial parameter space is only three-dimensional (x- and y-position and orientation).

Movement patterns of the virtual rat

As an approximation of a rat's trajectory during exploration in place field experiments we simulate Brownian motion on the three-dimensional parameter space of position and orientation. The virtual rat's position `pos(t)` at each time step `t` is updated by a weighted sum of the current velocity and Gaussian white noise `noise` with standard deviation `vr`. The momentum term `m` can assume values between zero (massless particle) and one (infinitely heavy particle), so that higher values of m lead to smoother trajectories and a more homogeneous sampling of the apparatus in limited time. When the virtual rat's movement would traverse the apparatus boundaries, the current velocity is halved and an alternative random velocity update is generated, until a new valid position is reached (see table 1).

```
currentVelocity = pos(t) - pos(t - 1);
repeat
   noise = GaussianWhiteNoise2d() * vr;
   pos(t + 1) = pos(t) + m * currentVelocity + (1 - m) * noise;
   if not isInsideApparatus(pos(t + 1))
      currentVelocity = currentVelocity / 2;
   end
until isInsideApparatus(pos(t + 1))
```

Table 1: Pseudocode for the computation of the translational movement for the virtual rat's path.

We call the standard deviation (normalized by room size L) of the noise term *translational speed* v_r and the standard deviation of head direction trajectory *rotational speed* v_ϕ. On long timescales and with finite room size this type of movement approximates homogeneous position and orientation probability densities, except at the apparatus boundaries where a high momentum term can increase the position probability. We call the ratio of rotational to translational speed v_ϕ/v_r the *relative rotational speed* v_{rel}.

The actual choice of v_{rel} is based on the rat's behavior in different environments and behavioral tasks. In linear track experiments the rat's movement is essentially one-dimensional and the animal rarely turns on mid-track but instead mostly at the track ends. Accordingly, we use a large momentum term, so that the virtual rat often translates smoothly between track ends and rarely turns on mid-track. In the open field, on the other hand, full two-dimensional movement and rotation is possible, but the actual statistics depends on the behavioral task at hand. We mimick the common pellet-chasing experiment (Markus et al., 1995) by using isotropic two-dimensional translational speed and setting v_{rel} to a relatively high value. Three different movement paradigms are explored in the following: *simple movement*, *restricted head movement* and *spatial view*. In the *simple movement* paradigm head orientation and body movement are completely independent, so that head direction can be modeled

with unrestricted Brownian motion. In the *restricted head movement* paradigm the head direction is enforced to be within 90 degrees from the direction of body movement (see table 2).

```
repeat
   noise = GaussianWhiteNoise1d() * vphi;
   phi(t + 1) = phi(t) + m * (phi(t) - phi(t - 1)) + (1 - m) * noise;
until isHeadDirWithin90DegOfMovementDir(pos(t + 1) - pos(t), phi(t + 1))
```

Table 2: Pseudocode for the computation of the head direction on the virtual rat's path in the restricted head movement paradigm.

This constraint implicitly restricts the range of possible relative speeds: while it is still possible to have arbitrarily high relative rotational speed by turning often or quickly, very low relative rotational speed cannot be achieved anymore in finite rooms. Typically, if the rat reaches a wall, it has to turn. Thus the maximum travel length for a full turn is roughly the circumference of the apparatus, resulting in a lower bound for the relative rotational speed v_{rel}. In order to generate input sequences with lower v_{rel} one would have to discard periods with dominant rotations from the input sequence. For a biological implementation of such a mechanism the rat's limbic system could access the vestibular rotational acceleration signal in order to downregulate the learning rate during quick turns. We will refer to this mechanism as *learning rate adaptation* (LRA). A third movement statistics can be generated if we assume that an animal fixates objects or locations in the room for some time while moving around. During this period the animal fixates a specific location L in the room, i.e. it always turns its head into the direction of L, independent of its position. We implement L as a fixation point on the wall and change its position with a similar statistics (and low v_{rel}) as the head direction in the other paradigms. In this paradigm both position and orientation are dependent and vary rather quickly, while the position of L changes slowly. We call this movement pattern *spatial view* paradigm and suggest that it is a more appropriate description of a primate's movement pattern than the previous two.

Model architecture

Our computational model consists of a converging hierarchy of layers of SFA nodes and a single final sparse coding node (see figure 2C). Each SFA node finds the slowest output features from its input according to the SFA learning rule given in section 2 and performs the following sequence of operations: linear SFA for dimensionality reduction, quadratic expansion with additive Gaussian white noise, another linear SFA step for slow-feature extraction, and clipping of extreme values at ± 4 (see figure 2D). Effectively, a node implements a subset of full quadratic SFA. The clipping removes extreme values that can occur on test data very different from training data.

In the following, the part of the input image that influences a node's output will be denoted as its *receptive field*. On the lowest layer the receptive field of each node consists of an image patch of 10 by 10 pixels with 3 color dimensions each. The nodes form a regular (i.e. non-foveated) 7 by 63 grid with partially overlapping receptive fields that jointly cover the input image of 40 by 320 pixels. The second layer contains 2 by 15 nodes where each receives input from 3 by 8 layer 1 nodes with neighboring receptive fields, resembling a retinotopical layout. All layer 2 output converges onto a single node in layer 3, whose output we call *SFA-output*. Thus the hierarchical organization of the model captures two important aspects of cortical visual processing: increasing receptive field sizes and accumulating computational power at higher layers.

The network's SFA-output is subsequently fed into a final computational node that performs linear sparse coding, either by applying independent component analysis (we use CuBICA which is based on the diagonalization of third and fourth order cumulants (Blaschke and Wiskott, 2004)) or by performing competitive learning (CL). The top-layer output will be called *ICA-output*, or *CL-output*. ICA applied to non-localized grid-cell inputs finds sparser codes than CL, but the latter is biologically more realistic. More details on different approaches for sparse coding of grid-cell input can be found in (Franzius et al., 2007).

The network is implemented in Python using the MDP toolbox (Berkes and Zito, 2005) and the code is available upon request.

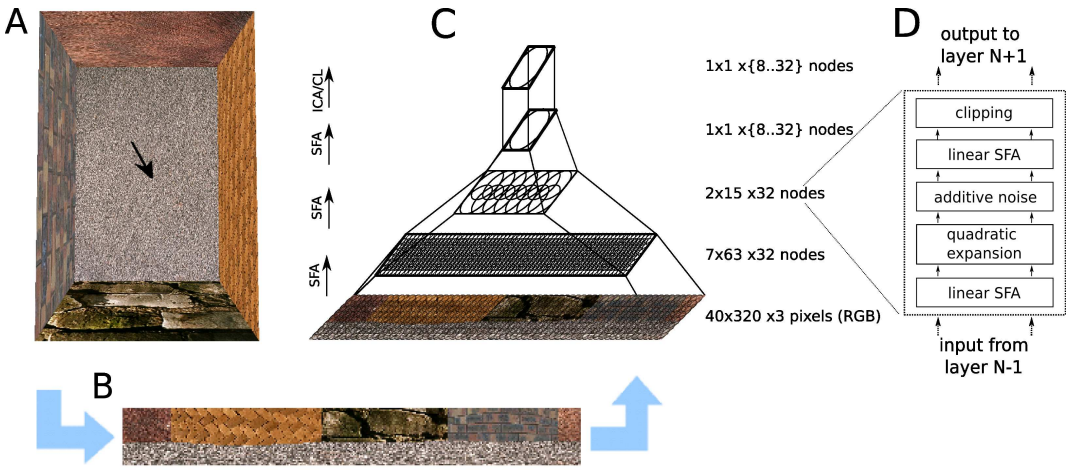

Figure 2: **Model Architecture.** At a given position and orientation of the virtual rat (arrow) in the naturally textured virtual-reality environment (A), input views are generated (B), and processed in a hierarchical network (C). The lower 3 layers perform the same sequence (D) of linear SFA (for dimensionality reduction), expansion, additive noise, linear SFA (for feature extraction), and clipping, the last layer performs sparse coding (either ICA or CL).

Model training

The layers are trained subsequently from bottom to top on different trajectories through one of the simulated environments. For computational efficiency we train only one node with stimuli from all node locations in its layer and replicate this node throughout the layer. This mechanism effectively implements a weight sharing constraint. However, the system performance does not critically depend on this mechanism. To the contrary, individually learned nodes *improve* the overall performance.

In analogy to a rat's brain, the lower two layers are trained only once and are kept fixed for all simulations presented here (like the visual system, which remains rather stable for adult animals). Only the top SFA and ICA layer are retrained for different movement statistics and environments. For our simulations we use 100.000 time points for the training of each layer. Since training time of the entire model on a single PC is on the order of multiple days, the implementation is parallelized and training times thus reduced to hours. The simulated rat's views are generated from its configuration (position and orientation) with floating point precision and are not artificially discretized to a smaller configuration set.

Analysis methods

The highly nonlinear functions learned by the hierarchical model can be characterized by their outputs on the three-dimensional configuration space of position and head direction. We will call two-dimensional sections of the output with constant (or averaged) head direction *spatial firing maps* and one-dimensional sections of the output with constant (or averaged) position *orientation tuning curves*. For the sparse coding results with ICA the otherwise arbitrary signs are chosen such that the largest absolute response is positive.

The sensitivity of a function f to spatial position r will be characterized by its mean positional variance η_r, which is the variance of $f(r,\phi)$ with respect to r averaged over all head directions ϕ: $\eta_r(f) = \langle \mathrm{var}_r(f(r,\phi))\rangle_\phi$. Correspondingly, the sensitivity of a function f to head direction ϕ will be characterized by its directional variance η_ϕ averaged over all spatial positions r: $\eta_\phi(f) = \langle \mathrm{var}_\phi(f(r,\phi))\rangle_r$. A perfect head-direction cell has no spatial structure and thus a vanishing η_r and positive η_ϕ, while a perfect place cell has positive η_r due to its spatial structure but no orientation dependence and thus a vanishing η_ϕ.

4 Theoretical methods

Considering the complexity of the computational model presented in the last section, one might expect that it would be impossible to make any analytical statement about the model's behavior. However, in this section we introduce a mathematical framework that actually allows us to make detailed predictions depending on the movement statistics of the simulated rat. The theoretically less inclined reader should feel free to skip all sections marked by a * without loss of the general understanding of our model and the results.

4.1 The modified optimization problem*

Consider a rat in an environment that is kept unchanged for the duration of the experiment. The visual input the rat perceives during the experiment is the input signal for the learning task stated above. This section addresses the following question: Can we predict the functions learnt in such an experiment and, in particular, will they encode the rat's position in a structured way?

As the rat's environment remains unchanged for the duration of the experiment, its visual input cannot cover the full range of natural images but only the relatively small subset that can be realized in our setup. Given the environment, the rat's visual input can at all times be uniquely characterized by the rat's position and its head direction. We combine these parameters in a single *configuration vector* $\mathbf{s}$ and denote the image the rat perceives when it is in a particular configuration $\mathbf{s}$ as $\mathbf{x}(\mathbf{s})$. We refer to the manifold of possible configurations as *configuration space* V. Note, that V in general does not have the structure of a vector space.

In a sufficiently complex environment we cannot only infer the image from the configuration but also the configuration from the image, so that there is a one-to-one correspondence between the configurations and the images. If we are not interested in how the functions the system learns respond to images other than those possible in the experiment, we can think of them as functions of the configuration $\mathbf{s}$, since for any function $\tilde{g}(\mathbf{x})$ of the images, we can immediately define an equivalent function $g(\mathbf{s})$ of the configuration:

$$g(\mathbf{s}) := \tilde{g}(\mathbf{x}(\mathbf{s})). \tag{5}$$

This leads to a simplified version of our problem. Instead of using the images $\mathbf{x}(t)$ we use the configuration $\mathbf{s}(t)$ as an input signal for our learning task.

