A Notation

For $N, K \geq 2$, let $\mathbb{R}^N$ denote N-dimensional Euclidean space equipped with the usual Euclidean norm $\|\cdot\|$ and let $\mathbb{R}^N_+$ denote the non-negative orthant. Let $\mathbf{0} = [0, \ldots, 0]^\top$ and $\mathbf{1} = [1, \ldots, 1]^\top$ respectively denote the vectors of zeros and ones, whose dimensions should be clear from context.

Let $\mathbb{R}^{N\times K}$ denote the set of $N \times K$ real-valued matrices. Let $\|\cdot\|_F$ denote the Frobenius norm and $\|\cdot\|_{\text{op}}$ denote the operator norm. Let $O(N)$ denote the set of $N \times N$ orthogonal matrices. Let $\mathbb{S}^N$ (resp. $\mathbb{S}^N_+$, $\mathbb{S}^N_{++}$) denote the set of $N \times N$ symmetric (resp. positive semidefinite, positive definite) matrices. Let $\mathbf{I}_N$ denote the $N \times N$ identity matrix.

Given vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$, let $\mathbf{u} \circ \mathbf{v} = [u_1 v_1, \ldots, u_N v_N]^\top \in \mathbb{R}^N$ denote the Hadamard (elementwise) product of $\mathbf{u}$ and $\mathbf{v}$. Let $\text{diag}(\mathbf{u})$ denote the $N \times N$ diagonal matrix whose $(i,i)^{\text{th}}$ entry is u_i, for $i = 1, \ldots, N$. Given a matrix $\mathbf{M} \in \mathbb{R}^{N\times N}$, we also let $\text{diag}(\mathbf{M})$ denote the N-dimensional vector whose i^{th} entry is M_{ii}.

B Separation of timescales for the gain and synaptic weight updates

In this section, we consider an algorithm where we directly optimize the objective in equation 3. In particular, for each context c, we first optimize over the gains $\mathbf{g}$ and then take a gradient descent step with respect to $\mathbf{W}$.

We first compute the gains using the formula for the optimal gains derived in [25, equation 18]:

$$\mathbf{g} = \left[(\mathbf{W}^\top \mathbf{W})^{\circ 2}\right]^\dagger \text{diag}\left(\mathbf{W}^\top \mathbf{C}_{ss}^{1/2}(c)\mathbf{W} - \mathbf{W}^\top \mathbf{W}\right).$$

We then update the synaptic weights by taking the following gradient descent step:

$$\Delta \mathbf{w}_i = \mathbb{E}_{\mathbf{s}\sim p(\mathbf{s}|c)}\left[\eta_w(\mathbf{r}\mathbf{r}^\top \mathbf{w}_i g_i - \mathbf{w}_i g_i)\right] = \eta_w \left(\mathbf{M}(\mathbf{W}, \mathbf{g})^{-1}\mathbf{C}_{ss}(c)\mathbf{M}(\mathbf{W}, \mathbf{g})^{-1} - \mathbf{I}_N\right)\mathbf{w}_i g_i,$$

where $\mathbf{M}(\mathbf{W}, \mathbf{g}) := \alpha \mathbf{I}_N + \mathbf{W}\text{diag}(\mathbf{g})\mathbf{W}^\top$. Combining these updates yields Algorithm 2, which takes context-dependent covariance matrices $\mathbf{C}_{ss}(c)$ as its input.

Algorithm 2: Adaptive whitening via synaptic plasticity and gain modulation

Input: Covariance matrices $\mathbf{C}_{ss}(1), \mathbf{C}_{ss}(2), \ldots$
Initialize: $\mathbf{W} \in \mathbb{R}^{N\times K}$; $\eta_w > 0$
for $c = 1, 2, \ldots$ **do**
  $\mathbf{g} \leftarrow \left[(\mathbf{W}^\top \mathbf{W})^{\circ 2}\right]^\dagger \text{diag}\left(\mathbf{W}^\top \mathbf{C}_{ss}^{1/2}(c)\mathbf{W} - \mathbf{W}^\top \mathbf{W}\right)$
  $\mathbf{G} \leftarrow \text{diag}(\mathbf{g})$
  $\mathbf{W} \leftarrow \mathbf{W} + \eta_w \left(\left(\mathbf{W}\mathbf{G}\mathbf{W}^\top\right)^{-1}\mathbf{C}_{ss}(c)\left(\mathbf{W}\mathbf{G}\mathbf{W}^\top\right)^{-1}\mathbf{W}\mathbf{G} - \mathbf{W}\mathbf{G}\right)$
end for

C Adaptive whitening of natural images

In this section, we elaborate on the converged structure of $\mathbf{W}_T$ using natural image patches. To better visualize the relationship between the learned columns of $\mathbf{W}$ and sinusoidal basis functions (e.g. DCT), we focus on 1-dimensional image patches (rows of pixels). The results are similar with 2D image patches.

