Supplementary Online Material (SOM)

for

Learning place cells, grid cells and invariances: A unifying model

Simon N. Weber[1], Henning Sprekeler[1*]

[1] Technische Universität Berlin, Germany

* Correspondence: h.sprekeler@tu-berlin.de

Please note: All supplementary figures are labeled with an S (e.g., Fig. **S1**). All figures from the main manuscript are labeled without an S.

Contents

1 Mathematical analysis of the learning rules

In the following, we derive the spacing of periodic firing patterns as a function of the simulation parameters for the linear track. To this end, we first show that homogeneous weights, chosen such that the output neuron fires at the target rate, are a fixed point for the time evolution of excitatory and inhibitory weights under the assumption of slow learning. We then perturb this fixed point and derive a two dimensional linear dynamical system for the Fourier modes of the excitatory and inhibitory weight perturbations. The translational invariance of the input overlap leads to decoupling of spatial frequencies in Fourier space. For smoother spatial tuning of inhibitory input than excitatory input, the eigenvalue spectrum of the dynamical system has a unique maximum, which indicates the most unstable spatial frequency. This frequency accurately predicts the grid spacing. We first consider place cell-like input (Gaussians) and then non-localized input (Gaussians convolved with white noise).

See Section **4** for a glossary of the notation. Whenever we use P as a sub- or superscript instead of E or I, this implies that the equation holds for neurons of the excitatory and the inhibitory population, respectively.

1.1 Assumption of slow learning

The firing rate of the output neuron is the weighted sum of excitatory and inhibitory input rates:

$$r^{\text{out}} = \left[\mathbf{w}^{\text{E}}\mathbf{r}^{\text{E}} - \mathbf{w}^{\text{I}}\mathbf{r}^{\text{I}}\right]_{+} . \tag{1}$$

where $[\ldots]_+$ indicates that negative firing rates are set to zero.

Written as a differential equation, the excitatory learning rule with quadratic multiplicative normalization is given by:

$$\frac{\mathrm{d}\mathbf{w}^{\mathrm{E}}}{\mathrm{d}t} = \eta_{\mathrm{E}} \left(\mathbb{1} - \frac{\mathbf{w}^{\mathrm{E}} \mathbf{w}^{\mathrm{E}T}}{\left\| \mathbf{w}^{\mathrm{E}} \right\|^2} \right) \mathbf{r}^{\mathrm{E}} r^{\mathrm{out}} \,. \tag{2}$$

The projection operator $\frac{\mathbf{w}^{\mathrm{E}} \mathbf{w}^{\mathrm{E}T}}{\left\| \mathbf{w}^{\mathrm{E}} \right\|^2}$ ensures that the weights are constrained to remain on the hypersphere whose radius is defined by the initial value of the sum over the squares of all excitatory weights [1]. The inhibitory learning rule is given by:

$$\frac{\mathrm{d}\mathbf{w}^{\mathrm{I}}}{\mathrm{d}t} = \eta_{\mathrm{I}} \mathbf{r}^{\mathrm{I}} \left(r^{\mathrm{out}} - \rho_0 \right) \,. \tag{3}$$

We assume the rat to learn slowly, such that it forages through the environment before significant learning (i.e., weight change) occurs. Therefore we can coarsen the time scale and rewrite Eqs. (2) and (3) as

$$\frac{\mathrm{d}\mathbf{w}^{\mathrm{E}}}{\mathrm{d}t} = \eta_{\mathrm{E}} \left\langle \left(\mathbb{1} - \frac{\mathbf{w}^{\mathrm{E}} \mathbf{w}^{\mathrm{E}T}}{\left\| \mathbf{w}^{\mathrm{E}} \right\|^2} \right) \mathbf{r}^{\mathrm{E}} r^{\mathrm{out}} \right\rangle_x \tag{4}$$

and

$$\frac{\mathrm{d}\mathbf{w}^{\mathrm{I}}}{\mathrm{d}t} = \eta_{\mathrm{I}} \left\langle \mathbf{r}^{\mathrm{I}} \left(r^{\mathrm{out}} - \rho_0 \right) \right\rangle_x \tag{5}$$

respectively, where the spatial average, $\langle \ldots \rangle_x$, is defined as

$$\langle (\ldots) \rangle_x = \frac{1}{L} \int_{-L/2}^{+L/2} (\ldots) \, \mathrm{d}x \tag{6}$$

and L is the length of the linear track.

1.2 High density assumption and continuum limit for place cell-like input

We assume a high density of input neurons and formulate the system in continuous variables. More precisely, we assume the distance between two neighboring firing fields to be much smaller than the width of the firing fields, i.e., $L/N_\mathrm{P} \ll \sigma_\mathrm{P}$. Furthermore, we assume that the linear track is very long compared to the width of the firing fields, i.e., $\sigma_\mathrm{P} \ll L$.

We replace the neuron index with the continuous variable μ and denote the weight w_μ^P and the tuning function $r^\mathrm{P}(\mu, x)$ associated with a place field that is centered at μ in the continuum limit as:

$$w_i^\mathrm{P} \to w^\mathrm{P}(\mu) \quad \text{and} \quad r_i^\mathrm{P}(x) \to r^\mathrm{P}(\mu, x)\,. \tag{7}$$

The distance between two neighboring place fields is given by $\Delta\mu = N_\mathrm{P}/L$. For sums over all neurons we thus get the following integral in the continuum limit:

$$\sum_{i=1}^{N_\mathrm{P}} f_i = \frac{1}{\Delta\mu} \sum_{i=1}^{N_\mathrm{P}} f_i\, \Delta\mu \to \frac{N_\mathrm{P}}{L} \int_{-L/2}^{+L/2} f(\mu)\, \mathrm{d}\mu\,. \tag{8}$$

In the following we will switch between the discrete and the continuous formulation and use whatever is more convenient.

For place cell-like input we take Gaussian tuning curves:

$$r_i^\mathrm{P}(x) = \alpha_\mathrm{P} \exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_\mathrm{P}^2}\right\}, \tag{9}$$

with height α_{P} and standard deviation σ_{P}. In the continuum limit we thus get:

$$r_i^{\mathrm{P}}(x) \to r^{\mathrm{P}}(\mu, x) = r^{\mathrm{P}}(|x-\mu|) = \alpha_{\mathrm{P}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma_{\mathrm{P}}^2}\right\}. \tag{10}$$

Because of the translational invariance of $r^{\mathrm{P}}(\mu, x)$, integration over space gives the same result as integration over all center locations and the mean of all inputs is the same:

$$\begin{aligned}
\left\langle r_i^{\mathrm{P}}(x)\right\rangle_x &= \left\langle r^{\mathrm{P}}(\mu, x)\right\rangle_x && (11)\\
&= \frac{1}{L}\int_{-L/2}^{+L/2} r^{\mathrm{P}}(\mu, x)\,\mathrm{d}x && (12)\\
&= \frac{1}{L}\int_{-L/2}^{+L/2} r^{\mathrm{P}}(\mu, x)\,\mathrm{d}\mu \approx \frac{\alpha_{\mathrm{P}}}{L}\sqrt{2\pi\sigma_{\mathrm{P}}^2} = M_{\mathrm{P}}/L && (13)
\end{aligned}$$

where we introduced $M_{\mathrm{P}} := \alpha_{\mathrm{P}}\sqrt{2\pi\sigma_{\mathrm{P}}^2}$ for the area under the tuning curves. Accordingly, we get a summarized input activity that is independent of location:

$$\sum_{i=1}^{N_{\mathrm{P}}} r_i^{\mathrm{P}}(x) = \frac{N_{\mathrm{P}}}{L}\int_{-L/2}^{+L/2} r^{\mathrm{P}}(\mu, x)\,\mathrm{d}\mu \approx \frac{N_{\mathrm{P}}}{L} M_{\mathrm{P}}\,. \tag{14}$$

1.3 Equal weights form a fixed point

In the following, we will show that equal weights $w^{\mathrm{E}}(\mu) = w_0^{\mathrm{E}}$ and $w^{\mathrm{I}}(\mu') = w_0^{\mathrm{I}}$, $\forall \mu, \mu'$ form a fixed point if w_0^{I} is chosen such that the output neuron fires at the target rate, ρ_0, throughout the arena.

With equal weights we get a constant firing rate r_0^{out},

$$r^{\mathrm{out}}(x) = r_0^{\mathrm{out}} = \left[w_0^{\mathrm{E}}\sum_i r_i^{\mathrm{E}}(x) - w_0^{\mathrm{I}}\sum_i r_i^{\mathrm{I}}(x)\right]_+, \tag{15}$$

which according to Eq. (14) does not depend on x. Furthermore, according to Eq. (11), $\left\langle r_i^{\mathrm{P}}(x)\right\rangle_x$ does not depend on the neuron index i. Now the stationarity of the excitatory weight evolution follows from Eq. (4):

$$\frac{\mathrm{d}w_i^{\mathrm{E}}}{\mathrm{d}t} = \eta_{\mathrm{E}} \left\langle r^{\mathrm{out}} \sum_j r_j^{\mathrm{E}} \left(\delta_{ij} - \frac{w_i^{\mathrm{E}} w_j^{\mathrm{E}}}{\sum_k w_k^{\mathrm{E}2}} \right) \right\rangle_x \tag{16}$$

$$= \eta_{\mathrm{E}} r_0^{\mathrm{out}} \sum_j \left[\left\langle r_j^{\mathrm{E}} \right\rangle_x \left(\delta_{ij} - \frac{w_0^{\mathrm{E}2}}{N_{\mathrm{E}} w_0^{\mathrm{E}2}} \right) \right] \tag{17}$$

$$= \frac{r_0^{\mathrm{out}} \eta_{\mathrm{E}} M_{\mathrm{E}}}{L} \sum_{j=1}^{N_{\mathrm{E}}} \left(\delta_{ij} - \frac{1}{N_{\mathrm{E}}} \right) = 0\,, \tag{18}$$

i.e., excitatory weights are stationary for all values of w_0^{E} and w_0^{I} (here δ_{ij} denotes the Kronecker delta which is 1 if $i = j$ and 0 otherwise). This holds for all input functions for which $\left\langle r_j^{\mathrm{E}}(x)\right\rangle_x$ is independent of j. If $r^{\mathrm{out}} = \rho_0$ it immediately follows from Eq. (3) that $\frac{\mathrm{d}w^{\mathrm{I}}}{\mathrm{d}t} = 0$, so the inhibitory weights are stationary if

$$\rho_0 = \mathbf{w}^{\mathrm{E}}\mathbf{r}^{\mathrm{E}} - \mathbf{w}^{\mathrm{I}}\mathbf{r}^{\mathrm{I}} = w_0^{\mathrm{E}} \sum_i r_i^{\mathrm{E}} - w_0^{\mathrm{I}} \sum_i r_i^{\mathrm{I}}\,, \tag{19}$$

which is fulfilled if

$$w_0^{\mathrm{I}} = \frac{w_0^{\mathrm{E}} \sum_i r_i^{\mathrm{E}} - \rho_0}{\sum_i r_i^{\mathrm{I}}} = \frac{w_0^{\mathrm{E}} N_{\mathrm{E}} M_{\mathrm{E}} - \rho_0}{N_{\mathrm{I}} M_{\mathrm{I}}}\,. \tag{20}$$

1.4 Linear stability analysis

In the following, we will show that the fixed point of equal weights, the *homogeneous steady state*, is unstable, when the spatial tuning of inhibitory inputs is broader than that of the excitatory inputs. In this case, perturbations of the fixed point will grow and a particular spatial frequency will grow fastest. We will show that this spatial

frequency predicts the spacing of the resulting periodic pattern (Fig. **1g**).

We disturb the fixed point

$$w^{\mathrm{E}}(\mu) = w_0^{\mathrm{E}} + \delta w^{\mathrm{E}}(\mu), \quad w^{\mathrm{I}}(\mu) = w_0^{\mathrm{I}} + \delta w^{\mathrm{I}}(\mu) \tag{21}$$

and look at the time evolution of the perturbations $\frac{\mathrm{d}\delta w^{\mathrm{E}}}{\mathrm{d}t}$ and $\frac{\mathrm{d}\delta w^{\mathrm{I}}}{\mathrm{d}t}$ of the excitatory and inhibitory weights around the fixed point.

Close to the fixed point the output neuron fires around the target rate ρ_0. We thus ignore the rectification in Eq. (1), i.e., $r^{\mathrm{out}} = \rho_0 + \delta r^{\mathrm{out}}$, with $\delta r^{\mathrm{out}} = \sum_k \delta w_k^{\mathrm{E}} r_k^{\mathrm{E}} - \sum_{k'} \delta w_{k'}^{\mathrm{I}} r_{k'}^{\mathrm{I}}$.