It is intuitively clear that functions that vary slowly with respect to the configuration $\mathbf{s}$ will create slowly varying output when applied to $\mathbf{s}(t)$ as an input signal, because $\mathbf{s}(t)$ is continuous in time. Mathematically, this is reflected by the chain rule:

$$\dot{y}_j = \frac{\mathrm{d}}{\mathrm{d}t} g_j(\mathbf{s}(t)) = \nabla g_j(\mathbf{s}) \cdot \dot{\mathbf{s}} =: \nabla g_j(\mathbf{s}) \cdot \mathbf{v} \tag{6}$$

where ∇g_j is the gradient of g_j and $\mathbf{v} = \dot{\mathbf{s}}$ is the velocity in configuration space (note the difference in notation to $\nabla \cdot \mathbf{A}(\mathbf{s})$, which denotes the divergence of a vector-valued function $\mathbf{A}$).

In order to generate slowly varying output, g_j should vary slowly with $\mathbf{s}$ in configuration regions with large velocities $\mathbf{v}$ and reserve stronger gradients for regions with small velocities. Thus, the optimal functions depend on the velocity statistics of the input signal. As their dependence on the detailed time-course of the input signal $\mathbf{s}(t)$ is inconvenient to handle mathematically, we assume that the duration of the experiment is long enough to do statistics on the behavior of the rat. Its motion can then be described by means of a joint probability density function $p_{\mathbf{s},\mathbf{v}}(\mathbf{s},\mathbf{v})$, which quantifies how often the rat is found in a particular configuration $\mathbf{s}$ and moves with velocity $\mathbf{v}$. We may then equivalently replace the temporal averages in the original formulation of the learning task by weighted averages over all configurations and velocities:

$$\langle\cdot\rangle_t \to \langle\cdot\rangle_{\mathbf{s},\mathbf{v}} = \int \cdot(\mathbf{s},\mathbf{v})\, p_{\mathbf{s},\mathbf{v}}(\mathbf{s},\mathbf{v})\, \mathrm{d}\mathbf{s}\, \mathrm{d}\mathbf{v} \tag{7}$$

If we take the average of a function that does not explicitly depend on the velocity $\mathbf{v}$, we can simplify the average $\langle\cdot\rangle_{\mathbf{s},\mathbf{v}}$ by integrating over the velocity:

$$\langle\cdot\rangle_{\mathbf{s},\mathbf{v}} = \int \cdot(\mathbf{s}) p_{\mathbf{s},\mathbf{v}}(\mathbf{s},\mathbf{v})\, \mathrm{d}\mathbf{s}\, \mathrm{d}\mathbf{v} = \int \cdot(\mathbf{s}) \underbrace{\left[\int p_{\mathbf{s},\mathbf{v}}(\mathbf{s},\mathbf{v})\, \mathrm{d}\mathbf{v}\right]}_{=:p_{\mathbf{s}}(\mathbf{s})} \mathrm{d}\mathbf{s} =: \langle\cdot\rangle_{\mathbf{s}} \tag{8}$$

Here $p_{\mathbf{s}}$ is the marginal probability of finding the rat in configuration $\mathbf{s}$, irrespective of its velocity.

Making use of (5-8) we can now state an equivalent alternative formulation of the learning task:
Optimization problem 2: *Given a function space $\mathcal{F}$ on a configuration space V, which is sampled with probability density $P(\mathbf{s},\mathbf{v})$, find a set of J functions $g_j(\mathbf{s}) \in \mathcal{F}$ that*

$$\textit{minimize} \quad \Delta(g_j) := \langle(\nabla g_j(\mathbf{s}) \cdot \mathbf{v})^2\rangle_{\mathbf{s},\mathbf{v}} \tag{9}$$

under the constraints

$$\langle g_j(\mathbf{s})\rangle_{\mathbf{s}} = 0 \quad \textit{(zero mean)}, \tag{10}$$

$$\langle g_j(\mathbf{s})^2\rangle_{\mathbf{s}} = 1 \quad \textit{(unit variance)}\ , \tag{11}$$

$$\forall i < j : \langle g_i(\mathbf{s}) g_j(\mathbf{s})\rangle_{\mathbf{s}} = 0 \quad \textit{(decorrelation and order)}\ . \tag{12}$$

If we do not impose any restriction on the function space $\mathcal{F}$ (apart from sufficient differentiability and integrability), this modified optimization problem can be solved analytically for a number of cases. Following a previous analytical treatment (Wiskott, 2003) we refer to the optimal functions in the unrestricted function space as Δ-*optimal functions*; they are shown in section 5 together with the numerical simulations.

4.2 A differential equation for the optimal functions*

In this section we apply variational calculus to optimization problem 2 and derive a partial differential equation for the optimal functions g_j. We prove that the optimization problem can be simplified to an eigenvalue problem of a partial differential operator $\mathcal{D}$ whose eigenfunctions and eigenvalues form the Δ-optimal functions and their Δ-values, respectively. For the sake of brevity we shift the proofs to the appendix, so that the reader can focus on the main theorems.

Using Lagrange multipliers we get an objective function for the functions g_j that incorporates the constraints:

$$\Psi(g_j) = \frac{1}{2}\Delta(g_j) - \lambda_{j0}\langle g_j(\mathbf{s})\rangle_{\mathbf{s}} - \frac{1}{2}\lambda_{jj}\langle g_j(\mathbf{s})^2\rangle_{\mathbf{s}} - \sum_{i<j}\lambda_{ji}\langle g_i(\mathbf{s})g_j(\mathbf{s})\rangle_{\mathbf{s}}. \tag{13}$$

Here, factors $\frac{1}{2}$ have been introduced for mathematical convenience and have no influence on the results.

In the following we will not need the full dependence of the probability density $p_{\mathbf{s},\mathbf{v}}$ on the velocity, but only the following function:

$$\mathbf{K}(\mathbf{s}) := \frac{1}{p_{\mathbf{s}}(\mathbf{s})}\int \mathbf{v}\mathbf{v}^T p_{\mathbf{s},\mathbf{v}}(\mathbf{s},\mathbf{v})\,\mathrm{d}\mathbf{v} = \int \mathbf{v}\mathbf{v}^T p_{\mathbf{v}|\mathbf{s}}(\mathbf{v}|\mathbf{s})\,\mathrm{d}\mathbf{v} = \langle \mathbf{v}\mathbf{v}^T\rangle_{\mathbf{v}|\mathbf{s}}\,. \tag{14}$$

$\mathbf{K}$ is the matrix containing the second-order moments of the conditional velocity distribution $P(\mathbf{v}|\mathbf{s}) = \frac{P(\mathbf{s},\mathbf{v})}{P(\mathbf{s})}$. It contains information on how fast and in which direction the rat typically moves given it is in configuration $\mathbf{s}$.

Applying variational calculus to the objective function (13), we can derive a necessary condition for the solutions of optimization problem 2.

Theorem 1 *For a particular choice of the parameters λ_{ij}, the solutions g_j of optimization problem 2 obey the Euler-Lagrange equation*

$$\mathcal{D}g_j(\mathbf{s}) - \lambda_{j0} - \lambda_{jj}g_j(\mathbf{s}) - \sum_{i<j}\lambda_{ji}g_i(\mathbf{s}) = 0 \tag{15}$$

with the boundary condition

$$\mathbf{n}(\mathbf{s})^T\mathbf{K}(\mathbf{s})\nabla g_j(\mathbf{s}) = 0 \quad \textit{for } \mathbf{s}\in\partial V. \tag{16}$$

Here, the partial differential operator $\mathcal{D}$ is defined as

$$\mathcal{D} := -\frac{1}{p_{\mathbf{s}}(\mathbf{s})}\nabla\cdot p_{\mathbf{s}}(\mathbf{s})\mathbf{K}(\mathbf{s})\nabla \tag{17}$$

and $\mathbf{n}(\mathbf{s})$ is the unit normal vector on the boundary ∂V of the configuration space V.

We now show that the solutions of optimization problem 2 are given by the eigenfunctions of the operator $\mathcal{D}$. The essential observation we need is stated in

Theorem 2 *Let $\mathcal{F}_b \subset \mathcal{F}$ be the space of functions that obey the boundary condition (16). Then $\mathcal{D}$ is self-adjoint on $\mathcal{F}_b$ with respect to the scalar product*

$$(f,g) := \langle f(\mathbf{s})g(\mathbf{s})\rangle_{\mathbf{s}}, \tag{18}$$

i.e.

$$\forall f,g\in\mathcal{F}_b : (\mathcal{D}f,g) = (f,\mathcal{D}g). \tag{19}$$

This property is useful, as it allows the application of the spectral theorem known from functional analysis, which states that any self-adjoint operator possesses a complete set of eigenfunctions $f_j(\mathbf{s}) \in \mathcal{F}_b$ with real eigenvalues Δ_j, which are pairwise orthogonal, i.e. a set of functions that fulfills the following conditions:

$$\mathcal{D}f_j = \Delta_j f_j \quad \text{with } \Delta_j \in \mathbb{R} \qquad \text{(eigenvalue equation)}\,, \tag{20}$$

$$(f_i, f_j) = \delta_{ij} \qquad \text{(orthonormality)}\,, \tag{21}$$

$$\forall f \in \mathcal{F}_b \, \exists \, \alpha_k : f = \sum_{k=0}^{\infty} \alpha_k f_k \qquad \text{(completeness)}\,. \tag{22}$$

Because the weighted average over configurations is equivalent to a temporal average, the scalar product (18) is essentially the covariance of the output of the functions f and g (if they have zero mean). The orthonormality (21) of the eigenfunctions thus implies that the eigenfunctions fulfill the unit variance and decorrelation constraint. This is stated in

Theorem 3 *Apart from the constant function, which is always an eigenfunction, the (adequately normalized) eigenfunctions $f_j \in \mathcal{F}_b$ of the operator $\mathcal{D}$ fulfill the constraints (10-12).*

If we set $\lambda_{0j} = \lambda_{ji} = 0$ for $i \neq j$, the eigenfunctions also solve eqn. (15), making them good candidates for the solution of optimization problem 2. To show that they indeed minimize the Δ-value we need

Theorem 4 *The Δ-value of the normalized eigenfunctions f_j is given by their eigenvalue Δ_j.*

At this point, it is intuitively clear that the eigenfunctions with the smallest eigenvalues form the solution to optimization problem 2. This is stated in

Theorem 5 *The J eigenfunctions with the smallest eigenvalues $\Delta_j \neq 0$ are a solution of optimization problem 2.*

The advantage of this approach is that it transfers the original optimization problem to that of finding the eigenfunctions of a partial differential operator. This type of problem is encountered frequently in other contexts and has been studied extensively.

It is worth noting that the formalism described here is not restricted to the example used here. As it is independent of the concrete nature of the configuration space, it can be applied to more complicated problems, e.g. to a rat moving in an environment with moving objects, whose positions would then be additional components of the configuration $\mathbf{s}$.

5 Results

We apply our theoretical framework and computer simulations to a number of environments and movement patterns that resemble typical place cell experiments. In section 5.1, we show results for the open field, beginning with the mathematical analysis and simulation results for the simple movement paradigms with high and low relative speeds. Subsequently, the simulation results for the restricted head movement paradigm, including learning rate adaptation, and the spatial view paradigm are shown. In section 5.2 the results for the linear track with its two-dimensional configuration space are shown.

5.1 Open field

One of the most common environments for place cell experiments is an open field apparatus of rectangular or circular shape. Here, the most typical experimental paradigm is to throw food pellets randomly into the apparatus at regular intervals leading to a random search behavior of the rat. For this case the rat's oriospatial configuration space comprises the full three dimensional manifold of position and orientation. In this section, we present results from experiments with simulated rat trajectories at either high or low relative rotational speeds leading to undirected place cells or position-invariant head-direction cell type results, respectively.