It is well known that eigenvectors of natural images are well-approximated by sinusoidal basis functions [e.g. the DCT; 33; 34]. Using the same images from the main text [32], we generated 56 contexts by sampling 16×1 pixel patches from separate images, with 2E4 samples each. We train Algorithm 2 with $K = N = 16$, $\eta_w =$ 5E-2, and random $\mathbf{W}_0 \in O(16)$ on a training set of X of the images, presented uniformly at random $T =$ 1E5 times. Fig C.1A,B shows that $\mathbf{W}_T$ approximates the principal components of the aggregated context-dependent covariance, $\mathbb{E}_{c\sim p(c)}[\mathbf{C}_{ss}(c)]$, which are closely aligned with the DCT. To show that this structure is inherent in the spatial statistics of natural images, we generated control contexts, $\mathbf{C}_{ss}(c)$, by forming covariance matrices with matching

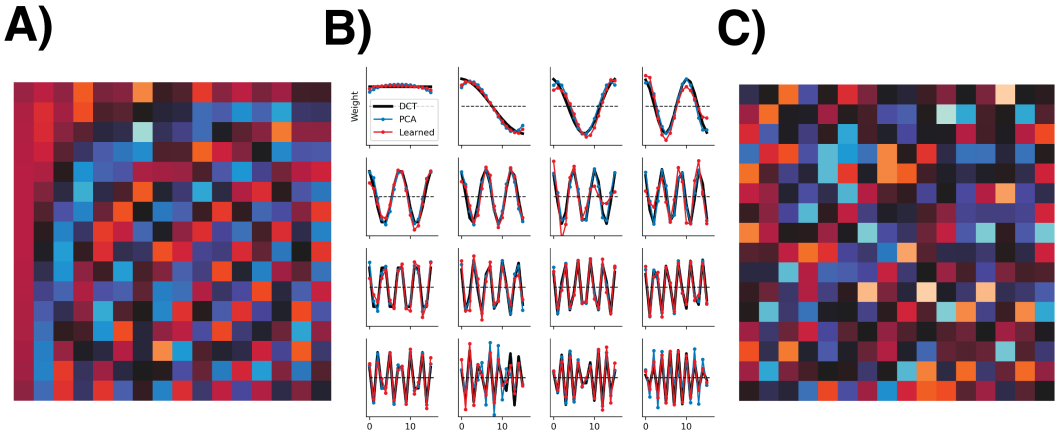

Figure C.1: Control experiment accompanying Sec. 5.2. **A)** $\mathbf{W}_T$ learned from natural image patches. **B)** Basis vectors from **A** displayed as line plots, compared to the 1D DCT, and principal components of $\mathbb{E}_{c\sim p(c)}[\mathbf{C}_{ss}(c)]$. **C)** Control condition. $\mathbf{W}_T$ learned from spectrally-matched image patches with random eigenvectors.

eigenspectra, but each with *random* and distinct eigenvectors. This destroys the structure induced by natural image statistics. Consequently, the learned vectors in $\mathbf{W}_T$ are no longer sinusoidal (Fig C.1C). As a result, whitening error with $\mathbf{W}_T$ is much higher on the training set, with 0.3 ± 0.02 error (mean $\pm$ standard error over 10 random initializations; Eq. 6) on natural image contexts and 2.7 ± 0.1 on the control contexts. While for the natural images, a basis approximating the DCT was sufficient to adaptively whiten all contexts in the ensemble, this is not the case for the generated control contexts.

Finally, we find that as K increases from $K = 1$ to $K = 16$, the basis vectors in $\mathbf{W}_T$ *progressively* learn higher frequency components of the DCT (Fig. C.2). This is a sensible solution, due to the ℓ_2 reconstruction error of our objective, and the $1/f$ spectral content of natural image statistics. With more flexibility, as K increases past N (i.e. the overcomplete regime), the network continues to improve its whitening error (Fig. C.3A) by learning a basis, $\mathbf{W}_T$, that can account for within-context information that is insufficiently captured by the DCT (Fig. C.3B). Taken together, our model successfully learns a basis $\mathbf{W}_T$ that exploits the spatial structure present in natural images.