1.4.1 Time evolution of perturbations of the inhibitory weights

We start with the time evolution of the inhibitory weight perturbations:

$$\begin{aligned}
\frac{\mathrm{d}\delta w_i^{\mathrm{I}}}{\mathrm{d}t} = \frac{\mathrm{d}w_i^{\mathrm{I}}}{\mathrm{d}t} &= \eta_{\mathrm{I}} \left\langle \left(r^{\mathrm{out}} - \rho_0\right) r_i^{\mathrm{I}} \right\rangle_x && (22)\\
&= \eta_{\mathrm{I}} \left\langle \left(\rho_0 + \delta r^{\mathrm{out}} - \rho_0\right) r_i^{\mathrm{I}} \right\rangle_x && (23)\\
&= \eta_{\mathrm{I}} \left\langle r_i^{\mathrm{I}} \delta r^{\mathrm{out}} \right\rangle_x && (24)\\
&= \eta_{\mathrm{I}} \left\langle r_i^{\mathrm{I}} \left(\sum_k \delta w_k^{\mathrm{E}} r_k^{\mathrm{E}} - \sum_{k'} \delta w_{k'}^{\mathrm{I}} r_{k'}^{\mathrm{I}} \right) \right\rangle_x && (25)\\
&= \eta_{\mathrm{I}} \left(\sum_{k=1}^{N_{\mathrm{E}}} \left\langle r_i^{\mathrm{I}} r_k^{\mathrm{E}} \right\rangle_x \delta w_k^{\mathrm{E}} - \sum_{k'=1}^{N_{\mathrm{I}}} \left\langle r_i^{\mathrm{I}} r_{k'}^{\mathrm{I}} \right\rangle_x \delta w_{k'}^{\mathrm{I}} \right) , && (26)
\end{aligned}$$

where we used that only the rates $\mathbf{r}^{\mathrm{P}}$ depend on x. Intuitively, the first term in Eq. (26) means that the rate of change of the inhibitory weight perturbation of the weight associated to one location depends on the excitatory perturbations of the weights associated to every other location, weighted with the overlap of the two

associated tuning functions (analogous for inhibitory weight perturbations). In the continuum limit, the sums are:

$$\eta^{\mathrm{P}} \sum_{k=1}^{N_{\mathrm{P}'}} \left\langle r_i^{\mathrm{P}} r_k^{\mathrm{P}'} \right\rangle_x \delta w_k^{\mathrm{P}'} \to \eta^{\mathrm{P}} \frac{N_{\mathrm{P}'}}{L} \int_{-L/2}^{+L/2} \left\langle r^{\mathrm{P}}(\mu) r^{\mathrm{P}'}(\mu') \right\rangle_x \delta w^{\mathrm{P}'}(\mu') \, \mathrm{d}\mu' \tag{27}$$

$$= \int_{-L/2}^{+L/2} K^{\mathrm{PP}'}(\mu, \mu') \delta w^{\mathrm{P}'}(\mu') \, \mathrm{d}\mu' \,, \tag{28}$$

where we introduced overlap kernels

$$K^{\mathrm{PP}'}(\mu, \mu') := \eta^{\mathrm{P}} \frac{N_{\mathrm{P}'}}{L} \left\langle r^{\mathrm{P}}(\mu) r^{\mathrm{P}'}(\mu') \right\rangle_x \quad \mathrm{P}, \mathrm{P}' \in \{\mathrm{E}, \mathrm{I}\} \ . \tag{29}$$

The overlap $\left\langle r^{\mathrm{X}}(\mu) r^{\mathrm{Y}}(\mu') \right\rangle_x$ only depends on the distance of the Gaussian fields, i.e.,

$$K^{\mathrm{XY}}(\mu, \mu') = K^{\mathrm{XY}}(\mu - \mu') \,. \tag{30}$$

Taking $L \to \infty$, the time evolution of the perturbations of the inhibitory weights can thus be written as convolutions:

$$\frac{\mathrm{d}\delta w^{\mathrm{I}}(\mu)}{\mathrm{d}t} = (K^{\mathrm{IE}} * \delta w^{\mathrm{E}})(\mu) - (K^{\mathrm{II}} * \delta w^{\mathrm{I}})(\mu) \,, \tag{31}$$

where $*$ denotes a convolution.

1.4.2 Time evolution of perturbations of the excitatory weights

To derive the time evolution of the excitatory weights, we first show that the weight normalization term in Eq. (4), expressed through the projection operator $P_{ij} = \frac{w_i w_j}{\sum_k w_k^2}$, leads to a term that balances homogeneous weight perturbations and a term that can be neglected in the continuum limit.

Let P be the projection operator responsible for the normalization of the excitatory weights by projecting a weight update onto a vector that is orthogonal to the hypersphere of constant $\sum_{i=1}^{N_\mathrm{E}}(w_i^\mathrm{E})^2$. We now determine the projection operator around the fixed point (We drop the index 'E' in the following, to improve readability):

$$P_{ij} = \frac{(w_0 + \delta w_i)(w_0 + \delta w_j)}{\sum_k (w_0 + \delta w_k)^2} \equiv P_{ij}(\mathbf{w} + \delta\mathbf{w})\,. \tag{32}$$

Using Taylor's theorem

$$P_{ij}(\mathbf{w} + \delta\mathbf{w}) = P_{ij}(\mathbf{w}) + \sum_{l=1}^{N} \delta w_l \frac{\mathrm{d}P_{ij}(\mathbf{w})}{\mathrm{d}w_l} + \mathcal{O}(\delta\mathbf{w}^2) \tag{33}$$

and $w_l = w_0 \;\forall\; l$ we get

$$P_{ij}(\mathbf{w}) = \frac{w_i w_j}{\sum_k w_k^2} = 1/N\,, \tag{34}$$

$$\frac{\mathrm{d}P_{ij}(\mathbf{w})}{\mathrm{d}w_l} = \frac{\delta_{il} w_j}{\sum_k w_k^2} + \frac{\delta_{jl} w_i}{\sum_k w_k^2} - \frac{w_i w_j 2 w_l}{(\sum_k w_k^2)^2} = \frac{\delta_{il}}{N w_0} + \frac{\delta_{jl}}{N w_0} - \frac{2}{N^2 w_0}\,. \tag{35}$$

In summary this gives:

$$P_{ij} = \underbrace{\frac{1}{N_\mathrm{E}}}_{\equiv P_0 \propto \mathcal{O}(1)} + \underbrace{\frac{1}{N_\mathrm{E} w_0^\mathrm{E}} \left(\delta w_i^\mathrm{E} + \delta w_j^\mathrm{E} - \frac{2\sum_{l=1}^{N_\mathrm{E}} \delta w_l^\mathrm{E}}{N_\mathrm{E}} \right)}_{\equiv \delta P_{ij} \propto \mathcal{O}(\delta\mathbf{w})} + \mathcal{O}(\delta\mathbf{w}^2)\,. \tag{36}$$

Using the perturbed projection operator Eq. (36) with Eq. (4) we obtain the time

evolution of the excitatory weight perturbation to linear order:

$$\frac{\mathrm{d}\delta w_i^{\mathrm{E}}}{\mathrm{d}t} = \frac{\mathrm{d}w_i^{\mathrm{E}}}{\mathrm{d}t} \tag{37}$$

$$= \eta_{\mathrm{E}} \left\langle r^{\mathrm{out}} \sum_j (\delta_{ij} - P_{ij}) r_j^{\mathrm{E}} \right\rangle_x \tag{38}$$

$$= \eta_{\mathrm{E}} \left\langle (\rho_0 + \delta r^{\mathrm{out}}) \sum_j (\delta_{ij} - P_0 - \delta P_{ij}) r_j^{\mathrm{E}} \right\rangle_x \tag{39}$$

$$= \eta_{\mathrm{E}} \underbrace{\left\langle \rho_0 \sum_j (\delta_{ij} - P_0) r_j^{\mathrm{E}} \right\rangle_x}_{=0, \text{cf. Eq. (16)}} + \left\langle \delta r^{\mathrm{out}} \sum_j (\delta_{ij} - P_0) r_j^{\mathrm{E}} \right\rangle_x - \left\langle \rho_0 \sum_j \delta P_{ij} r_j^{\mathrm{E}} \right\rangle_x + \mathcal{O}(\delta \mathbf{w}^2) \tag{40}$$

$$= \eta_{\mathrm{E}} \left(\underbrace{\left\langle r_i^{\mathrm{E}} \delta r^{\mathrm{out}} \right\rangle_x}_{(1)} - \underbrace{P_0 \left\langle \delta r^{\mathrm{out}} \sum_j r_j^{\mathrm{E}} \right\rangle_x}_{(2)} - \underbrace{\rho_0 \left\langle \sum_j \delta P_{ij} r_j^{\mathrm{E}} \right\rangle_x}_{(3)} \right) + \mathcal{O}(\delta \mathbf{w}^2) \tag{41}$$

Term (1) in Eq. (41) has a similar structure as in the inhibitory case (Eq. (24)) and will lead to analogous convolutions. In the continuum limit the second term is given

by

$$(2) = \frac{1}{N_\mathrm{E}} \left\langle \left(\sum_k r_k^\mathrm{E} \delta w_k^\mathrm{E} - \sum_{k'} r_{k'}^\mathrm{I} \delta w_{k'}^\mathrm{I} \right) \sum_j r_j^\mathrm{E} \right\rangle_x \tag{42}$$

$$= \frac{M_\mathrm{E}}{L} \left\langle \sum_k r_k^\mathrm{E} \delta w_k^\mathrm{E} - \sum_{k'} r_{k'}^\mathrm{I} \delta w_{k'}^\mathrm{I} \right\rangle_x \tag{43}$$

$$= \frac{M_\mathrm{E}}{L} \left(\sum_k \left\langle r_k^\mathrm{E} \right\rangle_x \delta w_k^\mathrm{E} - \sum_{k'} \left\langle r_{k'}^\mathrm{I} \right\rangle_x \delta w_{k'}^\mathrm{I} \right) \tag{44}$$

$$= \frac{M_\mathrm{E}}{L^2} \left(M_\mathrm{E} \sum_k \delta w_k^\mathrm{E} - M_\mathrm{I} \sum_{k'} \delta w_{k'}^\mathrm{I} \right) \tag{45}$$

$$\text{cont. limit} \rightarrow \frac{M_\mathrm{E}}{L^3} \left(N_\mathrm{E} M_\mathrm{E} \int_{-L/2}^{+L/2} \delta w^\mathrm{E}(\mu') \, \mathrm{d}\mu' - N_\mathrm{I} M_\mathrm{I} \int_{-L/2}^{+L/2} \delta w^\mathrm{I}(\mu'') \, \mathrm{d}\mu'' \right) \tag{46}$$

and the third term by

$$(3) = \frac{\rho_0}{N_\mathrm{E} w_0^\mathrm{E}} \left\langle \sum_j r_j^\mathrm{E} \left(\delta w_i^\mathrm{E} + \delta w_j^\mathrm{E} - \frac{2}{N_\mathrm{E}} \sum_l \delta w_l^\mathrm{E} \right) \right\rangle_x \tag{47}$$

$$= \frac{\rho_0}{N_\mathrm{E} w_0^\mathrm{E}} \sum_j \left\langle r_j^\mathrm{E} \right\rangle_x \left(\delta w_i^\mathrm{E} + \delta w_j^\mathrm{E} - \frac{2}{N_\mathrm{E}} \sum_l \delta w_l^\mathrm{E} \right) \tag{48}$$

$$= \frac{\rho_0 M_\mathrm{E}}{N_\mathrm{E} w_0^\mathrm{E} L} \sum_j \left(\delta w_i^\mathrm{E} + \delta w_j^\mathrm{E} - \frac{2}{N_\mathrm{E}} \sum_l \delta w_l^\mathrm{E} \right) \tag{49}$$

$$= \frac{\rho_0 M_\mathrm{E}}{w_0^\mathrm{E} L} \left(\delta w_i^\mathrm{E} + \frac{1}{N_\mathrm{E}} \sum_j \delta w_j^\mathrm{E} - \frac{2}{N_\mathrm{E}} \sum_l \delta w_l^\mathrm{E} \right) \tag{50}$$

$$= \frac{\rho_0 M_\mathrm{E}}{w_0^\mathrm{E} L} \left(\delta w_i^\mathrm{E} - \frac{1}{N_\mathrm{E}} \sum_j \delta w_j^\mathrm{E} \right) \tag{51}$$

$$\text{cont. limit} \rightarrow \frac{\rho_0 M_\mathrm{E}}{w_0^\mathrm{E} L} \left(\delta w^\mathrm{E}(\mu) - \frac{1}{L} \int_{-L/2}^{+L/2} \delta w^\mathrm{E}(\mu') \, \mathrm{d}\mu' \right) \tag{52}$$

$$= \frac{\rho_0 M_\mathrm{E}}{w_0^\mathrm{E} L} \int_{-L/2}^{+L/2} d\mu' \delta w^\mathrm{E}(\mu') \left[\delta(\mu - \mu') - \frac{1}{L} \right] , \tag{53}$$

where $\delta(\mu - \mu')$ denotes the Dirac delta function. Together this leads to the time

evolution of the excitatory weight perturbations:

$$\frac{\mathrm{d}\delta w^{\mathrm{E}}(\mu)}{\mathrm{d}t} = \int_{-L/2}^{+L/2} d\mu' \delta w^{\mathrm{E}}(\mu') \Bigg[K^{\mathrm{EE}}(\mu-\mu') - \frac{\eta_{\mathrm{E}}\rho_0 M_{\mathrm{E}}}{w_0^{\mathrm{E}} L}\delta(\mu-\mu') \tag{54}$$

$$+ \frac{\eta_{\mathrm{E}} M_{\mathrm{E}}}{L^2}\left(\frac{\rho_0}{w_0^{\mathrm{E}}} - \frac{N_{\mathrm{E}} M_{\mathrm{E}}}{L}\right)\Bigg] \tag{55}$$

$$- \int_{-L/2}^{+L/2} d\mu'' \delta w^{\mathrm{I}}(\mu'') \left[K^{\mathrm{EI}}(\mu-\mu'') - \frac{\eta_{\mathrm{E}} N_{\mathrm{I}} M_{\mathrm{E}} M_{\mathrm{I}}}{L^3}\right] . \tag{56}$$