5.1.1 Theoretical predictions for the simple movement paradigm*

In a rectangular open field the configuration space can be parametrized by the animals position, indicated by the coordinates x and y, and its head direction ϕ. The total configuration space is then given by $\mathbf{s} = (x, y, \phi) \in [0, L_x] \times [0, L_y] \times [0, 2\pi[$. L_x and L_y denote the size of the room in x- and y-direction, respectively. We choose the origin of the head direction ϕ such that $\phi = \pi/2$ corresponds to the rat looking to the north. The velocity vector is given by $\mathbf{v} = (v_x, v_y, \omega)$, where v_x, v_y denote the translation velocities and ω is the rotation velocity. For the typical pellet-throwing experiment we make the approximation that the velocities in the three different directions are decorrelated and that the rat's position and head direction are homogeneously distributed in configuration space. Moreover, in an open field there is no reason why the variance of the velocity should be different in x- and y-direction. The covariance matrix of the velocities then takes the form

$$\mathbf{K} = \begin{pmatrix} \langle v^2 \rangle & 0 & 0 \\ 0 & \langle v^2 \rangle & 0 \\ 0 & 0 & \langle \omega^2 \rangle \end{pmatrix} \tag{23}$$

and the probability density $p(x, y, \phi)$ is a constant.

In this case the eigenvalue problem (20) for the operator $\mathcal{D}$ takes the following form:

$$-\left[\langle v^2 \rangle \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) + \langle \omega^2 \rangle \frac{\partial^2}{\partial \phi^2}\right] g(x, y, \phi) = \Delta g(x, y, \phi) \tag{24}$$

with the boundary conditions (16) yielding

$$\frac{\partial}{\partial x} g(x, y, \phi) = 0 \quad \text{for} \quad x \in \{0, L_x\} \tag{25}$$

$$\frac{\partial}{\partial y} g(x, y, \phi) = 0 \quad \text{for} \quad x \in \{0, L_y\} \tag{26}$$

and cyclic boundary conditions in the angular direction.

It is easy to check that the eigenfunctions and the corresponding Δ-values are given by

$$g_{lmn}(x, y, \phi) = \begin{cases} \sqrt{2}^3 \cos(l\pi \frac{x}{L_x}) \cos(m\pi \frac{y}{L_y}) \sin(\frac{n+1}{2}\phi) & \text{for l odd} \\ \sqrt{2}^3 \cos(l\pi \frac{x}{L_x}) \cos(m\pi \frac{y}{L_y}) \cos(\frac{n}{2}\phi) & \text{for l even} \end{cases} \tag{27}$$

$$\Delta_{lmn} = \begin{cases} \pi^2 \langle v^2 \rangle \left(\frac{l^2}{L_x^2} + \frac{m^2}{L_y^2}\right) + \langle \omega^2 \rangle \frac{(n+1)^2}{4} & \text{for l odd} \\ \pi^2 \langle v^2 \rangle \left(\frac{l^2}{L_x^2} + \frac{m^2}{L_y^2}\right) + \langle \omega^2 \rangle \frac{n^2}{4} & \text{for l even,} \end{cases} \tag{28}$$

with l, m, and n being nonnegative natural numbers. Only $l = m = n = 0$ is not allowed, as this case corresponds to the constant solution, which violates the unit variance constraint.

To predict the actual outcome of the simulations we need to order these solutions by their Δ-values. For better comparability with the simulation results it is convenient to rewrite the Δ-values in the following form:

$$\Delta_{lmn} = \frac{\pi^2 \langle v^2 \rangle}{L_x^2} \begin{cases} l^2 + \frac{L_x^2}{L_y^2} m^2 + v_{rel}^2 (n+1)^2 & \text{for l odd} \\ l^2 + \frac{L_x^2}{L_y^2} m^2 + v_{rel}^2 n^2 & \text{for l even,} \end{cases} \tag{29}$$

where

$$v_{rel}^2 = \frac{\langle (\frac{\omega}{2\pi})^2 \rangle}{\langle (\frac{v}{L_x})^2 \rangle} \tag{30}$$

denotes the relative rotational speed, i.e. the ratio of the root mean square of rotational and translational velocity, if translational velocity is measured in room size in x-direction per second and rotational velocity is measured in full circles per second.

We can now discuss two limit cases in terms of the relative velocity v_{rel}. Let us first consider the case where the rat moves at small velocities while making a lot of quick turns, i.e. $v_{rel} \gg 1$. In this case, the smallest Δ-values can be reached by setting $n = 0$ unless $l^2 + \frac{L_x^2}{L_y^2} m^2 > v_{rel}^2$. Since for $n = 0$ the functions g_{lmn} do not depend on the angle ϕ, the slowest functions for this case are invariant with respect to head direction and lead to place cells, see below. The behavior of the solutions and the respective simulation results are depicted in figure 3A and B.

In the other extreme, v_{rel} is much smaller than one, i.e. the rat runs relatively fast while making few or slow turns. The smallest Δ-values can then be reached by choosing $l = m = 0$ unless $n^2 > \min(1, \frac{L_x^2}{L_y^2})/v_{rel}^2$. The corresponding functions are invariant with respect to position while being selective to head direction, a feature that is characteristic for head-direction cells. A comparison of these theoretically predicted functions with simulation results are shown in figure 3D and E.

5.1.2 Simulation results for the simple movement paradigm

It is intuitively clear and has been shown in the last section that for high relative orientational speed v_{rel} the system output becomes slowest if it is invariant to head direction and only codes for spatial position. For low v_{rel} on the other hand invariance for position while coding for head orientation is the best solution to the optimization problem.

In figure 3B the spatial firing maps of SFA output units from the simulation with high v_{rel} are shown. Here, all units are almost completely orientation-invariant and resemble the theoretical predictions from figure 3A. The first unit is not active when the simulated rat is in the south of the apparatus, most active in the north, and shows a gradual increase in the shape of a half cosine wave in between. The unit is invariant to movements in east-west direction. The second unit behaves similarly, but its activity pattern is rotated by 90 degrees. The following units have more spatial oscillations and somewhat resemble grid cells which are not localized.

Figure 3C shows ICA output units from the same simulation as in figure 3B. All units are orientation-invariant, just as their input from the first 16 SFA units, but most have only a single peak of activity and each at a different position. The sparser units are more localized in space while less sparse units have larger firing fields or multiple peaks. These results closely resemble place cells from rodent's hippocampal areas CA1 and CA3.

In figure 3E SFA output units from the simulation with low relative rotational speed v_{rel} are shown. In this case, all units are almost completely position-invariant but their response oscillates with the orientation of the rat. The first unit changes activity with the sine of orientation and the second unit is modulated like a cosine. Unit #3 has twice the frequency, unit #5 has a frequency of three, and unit #8 a frequency of eight. Figure 3F shows ICA output units from the same simulation as in figure 3E. All units are position-invariant like their inputs from the first 8 SFA units, but most have only a single peak of activity and each at a different orientation. The sparser units are more localized in orientation while later ones have broader tuning curves. These results closely resemble head-direction cells from rodent's subicular areas.

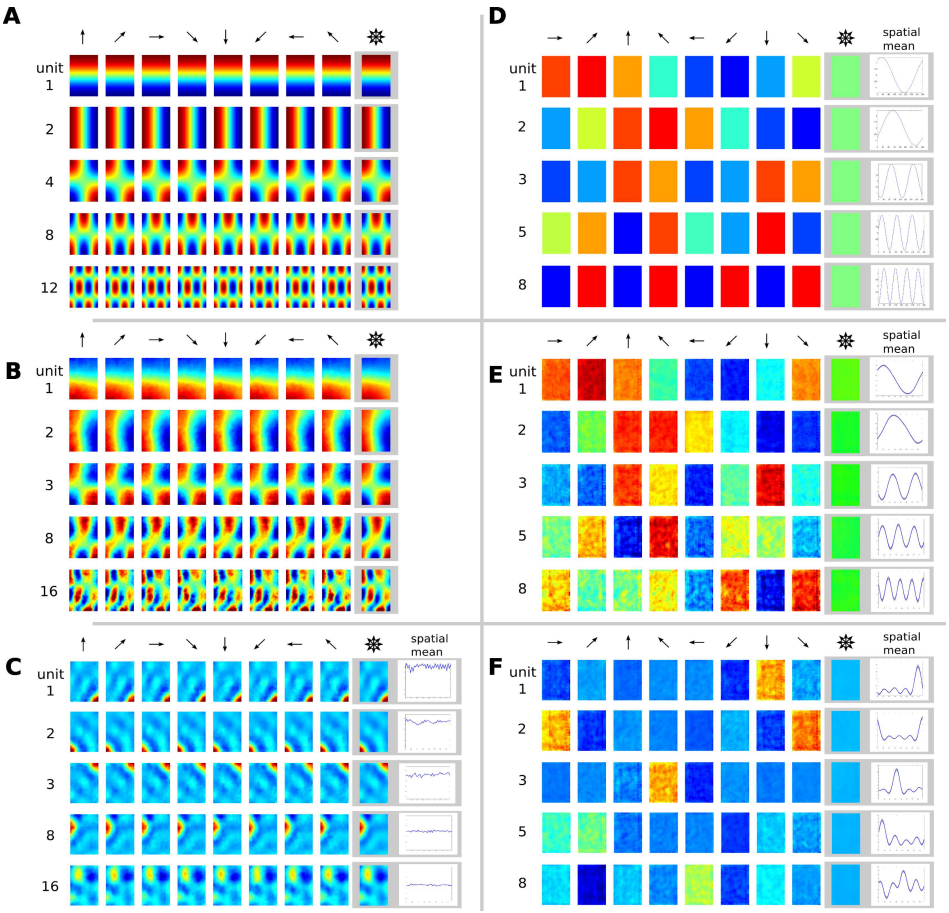

Figure 3: **Theoretical predictions and simulation results for the open field with the simple movement paradigm (independent translation and head direction), separately learned place cells and head-direction cells, and ICA for sparsification.** Each row within each panel shows the response of one unit as a function of position for different head directions, as well as the mean value averaged over all head directions (indicated by the superimposed arrows). Panel D also shows orientation tuning curves (at the position of a unit's maximal activity). Panels D-F also show orientation tuning curves (averaged over all positions) ± 1 standard deviation.

A: Theoretical prediction for the SFA layer with relatively quick rotational speed compared to translational speed. Solutions are ordered by slowness. All solutions are head direction invariant and form regular rectangular grid structures.

B: Simulation results for the SFA layer for the same settings as in A, ordered by slowness. The results are similar to the theoretical predictions up to mirroring, sign, and mixing of almost equally slow solutions. All units are head direction invariant and code for spatial position but are not localized in space.

C: Simulation results for the ICA layer for the same simulation as in B, ordered by sparseness (kurtosis). Firing patterns of all units are head direction invariant and localized in space, resembling hippocampal place cells.

D: Theoretical prediction for the SFA layer for relatively slow rotational speed compared to translational speed. Solutions are ordered by slowness. All solutions are position invariant and constitute a Fourier basis in head direction space.

E: Simulation results for the SFA layer for the same settings as in D, ordered by slowness. The results are similar to the theoretical predictions up to phase shift and sign. All units are position invariant and head direction specific but not localized in head direction space, i.e. all units except #1 and #2 have multiple peaks.

F: Simulation results for the ICA layer for the same simulation as in E ordered by sparseness (kurtosis). Firing patterns of all units are position invariant and localized in head direction space resembling subicular head-direction cells.

5.1.3 Simulation results for the restricted head movement paradigm

In the previous section we used independent head direction and body movement and used different movement statistics for different cell types, such as fast rotational speed for place cells and slow rotational speed for head-direction cells. This allowed us to obtain nearly ideal simulation results that match closely the theoretical predictions, but it is unrealistic for two reasons. Firstly, in a real rat head-direction and movement direction are correlated. Secondly, in a real rat place cells and head-direction cells have to be learned simultaneously and thus with the same movement pattern.

In this section we introduce three changes for higher realism. Firstly, a more realistic movement pattern is used, where the rat's head is enforced to be within 90° of the current body movement (see methods). Secondly, place cells and head-direction cells are learned on the same input statistics and learning rate adaptation (LRA) is used in the top SFA layer for the head-direction cell population (see methods). Thirdly, ICA for sparse coding in the last layer is replaced by competitive learning (CL). Simulation results are shown in figure 4.