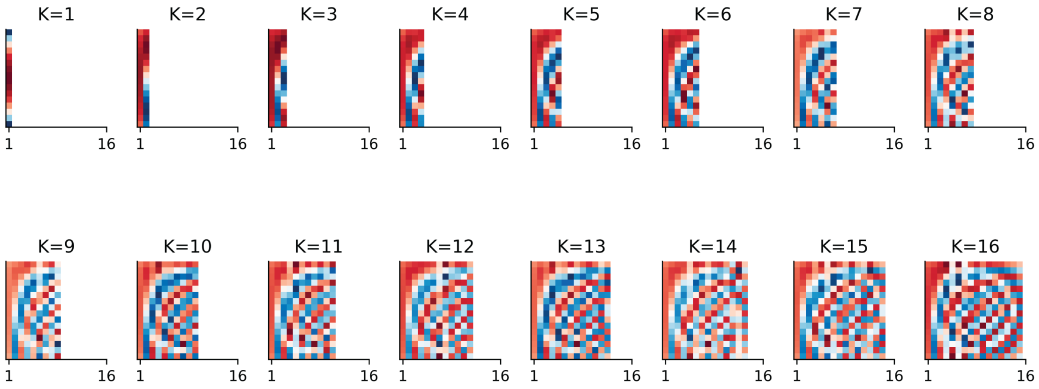

Figure C.2: As K increases, columns of $\mathbf{W}$ progressively learn higher frequency components of the DCT.

D Modifications for increased biological realism

In this section, we modify Algorithm 1 to be more biologically realistic.

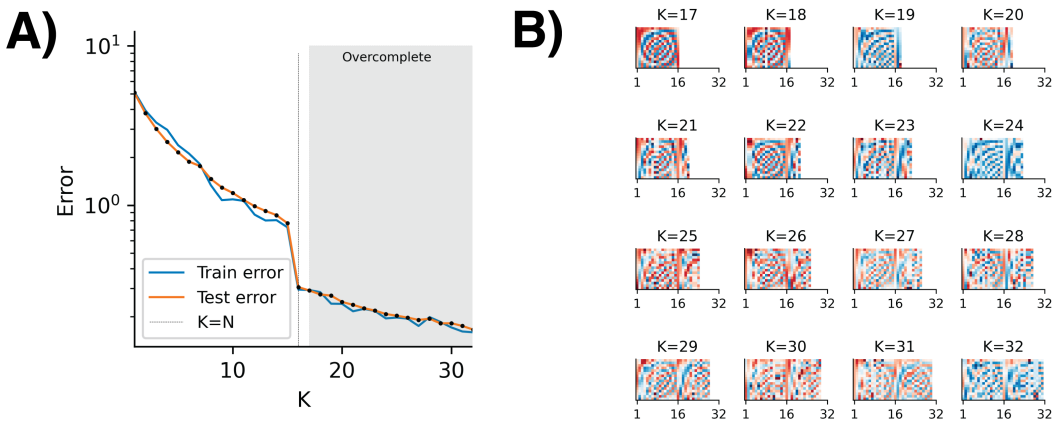

Figure C.3: **A)** Error on training and test set as a function of K. **B)** In the overcomplete regime, the network converges to a $\mathbf{W}_T$ that helps to improve error compared to the the $K \leq N$ regime.

D.1 Enforcing unit norm basis vectors

In our algorithm, there is no constraint on the magnitude of the column vectors of $\mathbf{W}$. We can enforce a unit norm (here measured using the Euclidean norm) constraint by adding Lagrange multipliers to the objective in equation 3:

$$\min_{\mathbf{W}\in\mathbb{R}^{N\times K}} \max_{\mathbf{m}\in\mathbb{R}^K} \mathbb{E}_{c\sim p(c)} \left[\min_{\mathbf{g}\in\mathbb{R}^K} \mathbb{E}_{\mathbf{s}\sim p(\mathbf{s}|c)} \left[g\left(\mathbf{W}, \mathbf{g}, \mathbf{r}, \mathbf{s}\right)\right]\right], \tag{D.1}$$

where

$$g(\mathbf{W}, \mathbf{g}, \mathbf{r}, \mathbf{s}) = \ell(\mathbf{W}, \mathbf{g}, \mathbf{r}, \mathbf{s}) + \sum_{i=1}^{K} m_i(\|\mathbf{w}_i\|^2 - 1).$$

Taking partial derivatives with respect to $\mathbf{w}_i$ and $\mathbf{m}_i$ results in the updates:

$$\begin{aligned}\Delta\mathbf{w}_i &= \eta_w(n_i\mathbf{r} - (g_i + m_i)\mathbf{w}_i)\\ \Delta m_i &= \|\mathbf{w}_i\|^2 - 1.\end{aligned}$$