We now take $L \to \infty$ and write everything as convolutions, also trivial ones:

$$\frac{\mathrm{d}\delta w^{\mathrm{E}}(\mu)}{\mathrm{d}t} = \left(\left[K^{\mathrm{EE}} - \frac{\eta_{\mathrm{E}}\rho_0 M_{\mathrm{E}}}{w_0^{\mathrm{E}} L}\delta + \frac{\eta_{\mathrm{E}} M_{\mathrm{E}}}{L^2}\left(\frac{\rho_0}{w_0^{\mathrm{E}}} - \frac{N_{\mathrm{E}} M_{\mathrm{E}}}{L}\right)\right] * \delta w^{\mathrm{E}} \right)(\mu)$$

$$- \left(\left[K^{\mathrm{EI}} - \frac{\eta_{\mathrm{E}} N_{\mathrm{I}} M_{\mathrm{E}} M_{\mathrm{I}}}{L^3}\right] * \delta w^{\mathrm{I}} \right)(\mu) . \tag{57}$$

1.4.3 Decoupling of spatial frequencies

The convolutions in Eqs. (26) and (57) show how the excitatory and inhibitory weight perturbations at one location influence the time evolution of weights at every other location. Transforming the system to frequency space leads to a drastic simplification: The time evolution of a perturbation of a particular spatial frequency only depends on the excitatory and inhibitory perturbation of the same spatial frequency, i.e., the Fourier components decouple. We define the Fourier transform $f(k) \equiv \mathcal{F}[f(\mu)]$ with wavevector k of a function $f(\mu)$ of locations μ as:

$$f(k) \equiv \int_{-\infty}^{+\infty} f(\mu) e^{-ik\mu}\, \mathrm{d}\mu \tag{58}$$

and note that

$$\int_{-\infty}^{+\infty} e^{-ik\mu}\,\mathrm{d}\mu = 2\pi\delta(k)\,. \tag{59}$$

Using the Convolution theorem and the linearity of the Fourier transform we get

$$\begin{aligned}\frac{\mathrm{d}\delta w^{\mathrm{E}}(k)}{\mathrm{d}t} = &\left[\frac{\eta_{\mathrm{E}}M_{\mathrm{E}}}{L^2}\left(\frac{\rho_0}{w_0^{\mathrm{E}}} - \frac{N_{\mathrm{E}}M_{\mathrm{E}}}{L}\right)\delta w^{\mathrm{E}}(k) + \frac{\eta_{\mathrm{E}}N_{\mathrm{I}}M_{\mathrm{E}}M_{\mathrm{I}}}{L^3}\delta w^{\mathrm{I}}(k)\right]2\pi\delta(k) \\ &- \frac{\eta_{\mathrm{E}}\rho_0 M_{\mathrm{E}}}{w_0^{\mathrm{E}}L}\delta w^{\mathrm{E}}(k) + \left[K^{\mathrm{EE}}(k)\delta w^{\mathrm{E}}(k) - K^{\mathrm{EI}}(k)\delta w^{\mathrm{I}}(k)\right]\end{aligned} \tag{60}$$

and

$$\frac{\mathrm{d}\delta w^{\mathrm{I}}(k)}{\mathrm{d}t} = K^{\mathrm{IE}}(k)\delta w^{\mathrm{E}}(k) - K^{\mathrm{II}}(k)\delta w^{\mathrm{I}}(k)\,. \tag{61}$$

The $\delta(k)$ term in Eq. (60) balances homogeneous perturbations in such a way that the output neuron would still fire at the target rate, if not for permutations at other frequencies. In the following, we drop this term, because we are not interested in spatially homogeneous perturbations. Moreover, the continuum limit is only valid for high densities: $N_{\mathrm{P}}/L \to \infty$. We can thus drop terms of lower order than N_{X}/L, which eliminates the $\frac{\eta_{\mathrm{E}}\rho_0 M_{\mathrm{E}}}{w_0^{\mathrm{E}}L}$ term. Writing the remaining terms of Eqs. (60) and (61) as a matrix leads to:

$$\begin{bmatrix}\dot{\delta w}^{\mathrm{E}} \\ \dot{\delta w}^{\mathrm{I}}\end{bmatrix}(k) = \begin{bmatrix}K^{\mathrm{EE}}(k) & -K^{\mathrm{EI}}(k) \\ K^{\mathrm{IE}}(k) & -K^{\mathrm{II}}(k)\end{bmatrix}\begin{bmatrix}\delta w^{\mathrm{E}} \\ \delta w^{\mathrm{I}}\end{bmatrix}(k)\,, \tag{62}$$

which contains no terms from the weight normalization anymore. The characteristic polynomial of the above matrix is:

$$\lambda^2 + \lambda \left(K^{\mathrm{II}} - K^{\mathrm{EE}}\right) + K^{\mathrm{EI}} K^{\mathrm{IE}} - K^{\mathrm{EE}} K^{\mathrm{II}} = 0 \tag{63}$$

The difference, $K^{\mathrm{EI}} K^{\mathrm{IE}} - K^{\mathrm{EE}} K^{\mathrm{II}}$, vanishes for Gaussian input, because:

$$\begin{aligned} K^{\mathrm{XY}}(\mu, \mu' = 0) &= \frac{\eta^{\mathrm{X}} N_{\mathrm{Y}}}{L} \left\langle r^{\mathrm{X}}(\mu) r^{\mathrm{Y}}(0) \right\rangle_x & (64) \\ &= \frac{\alpha_{\mathrm{X}} \alpha_{\mathrm{Y}} \eta^{\mathrm{X}} N_{\mathrm{Y}}}{L^2} \int_{-L/2}^{+L/2} \mathrm{d}x \exp\left\{ -\frac{(x-\mu)^2}{2\sigma_{\mathrm{X}}^2} - \frac{x^2}{2\sigma_{\mathrm{Y}}^2} \right\} & (65) \\ &\approx \frac{\alpha_{\mathrm{X}} \alpha_{\mathrm{Y}} \eta^{\mathrm{X}} N_{\mathrm{Y}}}{L^2} \sqrt{\frac{2\pi}{\frac{1}{\sigma_{\mathrm{X}}^2} + \frac{1}{\sigma_{\mathrm{Y}}^2}}} \exp\left\{ -\frac{\mu^2}{2(\sigma_{\mathrm{X}}^2 + \sigma_{\mathrm{Y}}^2)} \right\}, & (66) \end{aligned}$$

where we completed the square and used $\int_{-\infty}^{+\infty} e^{-ax^2} = \sqrt{\frac{\pi}{a}}$. Taking the Fourier transform and completing the square again gives

$$K^{\mathrm{XY}}(k) = \frac{\eta^{\mathrm{X}} N_{\mathrm{Y}} M_{\mathrm{X}} M_{\mathrm{Y}}}{L^2} \exp\left\{ -\frac{k^2}{2} (\sigma_{\mathrm{X}}^2 + \sigma_{\mathrm{Y}}^2) \right\}. \tag{67}$$

and thus $K^{\mathrm{EI}} K^{\mathrm{IE}} - K^{\mathrm{EE}} K^{\mathrm{II}} = 0$.

For $X = Y$ Eq. (67) simplifies to:

$$K^{\mathrm{XX}}(k) = \frac{\eta^{\mathrm{X}} N_{\mathrm{X}} M_{\mathrm{X}}^2}{L^2} \exp\left\{ -k^2 \sigma_{\mathrm{X}}^2 \right\}. \tag{68}$$

This leads to the eigenvalues:

$$\lambda_0(k) = 0 \tag{69}$$

$$\lambda_1(k) = K^{\mathrm{EE}}(k) - K^{\mathrm{II}}(k) \tag{70}$$

$$= \frac{1}{L^2}\left(\eta_{\mathrm{E}} M_{\mathrm{E}}^2 N_{\mathrm{E}} \exp\left\{-k^2\sigma_{\mathrm{E}}^2\right\} - \eta_{\mathrm{I}} M_{\mathrm{I}}^2 N_{\mathrm{I}} \exp\left\{-k^2\sigma_{\mathrm{I}}^2\right\}\right) , \tag{71}$$

which are shown in Fig. **S1**. Perturbations with spatial frequencies for which $\lambda_1(k)$ is positive will grow. Setting $\frac{\mathrm{d}\lambda_1(k)}{\mathrm{d}k} = 0$ gives the wavevector $k_{\max}$ of the Fourier

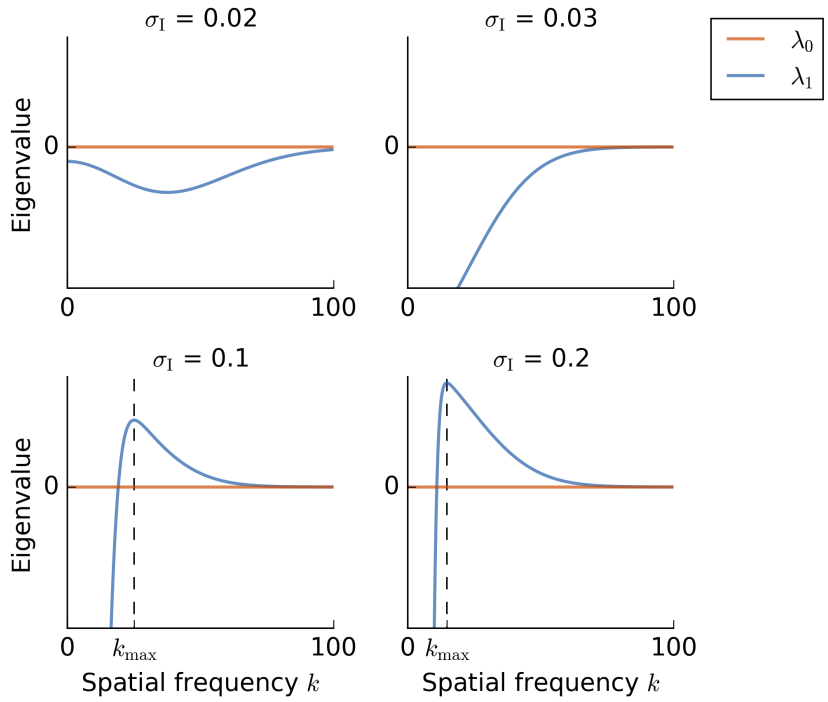

Figure S1: The eigenvalue spectrum for the eigenvalues of Eq. (69) for an excitatory tuning of width $\sigma_{\mathrm{E}} = 0.03$. The first eigenvalue λ_0 is always 0. If the inhibitory tuning is more narrow than the excitatory tuning, i.e., $\sigma_{\mathrm{I}} < \sigma_{\mathrm{E}}$, the second eigenvalue λ_1 is negative for every wavevector k. For $\sigma_{\mathrm{I}} > \sigma_{\mathrm{E}}$ the eigenvalue spectrum has a unique positive maximum $k_{\max}$, i.e., the most unstable spatial frequency. The wavevector $k_{\max}$ at which λ_1 is maximal is obtained from Eq. (75) and marked with a dashed line.

component that grows fastest:

$$\frac{2}{L^2}\left(\eta_{\mathrm{I}} M_{\mathrm{I}}^2 N_{\mathrm{I}} \sigma_{\mathrm{I}}^2 k_{\max} \exp\left\{-k_{\max}^2 \sigma_{\mathrm{I}}^2\right\} - \eta_{\mathrm{E}} M_{\mathrm{E}}^2 N_{\mathrm{E}} \sigma_{\mathrm{E}}^2 k_{\max} \exp\left\{-k_{\max}^2 \sigma_{\mathrm{E}}^2\right\}\right) = 0 \tag{72}$$

$$\Rightarrow \ln(\eta_{\mathrm{I}} M_{\mathrm{I}}^2 N_{\mathrm{I}} \sigma_{\mathrm{I}}^2) - k_{\max}^2 \sigma_{\mathrm{I}}^2 = \ln(\eta_{\mathrm{E}} M_{\mathrm{E}}^2 N_{\mathrm{E}} \sigma_{\mathrm{E}}^2) - k_{\max}^2 \sigma_{\mathrm{E}}^2 \tag{73}$$

$$\Rightarrow k_{\max} = \sqrt{\frac{\ln(\frac{\eta_{\mathrm{I}} M_{\mathrm{I}}^2 N_{\mathrm{I}} \sigma_{\mathrm{I}}^2}{\eta_{\mathrm{E}} M_{\mathrm{E}}^2 N_{\mathrm{E}} \sigma_{\mathrm{E}}^2})}{\sigma_{\mathrm{I}}^2 - \sigma_{\mathrm{E}}^2}}\,. \tag{74}$$

Assuming that the fastest-growing spatial frequency from the linearized system will prevail, the final spacing of the periodic pattern, ℓ, is determined by:

$$\ell = 2\pi/k_{\max} = 2\pi \sqrt{\frac{\sigma_{\mathrm{I}}^2 - \sigma_{\mathrm{E}}^2}{\ln\left(\frac{\eta_{\mathrm{I}} \sigma_{\mathrm{I}}^4 N_{\mathrm{I}} \alpha_{\mathrm{I}}^2}{\eta_{\mathrm{E}} \sigma_{\mathrm{E}}^4 N_{\mathrm{E}} \alpha_{\mathrm{E}}^2}\right)}}\,. \tag{75}$$

Eq. (75) is in exact agreement with the grid spacing obtained in simulations (Fig. **1g**). Moreover, it indicates the bifurcation point: When excitation is as smooth as inhibition ($\sigma_{\mathrm{E}} = \sigma_{\mathrm{I}}$), there is no unstable spatial frequency anymore and every perturbation gets balanced (Fig. **1g**, compare Eq. (98)). The grid spacing also depends on the ratio of the inhibitory and excitatory parameters $\eta^{\mathrm{P}}, N_{\mathrm{P}}, \alpha_{\mathrm{P}}$ (logarithmic term in Eq. (75)). We confirmed this dependence with simulations on the linear track where we increased either η_{I} or N_{I} or α_{I}^2 such that the product $\gamma = \eta_{\mathrm{I}} N_{\mathrm{I}} \alpha_{\mathrm{I}}^2$ increases with respect to the initial product γ_0. We find a good agreement with the theoretical prediction for all three variations (Fig. **S2**).