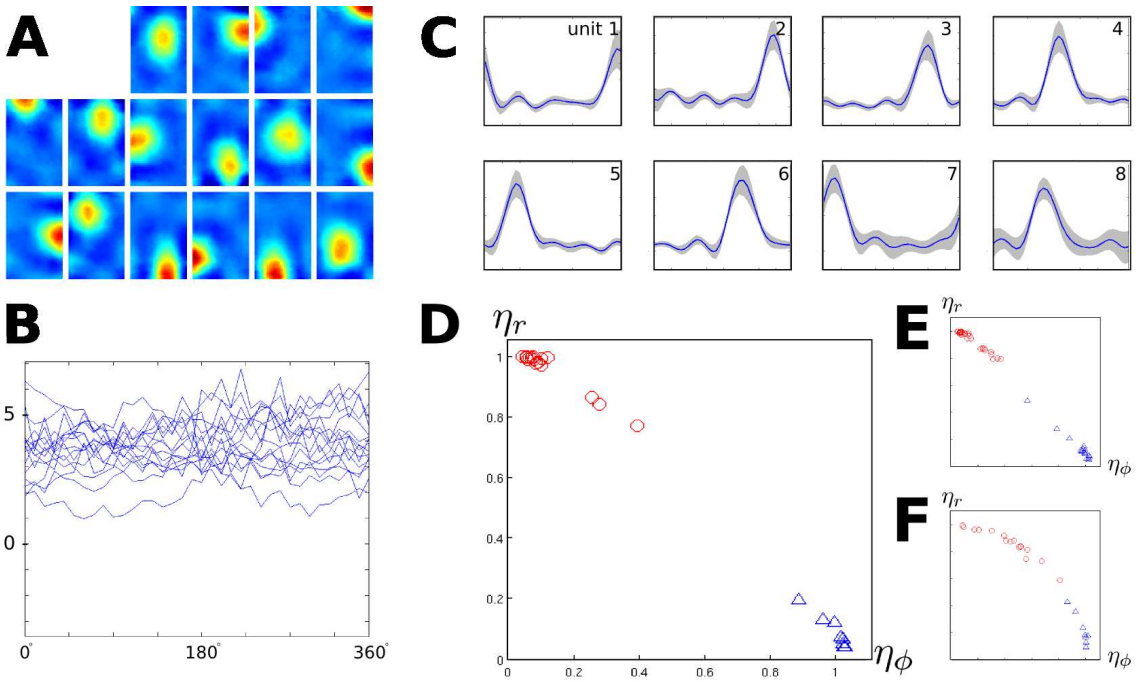

Figure 4: **Simulation results for the open field with more realistic movement patterns and competitive learning (CL)** for sparsification in the last layer.
The network was trained with a movement pattern of relatively high rotational speed. Two distinct populations of cells were trained, one as before, the other was trained with learning rate adaptation (LRA) in the top SFA layer, reducing the impact of periods with high rotational speed.
A: Simulation results for the top layer CL units without LRA. Each subplot shows the mean spatial firing rate of one output unit averaged over all orientations. The slowest 16 SFA outputs were used for CL, and 16 CL units were trained. All units are localized in space, closely resembling hippocampal place cells.
B: Orientation tuning of the units shown in A. Firing patterns of all units are mostly head direction invariant.
C: Simulation results for the top layer CL units with LRA in the top SFA layer. Each subplot shows the mean orientation tuning curve in blue and a grey area indicating ± 1 standard deviation. The slowest 8 SFA-outputs were used for CL, and 8 CL units were trained. Firing patterns of all units are mostly position invariant and localized in head direction space closely resembling subicular head-direction cells.
D: Scatterplot of mean directional variance η_ϕ and mean positional variance η_r for the results shown in A (red circles) and C (blue triangles). Units from A cluster in an area with high positional variance η_r and low orientational variance η_ϕ, while units from C cluster in an area with low positional variance η_r and high orientational variance η_ϕ.
E: Scatterplot of η_ϕ and η_r for the same simulation parameters as in A-D but with more CL output units. 32 units were trained without LRA (red circles) and 16 with LRA (blue triangles). The solutions lie in similar areas as in D.
F: Scatterplot of η_ϕ and η_r for the same simulation parameters as in A-D, but with more SFA outputs used for CL. 32 SFA units were used without LRA (red circles) and 16 with LRA (blue triangles circles). The solutions show mixed dependence on position and head direction but are still clearly divided into a mostly head direction-invariant population (red) and a mostly position-invariant population (blue).

As the relative rotational speed is smaller than in the previous section some SFA solutions (not shown) change with head direction: unit #16 of 32 is the first unit with noticeable head direction dependence here while none of the first 32 SFA solutions in the last section was head direction dependent. In figure 4A the spatial firing maps for all trained units without LRA are shown averaged over all orientations. The corresponding orientation tuning curves (measured at the peak of the place field) are given in panel B. All units are localized in space and largely independent of orientation with activity centers distributed evenly in the room.

Figure 4C shows the simulation results with identical movement statistics but with LRA turned on in the top SFA layer, so that learning is downregulated at timepoints with rapid head direction changes. Tuning curves of all units are shown together with the spatial standard deviation of activity, which is generally very small. All units are localized in head direction space and mostly position independent with approximately even spacing of directions of maximum activity. The LRA can eliminate the effect of head rotation only to some extent and thus SFA units #7 and #8 show significant dependence on position while the slowest unit affected by position in the previous section was #15.

A scatterplot of the mean positional variance η_r versus mean orientational variance η_ϕ (see methods) of the units from A and C is shown in figure 4D. Perfect head-direction cells would be located in the bottom right while perfect place cells would be located in the top left. Red circles denote the simulated place cells from panel A; the blue triangles denote the simulated head-direction cells from panel C. Both populations cluster near the positions of optimal solutions in the corners.

How does the number of inputs to the last layer (i.e. the number of SFA-outputs used) and the number of CL outputs influence the results? Panel E shows the same analysis for a simulation with identical settings except the number of CL-output units was doubled to 32 without LRA and 16 with LRA, respectively. Most units lie in a similar area as in D, but the clusters are denser, since the number of units has doubled. In panel F, the number of output units is again the same as in D, but the number of SFA outputs for the last layer is doubled to 32 for the simulation without LRA and 16 for the simulation with LRA. The output units now get inputs from higher, i.e. quicker, SFA units which tend to have stronger influence of both position and orientation. As a result, the CL units span the entire spectrum of completely position invariant to complete orientation invariant solutions, with the less position-dependent solutions coming from the simulations without LRA, and the less head direction dependent solutions coming from the LRA simulation. We conclude that the number of CL-output units mostly determines the density of place cells but not the qualitative behavior of the solutions while the number of SFA-outputs directly affects the invariance properties of the solutions.

5.1.4 Simulation results for the spatial view paradigm

The previous sections have shown that the same learning mechanism in the same environment, just with different movement statistics, results in either head-direction or place-cell like representations. Although the last section introduced certain restrictions on the head direction, body position and head direction remained mostly independent.

In the following simulation, the virtual animal fixates a location L on a wall while it moves through the room. The position of L changes with the same statistics as for the head direction simulation above (see methods). A visualization of the simulation results by plotting the activity of a unit at a given position vs. "global" orientation, as in the previous figures, looks inconclusive (figure 5A). Plotting the activity of a unit such that at each position the orientation is chosen to face a fixed specific location marked by an '$\times$' shows spatially homogeneous activities (figure 5C; cf. figure 1). These cells jointly code for the 'view space' but as before the SFA results are not localized. Figure 5B and D show the

results of the ICA layer. The 'global direction' plot in B is as inadequate as in A while plot D clearly illustrates the behavior of these cells. Unit #2, for example, is active only when looking at the bottom left corner of the rectangular room, independently of the animal's position. This cell type resembles spatial view cells found in the primate hippocampal formation (e.g. Rolls et al., 2005).

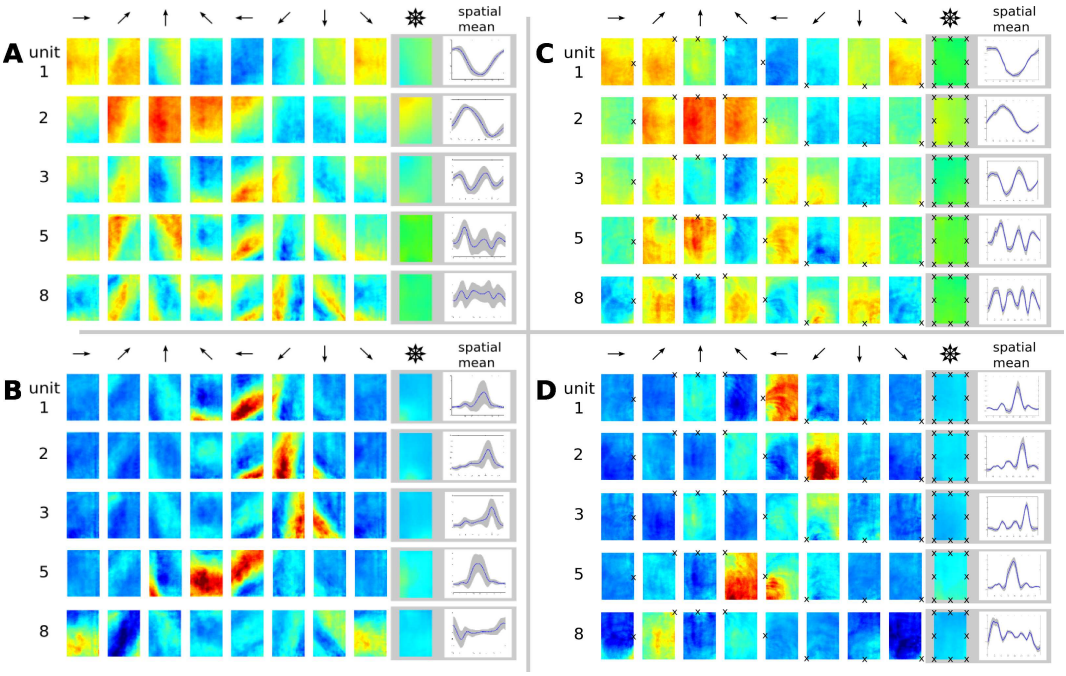

Figure 5: **Simulation results for the open field with trajectories where spots on the wall were fixated.** A: Spatial firing map of five representative SFA output units for different 'global head directions' (indicated by arrows) and averages over orientations and space. No unit shows spatial or orientation invariance when plotting position and 'global head direction' as in previous figures. C: Same results as in A but plotted with 'local head direction' (at each position oriented towards fixation point '×'). B: ICA results plotted with 'global head direction'. D: Same results as in B but using the plot method from C. All units code for a specific view closely resembling primate spatial view cells.

5.2 Linear track

In a linear track the rat's movement is essentially restricted to two degrees of freedom, a spatial and an orientational one. In experimental measurements the orientational dimension is often collapsed into a binary variable indicating only the direction of movement. In the linear track these two dimensions are thus experimentally much easier to sample smoothly than the full three dimensional parameter space of the open field.

5.2.1 Theoretical predictions for the linear track*

In principle the configuration space for the linear track is the same as for the open field, only with small side length L_x in one direction. Equation (28) shows that for small L_x the solutions that are not

constant in the x-direction, i.e. the solutions with $k \neq 0$, have large Δ-values and thus vary quickly. Because slow functions will thus be independent of x, we will neglect this dimension and restrict the configuration space to position in x-direction and head direction ϕ.

