Supplementary Information

Network equations

The network we used consisted of three input layers and an intermediate layer. The three input layers – an eye-centered layer, an eye position layer and a head-centered layer – are also output layers; the final estimates of the network are read from these layers after relaxation. Below we describe how that network was constructed and how it evolved in time. We also derive expressions for the maximum likelihood variances, which are used to evaluate network performance.

The three input layers consist of three topographic layers of N units. The units are indexed by their position, j, where $j = 1 \ldots N$. Similarly, the hidden layer is a topographic 2D map of $\frac{N}{2}$x$\frac{N}{2}$ units indexed by their position l, m, where $l = 2, 4, \ldots, N$ and $m = 2, 4, \ldots, N$. The intermediate layer is sampled more coarsely than the input layers; this was done solely to increase simulation speed.

The input units are symmetrically interconnected with the hidden layer, and the corresponding matrices of connection weights are denoted W^r, W^e and W^a for, respectively, the eye-centered, eye position and head-centered layers. The (dimensionless) connection strengths between unit j in each input layer and unit (l, m) in the intermediate layer are given by

$$W^r_{jlm} = K_w \exp\left[\frac{\cos[(2\pi/N)(j-l)] - 1}{\sigma_w^2}\right] \quad (1)$$

$$W^e_{jlm} = K_w \exp\left[\frac{\cos[(2\pi/N)(j-m)] - 1}{\sigma_w^2}\right] \quad (2)$$

$$W^a_{jlm} = K_w \exp\left[\frac{\cos[(2\pi/N)(j-l-m)] - 1}{\sigma_w^2}\right]. \quad (3)$$

The variable σ_w represents lateral spread: units j and l are strongly connected if $|j-l|/N \lesssim \sigma_w/2\pi$. As such, it has dimensions of radians.

Note that with these connection matrices, unit (l, m) in the intermediate layer is most strongly interconnected with unit $j = l$ in the eye-centered layer, $j = m$ in the eye position layer and $j = l+m$ in the head-centered layer. Unit (l, m) is connected more weakly to neighboring units in each layer, with the spatial extent of these connections is controlled by σ_w.

The evolution of the activities (chosen to be firing rates) in the recurrent network are described by a set of coupled nonlinear equations. Denoting $A_{lm}(t)$ as the activity of unit (l, m) in the intermediate layer at time t, and $R_{rj}(t)$, $R_{ej}(t)$, and $R_{aj}(t)$ as the activity of unit j in the eye-centered, eye-position and head-centered layer at time t, the evolution equations are written

$$A_{lm}(t+1) = \frac{L_{lm}(t)^2}{S + \mu \sum_{l'm'} L_{l'm'}(t)^2} \quad (4)$$

$$R_{rj}(t+1) = \frac{\left[\sum_{lm} W^r_{jlm} A_{lm}(t+1)\right]^2}{S + \mu \sum_j \left[\sum_{lm} W^r_{jlm} A_{lm}(t+1)\right]^2} \quad (5)$$

$$R_{ej}(t+1) = \frac{\left[\sum_{lm} W^e_{jlm} A_{lm}(t+1)\right]^2}{S + \mu \sum_j \left[\sum_{lm} W^e_{jlm} A_{lm}(t+1)\right]^2} \quad (6)$$

$$R_{aj}(t+1) = \frac{\left[\sum_{lm} W^a_{jlm} A_{lm}(t+1)\right]^2}{S + \mu \sum_j \left[\sum_{lm} W^a_{jlm} A_{lm}(t+1)\right]^2} \quad (7)$$

where $L_{lm}(t)$ represents a linear pooling of activities from the three input layers,

$$L_{lm}(t) = \sum_j W^r_{jlm} R_{rj}(t) + \sum_j W^e_{jlm} R_{ej}(t) + \sum_j W^a_{jlm} R_{aj}(t). \tag{8}$$

The activation functions represented by Eqs. (4-8) implement a quadratic nonlinearity coupled with a divisive normalization. This choice of activation function is consistent with neurophysiological data [1, 2].

In all simulation, we used $N = 40$, corresponding to 40 units in the input layers and a 20×20 array of units for the intermediate layer.

Network initialization and parameters

Although activity is a continuous variable, for convenience we sampled the initial activity from a discrete distribution. We used a Poisson distribution, for which calculation of the Fisher information is especially straightforward (see Eq. (14) below). For eye-centered position, the probability distribution for the initial activity, denoted $R_{rj}(0)$, is given by

$$P(R_{rj}(0)|x_r) = \frac{f_j(x_r)^{R_{rj}(0)} e^{-f_j(x_r)}}{R_{rj}(0)!}. \tag{9}$$

The expressions for $P(R_{ej}(0)|x_e)$ and $P(R_{aj}(0)|x_a)$ are identical, except that r is replaced by e or a. The activity in the intermediate layer, $A_{lm}(0)$, is initialized to 0: $A_{lm}(0) = 0 \ \forall l, m$.