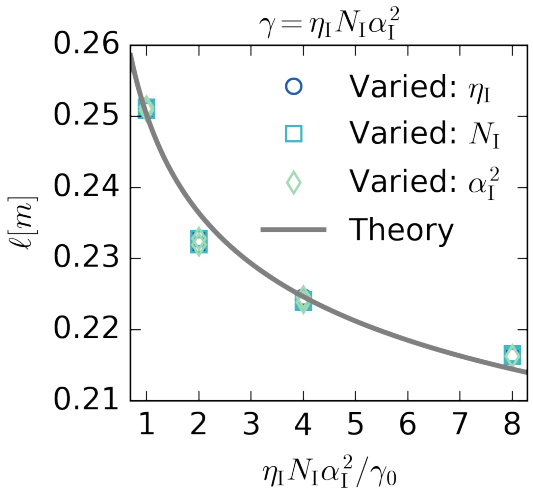

Figure S2: Dependence of grid spacing on learning rate η_I, number of input neurons N_I and input height α_I is accurately predicted by the theory. The gray line shows the grid spacing obtained from Eq. (75). We vary either the inhibitory learning rate, η_I (circles), the number of inhibitory input neurons, N_I (squares), or the square of the height of the inhibitory input place fields, α_I^2 (diamonds). The horizontal axis shows the ratio of the product $\eta_\mathrm{I} N_\mathrm{I} \alpha_\mathrm{I}^2$ to the initial value of the product γ_0. We keep $\eta_\mathrm{E} = 3.3 \times 10^{-5}$, $N_\mathrm{E} = 800$ and $\alpha_\mathrm{E} = 1$ in each simulation and the γ_0 parameters are: $\eta_\mathrm{I} = 3.3 \times 10^{-4}$, $N_\mathrm{I} = 200$, $\alpha_\mathrm{I} = 1$.

1.5 Analysis for non-localized input (Gaussian random fields)

Above, we derived the time evolution of perturbations of excitatory and inhibitory weights for place field-like input, i.e., Gaussian tuning curves. In the following we conduct a similar analysis, using non-localized input, i.e., random functions with a given spatial autocorrelation length. We show that the grid spacing is predicted by an equation that is equivalent to Eq. (75).

The non-localized input r_i^P for input neuron i of population P was obtained by rescaling a Gaussian random field (GRF) g_i^P to mean 1/2 and minimum 0:

$$r_i^\mathrm{P}(x) = \frac{g_i^\mathrm{P}(x) - \min_x g_i^\mathrm{P}(x)}{2 \left\langle g_i^\mathrm{P}(x) - \min_x g_i^\mathrm{P}(x) \right\rangle_x}, \tag{76}$$

where $\min_x$ denotes the minimum over all locations and the GRF g_i^{P} is obtained by convolving a Gaussian $\mathcal{G}^{\mathrm{P}}(x) = \exp(-x^2/2\sigma_{\mathrm{P}}^2)$ with white noise ξ_i from a uniform distribution between -0.5 an 0.5:

$$g_i^{\mathrm{P}}(x) = \int \mathcal{G}^{\mathrm{P}}(x - x')\, \xi_i^{\mathrm{P}}(x')\, \mathrm{d}x' \,. \tag{77}$$

94 Again we consider infinitely large systems $L \to \infty$ with infinite density $N_{\mathrm{P}}/L \to \infty$.

The mean of the distribution of GRF minima of different input neurons scales logarithmically with the number of samples [2]. Here the number of samples corresponds to the number of minima in a GRF which scales inversely with the width of the convolution kernel that was used to obtain the GRF:

$$\text{Number of minima} \propto L/\sigma_{\mathrm{P}} \,. \tag{78}$$

In the continuum limit the variance of the minima distribution over cells decreases and the relative difference between the mean minimum value of excitation and inhibition vanishes[1] (Fig. **S3**). We thus take the minimum value as a constant m, which does not depend on the population nor on the input neuron. This leads to the simplified expression of the input tuning functions:

$$r_i^{\mathrm{P}} = \frac{1}{2}\left(1 - \frac{g_i^{\mathrm{P}}(x)}{m}\right) . \tag{80}$$

[1]

$$\frac{\log(L/\sigma_{\mathrm{E}}) - \log(L/\sigma_{\mathrm{I}})}{\log(L/\sigma_{\mathrm{E}})} = \frac{\log(\sigma_{\mathrm{I}}/\sigma_{\mathrm{E}})}{\log(L/\sigma_{\mathrm{E}})} \to 0 \,. \tag{79}$$

For the argument it doesn't matter if it scales purely logarithmically or with $\log^\gamma$, where γ is any exponent.

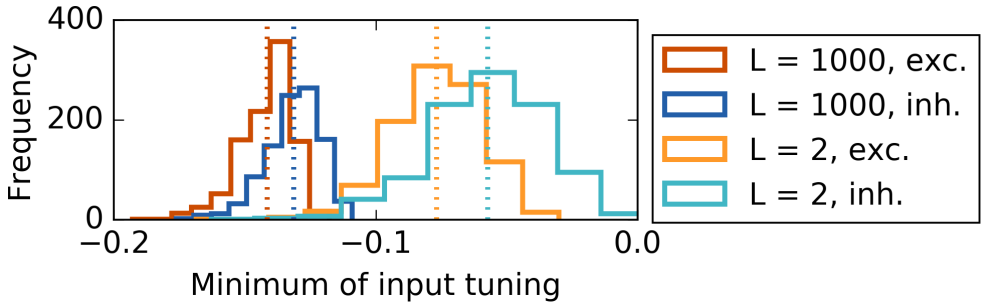

Figure S3: Distribution of minimal values of GRF input. Histograms show the distribution of the minimal values of 1000 input neurons for a small linear track, $L = 2$, and a large linear track $L = 1000$. Red and blue colors correspond to the tuning of excitatory and inhibitory input neurons, respectively. Each dotted line indicates the mean of the histogram of the same color. For larger systems, the distribution of the minimum values gets more narrow and the relative distance between the minima of excitatory and inhibitory neurons decreases.

Since $\left\langle r_i^{\mathrm{P}} \right\rangle_x = 0.5$ is independent of i, equal excitatory weights are a fixed point for the excitatory learning rule Eq. (4) as described in Eq. (16). Moreover, the sum over all input neurons does not depend on the location:

$$\sum_{i=1}^{N_{\mathrm{P}}} r_i^{\mathrm{P}}(x) = \frac{1}{2}\left(\sum_{i=1}^{N_{\mathrm{P}}} 1 - \sum_{i=1}^{N_{\mathrm{P}}} g_i^{\mathrm{P}}(x)\right) = \frac{N_{\mathrm{P}}}{2} - \frac{1}{2}\int \mathcal{G}^{\mathrm{P}}\left(x - x'\right) \underbrace{\sum_{i=1}^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(x')}_{=0 \text{ in cont. limit}} \mathrm{d}x' = \frac{N_{\mathrm{P}}}{2}\,. \tag{81}$$

Therefore, given constant excitatory weights, all inhibitory weights can be set to a value w_0^{I} such that the output neuron fires at the target rate, i.e., homogeneous weights are a fixed point of the learning rules, as in the scenario with Gaussian input. Moreover, Eq. (26) holds also for GRF input. The analysis of the projection operator of the weight normalization lead to a term of homogeneous weight perturbations and a term that could be neglected in the high density limit. We now omit these terms a priori. The time evolution of excitatory and inhibitory weight perturbations can

thus be summarized as (compare Eqs. (26) and (41)):

$$\frac{\mathrm{d}\delta w_i^{\mathrm{P}}}{\mathrm{d}t} = \eta^{\mathrm{P}} \left(\sum_{k=1}^{N_{\mathrm{E}}} \left\langle r_i^{\mathrm{P}}(x) r_k^{\mathrm{E}}(x) \right\rangle_x \delta w_k^{\mathrm{E}} - \sum_{k'=1}^{N_{\mathrm{I}}} \left\langle r_i^{\mathrm{P}}(x) r_{k'}^{\mathrm{I}}(x) \right\rangle_x \delta w_{k'}^{\mathrm{I}} \right) . \tag{82}$$

The above equation describes the time evolution of each synaptic weight. For the Gaussian input of the earlier sections, each synaptic weight is associated with one location. In the continuum limit we thus identified the synaptic weight associated to location μ with $w^{\mathrm{P}}(\mu)$. An increase of $w^{\mathrm{E}}(\mu)$ corresponded to an increase in firing at location μ (and the surrounding given by the Gaussian of the excitatory tuning). Analogously, an increase of $w^{\mathrm{I}}(\mu)$ caused a decrease in firing at location μ (and the surrounding given by the Gaussian of the inhibitory tuning). Because of the non-localized tuning of GRF input, each synaptic weight has an influence on the firing rate at many locations. The influence of neuron i of population P at location μ is expressed by $\xi_i^{\mathrm{P}}(\mu)$. If one wanted to increase the firing rate at location μ one would thus increase all excitatory weights that have a high $\xi_i^{\mathrm{P}}(\mu)$ and decrease all excitatory weights that have a low $\xi_i^{\mathrm{P}}(\mu)$ (note that ξ^{P} can also be negative). The 'weight' that corresponds to location μ is thus expressed as:

$$w^{\mathrm{P}}(\mu) := \sum_{i}^{N_{\mathrm{P}}} w_i^{\mathrm{P}} \xi_i^{\mathrm{P}}(\mu) \,, \tag{83}$$

where we weighted each synaptic weight with the value of the corresponding white noise at location μ. This corresponds to expressing the weights in a basis that is associated with the location and not with the individual input neurons. Combining Eq. (83) and Eq. (82) gives the time evolution of the weight perturbations associated

with location μ:

$$\frac{\mathrm{d}\delta w^{\mathrm{P}}(\mu)}{\mathrm{d}t} = \sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) \frac{\mathrm{d}\delta w_i^{\mathrm{P}}}{\mathrm{d}t} \tag{84}$$

$$= \eta^{\mathrm{P}} \sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) \left(\sum_{k=1}^{N_{\mathrm{E}}} \left\langle r_i^{\mathrm{P}}(x) r_k^{\mathrm{E}}(x) \right\rangle_x \delta w_k^{\mathrm{E}} - \sum_{k'=1}^{N_{\mathrm{I}}} \left\langle r_i^{\mathrm{P}}(x) r_{k'}^{\mathrm{I}}(x) \right\rangle_x \delta w_{k'}^{\mathrm{I}} \right) . \tag{85}$$

We now look at the first term of the above equation, the second term will be treated analogously:

$$\sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) \sum_{k=1}^{N_{\mathrm{E}}} \left\langle r_i^{\mathrm{P}}(x) r_k^{\mathrm{E}}(x) \right\rangle_x \delta w_k^{\mathrm{E}} = \left\langle \left(\sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) r_i^{\mathrm{P}}(x) \right) \left(\sum_{k=1}^{N_{\mathrm{E}}} \delta w_k^{\mathrm{E}} r_k^{\mathrm{E}}(x) \right) \right\rangle_x . \tag{86}$$

The sum containing the white noise can be simplified using the zero mean and the expression for the variance of the uniform white noise:

$$\sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) r_i^{\mathrm{P}}(x) = \frac{1}{2} \left(\underbrace{\sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu)}_{=0} - \frac{1}{m} \sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) g_i^{\mathrm{P}}(x) \right) \tag{87}$$

$$= -\frac{1}{2m} \sum_i^{N_{\mathrm{P}}} \int \mathcal{G}^{\mathrm{P}}(x - x') \underbrace{\sum_i^{N_{\mathrm{P}}} \xi_i^{\mathrm{P}}(\mu) \xi_i^{\mathrm{P}}(x')}_{=\beta N_{\mathrm{P}} \delta(x' - \mu) \text{ in cont. limit}} \mathrm{d}x' \tag{88}$$

$$= -\frac{\beta N_{\mathrm{P}}}{2m} \mathcal{G}^{\mathrm{P}}(x - \mu) , \tag{89}$$

where β is a proportionality constant that does not depend on the population type P. The Dirac delta $\delta(x' - \mu)$ occurs, because the white noise at different locations is uncorrelated. The sum of the product of weight perturbations and input rates can

be rewritten as:

$$\sum_{k=1}^{N_\mathrm{E}} \delta w_k^\mathrm{E} r_k^\mathrm{E}(x) = \frac{1}{2}\left(\underbrace{\sum_{k=1}^{N_\mathrm{E}} \delta w_k^\mathrm{E}}_{\text{homog. pert.}} - \frac{1}{m}\int \mathcal{G}^\mathrm{E}(x-\mu') \underbrace{\sum_{k=1}^{N_\mathrm{E}} \delta w_k^\mathrm{E} \xi_k^\mathrm{E}(\mu')}_{=:\delta w^\mathrm{E}(\mu');\ Eq.\ (83)} \mathrm{d}\mu' \right). \tag{90}$$