Another difference between the simulation setup for the open field and the linear track lies in the movement statistics of the rat. Due to the momentum of the Brownian motion the rat rarely turns on mid-track. In combination with the coupling between head direction and body motion this implies that given the sign of the velocity in y-direction the head direction is restricted to angles between either 0 and π (positive velocity) or between π and 2π (negative velocity). If, in addition, the rat makes a lot of quick head rotations, the resulting functions can only be slowly varying if they are invariant with respect to head direction within these ranges. This leaves us with a reduced configuration space that contains the position y and a binary value $d \in \{\text{North, South}\}$ that determines whether $0 \leq \phi < \pi$ (positive velocity in y-direction, north) or $\pi \leq \phi < 2\pi$ (negative velocity in y-direction, south).

We assume that the rat only switches between north and south at the ends of the track. Because discontinuities in the functions lead to large Δ-values, slow functions $g(y, d)$ should fulfill the continuity condition that $g(0, \text{North}) = g(0, \text{South})$ and $g(L_y, \text{North}) = g(L_y, \text{South})$. This means that the configuration space has the topology of a circle, where one half of the circle represents all positions with the rat facing north and the other half the positions with the rat facing south. It is thus convenient to introduce a different variable $\xi \in [0, 2L_y]$ that labels the configurations in the following way:

$$(x(\xi), d(\xi)) = \begin{cases} (\xi, \text{North}) & \text{for} \quad \xi < L_y \\ (2L_y - \xi, \text{South}) & \text{for} \quad \xi \geq L_y \end{cases} . \tag{31}$$

The topology of the configuration space is then captured by cyclic boundary conditions for the functions $g(\xi)$.

For simplicity we assume that there are no preferred positions or head directions, i.e. that both the variance of the velocity $K = \langle \dot{\xi}^2 \rangle$ and the probability distribution $p(\xi)$ is independent of ξ. The equation for the optimal function is then given by

$$-\langle \dot{\xi}^2 \rangle \frac{\partial^2}{\partial \xi^2} g(\xi) = \Delta g(\xi) \tag{32}$$

The solutions that satisfy the cyclic boundary condition and their Δ-values are given by

$$g_j(\xi) = \begin{cases} \sqrt{2} \sin(j\pi \frac{\xi}{2L_y}) & \text{for j even} \\ \sqrt{2} \cos((j+1)\pi \frac{\xi}{2L_y}) & \text{for j odd} \end{cases} , \tag{33}$$

$$\Delta_j(\xi) = \begin{cases} \pi^2 \frac{\langle \dot{\xi}^2 \rangle}{4L_y^2} j^2 & \text{for j even} \\ \pi^2 \frac{\langle \dot{\xi}^2 \rangle}{4L_y} (j+1)^2 & \text{for j odd} \end{cases} . \tag{34}$$

Note that there are always two functions with the same Δ-value. Theoretically, any linear combination of these functions has the same Δ-value and is thus also a possible solution. In the simulation, this degeneracy does not occur, because mid-track turns do occur occasionally, so those functions that are head-direction-dependent on mid-track (i.e. even j) will have higher Δ-values than theoretically predicted. This avoids mixed solutions and changes the order of the functions when ordered by slowness. Figure 6A shows seven of the slowest functions $g_,$.

5.2.2 Simulation results for the linear track

For simulations in the linear track we use the more realistic movement paradigm similar to the open field experiment from section 5.1.3. A similar relative speed is assumed and sparse coding in the last

layer is performed with ICA.

Figure 6B and C shows the simulation results for the linear track. The spatial firing maps of the slowest seven SFA outputs out of 10 are shown in figure 6B. Units #1–6 are mostly head direction invariant ($\eta_\phi \leq 0.1$), and code for spatial position in the form of sine waves with respective frequencies of $\frac{1}{2}$, 1, $1\frac{1}{2}$, 2, $2\frac{1}{2}$, and 3, as theoretically predicted. Unit #7 codes for position and orientation. At track ends, where most rotation occurs, all units are head-direction invariant and the spatial modulation is compressed due to slower mean translational speeds compared to mid-track (cf. appendix). As expected, none of these units are localized in space or orientation.

The spatial firing maps of the first seven out of ten ICA outputs for different head directions are shown in figure 6C. Units #1 and #6 are only active at the southern track end independently of head direction. The other five units are localized in the joint position-head-direction space meaning that they fire only at specific positions on the track when the rat faces a specific direction. These results are similar to place cell recordings from rats in linear tracks where most cells only fire when the rat moves in one direction (Muller et al., 1994).

Changing the movement pattern to yield much higher or much lower mean relative rotational speeds, respectively, leads to very different results resembling those presented earlier for the open field, namely head-direction cells and head-direction invariant place cells.

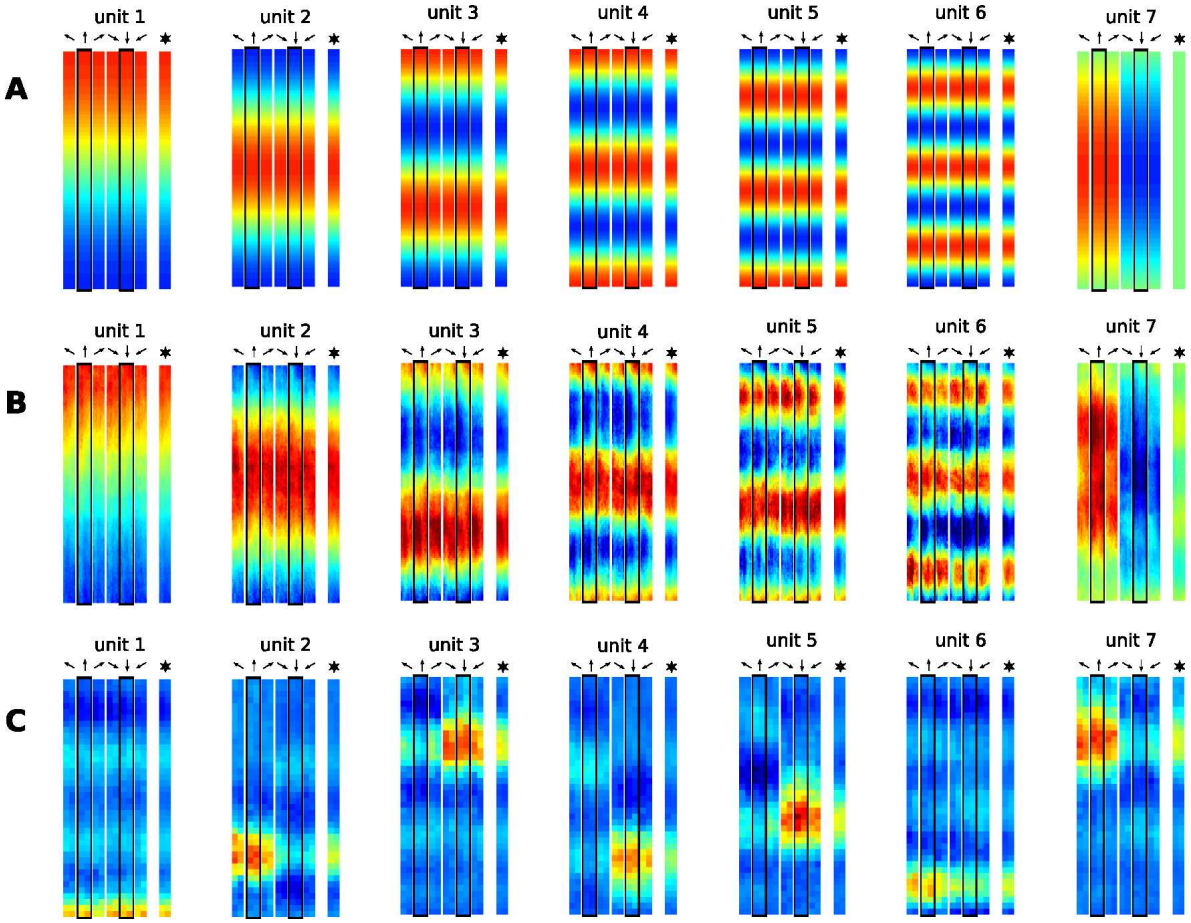

Figure 6: **Theoretical predictions and simulation results for the linear track.** Head directions are indicated by arrows, orientation averages are indicated by superimposed arrows, and principal directions (north, south) are emphasized with a dark border. A: Theoretical predictions. B: Spatial firing maps of the first (slowest) seven SFA output units out of 10. Units #1-#6 are mostly head direction invariant, whereas unit #7 responds differently to north and south views. C: Spatial firing maps of the first (most kurtotic) seven out of 10 ICA output units. All units are localized in space and most are only active for either north or south views closely resembling place fields recorded from rats in linear track experiments.

5.3 Model parameters

Although most of the parameters in our model (i.e. all the weights in the SFA and ICA steps) are learned in an unsupervised manner a number of parameters were chosen manually. These parameters include the input picture size, receptive field sizes, receptive field positions and overlaps in all layers, the room shape and textures, the expansion function space, number of layers, choice of sparsification algorithm, movement pattern, field of view, and number of training steps. We cannot explore the entire parameter space here and show instead that the model performance is very robust with respect

to most of these parameters. The fact that the presented simulation results are very similar to the analytical solutions also indicates that the results are generic and not a mere artifact of a specific parameter set.

We use high-resolution input pictures of 40 by 320 RGB pixels showing the capability of the model to handle high-dimensional sensory data. Nevertheless, it could be argued that the rat's vision is rather blurred and has little color sensitivity. However, we find that smaller and/or grayscale input pictures yield similar results, which degrade only below a dimensionality of a few hundred input pixels.

The model's field of view (FOV) has been modeled to represent the 320° of a rat's FOV. Smaller FOVs below 90° still reproduce our results and especially rotation invariance is not an effect of a large FOV. Nevertheless, the views have to contain enough visual information in order to fulfill the one-to-one correspondence between stimulus and oriospatial configuration. For smaller FOV values and symmetrical environments the model's representations become symmetrical as well.

The receptive fields are restricted to about 100 input dimensions (before quadratic expansion) due to computational limitations. Larger receptive fields tend to yield better solutions, since the available total function space increases. Position and overlap of receptive fields have been varied to some extent but have no noticeable impact on the result unless too many of the inputs are discarded.

The room shape has a strong impact on the SFA solutions, which can be predicted analytically. We show here only results from convex rooms, but experiments with radial mazes and multiple rooms have been performed and these results are similar to experimental data, too. Choice of specific textures was irrelevant for the model's performance except when multiple walls are textured with similar or identical textures, which leads to degraded results due to visual ambiguities.

The expansion function was chosen as all monomials up to degree 2, but alternative function spaces like linear random mixtures passed through sigmoidals with different offsets were successful, too; however, the size of the function space is limited by computational constraints and monomials have proven to be particularly efficient.

The number of layers is determined by receptive field sizes and overlaps. An increased number of layers also increases the function space and can thus improve performance. We did not see any effect of overfitting for larger numbers of layers. Additional top layers simply reproduced the output of earlier layers.

As for the choice of the sparse coding algorithm, we found no large qualitative difference for different techniques including CuBICA, fastICA, competitive learning, or just finding rotations of the SFA output with maximal kurtosis (Franzius et al., 2007).

The choice of movement pattern has a clear impact on the optimal solutions of SFA. The theoretical analysis presented here can in principle predict the solutions for arbitrary movement patterns but for the predictions presented here we made simplifying assumptions to obtain closed form solutions. In spite of these simplifications, the theoretical predictions are still close to the simulation results, e.g. in section 5.1.3, where the head orientation is restricted to an angular range with respect to the direction of body motion. simulation results are still similar to the theoretical predictions.

More training steps result in a smoother sampling of the virtual reality environment and yield better approximations to the theoretical predictions. We found that a few laps crossing and spanning the whole room within a few thousand training samples were sufficient for the qualitative results already. For too little training data and too few crossings of paths an overfitting effect occurs resulting in a slowly varying activity of the outputs on the training path but not on other (test) paths.