The input tuning curve for eye-centered position, $f_j(x_r)$, which describes the mean response to position x_r in the input layer *before* relaxation, is taken to be a circular normal function with spontaneous activity, ν,

$$f_j(x_r) = C_r \left(K \exp\left[\frac{\cos[x_r - (2\pi/N)j] - 1}{\sigma^2} \right] + \nu \right). \tag{10}$$

The expressions for $f_j(x_a)$ and $f_j(x_e)$ are identical, except that r is replaced by a or e. The tuning curve, f_j, is a dimensionless parameter (it has dimensions of firing rate times time). Since ν has dimensions of firing rate, so does K; this forces C_r, C_e and C_a, the input gains for the eye-centered, eye, and head-centered layers, respectively, to have dimensions of time. Like σ_w, σ has dimensions of radians.

The parameters K_w and σ_w (Eqs. (1-3)), S and μ (Eqs. (4-7)), and K, σ and ν (Eq. (10)) were fixed and identical in the three input layers. C_r and C_e were always fixed at 1 s; C_a was varied from 0 to 2 s. The remaining parameters were chosen as follows: $K = 20$ Hz, $\nu = 1$ Hz, $\sigma = 0.40$ radians (corresponding to a full width at half maximum of ∼55°), $K_w = 1$, $\mu = 0.002$ s, and $S = 0.1$ Hz. The spatial extent of the weights, σ_w, was optimized with respect to the variance of the estimator (Eq. (12)); the optimum value was $\sigma_w = 0.37$ radians (corresponding a lateral spread of ∼50°, measured as the full width at half maximum). σ_w was optimized for $C_r = C_e = C_a = 1$ s, and not re-optimized when C_a was changed.

Network evolution

The network equations, (4-8), were initialized as described in the previous section, then iterated. To get close to the attractor, the equations need to be iterated a large number of times although, for this network, we found that "large" was three – the network estimates (Eq. (11)) and variance (Eq. (12)) changed by less than 1% when the number of iterations increased from three to six. Thus, in all simulations we iterated Eq. (4-8) three times.

Network estimates and their errors

After iterating the network equations three times, we read out the position of each of the three smooth hills using a complex estimator. For instance, the network estimate of eye-centered position, denoted $\hat{x}_r$, was given by

$$\hat{x}_r = \text{phase}\left(\sum_{j=1}^{N} R_j(3) e^{i(2\pi/N)j}\right) \tag{11}$$

where $i \equiv \sqrt{-1}$. The network estimates of eye position and head-centered position, x_e and x_a, respectively, are identical to Eq. (11) except that r is replaced by e or a.

To compute the variance of the network estimate of eye-centered position, we used the standard formula

$$\langle (x_r - \hat{x})^2 \rangle_{\text{network}} = \frac{1}{M-1} \sum_{k=1}^{M} (\hat{x}_{kr} - x_r)^2 \tag{12}$$

where M is the number of trials (we used $M = 100,000$) and $\hat{x}_{kr}$ is the network estimate for the kth trial, found using Eq. (11). The network estimates of the variance of eye position and head-centered position are identical to Eq. (12) except that r is replaced by e or a.

To avoid edge effects, we used an architecture with periodic boundary conditions. Our approach, however, is not limited to periodic functions: we can compute non-periodic functions by using arrays of units with Gaussian tuning curves. This type of network also achieves maximum likelihood, so long as the hills of activity are kept away from the edges of the neuronal arrays.

Maximum likelihood variance

To determine how well the network performed compared to how well it could perform in principle, we compared the variances of the network estimates to the variances of the maximum likelihood estimates. In the large N limit, the latter are given by the Cramér-Rao bound [3]. Here we compute explicitly the Cramér-Rao bound for eye-centered position, x_r, then use symmetry to write down the bounds for x_e and x_a.

Because of the constraint $x_a = x_e + x_r$, the conditional probability of observing a set of initial conditions, $\mathbf{R}_r(0)$, $\mathbf{R}_e(0)$ and $\mathbf{R}_a(0)$, given x_r and x_e, is written

$$P\left(\mathbf{R}_r, \mathbf{R}_e, \mathbf{R}_a | x_r, x_e\right) = P\left(\mathbf{R}_r | x_r\right) P\left(\mathbf{R}_e | x_e\right) P\left(\mathbf{R}_a | x_r + x_e\right) \tag{13}$$

where we are assuming the noise is independent in each layer and $\mathbf{R}_r$, $\mathbf{R}_e$ and $\mathbf{R}_a$ are shorthand for $\mathbf{R}_r(0)$, $\mathbf{R}_e(0)$ and $\mathbf{R}_a(0)$, respectively. (The substitution $x_a = x_r + x_e$ is arbitrary; our answer does not depend on whether we replace x_a with $x_r + x_e$ or x_e with $x_a - x_r$.)