The first term is independent of location x and will thus only lead to spatially homogeneous perturbations which we do not consider in the following. Inserting Eqs. (89) and (90) and the analogous terms for inhibition in Eq. (86) leads to:

$$\sum_i^{N_\mathrm{P}} \xi_i^\mathrm{P}(\mu) \sum_{k=1}^{N_\mathrm{E}} \left\langle r_i^\mathrm{P}(x) r_k^\mathrm{E}(x) \right\rangle_x \delta w_k^\mathrm{E} = \frac{\beta N_\mathrm{P}}{4m^2} \int \left\langle \mathcal{G}^\mathrm{P}(x-\mu)\, \mathcal{G}^\mathrm{E}(x-\mu') \right\rangle_x \delta w^\mathrm{E}(\mu')\, \mathrm{d}\mu' \tag{91}$$

$$= \frac{1}{\eta^\mathrm{P}} \int \hat{K}^\mathrm{PE}(\mu-\mu') \delta w^\mathrm{E}(\mu')\, \mathrm{d}\mu' \tag{92}$$

$$= \frac{1}{\eta^\mathrm{P}} (\hat{K}^\mathrm{PE} * \delta w^\mathrm{E})(\mu)\,, \tag{93}$$

where we introduced kernels for the translation invariant overlap between two Gaussians with different centers (similar to Eq. (29)):

$$\hat{K}^{\mathrm{PP}'}(\mu-\mu') := \frac{\beta \eta^\mathrm{P} N_\mathrm{P}}{4m^2} \left\langle \mathcal{G}^\mathrm{P}(\mu)\, \mathcal{G}^{\mathrm{P}'}(\mu') \right\rangle_x = \frac{\beta \eta^\mathrm{P} N_\mathrm{P}}{4m^2} \left\langle \mathcal{G}^\mathrm{P}(0)\, \mathcal{G}^{\mathrm{P}'}(|\mu-\mu'|) \right\rangle_x \tag{94}$$

Eq. (84) can thus be written as:

$$\frac{\mathrm{d}\delta w^\mathrm{P}(\mu)}{\mathrm{d}t} = (\hat{K}^\mathrm{PE} * \delta w^\mathrm{E})(\mu) - (\hat{K}^\mathrm{PI} * \delta w^\mathrm{I})(\mu)\,, \tag{95}$$

which leads to a dynamical system for the Fourier components of the weight pertur-

bations that is equivalent to Eq. (62) with eigenvalues:

$$\lambda_0(k) = 0 \tag{96}$$

$$\lambda_1(k) = \hat{K}^{\mathrm{EE}}(k) - \hat{K}^{\mathrm{II}}(k) \tag{97}$$

$$= \frac{\beta}{4m^2}\left(\eta_{\mathrm{E}} M_{\mathrm{E}}^2 N_{\mathrm{E}} \exp\left\{-k^2\sigma_{\mathrm{E}}^2\right\} - \eta_{\mathrm{I}} M_{\mathrm{I}}^2 N_{\mathrm{I}} \exp\left\{-k^2\sigma_{\mathrm{I}}^2\right\}\right) . \tag{98}$$

We thus get the same expression for the grid spacing as in the scenario of Gaussian input (with $\alpha_{\mathrm{E}} = \alpha_{\mathrm{I}} = 1$):

$$\ell = \sqrt{\frac{\sigma_{\mathrm{I}}^2 - \sigma_{\mathrm{E}}^2}{\ln\left(\frac{\eta_{\mathrm{I}}\sigma_{\mathrm{I}}^4 N_{\mathrm{I}}}{\eta_{\mathrm{E}}\sigma_{\mathrm{E}}^4 N_{\mathrm{E}}}\right)}} . \tag{99}$$

2 Additional methods

2.1 Simulation parameters

	$[\sigma_{\mathrm{E},x}, \sigma_{\mathrm{E},y}, \sigma_{\mathrm{E},z}]$	N_{E}	η_{E}	$w^{\mathrm{E,init}}$	$N_{\mathrm{E}}^{\mathrm{f}}$
Fig. **1b**	0.05	2000	10^{-5}	1	∞
Fig. **1c**	0.08	2000	10^{-5}	1	∞
Fig. **1d**	0.06	2000	2×10^{-6}	1	∞
Fig. **1f**	0.04	160	1×10^{-3}	1	1
Fig. **1g**	0.03	1600	3.6×10^{-5}	1	1
Fig. **1h**	0.03	10000	3.5×10^{-7}	1	∞
Fig. **2a**	[0.05, 0.05]	4900	6.7×10^{-5}	1	1
Fig. **2b**	[0.05, 0.05]	4900	2×10^{-6}	1	100
Fig. **2c**	[0.05, 0.05]	4900	6×10^{-6}	1	∞
Fig. **3a-d**	[0.05, 0.05]	4900	2×10^{-4}	1	1
Fig. **4a**	[0.07, 0.07]	4900	6×10^{-6}	0.5	∞
Fig. **4b**	[0.05, 0.05]	4900	1.1×10^{-6}	0.0455	∞
Fig. **4c**	[0.08, 0.08]	4900	6×10^{-6}	0.5	∞
Fig. **4d**	[0.05, 0.05]	4900	6.7×10^{-5}	1	1
Fig. **5a**	[0.07, 0.07, 0.2]	37500	1.5×10^{-5}	1	1
Fig. **5b**	[0.08, 0.08, 0.2]	50000	10^{-5}	1	1
Fig. **5c**	[0.1, 0.1, 0.2]	50000	10^{-5}	1	1
Fig. **6a**	[0.05, 0.05]	4900	6.7×10^{-5}	1	1
Fig. **6b**	0.04	2000	5×10^{-5}	1	1
	0.04	2000	5×10^{-7}	1.0	100
	0.05	2000	5×10^{-6}	0.5	∞
Fig. **6c**	[0.05, 0.05]	4900	2×10^{-6}	1	100

Table 1: Parameters for excitatory inputs for all figures in the manuscript. $N_{\mathrm{E}} = \infty$ indicates that the excitatory input is a Gaussian random field.

	$[\sigma_{\mathrm{I},x}, \sigma_{\mathrm{I},y}, \sigma_{\mathrm{I},z}]$	N_{I}	η_{I}	$w^{\mathrm{I,init}}$	$N_{\mathrm{I}}^{\mathrm{f}}$
Fig. **1b**	0.12	500	10^{-4}	4.4	∞
Fig. **1c**	0.07	2000	1×10^{-4}	1.1	∞
Fig. **1d**	∞	500	2×10^{-5}	4.39	∞
Fig. **1f**	0.13	40	1×10^{-2}	1.31	1
Fig. **1g**	From 0.08 to 0.3 in 0.02 steps	400	3.6×10^{-4}	Eq. (106)	1
Fig. **1h**	From 0.08 to 0.3 in 0.02 steps	2500	7×10^{-6}	4.03	∞
Fig. **2a**	[0.1, 0.1]	1225	2.7×10^{-4}	1.5	1
Fig. **2b**	[0.1, 0.1]	1225	8×10^{-6}	1.52	100
Fig. **2c**	[0.1, 0.1]	1225	6×10^{-5}	4.0	∞
Fig. **3a-d**	[0.1, 0.1]	1225	8×10^{-4}	1.5	1
Fig. **4a**	$[\infty, \infty]$	1225	6×10^{-5}	2	∞
Fig. **4b**	[0.049, 0.049]	1225	4.4×10^{-5}	0.175	∞
Fig. **4c**	[0.3, 0.07]	1225	6×10^{-5}	2	∞
Fig. **4d**	[0.049, 0.049]	4900	2.7×10^{-4}	1.02	1
	[0.2, 0.1]; [0.1, 0.2]	1225	2.7×10^{-4}	1.04	1
	[2, 0.1]; [0.1, 2]	1225	2.7×10^{-4}	2.74	1
	[2, 0.2]; [0.2, 2]	1225	2.7×10^{-4}	1.38	1
	[0.1, 0.1]	1225	2.7×10^{-4}	1.5	1
	[0.2, 0.2]	1225	2.7×10^{-4}	0.709	1
	[2, 2]	1225	2.7×10^{-4}	0.259	1
	[0.1, 0.049]; [0.049, 0.1]	1225	2.7×10^{-4}	2.48	1
	[0.2, 0.049]; [0.049, 0.2]	1225	2.7×10^{-4}	1.74	1
	[2, 0.049]; [0.049, 2]	1225	2.7×10^{-4}	5.56	1
Fig. **5a**	[0.15, 0.15, 0.2]	9375	1.5×10^{-4}	1.55	1
Fig. **5b**	[0.12, 0.12, 1.5]	3125	1×10^{-4}	5.68	1
Fig. **5c**	[0.09, 0.09, 1.5]	12500	1×10^{-4}	2.71	1
Fig. **5d**	Same as	Fig. **5a,b,c**			
Fig. **6a**	[0.1, 0.1]	1225	2.7×10^{-4}	1.5	1
Fig. **6b**	0.12	500	5×10^{-4}	1.6	1
	0.12	500	5×10^{-6}	1.62	100
	0.12	500	5×10^{-5}	1.99	∞
Fig. **6c**	[0.1, 0.1]	1225	8×10^{-6}	1.52	100

Table 2: Parameters for inhibitory inputs for all figures in the manuscript. $N_{\mathrm{I}} = \infty$ indicates that the inhibitory input is a Gaussian random field. We denote spatially untuned inhibition with: $\sigma_{\mathrm{I}} = \infty$.

	t_{sim}	L
Fig. **1b**,**c**,**d**,**f**	4.0×10^5	2
Fig. **1g**	8.0×10^7	14
Fig. **1h**	4.0×10^7	10
Fig. **2a**,**b**,**c**	1.8×10^6	1
Fig. **3a,b,c,d**	5.4×10^5	1
Fig. **4a,b,c,d**	1.8×10^6	1
Fig. **5a,b,c,d**	1.8×10^6	1
Fig. **6a**,**c**	1.8×10^6	1
Fig. **6b**	4.0×10^5	2

Table 3: Simulation time t_{sim} and system size L for all figures in the manuscript.

2.2 Rat trajectory

In the linear track model (one dimension, Fig. **1**) we create artificial trajectories $x(t)$. The rat moves along a line of length L with constant velocity $v = 1\,\mathrm{cm}$ per unit time step $\Delta t = 1$. The rat always inverts its direction of motion when it hits either end of the enclosure at $-L/2$ or $L/2$. Additionally, in each unit time step it inverts its direction with a probability of $2v\Delta t/L$, resulting in a typical persistence length of $L/2$. Assuming that the rat moves with $3.0 \times 10^1\,\mathrm{cm/s}$, the 4×10^5 time steps shown in Fig. **1b** correspond to 3.6 hours.

In the open arena model (two dimensions, Figs. **2** to **4**), we take trajectories $\mathbf{x}(t)$ from behavioral data [3] of a rat that moved in a $1\,\mathrm{m} \times 1\,\mathrm{m}$ quadratic enclosure. The data provides coherent trajectories in intervals of 10 minutes. To get a 10 hours trajectory we concatenate 60 individual trajectories. Different trajectories in our simulations correspond to different random orders of concatenation. A 10 minute trajectory contains 3.0×10^4 locations. We update the location in every unit time step. A time step thus corresponds to $2.0 \times 10^1\,\mathrm{ms}$.

In the model for neurons with head direction tuning (three dimensions, Fig. **5**) we use the same behavioral trajectories as in two dimensions. We model the head direction of the animal as the direction of motion plus a random angle that is drawn in each unit time step from a normal distribution with standard deviation $\pi/6$.

In all dimensions, we find that the precise trajectory of the rat only has a small influence on the results for the learning rates under consideration (see also Fig. **S9**).

2.3 Spatially tuned inputs

The firing rates of excitatory and inhibitory synaptic inputs $r_i^{\mathrm{E}}, r_j^{\mathrm{I}}$ are tuned to the location $\mathbf{x}$ of the animal. In the following, we use x and y for the first and second spatial dimensions and z for the head direction. The value of x, y, z are in the range $[-L/2, L/2]$.

We analyzed three different kinds of input tuning functions. Place cells (single Gaussians), several place fields (sum of multiple Gaussians) and non-localized input (Gaussians convolved with white noise). We summarize the tuning functions of neurons from the excitatory and the inhibitory population by referring to them as population P where $\mathrm{P} \in \{\mathrm{E}, \mathrm{I}\}$.

For readability we define a Gaussian of height one with standard deviation σ_{P}:

$$\mathcal{G}^{\mathrm{P}}(x) := \exp\left\{-\frac{x^2}{2\sigma_{\mathrm{P}}^2}\right\} \tag{100}$$

The input function of the i-th neuron of population P with $N_{\mathrm{P}}^{\mathrm{f}}$ place fields per input neuron in one dimension is then given by:

$$r_i^{\mathrm{P}}(x) = \sum_{\beta=1}^{N_{\mathrm{P}}^{\mathrm{f}}} \mathcal{G}^{\mathrm{P}}\left(x - \mu_{i,\beta}^{\mathrm{P}}\right) , \tag{101}$$

where $\mu_{i,\beta}^{\mathrm{P}}$ denotes the center location of field number β of input neuron i of population P. The scenario of place cell-like inputs is obtained by setting $N_{\mathrm{E}}^{\mathrm{f}} = N_{\mathrm{I}}^{\mathrm{f}} = 1$.