6 Discussion

We have presented a model for the formation of oriospatial cells based on the unsupervised learning principles of slowness and sparseness. The model is feed-forward, instantaneous, and purely sensory driven. The architecture of the model is inspired by the hierarchical organization of the visual system and applies the identical learning rule, Slow Feature Analysis, on all but the last layer, which performs sparse coding. Our results show that all major oriospatial cell types - place cells, head-direction cells, spatial view cells, and to some extent even grid cells - can be learned with this approach. We have shown that this model is capable of extracting cognitive information such as an animal's position from complex high-dimensional visual stimuli, which we simulated as views in a virtual environment. The generated representations were coding specifically for some information (e.g. position) and were invariant to the others (e.g. orientation). These invariant representations are not explicitly built into the model but induced by the input statistics, which in turn is determined by the room shape and a specific movement paradigm. Nevertheless, the type of learned invariance can be influenced by a temporal adaptation of the learning rate. Control experiments show that the model performance is robust to noise and architectural details. This robustness is also supported by a general mathematical framework that allows exact analytical predictions of the system behavior at the top SFA level.

Our model comprises sensory processing stages that mimic parts of the visual cortices and the hippocampal formation. The model layers cannot be exactly associated with specific brain areas, but we suggest some relations. The behavior of the lower two layers are primarily determined by the visual environment and mostly independent of the spatial movement pattern. In the simulations presented here, we trained the two lower layers only once and only adapted the higher layers for different environments and movement patterns. The first layer could be associated with V1 (Berkes and Wiskott, 2005), the second layer with higher visual cortices. Units in the third layer, whose non-localized spatial activity pattern resembles grid cells, strongly depend on the movement pattern and might be associated with grid cells in EC. Recent results from EC (Sargolini et al., 2006) show that grid cells in MEC exhibit some head-direction dependency, similar to our model for the case of the intermediate relative translational speed in the open field. Depending on the movement statistics during learning, representations in the sparse coding layer resemble either place cells as found in hippocampal areas CA1 and CA3 or head direction cells as found in many areas of the hippocampal formation or spatial view cells as found in the hippocampal formation of monkeys.

For the case of approximately uncorrelated body movement and head direction, the model learns either place or head-direction cells, depending on the relative speed of translation and rotation. For much quicker rotation than translation the model develops orientation-invariant place fields while for much quicker translation than rotation the model develops position-invariant head direction codes. In intermediate cases, e.g. for the linear track, mixed representations such as direction-dependent place fields emerge. In the case of correlated body movement and head direction caused by elongated fixations of objects or positions, the model learns view-specific codes, similar to spatial view cells in primates.

Although the model is capable of learning place cells and head direction cells if it learns on distinct adequate movement statistics, a model rat should obviously not have to traverse its environment once with high relative translational speed to learn head-direction cells and once more with low relative translational speed to learn place cells. How can both populations be trained with a single given input statistics? For this problem we consider output from the rat's vestibular system as a possible solution. This system is essential for the oriospatial specificity of head direction cells and place cells (Stackman and Zugaro, 2005). Other models like the well established ring attractor model by Skaggs et al. (1995) assume that the head direction system performs angular integration of body motion based on vestibular

velocity signals. We hypothesize that these signals could also be used to influence the learning rate of two populations of cells that learn according to our model. One of these populations learns more strongly at periods with high relative translational speed (as signalled by the vestibular velocity signals) and the other adapts more strongly for low relative translational speed. The former should develop head-direction cell characteristics and the latter place cell characteristics. In our simulations the model successfully learned both populations with the same input data, one population without learning rate adaptation, and one population with reduced learning rate during quick turns. Once the model has been trained, the vestibular acceleration signal is no longer needed for the model behavior. With learning rate adaptation the model neurons effectively learn on a different movement statistics, e.g. head direction cells learn more strongly at times with relatively high translational speed. Nevertheless, if the real movement statistics contains very few episodes of relatively quick translation at all, the mechanism fails and head direction cells cannot become position invariant.

Our implementation of the slowness principle involves solving an eigenvalue problem and cannot be considered biologically plausible. However, more plausible equivalent formulations of the slowness principle exist in the form of gradient-descent learning rules (Hashimoto, 2003; Kayser et al., 2001) and as spike based learning mechanisms (Sprekeler et al., 2006). The choice of ICA to generate localized representations from nonlocalized codes is also biologically unrealistic, whereas a formulation in the form of Hebbian learning (Oja and Karhunen, 1995) or competitive learning seems more plausible. An in-depth discussion of this topic can be found in Franzius et al. (2007).

Related work

According to Redish's classification, our model is a *local view model,* for it "only depends on the local view to explain place cell firing" (Redish, 1999). Models of this class usually extract a number of features from sensory inputs in order to obtain a lower-dimensional representation that still carries information about spatial position in the environment but is invariant to everything else. Pure local view models do not comprise a path integration system and thus cannot fully explain oriospatial firing properties, e.g. in darkness. Pure path integration systems without external sensory input on the other hand inherently accumulate errors, and hence a sensory coding mechanism, as proposed here, is necessary to complement any such model. Therefore many models combine local view and path integration mechanisms (McNaughton et al., 2006; Redish, 1999).

The model by Wyss et al. (2006) is based on similar principles as our model. It applies a learning rule based on temporal stability to natural stimuli, some of which are obtained from a robot. The resulting spatial representations are localized, resembling hippocampal place fields. The learning rule involves local memory and no explicit sparsification method is applied. The fact that the resulting representations are localized is somewhat surprising, since by itself temporal stability does not lead to localized representations (Franzius et al., 2007). We hypothesize that the decorrelation of the non-negative activities in the model implicitly leads to a sparsification because it favors a code where at any given time only one single unit is active. The article does not investigate head-direction-dependency of the learned representations or dependencies on the movement statistics.

The model by Sharp (1991) assumes abstract sensory inputs and acquires a place code by competitive learning, resulting in units that code for views with similar input features. Thus, this model is similar to our model's last layer performing sparsification. Similarly to our results, cells become less orientation-dependent if more rotations occur in the training trajectory.

The work by Fuhs et al. (1998) uses realistic natural stimuli obtained by a robot and extracts "blobs" of uniform intensity with rectangular or oval shape from these images. Radial basis functions are tuned to blob parameters at specific views, and a competitive learning scheme on these yields

place-cell-like representations. Our model agrees with their conclusion that rodents need no explicit object recognition in order to extract spatial information from natural visual stimuli.

The model by Brunel and Trullier (1998) investigates the head-direction dependency of simulated place fields using abstract local views as inputs. A recurrent network learns with an unsupervised Hebbian rule, associating local views with each other, such that their intrinsically directional place cells can become head-direction invariant for maze positions with many rotations. The article also conjectures that movement patterns determine head-direction dependence of place cells, which is consistent with our results.

The results by de Araujo et al. (2001) suggest that the size of the rat's field of view (FOV) is important for the distinction between spatial view cells and place cells. With a large FOV (as for rats) the animal can see most landmarks from all orientations while an animal with a small FOV (like a monkey) can only see a subset of all landmarks at each timepoint. We find no dependence of our results on the FOV size for values between 30 and 320 degree as long as the environment is rich enough (i.e. diverse textures, not a single cue card). Instead, our results suggest that differences in the movement statistics play a key role for establishing this difference.

To our knowledge, no prior model allows the learning of place cells, head-direction cells, and spatial view cells with the same learning rule. Furthermore there are only few models that allow clear theoretical predictions, learn oriospatial cells from (quasi) natural stimuli, and are based on a learning rule that is also known to model early visual processing well.

Future perspectives

Our model is not limited to processing visual stimuli, as presented here, but can integrate other modalities as well. The integration of olfactory cues, for example, might lead to even more accurate representations and possibly to an independence of the model of visual stimuli (simulated darkness).

Our simulated visual stimuli come from a virtual reality environment which is completely static during the training of the virtual rat. In this case the slowest features are position, orientation, or view direction as shown before. However, the assumption that the environment remains unchanged during oriospatial cell learning certainly does not hold for the real world. A more realistic environment will include other changing variables like lighting direction, pitch and roll of the head etc. The impact of these variables on the model representations depends on the timescale of the variable changes: e.g. the additional white noise in all SFA layers of the model is ignored since it varies much quicker than position and orientation, but the direction of sunlight might become the slowest feature. Generally, the SFA solutions will depend on any variable whose timescale is equal or slower than the position and orientation changes of the animal. After the sparse coding step representations will become not only localized in position and/or head direction but in the other variables as well. This behavior is not consistent with the definition of an ideal place or head-direction cell. However, many experiments show correlations of place cell firing with nonspatial variables as well (Redish, 1999). One particularly interesting instance of such a variable is 'room identity'. If a rat experiences multiple environments, usually transitions between these will be seldom, i.e. the rat will more often turn and traverse a single room than switch rooms. In this case room identity will be encoded by the SFA outputs. For n rooms at most $n-1$ decorrelated SFA outputs can code for the room identity. The following outputs will then code for a joint representation of space and room identity. After sparse coding, many output units will fire in one room only (the less sparse ones in few rooms), and possibly in a completely unrelated fashion to their spatial firing patterns in another room. This behavior is consistent with the 'remapping' phenomenon in place cells (e.g. Muller and Kubie, 1987).

A great amount of work has been done investigating the impact of environmental manipulations

on oriospatial cell firing in *known* rooms, e.g. shifts and rotations of landmarks relative to each other (Redish, 1999). How would our model behave after such changes to the learned environment? Such transformations effectively lead to visual input stimuli outside the set of all possible views in the training environment. In this case, we expect the system's performance to deteriorate unless a new representation is learned, but more work is necessary to investigate this question.

Our approach predicts increasing slowness (i.e. decreasing eta-values of firing rates) in the processing hierarchy between retina and hippocampus. Additionally, place cell and head direction cell output should be significantly sparser than their inputs. Our main prediction is that changing movement statistics directly influences the invariance properties of oriospatial cells: e.g. an experiment in a linear track where the rat more often turns on mid-track should yield less head-direction dependent place cells.

Experimentally, the joint positional and orientational dependence of oriospatial cells is hard to measure due to the size of the three-dimensional parameter space, and even more so if the development over time is to be measured. Furthermore, precise data on movement trajectories is rare in the existing literature on oriospatial cells. Accordingly, little data is available to verify or falsify our prediction how the brain's oriospatial codes depend on the movement statistics. As an alternative to determining the movement statistics in behavioral tasks, some work has been done on passive movement of rats, where the movement statistics is completely controlled by the experimenter (e.g. Gavrilov et al. 1998), but these results might not be representative for voluntary motion (Song et al., 2005). Markus et al. find directional place fields in the center of a plus maze although in the center of the maze more rotations occur than in the arms (Markus et al., 1995). This could be a contradiction to our model, although not the frequency but the relative speed, which was not measured in (Markus et al., 1995), determines head direction invariance in our model. Overall, the dependence of oriospatial cells on the animal's movement statistics as proposed here remains to be tested experimentally.

Conclusion

We conclude that a purely sensory driven unsupervised system can reproduce many properties of oriospatial cells in the rodent brain, including place cells, head-direction cells, spatial view cells, and to some extent even grid cells. These different cell types can be modeled with the same system, and the output characteristics solely depends on the movement statistics of the virtual rat. Furthermore, we showed that the integration of vestibular acceleration information can be used to learn place cells and head-direction cells with the same movement statistics and thus at the same time.

7 Acknowledgments

We thank Konrad Körding for discussions about the connection between slowness and place fields. This research was funded by the Volkswagen Foundation through a grant to LW for a junior research group.

8 Appendix*

8.1 Proofs of Theorems

Proof of Theorem 1

The technique of variational calculus can be illustrated by means of an expansion in the spirit of a Taylor expansion. Let us assume, we knew the function g_j that optimizes the objective function Ψ. The effect of a small change δg of g_j on the objective function Ψ can be written as

$$\Psi(g_j + \delta g) - \Psi(g_j) = \int \frac{\delta \Psi}{\delta g}(\mathbf{s})\, \delta g(\mathbf{s})\, \mathrm{d}\mathbf{s} + \dots, \tag{35}$$

where the ellipses stand for higher order terms in δg. The function $\frac{\delta \Psi}{\delta g}$ is the *variational derivative* of the functional Ψ and usually depends on the configuration, the optimal function g_j and possibly derivatives of g_j. Its analogue in finite-dimensional calculus is the gradient.