For the Cramér-Rao bound we use the diagonal elements of the inverse of the Fisher information [6]. The Fisher information is given by

$$I_{\alpha\beta} = \left\langle -\frac{\partial^2}{\partial x_\alpha \partial x_\beta} \left[\log P\left(\mathbf{R}_r | x_r\right) + \log P\left(\mathbf{R}_e | x_e\right) + \log P\left(\mathbf{R}_a | x_r + x_e\right)\right] \right\rangle$$

where α and β can take on the values e and r and the angle brackets indicate an average with respect to the probability distribution given in Eq. (13). Performing the derivatives and taking the averages, the latter with the aid of Eq. (9), we find that

$$\mathbf{I} = \begin{pmatrix} \sigma_r^{-2} + \sigma_a^{-2} & \sigma_a^{-2} \\ \sigma_a^{-2} & \sigma_e^{-2} + \sigma_a^{-2} \end{pmatrix} \tag{14}$$

where

$$\sigma_r^2 = \left[\sum_{i=1}^{N} \frac{f_i'^2(x_r)}{f_i(x_r)}\right]^{-1} = \left\langle -\frac{\partial^2 \log P\left(\mathbf{R}_r | x_r\right)}{\partial x_r^2} \right\rangle^{-1}$$

is the Cramér-Rao bound for the variance of eye-position *taken alone* (this is the standard Cramér-Rao bound for Poisson statistics; see [4, 5]). Analogous expressions apply for for σ_e^2 and σ_a^2.

Inverting the Fisher information, Eq. (14), yields

$$\begin{aligned} \mathbf{I}^{-1} &= \frac{1}{(\sigma_r^{-2} + \sigma_a^{-2})(\sigma_e^{-2} + \sigma_a^{-2}) - \sigma_a^{-4}} \begin{pmatrix} \sigma_e^{-2} + \sigma_a^{-2} & -\sigma_a^{-2} \\ -\sigma_a^{-2} & \sigma_r^{-2} + \sigma_a^{-2} \end{pmatrix} \\ &= \frac{1}{\sigma_r^2 + \sigma_e^2 + \sigma_a^2} \begin{pmatrix} \sigma_r^2(\sigma_e^2 + \sigma_a^2) & \sigma_r^2 + \sigma_e^2 \\ \sigma_r^2 + \sigma_e^2 & \sigma_e^2(\sigma_e^2 + \sigma_a^2) \end{pmatrix}. \end{aligned} \tag{15}$$

The Cramér-Rao bounds for ${\sigma_r^{ML}}^2$ and ${\sigma_e^{ML}}^2$ can be read off the diagonal elements of $\mathbf{I}^{-1}$, Eq. (15). A similar calculation (or a simple permutation of indices) yields an expression for ${\sigma_a^{ML}}^2$. The results are

$$\begin{aligned} {\sigma_r^{ML}}^2 &= \frac{\sigma_r^2(\sigma_a^2 + \sigma_e^2)}{\sigma_r^2 + \sigma_a^2 + \sigma_e^2} \\ {\sigma_e^{ML}}^2 &= \frac{\sigma_e^2(\sigma_r^2 + \sigma_a^2)}{\sigma_r^2 + \sigma_a^2 + \sigma_e^2} \\ {\sigma_a^{ML}}^2 &= \frac{\sigma_a^2(\sigma_e^2 + \sigma_r^2)}{\sigma_r^2 + \sigma_a^2 + \sigma_e^2}. \end{aligned}$$

References

[1] J. Anderson, I. Lampl, D. Gillespie, and D. Ferster. The contribution of noise to contrast invariance of orientation tuning in cat visual cortex. *Science*, 290:1908–9, 2000.

[2] D. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–197, 1992.

[3] A. Papoulis. *Probability, random variables, and stochastic process.* McGraw-Hill, inc., 1991.

[4] M. Paradiso. A theory of the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological Cybernetics*, 58:35–49, 1988.

[5] H. Seung and H. Sompolinsky. Simple model for reading neuronal population codes. *Proceedings of National Academy of Sciences. USA.*, 90:10749–10753, 1993.

[6] K. Zhang and T. Sejnowski. Neuronal tuning: to sharpen or broaden? *Neural Computation*, 11:75–84, 1999.