For higher dimensions we define the center components as $\boldsymbol{\mu}_{i,\beta}^{\mathrm{P}} = (\mu_{i,\beta,x}^{\mathrm{P}}, \mu_{i,\beta,y}^{\mathrm{P}}, \mu_{i,\beta,z}^{\mathrm{P}})$. In two dimensions, the tuning of the i-th neuron of population P with $N_{\mathrm{P}}^{\mathrm{f}}$ place fields

per input neurons is thus given by:

$$r_i^{\mathrm{P}}(\mathbf{x}) = \sum_{\beta=1}^{N_{\mathrm{P}}^{\mathrm{f}}} \mathcal{G}^{\mathrm{P}}\left(x - \mu_{i,\beta,x}^{\mathrm{P}}\right) \mathcal{G}^{\mathrm{P}}\left(y - \mu_{i,\beta,y}^{\mathrm{P}}\right) . \tag{102}$$

In three dimensions we also consider bell-shaped tuning functions along the z-direction. However, since the head direction component is periodic we take von Mises functions that are periodic in the interval $[-L/2, L/2]$:

$$\mathcal{M}^{\mathrm{P}}(x) := \exp\left\{ \left(\frac{L}{2\pi\sigma_{\mathrm{P},z}}\right)^2 \left[\cos\left(\frac{2\pi z}{L}\right) - 1\right] \right\} \tag{103}$$

In three dimensions, the tuning of the i-th neuron of population P with $N_{\mathrm{P}}^{\mathrm{f}}$ place fields per input neurons is thus given by:

$$r_i^{\mathrm{P}}(\mathbf{x}) = \sum_{\beta=1}^{N_{\mathrm{P}}^{\mathrm{f}}} \mathcal{G}^{\mathrm{P}}\left(x - \mu_{i,\beta,x}^{\mathrm{P}}\right) \mathcal{G}^{\mathrm{P}}\left(y - \mu_{i,\beta,y}^{\mathrm{P}}\right) \mathcal{M}^{\mathrm{P}}\left(z - \mu_{i,\beta,z}^{\mathrm{P}}\right) . \tag{104}$$

The center locations $\boldsymbol{\mu}^{\mathrm{P}}$ for neurons of type P in an enclosure of side length L are drawn from a randomly distorted lattice (Fig. **S4**). First the total number of input neurons is factorized in its dimensional components $N_{\mathrm{P}} = N_{\mathrm{P},x} N_{\mathrm{P},y} N_{\mathrm{P},z}$. Then, for example along the x dimension, center locations of neurons of population P are placed equidistantly in $[-\frac{L}{2} - 3\sigma_{\mathrm{P},x}, \frac{L}{2} + 3\sigma_{\mathrm{P},x}]$. Allowing the field centers to lie a multiple of their standard deviation outside the box reduces boundary effects. Each point on the equidistant lattice is subsequently distorted with noise drawn from a uniform distribution in $[-\frac{L}{2(N_{\mathrm{P},x}-1)}, \frac{L}{2(N_{\mathrm{P},x}-1)}]$; see Fig. **S4**. Other dimensions are treated analogously. This procedure ensures a random but still efficiently dense coverage of the arena with few place fields. A truly random distribution of centers

leads to similar results (not shown) but requires more input neurons in order to cover the arena densely. We create $N_{\mathrm{P}}^{\mathrm{f}}$ of such distorted lattices. To each input neuron we assign one center location from each of the $N_{\mathrm{P}}^{\mathrm{f}}$ lattices at random and without replacement. This guarantees that each input neuron has $N_{\mathrm{P}}^{\mathrm{f}}$ randomly located fields that together cover the arena densely.

We obtain dense non-localized input by convolving Gaussians as in Eqs. (100) and (102) (with $N_{\mathrm{P}}^{\mathrm{f}} = 1$) with uniform white noise between -0.5 and 0.5. For the discretization we choose $\sigma_{\mathrm{P}}/20$ and centered the Gaussian convolution kernel on an array of 8 times its standard deviation. We convolve this array with a sufficiently large array of white noise such that we only keep the values where the array of the convolution kernel is inside the array of the white noise. This way we avoid boundary effects at the edges. From the resulting function we subtract its minimum and then divide by twice the mean of the difference between the function and its minimum. This increases the signal to noise ratio and ensures that all of the inputs have a mean value of 0.5 across the arena and a minimum at 0. For each input neuron we take a different realization of white noise. This results in arbitrary tuning functions of the same autocorrelation length as the Gaussian convolution kernel (we define the autocorrelation length as the distance at which the autocorrelation has decayed to $1/e$ of its maximum, where e is Euler's number). The above mentioned also holds for circular enclosures, only that we drop all field centers outside of a circle of radius $L/2 + 3\sigma_{\mathrm{P}}$ because they never get activated. This is not necessary but it reduces simulation time.

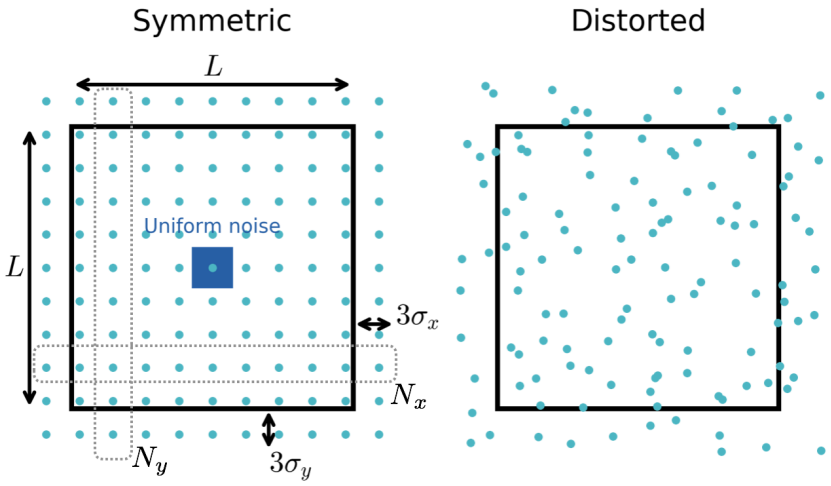

Figure S4: Distribution of input fields. Black square box: Arena in which the simulated rat can move. Blue circles: Locations of input firing fields. To create random place field locations that cover the space densely, we use locations from a distorted lattice. To this end we first create a symmetric lattice with N_x locations along the x-direction and N_y locations along the y-direction. These centers can lie a certain distance outside the boundary, to reduce boundary effects. We then add noise from a uniform distribution (blue square) to each location and obtain a distorted lattice (right).

2.4 Distribution of the initial synaptic weights

In order to start with reasonable firing rates, we take the initial weights close to the values that would correspond to the fixed point weights for Gaussian inputs. More precisely, initially all synaptic weights are chosen from a uniform distribution. For the spreading of the distribution we take $\pm 5\%$ of the mean value. For the mean value of the excitatory weights, w_0^{E}, we typically take $w_0^{\mathrm{E}} = 1$. We then determine the mean of the initial inhibitory weights, w_0^{I}, such that the output neuron fires on average around the target rate:

$$\mathbf{w}^{\mathrm{E}}\mathbf{r}^{\mathrm{E}} - \mathbf{w}^{\mathrm{I}}\mathbf{r}^{\mathrm{I}} = w_0^{\mathrm{E}} \sum_{i=1}^{N_{\mathrm{E}}} r_i^{\mathrm{E}} - w_0^{\mathrm{I}} \sum_{j=1}^{N_{\mathrm{I}}} r_j^{\mathrm{I}} \stackrel{!}{=} \rho_0 \,, \tag{105}$$

so

$$w_0^{\mathrm{I}} = \frac{w_0^{\mathrm{E}} \sum_{i=1}^{N_{\mathrm{E}}} r_i^{\mathrm{E}} - \rho_0}{\sum_{j=1}^{N_{\mathrm{I}}} r_j^{\mathrm{I}}} \,. \tag{106}$$

The sums are given by:

$$\sum_{i=1}^{N_{\mathrm{P}}} r_i^{\mathrm{P}} = \frac{N_{\mathrm{P}}}{A_{\mathrm{P}}} M_{\mathrm{P}} \,, \tag{107}$$

where N_{P} is the number of input neurons, M_{P} is the area under a tuning function and A_{P} the area in which the centers of the input tuning function can lie. For the fixed point weight relation Eq. (106) this leads to

$$w_0^{\mathrm{I}} = \frac{w_0^{\mathrm{E}} N_{\mathrm{E}} M_{\mathrm{E}} / A_{\mathrm{E}} - \rho_0}{N_{\mathrm{I}} M_{\mathrm{I}} / A_{\mathrm{I}}} \,. \tag{108}$$

The values for A_{P} and M_{P} depend on the dimensionality of the system.

2.4.1 One Dimension

For Gaussian input we have:

$$M_{\mathrm{P}} = \sqrt{2\pi} N_{\mathrm{P}}^{\mathrm{f}} \alpha_{\mathrm{P}} \sigma_{\mathrm{P}}, \quad A_{\mathrm{P}} = L + 6\sigma_{\mathrm{P}} \,. \tag{109}$$

For Gaussian random field input we have:

$$M_{\mathrm{P}} = \frac{A_{\mathrm{P}}}{2}, \quad A_{\mathrm{P}} = L \,. \tag{110}$$

2.4.2 Two Dimensions

For Gaussian input we have:

$$M_{\mathrm{P}} = \int\int r^{\mathrm{P}}(\mu_x, \mu_y)\,\mathrm{d}\mu_x\,\mathrm{d}\mu_y = 2\pi N_{\mathrm{P}}^{\mathrm{f}}\sigma_{\mathrm{P},x}\sigma_{\mathrm{P},y}, \quad A_{\mathrm{P}} = (L + 6\sigma_{\mathrm{P},x})(L + 6\sigma_{\mathrm{P},y})\,. \tag{111}$$

For Gaussian random field input we have:

$$M_{\mathrm{P}} = \frac{A_{\mathrm{P}}}{2}, \quad A_{\mathrm{P}} = L^2\,. \tag{112}$$

2.4.3 Three dimensions

In three dimensions we use a von Mises distribution along the third dimension to account for the periodicity of the head direction angle. We thus get

$$M_{\mathrm{P}} = \int\int\int r^{\mathrm{P}}(\mu_P, \mu_y, \mu_z)\,\mathrm{d}\mu_x\,\mathrm{d}\mu_y\,\mathrm{d}\mu_z \tag{113}$$

$$= N_{\mathrm{P}}^{\mathrm{f}} 2\pi\sigma_{\mathrm{P},x}\sigma_{\mathrm{P},y} L \frac{I_0\left[\left(\frac{L}{2\pi\sigma_{\mathrm{P},z}}\right)^2\right]}{\exp\left\{(\frac{L}{2\pi\sigma_{\mathrm{P},z}})^2\right\}} \tag{114}$$

where I_0 is the modified Bessel function. The area in which the function centers can lie is given by:

$$A_{\mathrm{P}} = (L + 6\sigma_{\mathrm{P},x})(L + 6\sigma_{\mathrm{P},y})L\,. \tag{115}$$

2.5 Measure for grid spacing on the linear track

We define the grid spacing of one dimensional grids as the location of the first non-centered peak in the autocorrelogram of the firing pattern (Fig. **1g**). For place cell-like input we obtain the grid spacing of from a single simulation.

For non-localized input the grids show defects, which results in misleading peaks in the correlogram. In this case, we used the first peak of the average of 50 correlograms to get the grid spacing (Fig. **1h**). The 50 correlograms were obtained from 50 realizations that differ only in the randomness of the input function. To avoid taking a fluctuation in the correlogram as the first peak – and thus get a misleading grid spacing – we take the maximum between $3\sigma_{\mathrm{E}}$ (to cut out the center of the correlogram) and 1 (a value larger than the largest grid spacing in Fig. **1h**).

For high values of the spatial smoothness of inhibition, σ_{I}, the simulation results deviate from the analytical solution. This is because for high σ_{I} but small σ_{E} the output neuron fires very sparsely, which impedes the learning. This can be readily overcome by increasing the tuning width, σ_{E}, of the excitatory input.

2.6 Measure for grid score

We use the grid score suggested in [3]. More precisely, we determine the grid score of a spatial autocorrelogram – the Pearson correlation coefficients for all spatial shifts of the firing rate maps against itself – in the following way: We crop a centered donut shape from the correlogram. To get the inner and the outer radius of the donut we clip all values in the correlogram with values smaller than 0.1 to 0. We obtain the resulting clusters of correlations that are larger than 0.1 using the `scipy.ndimage.measurements.label` function from the *SciPy* package for *Python*

with a quadratic filter structure, $((1,1,1),(1,1,1),(1,1,1))$, for a correlogram with 51×51 pixels. From the resulting clusters, we identify those seven with the most central center of mass. We use the distance of the outermost pixel of the inner most cluster from the center as the inner radius of the donut and the distance of the outermost pixel of the outermost cluster (of the seven most central clusters) from the center as the outer radius. We then rotate this donut around the center and correlate it with the unrotated donut. We determine the correlation for 30, 60, 90, 120 and 150 degrees. We define the grid score as the minimum of the correlation values at 60 and 120 degrees minus the maximum of the correlation values at 30, 90 and 150 degrees. A hexagonal symmetry thus leads to positive values whereas a quadratic symmetry leads to negative values.