We now derive an expression for the variational derivative of the objective function (13). To keep the calculations tidy, we split the objective in two parts and omit the dependence on the configuration $\mathbf{s}$.

$$\Psi(g_j) =: \frac{1}{2}\Delta(g_j) - \tilde{\Psi}(g_j) \tag{36}$$

The expansion of $\tilde{\Psi}$ is straightforward:

$$\begin{aligned} \tilde{\Psi}(g_j + \delta g) - \tilde{\Psi}(g_j) &= \langle \delta g\, [\lambda_{j0} + \lambda_{jj} g_j + \sum_{i<j} \lambda_{ji} g_i] \rangle_{\mathbf{s}} + \dots && (37) \\ &= \int \delta g\, p_{\mathbf{s}}\, [\lambda_{j0} + \lambda_{jj} g_j + \sum_{i<j} \lambda_{ji} g_i] \mathrm{d}\mathbf{s} + \dots && (38) \end{aligned}$$

For the expansion of $\Delta(g_j)$ we first simplify the expression by carrying out the velocity integration and using the velocity tensor $\mathbf{K}$:

$$\Delta(g_j) \overset{(9)}{=} \langle \nabla g_j^T \mathbf{v}\mathbf{v}^T \nabla g_j \rangle_{\mathbf{s},\mathbf{v}} = \langle \nabla g_j^T \langle \mathbf{v}\mathbf{v}^T \rangle_{\mathbf{v}|\mathbf{s}} \nabla g_j \rangle_{\mathbf{s}} \overset{(14)}{=} \langle \nabla g_j^T \mathbf{K} \nabla g_j \rangle_{\mathbf{s}} \tag{39}$$

We can now expand $\Delta(g_j)$ as follows

$$\begin{aligned} \frac{1}{2}\Delta(g_j + \delta g) - \frac{1}{2}\Delta(g_j) &\overset{(39)}{=} \frac{1}{2}\langle \nabla (g_j + \delta g)^T \mathbf{K} \nabla (g_j + \delta_g) \rangle_{\mathbf{s}} - \frac{1}{2}\langle \nabla g_j^T \mathbf{K} \nabla g_j \rangle_{\mathbf{s}} && (40) \\ &= \frac{1}{2}\langle \nabla g_j^T \mathbf{K} \nabla \delta g + \nabla \delta g^T \mathbf{K} \nabla g_j \rangle_{\mathbf{s}} + \dots && (41) \\ &= \langle \nabla \delta g \mathbf{K} \nabla g_j^T \rangle_{\mathbf{s}} + \dots && (42) \\ &\quad \text{(since } \mathbf{K} \text{ is symmetric)} \\ &\overset{(8)}{=} \int p_{\mathbf{s}} \nabla \delta g \mathbf{K} \nabla g_j^T \,\mathrm{d}\,\mathbf{s} && (43) \\ &= \int \nabla \cdot \left[\delta g p_{\mathbf{s}}\, \mathbf{n}^T \mathbf{K} \nabla g_j \right] \mathrm{d}\mathbf{s} - \int \delta g\, \nabla \cdot (p_{\mathbf{s}} \mathbf{K} \nabla g_j)\, \mathrm{d}\mathbf{s} + \dots && (44) \\ &= \int_{\partial V} \delta g p_{\mathbf{s}}\, \mathbf{n}^T \mathbf{K} \nabla g_j \,\mathrm{d}A - \int \delta g\, \nabla \cdot (p_{\mathbf{s}} \mathbf{K} \nabla g_j)\, \mathrm{d}\mathbf{s} + \dots && (45) \\ &\quad \text{(Gauss' theorem)} \\ &\overset{(17)}{=} \int_{\partial V} \delta g p_{\mathbf{s}}\, \mathbf{n}^T \mathbf{K} \nabla g_j \,\mathrm{d}A + \int \delta g p_{\mathbf{s}}\, (\mathcal{D} g_j)\, \mathrm{d}\mathbf{s} + \dots && (46) \end{aligned}$$

Here, $\mathrm{d}A$ is an infinitesimal surface element of the boundary ∂V of V and $\mathbf{n}$ is the normal vector on $\mathrm{d}A$. To get the expansion of the full objective function, we add (38) and (46):

$$\Psi(g_j + \delta g) - \Psi(g_j) = \int_{\partial V} \delta g p_{\mathbf{s}}\, \mathbf{n}^T \mathbf{K} \nabla g_j \,\mathrm{d}A + \int \delta g p_{\mathbf{s}} \left(\mathcal{D} g_j - \lambda_{j0} - \lambda_{jj} g_j - \sum_{i<j} \lambda_{ji} g_i\right) \mathrm{d}\mathbf{s} + \ldots \tag{47}$$

In analogy to the finite-dimensional case, g_j can only be an optimum of the objective function Ψ if any small change δg leaves the objective unchanged up to linear order. As we employ a Lagrange multiplier ansatz, we have an unrestricted optimization problem, so we are free in choosing δg. From this it is clear that the right hand side of (47) can only vanish if the integrands of both the boundary and the volume integral vanish separately. This leaves us with the differential equation (15) and the boundary condition (16).

Proof of Theorem 2

Proof: The proof can be carried out in a direct fashion. Again, we omit the explicit dependence on $\mathbf{s}$.

$$\begin{aligned}
(f, \mathcal{D} g) &\overset{(8,17,18)}{=} -\int p_{\mathbf{s}} f \frac{1}{p_{\mathbf{s}}} \nabla \cdot p_{\mathbf{s}} \mathbf{K} \nabla g \,\mathrm{d}\mathbf{s} && (48)\\
&= -\int \nabla \cdot \left[p_{\mathbf{s}} f \mathbf{n}^T \mathbf{K} \nabla g \right] \mathrm{d}\mathbf{s} + \int p_{\mathbf{s}} \nabla f^T \mathbf{K} \nabla g \mathrm{d}\mathbf{s} && (49)\\
&= -\int_{\partial V} p_{\mathbf{s}} f \underbrace{\mathbf{n}^T \mathbf{K} \nabla g}_{\overset{(16)}{=}} \,\mathrm{d}A + \int \nabla f^T p_{\mathbf{s}} \mathbf{K} \nabla g \mathrm{d}\mathbf{s} && (50)\\
&\qquad \text{(Gauss' theorem)}\\
&\overset{(16)}{=} \int p_{\mathbf{s}} \nabla f^T \mathbf{K} \nabla g \,\mathrm{d}\mathbf{s} && (51)\\
&= \int p_{\mathbf{s}} \nabla g^T \mathbf{K} \nabla f \,\mathrm{d}\mathbf{s} && (52)\\
&\qquad \text{(since } \mathbf{K} \text{ is symmetric)}\\
&\overset{(48-52)}{=} (\mathcal{D} f, g)\,. && (53)
\end{aligned}$$

Proof of Theorem 3

Zero mean: It is obvious that the constant function $f_0 = 1$ is always an eigenfunction of $\mathcal{D}$ for eigenvalue 0. As all other eigenfunctions are orthogonal to f_0, they must have zero mean:$f_0, f_j) = \langle f_j \rangle_{\mathbf{s}} = 0 \quad \forall\, j \neq 0$.
Decorrelation: For mean-free functions f and g the scalar product (f, g) is their covariance. The orthogonality of the eigenfunctions is thus equivalent to decorrelation.
Unit variance: Unit variance can easily be achieved by renormalizing the eigenfunctions such that $(f, f) = \langle f^2 \rangle_{\mathbf{s}} = 1$.

Proof of Theorem 4

$$\Delta(f_j) \overset{(39,52)}{=} (f_j, \mathcal{D} f_j) = (f_j, \Delta_j f_j) = \Delta_j \underbrace{(f_j, f_j)}_{=1} = \Delta_j\,. \tag{54}$$

Proof of Theorem 5

Without loss of generality we assume that the eigenfunctions f_j are ordered by increasing eigenvalue, starting with the constant $f_0 = 1$. There are no negative eigenvalues, because the eigenvalue is the Δ-value of the eigenfunction, which can only be positive by definition. According to Theorem 1, the optimal responses g_j obey the boundary condition (16) and are thus elements of the subspace $\mathcal{F}_b \subset \mathcal{F}$ defined in Theorem 2. Because of the completeness of the eigenfunctions on $\mathcal{F}_b$ we can do the expansion

$$g_j = \sum_{k=1}^{\infty} \alpha_{jk} f_k \tag{55}$$

where we may omit f_0 because of the zero mean constraint. We can now prove by complete induction that $g_j = f_j$ solves the optimization problem.

Basis (j=1): Inserting g_1 into eqn. (15) we find

$$\begin{aligned} 0 &= \mathcal{D}g_1 - \lambda_{10} - \lambda_{11} g_1 && (56)\\ &= -\lambda_{10} + \sum_{k=1}^{\infty} \alpha_{1k}(\Delta_k - \lambda_{11}) f_k && (57)\\ \Rightarrow & \quad \lambda_{10} = 0 \\ & \wedge \quad (\alpha_{1k} = 0 \vee \Delta_k = \lambda_{11}) \forall k\,, && (58) \end{aligned}$$

because f_k and the constant are linearly independent and (56) must be fulfilled for all $\mathbf{s}$. (58) implies that the optimal response g_1 must be an eigenfunction of $\mathcal{D}$. As the Δ-value of the eigenfunctions is given by their eigenvalue, it is obviously optimal to chose $g_1 = f_1$. Note that although this choice is optimal, it is not necessarily unique, since there may be several eigenfunctions with the same eigenvalue. In this case any linear combination of these functions is also optimal.

Induction step: Given that $g_i = f_i$ for $i < j$, we prove that $g_j = f_j$ is optimal. Because of the orthogonality of the eigenfunctions the decorrelation constraint (12) yields

$$0 \overset{(12)}{=} \langle g_i g_j \rangle_{\mathbf{s}} = (f_i, \sum_{k=1}^{\infty} \alpha_{jk} f_k) = \alpha_{ji} \quad \forall i < j\,. \tag{59}$$

Again inserting the expansion (55) into eqn. (15) yields

$$\begin{aligned} 0 &\overset{(15,55)}{=} (\mathcal{D} - \lambda_{jj}) \sum_{k=1}^{\infty} \alpha_{jk} f_k - \lambda_{j0} - \sum_{i<j} \lambda_{ji} f_i && (60)\\ &\overset{(59)}{=} (\mathcal{D} - \lambda_{jj}) \sum_{k=j}^{\infty} \alpha_{jk} f_k - \lambda_{j0} - \sum_{i<j} \lambda_{ji} f_i && (61)\\ &\overset{20}{=} \sum_{k=j}^{\infty} (\Delta_k - \lambda_{jj}) \alpha_{jk} f_k - \lambda_{j0} - \sum_{i<j} \lambda_{ji} f_i && (62)\\ \Rightarrow & \quad \lambda_{j0} = 0 \\ & \wedge \quad \lambda_{ji} = 0 \qquad \forall i < j \\ & \wedge \quad \alpha_{jk} = 0 \vee \Delta_k = \lambda_{jj} \quad \forall k \geq j\,, && (63) \end{aligned}$$

because the eigenfunctions f_i are linearly independent. The conditions (63) can only be fulfilled if g_j is an eigenfunction of $\mathcal{D}$. Because of Theorem 4 an optimal choice for minimizing the Δ-value without violating the decorrelation constraint is $g_j = f_j$.