2.7 Measure for head direction tuning

To quantify the head direction tuning of a cell we compare the head direction tuning to a uniform circular tuning, using Watson's U2 measure. We adopted the code from [4]. We created $10,000$ samples, `s_HD`, from a probability distribution created from the head direction tuning array and $10,000$ samples, `s_uniform`, from a uniform distribution and use `watson_u2(s_ uniform, s_HD)` from [4] to quantify the degree of non-circularity. The sharper the head direction tuning, the higher the resulting values.

3 Further results

3.1 Synaptic weight normalization does not influence the grids

In all simulations in the main text we used quadratic multiplicative normalization for the excitatory synaptic weights – a conventional normalization scheme. This choice was not crucial for the emergence of patterns. We observe very similar firing patterns without normalization or with different normalization schemes, such as linear multiplicative and linear subtractive normalization (Fig. **S5a,b**). Linear multiplicative normalization keeps the sum of all weights constant by multiplying each weight with a factor in each time step. Linear subtractive normalization keeps the sum of all weights roughly constant by adding or subtracting a factor from all weights and ensuring that negative weights are set to zero. The temporal evolution of individual synaptic weights looks very similar for all normalization schemes (Fig. **S5c**).

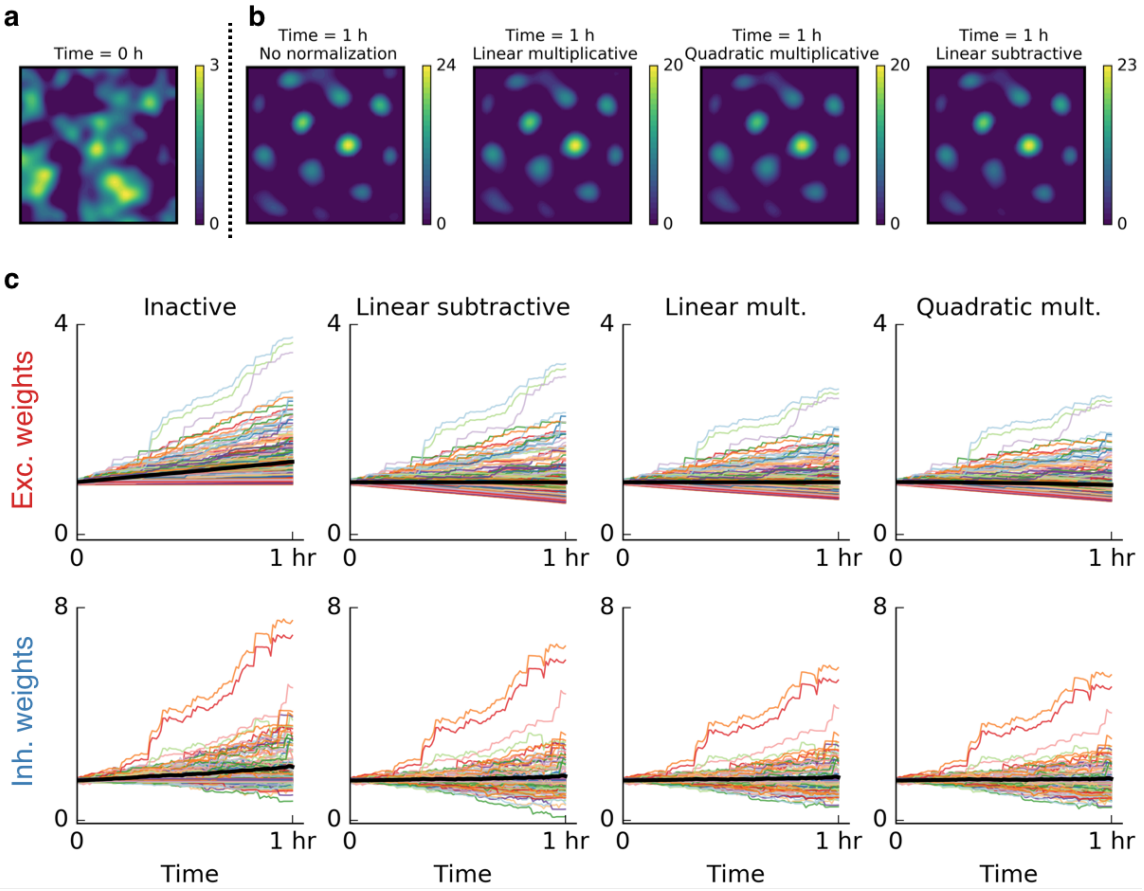

Figure S5: Weight normalization is not crucial for the emergence of grid cells. **a**): Firing rate map of a cell before it started exploring its surroundings. **b**): From left to right: firing rate of the output cell after one hour of spatial exploration for inactive, linear multiplicative, quadratic multiplicative and linear subtractive normalization. **c**): Time evolution of excitatory and inhibitory weights for different normalization schemes. Same simulations is in **a,b**. The colored lines show 200 individual weights. The black line shows the mean over all synaptic weights. From left to right: Inactive, linear multiplicative, quadratic multiplicative and linear subtractive normalization. Weights grow stronger without normalization, as is apparent from the growing mean.

3.2 Rapid emergence of grid cells with non-localized input

Grid cells emerge within minutes in experiments and in our simulation. For non-localized input the emergence of the final grid pattern typically takes longer. However, the grid fields that emerge early are still present in the final grid (Fig. **S6**), as has been observed in experiments [5].

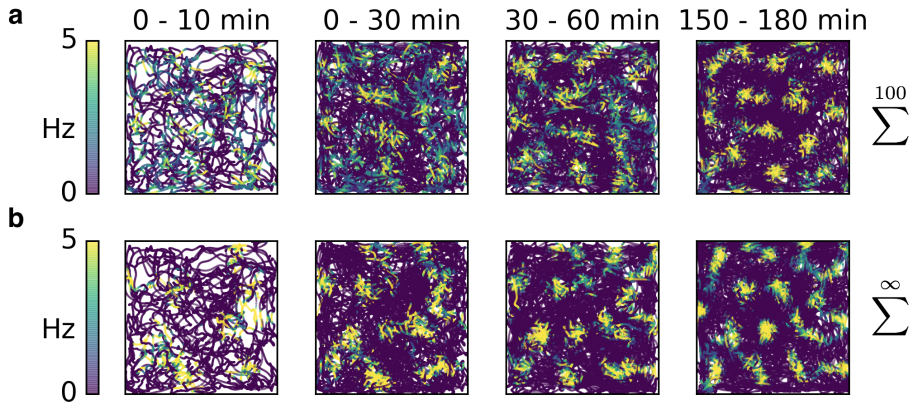

Figure S6: Development of a grid pattern from initially random weights for two different input scenarios. **a**) Sparse non-localized input (sum of 100 randomly located place fields) as in Fig. **2b**. **b**) Dense non-localized input (random function with fixed spatial smoothness) as in Fig. **2c**. While the emergence of the final patterns takes roughly an hour – and thus longer than for place cell-like inputs (Fig. **3**) – the early firing fields are still present in the final grid.

3.3 Conjunctive grid and head direction cells from non-localized input

Simulation results for combined spatial and head direction input that is non-localized and sparse is shown in Fig. **S7** and can be compared to Fig. **5** in the main text.

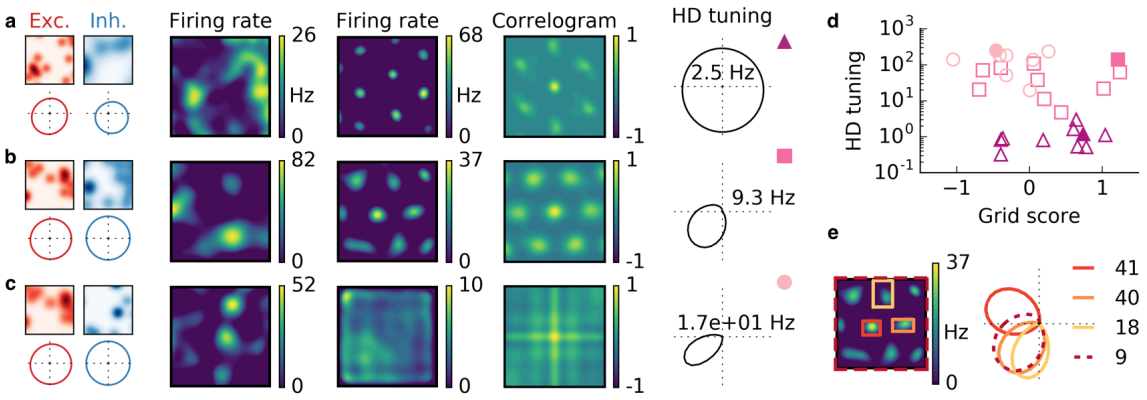

Figure S7: Cells with combined spatial and head direction tuning with input tuning that is given by the sum of 20 randomly placed Gaussian spheres; compare Fig. **5** in the main text. **a,b,c**) Columns from left to right: Spatial tuning and head direction tuning (polar plot) of excitatory and inhibitory input neurons (one example each); spatial firing rate map of the output neuron before learning and after spatial exploration of 10 hours with corresponding autocorrelogram; head direction tuning of the output neuron after learning. The numbers in the polar plots indicate the peak firing rate at the preferred head direction after averaging over space. **a**) Wider spatial tuning of inhibitory input neurons than of excitatory input neurons combined with narrower head direction tuning of inhibitory input neurons leads to a grid cell-like firing pattern in space with invariance to head direction, i.e. the output neuron fires like a pure grid cell. **b**) The same spatial input characteristics as in **a** combined with head direction-invariant inhibitory input neurons leads to grid cell-like activity in space and a preferred head direction, i.e. the output neuron fires like a conjunctive cell. **c**) If the spatial tuning of inhibitory input neurons is less smooth than that of excitatory neurons and the concurrent head direction tuning is wider for inhibitory than for excitatory neurons, the output neuron is not tuned to space but to a single head direction, i.e. the output neuron fires like a pure head direction cell. **d**) Head direction tuning and grid score of 10 simulations of the three cell types. Each symbol represents one realization with random input tuning. The markers correspond to the tuning properties of the input neurons as depicted in **a, b, c**: grid cell (triangles), conjunctive cell (squares), head direction cell (circles). The values that correspond to the output cells in **a, b, c** are shown as filled symbols. **e**) The head direction tuning of individual grid fields is sharper than the overall head direction tuning of the conjunctive cell. Depicted is a rate map of a conjunctive cell (left) and the corresponding head direction tuning (right, dashed). For three individual grid fields, indicated with colored squares, the head direction tuning is shown in the same polar plot. The overall tuning of the grid cell (dashed) is a superposition of the tuning of all grid fields. Numbers indicate the peak firing rate (in Hz) averaged individually within each of the four rectangles in the rate map.

3.4 Too fast learning leads to unstable grid patterns

We showed that stable grid patterns emerge within minutes of behavioral rat trajectories (Fig. **3**) for high learning rates. Our model requires thorough spatial exploration of the rat, before significant weight changes occur. Accordingly, no stable patterns should emerge, if the learning rate of the rat is too high. Indeed we observe flickering grids and a poorer development of hexagonal patterns if we double the learning rates compared to those in Fig. **3** (Fig. **S8**).

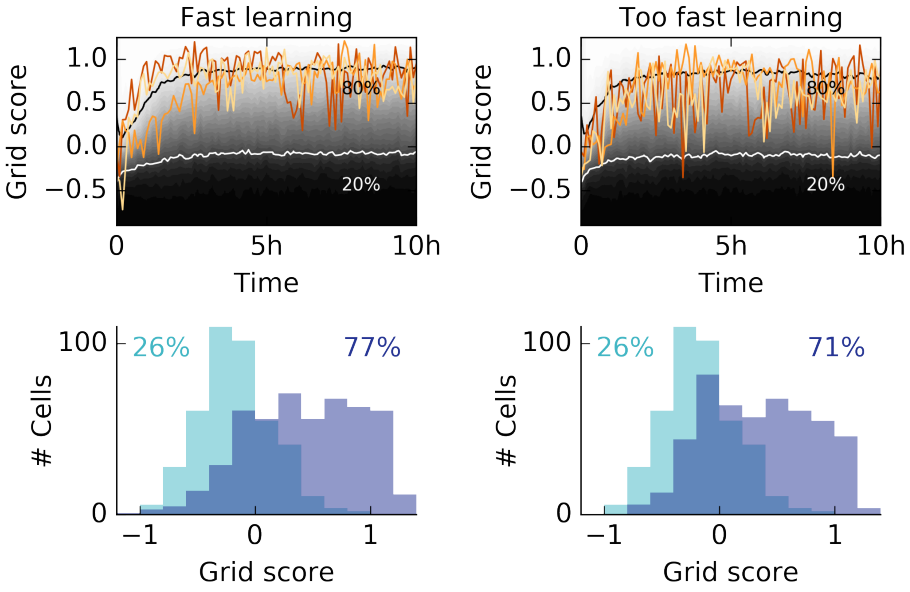

Figure S8: Left column: Same data as shown in Fig. **3c** with three different individual traces (top). Grid score histogram of 500 realizations before (light blue) and after 10 hours of spatial exploration (dark blue). Right column: The same simulations as shown on the left, but with twice the learning rates for excitatory and inhibitory synapses. The high learning rate leads to flickering unstable grids, which is expressed in the large fluctuations in the grid score. The histogram after 10 hours of spatial exploration shows that less cells develop a hexagonal pattern.