8.2 Qualitative Behavior of the Solutions for inhomogenous movement statistics

As seen in section 5.1.1 for the case where $p_{\mathbf{s}}$ and $\mathbf{K}$ are independent of $\mathbf{s}$, the solutions of the eigenvalue equation (20) generally show oscillations. A brief calculation for a 1-dimensional configuration space shows that their wavelength is given by $2\pi\sqrt{K/\Delta}$. It is reasonable to assume that this behavior will be preserved qualitatively if $p_{\mathbf{s}}$ and $\mathbf{K}$ are no longer homogeneous but depend weakly on the configuration. In particular, if the wavelength of the oscillation is much shorter than the typical scale on which $p_{\mathbf{s}}$ and $\mathbf{K}$ vary, it can be expected that the oscillation "does not notice" the change. Of course, we are not principally interested in quickly varying functions, but they can provide insights into the effect of variations in $p_{\mathbf{s}}$ and $\mathbf{K}$.

To examine this further, we consider the eigenvalue equation (20) for a 1-dimensional configuration space and multiply it by $p_{\mathbf{s}}$:

$$\frac{\mathrm{d}}{\mathrm{d}s}p_s(s)K(s)\frac{\mathrm{d}}{\mathrm{d}s}g(s) + \Delta p_s(s)g(s) \overset{(17,20)}{=} = 0 \tag{64}$$

We can derive an approximate solution of this equation by treating $\varepsilon := 1/\sqrt{\Delta}$ as a small but finite perturbation parameter. This corresponds to large Δ-values, i.e. quickly varying functions. For this case we can apply a perturbation theoretical approach that follows the scheme of Wentzel-Kramers-Brillouin (WKB) approximation used in quantum mechanics. Knowing that the solution shows oscillations, we start with the complex ansatz

$$g(s) = A\exp\left(\frac{i}{\varepsilon}\Phi(s)\right) \tag{65}$$

where $\Phi(s)$ is a complex function that needs to be determined. Treating ε as a small number, we can expand Φ in orders of ε

$$\Phi(s) = \Phi_0(s) + \varepsilon\Phi_1(s) + ... \tag{66}$$

where again the ellipses stand for higher order terms. We insert this expansion into equation (64) and collect terms of the same order in ε. Requiring each order to vanish separately and neglecting orders ε^2 and higher, we get equations for Φ_0 and Φ_1:

$$\Phi_0'^2 = \frac{1}{K} \tag{67}$$

$$\Phi_1' = \frac{i}{2}\frac{(p_sK\Phi_0')'}{p_sK\Phi_0'} \tag{68}$$

where the prime denotes the derivative with respect to s. These equations are solved by

$$\Phi_0(s) = \int_{s_0}^{s}\sqrt{\frac{1}{K(x)}}\mathrm{d}x \tag{69}$$

$$\Phi_1(s) = \frac{i}{2}\ln(p_sK^{1/2}) \tag{70}$$

where s_0 is an arbitrary reference point. Inserting this back into the ansatz (65), we get the approximate solution

$$g(s) = A(p_s^2K)^{-1/4}\exp\left(i\int_{s_0}^{s}\sqrt{\frac{\Delta}{K(x)}}\mathrm{d}x\right) \tag{71}$$

This shows, that the solutions with large Δ-values show oscillations with local frequency $\sqrt{\Delta/K(s)}$ and amplitude $\sim (p_s^2K)^{-1/4}$. As large values for K indicate that the rat moves quickly, this implies that the local frequency of the solutions is smaller in regions with larger velocities whereas small velocities, e.g. close to walls, lead to higher frequencies than expected for homogeneous movement. Intuitively this means that the functions compensate for quick movements with smaller spatial frequencies such that the effective temporal frequency of the output signal is kept constant.

Understanding the dependence of the amplitude on p_s and K is more subtle. Under the assumption that K is independent of s, the amplitude decreases where p_s is large and increases where p_s is small. Intuitively, this can be interpreted as an equalization of the fraction of the total variance that falls into a small interval of length $\Delta s \gg \sqrt{K/\Delta}$. This fraction is roughly given by the product of the probability $\mathrm{p}(s)\Delta s$ of being in this section times the squared amplitude $K(s)^{-1/2}/p(s)$ of the oscillation. For constant K, this fraction is also constant, so the amplitude is effectively rescaled to yield the same 'local variance' everywhere. If p is constant, on the other hand, the amplitude of the oscillation is small in places where the rat moves quickly and large where the rat moves slowly. This corresponds to the intuition that from the perspective of slowness there are two ways of treating places where the rat moves quickly: Decreasing the spatial frequency to generate slower output signals and/or decreasing the amplitude to 'pay less attention' to these regions. There is also a strong formal argument why the amplitude should depend on $p_s^2 K$. As the optimization problem is invariant under arbitrary invertible nonlinear coordinate changes, the amplitude of the oscillation should depend only on a function of p_s and K that is independent of the coordinate system. This constrains the amplitude to depend on $p_s^2 K$, as this is the only combination that is invariant under coordinate changes.

The key insight of this analysis is that the optimal functions show oscillations that are spatially compressed in regions where the rat moves with low velocities. This implies that the spatial resolution of the SFA solutions is higher in those regions. Consequently, the size of the place fields after sparse coding should be smaller in regions with small velocities.

References

Berkes, P. and Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, 5(6):579–602. `http://journalofvision.org/5/6/9/`, doi:10.1167/5.6.9.

Berkes, P. and Zito, T. (2005). Modular toolkit for data processing (version 2.0). `http://mdp-toolkit.sourceforge.net`.

Blaschke, T. and Wiskott, L. (2004). CuBICA: Independent component analysis by simultaneous third- and fourth-order cumulant diagonalization. *IEEE Transactions on Signal Processing*, 52(5):1250–1256.

Brunel, N. and Trullier, O. (1998). Plasticity of directional place fields in a model of rodent CA3. *Hippocampus*, 8:651–665.

de Araujo, I. E. T., Rolls, E. T., and Stringer, S. M. (2001). A view model which accounts for the spatial fields of hippocampal primate spatial view cells and rat place cells. *Hippocampus*, 11:699–706.

Földiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Franzius, M., Vollgraf, R., and Wiskott, L. (2007). From grids to places. *Journal of Computational Neuroscience*, in press.

Fuhs, M. C., Redish, A. D., and Touretzky, D. S. (1998). A visually driven hippocampal place cell model. *Proceedings of the Sixth Annual Conference on Computational Neuroscience : Trends in Research, 1998*, pages 101–106.

Gavrilov, V. V., Wiener, S. I., and Berthoz, A. (1998). Discharge correlates of hippocampal complex spike neurons in behaving rats passively displaced on a mobile robot. *Hippocampus*, 8(5):475–490.

Hafting, T., Fyhn, M., Molden, S., Moser, M., and Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806. doi:10.1038/nature03721.

Hashimoto, W. (2003). Quadratic forms in natural images. *Network: Computation in Neural Systems*, 14(4):765–788.

Hughes, A. (1978). A schematic eye for the rat. *Vision Research*, 19:569–588.

Jeffery, K. J. and O'Keefe, J. M. (1999). Learned interaction of visual and idiothetic cues in the control of place field orientation. *Experimental Brain Research*, 127:151–161.

Kayser, C., Einhäuser, W., Dümmer, O., König, P., and Körding, K. (2001). Extracting slow subspaces from natural videos leads to complex cells. *Artificial Neural Networks - ICANN 2001 Proceedings*, pages 1075–1080.

Knierim, J. J., Kudrimoti, H. S., and McNaughton, B. L. (1995). Place cells, head direction cells, and the learning of landmark stability. *Journal of Neuroscience*, 15(3 Pt 1):1648–1659.

Markus, E. J., Qin, Y. L., Leonard, B., Skaggs, W. E., McNaughton, B. L., and Barnes, C. A. (1995). Interactions between location and task affect the spatial and directional firing of hippocampal neurons. *Journal of Neuroscience*, 15(11):7079–7094.

McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., and Moser, M. B. (2006). Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience*, 7:663–678.

Mitchison, G. (1991). Removing time variation with the anti-Hebbian differential synapse. *Neural Computation*, 3:312–320.

Muller, R. U., Bostock, E., Taube, J. S., and Kubie, J. L. (1994). On the directional firing properties of hippocampal place cells. *Journal of Neuroscience*, 14(12):7235–7251.

Muller, R. U. and Kubie, J. L. (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *Journal of Neuroscience*, 7(7):1951–1968.

Oja, E. and Karhunen, J. (1995). Signal separation by nonlinear Hebbian learning. *Computational Intelligence: A Dynamic System Perspective*, pages 83–97.

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely moving rat. *Brain Research*, 34:171–5.

Picard, R., Graczyk, C., Mann, S., Wachman, J., Picard, L., and Campbell, L. (2002). Vision texture. Downloaded from `http://vismod.media.mit.edu/vismod/imagery/VisionTexture/vistex.html`.

Redish, A. D. (1999). *Beyond the Cognitive Map - From Place Cells to Episodic Memory*. MIT Press.

Rolls, E. T. (1999). Spatial view cells and the representation of place in the primate hippocampus. *Hippocampus*, 9:467–480.

Rolls, E. T. (2006). Neurophysiological and computational analyses of the primate presubiculum, subiculum and related areas. *Behavioral Brain Research*, 174:289–303.

Rolls, E. T., Xiang, J., and Franco, L. (2005). Object, space, and object-space representations in the primate hippocampus. *Journal of Neurophysiology*, 94:833–844.

Sargolini, F., Fyhn, M., Hafting, T., McNaughton, B. L., Witter, M. P., Moser, M., and Moser, E. I. (2006). Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, 312(5774):758–762.

Save, E., Nerad, L., and Poucet, B. (2000). Contribution of multiple sensory information to place field stability in hippocampal place cells. *Hippocampus*, 10(1):64–76.

Sharp, E. P. (1991). Computer simulation of hippocampal place cells. *Psychobiology*, 19(2):103–115.

Sharp, E. P., Blair, H. T., and Cho, J. (2001). The anatomical and computational basis of the rat head-direction cell signal. *Trends in Neurosciences*, 24(5):289–294.

Skaggs, W. E., Knierim, J. J., Kudrimoti, H. S., and McNaughton, B. L. (1995). A model of the neural basis of the rat's sense of direction. *Advances in neural information processing systems*, 7:173–80.

Song, E. Y., Kim, Y. B., Kim, Y. H., and Jung, M. W. (2005). Role of active movement in place-specific firing of hippocampal neurons. *Hippocampus*, 15:8–17.

Sprekeler, H., Michaelis, C., and Wiskott, L. (2006). Slowness: An objective for spike-timing dependent plasticity? In *Proc. 2nd Bernstein Symposium for Computational Neuroscience 2006, Berlin, October 1–3*, page 24. Bernstein Center for Computational Neuroscience (BCCN) Berlin.

Sprekeler, H., Michaelis, C., and Wiskott, L. (2007). Slowness: An objective for spike-timing-plasticity? *PLoS Computational Biology*. accepted.

Stackman, R. W. and Zugaro, M. B. (2005). Self-motion cues and resolving intermodality conflicts: Head direction cells, place cells, and behavior. In Wiener, S. I. and Taube, S., editors, *Head direction cells and the neural mechanisms of spatial orientation*, chapter 7, pages 137–162. MIT Press.

Stone, J. V. and Bray, A. (1995). A learning rule for extracting spatio-temporal invariances. *Network: Computation in Neural Systems*, 6:429–436.

Taube, J. S. and Bassett, J. P. (2003). Persistent neural activity in head direction cells. *Cerebral Cortex*, 13:1162–1172.

Taube, J. S., Muller, R. U., and Ranck, J. B. J. (1990). Head direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *Journal of Neuroscience*, 2(10):420–435.

Wiskott, L. (1998). Learning invariance manifolds. In Niklasson, L., Bodén, M., and Ziemke, T., editors, *Proc. 8th Intl. Conf. on Artificial Neural Networks, ICANN'98, Skövde*, Perspectives in Neural Computing, pages 555–560, London. Springer.

Wiskott, L. (2003). Slow feature analysis: A theoretical analysis of optimal free responses. *Neural Computation*, 15(9):2147–2177.

Wiskott, L. and Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770.

Wyss, R., König, P., and Verschure, P. (2006). A model of the ventral visual system based on temporal stability and local memory. *PLoS Biology*, 4(5):e120.