3.5 Influence of random initialization on final grids

Different random trials lead to different grid patterns. In the simulations in the manuscript, a different trial corresponds to different initial synaptic weights, different trajectories and different input functions (i.e., different center locations for sparse input and different sets of white noise for dense input). To test which of these random aspects has the largest influence on the final grid pattern, we ran simulations where we only varied either the initial synaptic weights, or the trajectories or the input functions. From the resulting final grid patterns (after simulating 10 hours of rat trajectory), we computed the cross correlation of all pairs of 500 grid cells within each set of simulations (Fig. **S9**). A high cross correlation between different trials indicates, that the influence of the varied parameter is small. We observed that the cross correlation is high, if only the initial weights are varied. It is lower if instead only the trajectories are varied and even lower if only the input functions are varied. As expected, the cross correlation between different trials is lowest, if initial synaptic weights as well as trajectories as well as input functions are varied together (Fig. **S9**, rightmost box).

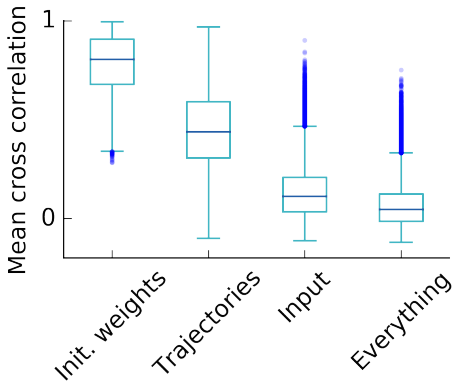

Figure S9: The influence of random simulation parameters on the final grid pattern. Box plot of the cross correlations of 500 simulations (i.e., $(500^2 - 500)/2 = 124750$ cross correlations) where only the parameter that is indicated on the $x-$axis was varied. A high cross correlation indicates that different simulations lead to similar grids and thus points towards a low influence of the varied parameter on the final grid pattern. We conclude that the influence on the final grid pattern in decreasing order is given by the parameters: Initial synaptic weights, trajectory of the rat, input tuning (i.e., locations of the randomly located input tuning curves). As expected, the correlation is lowest, if all parameters are different in each simulation (rightmost box). Each box extends from the first to the third quartile, with a dark blue line at the median. The whiskers extend from the first and third quartile by 1.5 the interquartile range. Dots show flier points.

3.6 Shape of stretched grids or band cells

As shown in Fig. **4** in the main text, an asymmetric autocorrelation structure of inhibitory inputs can lead to different symmetries in the output neuron. The possible patterns are illustrated in Fig. **S10**.

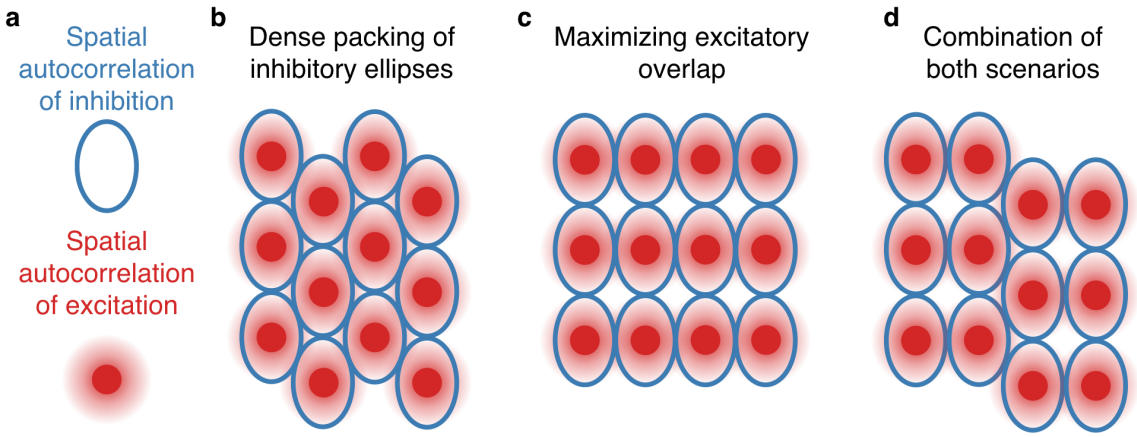

Figure S10: **a**) For ellipsoidal spatial autocorrelation structures of inhibitory input (blue line), we observed band cell-like firing patterns or stretched grids (Fig. **4d**). Interestingly, the resulting patterns alternate between two different symmetries. This can be understood by two competing arrangements of ellipsoids. **b**) A dense packing of ellipsoids maximizes the area with non-zero firing and is favored by the inhibitory learning rule. This leads to stretched grids. **c**) Maximizing the overlap between excitatory input fields is favored by the excitatory learning rule and leads to quadratic grids with different periodicities along different directions. **d**) Some simulations show a combination of both patterns. Compare Fig. **4d**.
The observed alignment of excitatory firing fields in **c** is particularly favored, if inhibition is very smooth along one direction. This could lead to the alignment of the head direction of individual grid fields in the simulations shown in Fig. **5**.

3.7 Boundary effects and stability of grids

The motion of the rat is not periodic. We constrained it to either a square or a circular box. The input tuning is not periodic either. Consequently, input neurons with tuning fields that lie partially outside the boundary receive less activation. This leads to boundary effects, for the following reason: Excitatory weights associated to fields at the boundaries grow less, because the Hebbian learning scales with the presynaptic activation. This leads to a smaller firing rate at the boundary. According to the inhibitory learning rule, the inhibitory weights of neurons that are tuned to boundary locations then also grow less. At a distance given by the width of excitatory

firing fields, the excitatory weights equally fast as those that are far away from the boundary. If inhibition is more broadly tuned than excitation, it is still reduced at these locations, though. Firing fields are thus favored at a distance from the boundary that is determined by the width of the excitatory tuning, because at this location the excitation will exceed the inhibition. This preference of firing at a certain distance from the boundary competes with the preference for hexagonal firing that is induced by the interaction of excitatory and inhibitory plasticity. For place field-like input that is arranged on a symmetric lattice, the alignment to the boundary can be seen in the alignment of one grid axis to the boundary in a square box (Fig. **S11a**). This alignment is not an artifact of the symmetric distribution of input fields, because it is not present in a circular arena (Fig. **S11b**). The strong tendency to align with the boundary can be overcome by using a random distribution of input fields (Fig. **S11c**) and in particular by using input with more than one place field per neuron, i.e., non-localized input. Nonetheless, we observed boundary effects in all simulations when simulating for very long times.

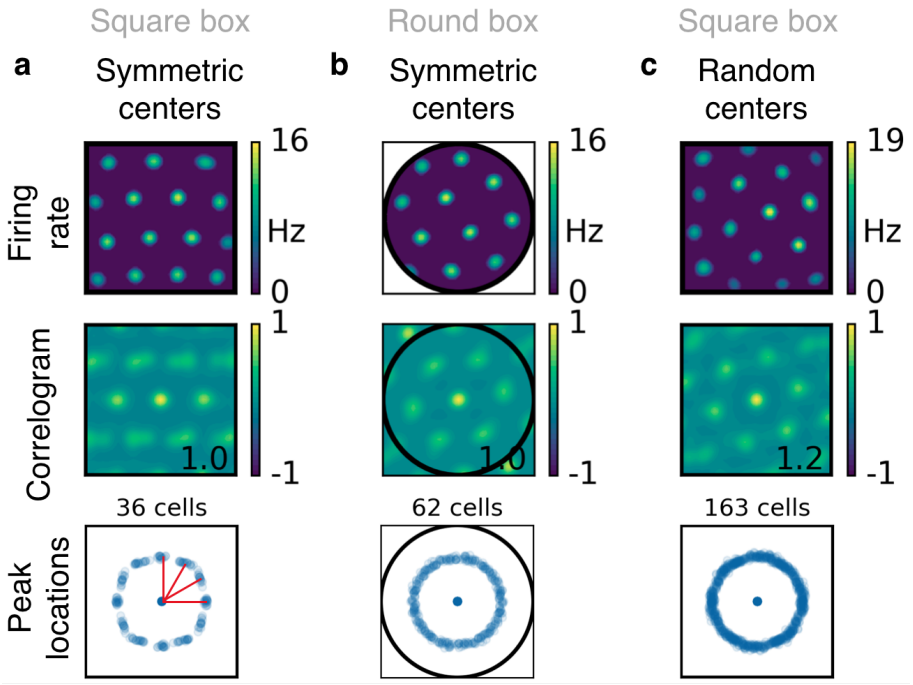

Figure S11: Boundary effects in simulations with place field-like input. **a**) Simulations in a square box with input place fields that are arranged on a symmetric grid. From top to bottom: Firing rate map and corresponding autocorrelogram for an example grid cell; peak locations of 36 grid cells. The clusters at orientation of 0, 30, 60 and 90 degrees (red lines) indicate that the grids tend to be aligned to the boundaries. **b**) Simulations in a circular box with input place fields that are arranged on a symmetric grid. Arrangement as in **a**. The grids show no orientation preference, indicating that the orientation preference in **a** is induced by the square shape of the box. **c**) Simulations in a square box with input place fields that are arranged on a distorted grid (see Fig. **S4**). Arrangement as in **a**. The grids show no orientation preference, indicating that the influence of the boundary on the grid orientation is small compared to the effect of randomness in the location of the input centers.

3.8 Non-localized input leads to decreased variance in synaptic weights

In the main text we have shown that non-localized input often leads to distorted grid patterns. Fig. **S12** shows that the coefficient of variation of synaptic weights is lower for non-localized than for localized input after a grid has formed. In other words, the structure of the grid is only weakly apparent in the structure of the weights. This renders the grid more susceptible to distortions.

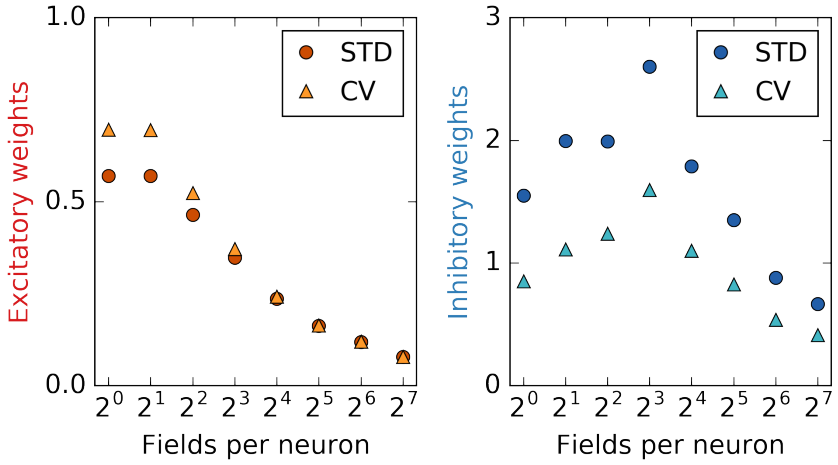

Figure S12: The standard deviation (STD) and the coefficient of variation (CV) of excitatory (left) and inhibitory (right) weights as a function of place fields per input neuron. The values are computed after the output neuron has established a stable grid pattern on a linear track. For excitatory weights, the CV decreases significantly with non-localized input. This indicates that different firing patterns in the output neuron are closer in weight space, the more non-localized the input is: An explanation for the defects in grids with non-localized input.

4 Glossary

A summary of notation:

The rat's position at time t : $\mathbf{x}(t)$

Spatial dimensions x, y and head direction z : $\mathbf{x} = (x, y, z)$

Population label; can be E (excitatory) or I (inhibitory) : P

Standard deviation of Gaussian tuning of popluation P : σ_{P}

Spatial autocorrelation length of input of population P : $\sigma_{\mathrm{P,corr}}$

Number of input neurons of population P : N_{P}

Number of place fields per input neuron of population P : $N_{\mathrm{P}}^{\mathrm{f}}$

Firing rate of output neuron : $r^{\mathrm{out}}(x)$

Firing rate of input neuron i of population P : $r_i^{\mathrm{P}}(x)$

Synaptic weight of input neuron i of population P to output neuron : $w_i^{\mathrm{P}}(t)$

Learning rates of excitation and inhibition : $\eta_{\mathrm{E}}, \eta_{\mathrm{I}}$

Target rate of the output neuron : ρ_0

Length of linear track : L

Height of the Gaussian input fields : $\alpha_{\mathrm{E}}, \alpha_{\mathrm{I}}$

Value of Gaussian with standard deviation σ_{P} at location x : $\mathcal{G}^{\mathrm{P}}(x)$

Von Mises distribution with width σ_{P} that is periodic in $[-L/2, L/2]$: $\mathcal{M}^{\mathrm{P}}(x)$

References

[1] K. D. Miller and D. J. MacKay, "The role of constraints in hebbian learning," *Neural Computation*, vol. 6, no. 1, pp. 100–126, 1994.

[2] A. Bovier, "Extreme values of random processes," 2005.

[3] F. Sargolini, M. Fyhn, T. Hafting, B. L. McNaughton, M. P. Witter, M.-B. Moser, and E. I. Moser, "http://www.ntnu.edu/kavli/research/grid-cell-data," 2016.

[4] P. Mégevand, "http://de.mathworks.com/matlabcentral/fileexchange/43543-watson-s-u2-statistic-based-permutation-test-for-circular-data," 2013.

[5] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," *Nature*, vol. 436, no. 7052, pp. 801–806, 2005.