Furthermore, since the weights are constrained to have unit norm, we can replace $\|\mathbf{w}_i\|^2$ with 1 in the gain update:

$$\Delta g_i = \eta_g(z_i^2 - 1).$$

D.2 Decoupling the feedforward and feedback weights

We replace the primary neuron-to-interneuron weight matrix $\mathbf{W}^\top$ (resp. interneuron-to-primary neuron weight matrix $-\mathbf{W}$) with $\mathbf{W}_{rn}$ (resp. $-\mathbf{W}_{nr}$). In this case, the update rules are

$$\begin{aligned}\mathbf{W}_{rn} &\leftarrow \mathbf{W}_{rn} + \eta_w\left(\mathbf{n}_t\mathbf{r}_t^\top - \text{diag}(\mathbf{g} + \mathbf{m})\mathbf{W}_{rn}\right)\\ \mathbf{W}_{nr} &\leftarrow \mathbf{W}_{nr} + \eta_w\left(\mathbf{r}_t\mathbf{n}_t^\top - \mathbf{W}_{nr}\text{diag}(\mathbf{g} + \mathbf{m})\right).\end{aligned}$$

Let $\mathbf{W}_{rn,t}$ and $\mathbf{W}_{nr,t}$ denote the values of the weights $\mathbf{W}_{rn}$ and $\mathbf{W}_{nr}$, respectively, after $t = 0, 1, \ldots$ iterates. Then for all $t = 0, 1, \ldots,$

$$\mathbf{W}_{rn,t}^\top - \mathbf{W}_{nr,t} = \left(\mathbf{W}_{rn,0}^\top - \mathbf{W}_{nr,0}\right)\left(\mathbf{I}_N - \eta_w\text{diag}(\mathbf{g} + \mathbf{m})\right)^t.$$

Thus, if $g_i + m_i \in (0, 2\eta_w^{-1})$ for all i (e.g., by enforcing non-negative g_i, m_i and choosing $\eta_w > 0$ sufficiently small), then the difference decays exponentially in t and the feedforward and feedback weights are asymptotically symmetric.

D.3 Sign-constraining the synaptic weights and gains

The synaptic weight matrix $\mathbf{W}$ and gains vector $\mathbf{g}$ are not sign-constrained in Algorithm 1, which is not consistent with biological evidence. We can modify the algorithm to enforce the sign constraints by rectifying the weights and gains at each step. Here $[\cdot]_+$ denote the elementwise rectification operation. This results in the updates

$$\begin{aligned}\mathbf{g} &\leftarrow [\mathbf{g} + \eta_g(\mathbf{z} \circ \mathbf{z} - \mathbf{1})]_+ \\ \mathbf{W}_{rn} &\leftarrow \left[\mathbf{W}_{rn} + \eta_w \left(\mathbf{n}_t \mathbf{r}_t^\top - \text{diag}(\mathbf{g} + \mathbf{m})\mathbf{W}_{rn}\right)\right]_+ \\ \mathbf{W}_{nr} &\leftarrow \left[\mathbf{W}_{nr} + \eta_w \left(\mathbf{r}_t \mathbf{n}_t^\top - \mathbf{W}_{nr}\text{diag}(\mathbf{g} + \mathbf{m})\right)\right]_+ .\end{aligned}$$

D.4 Online algorithm with improved biological realism

Combining these modifications yields our more biologically realistic multi-timescale online algorithm, Algorithm 3.

Algorithm 3: Biologically realistic multi-timescale adaptive whitening

Input: $\mathbf{s}_1, \mathbf{s}_2, \cdots \in \mathbb{R}^N$
Initialize: $\mathbf{W}_{nr} \in \mathbb{R}^{N \times K}$; $\mathbf{W}_{rn} \in \mathbb{R}^{K \times N}$; $\mathbf{m}, \mathbf{g} \in \mathbb{R}^K$; $\eta_r, \eta_m > 0$; $\eta_g \gg \eta_w > 0$
for $t = 1, 2, \ldots$ **do**
  $\mathbf{r}_t \leftarrow \mathbf{0}$
  while not converged **do**
    $\mathbf{z}_t \leftarrow \mathbf{W}_{rn}\mathbf{r}_t$; `// interneuron inputs`
    $\mathbf{n}_t \leftarrow \mathbf{g} \circ \mathbf{z}_t$; `// gain-modulated interneuron outputs`
    $\mathbf{r}_t \leftarrow \mathbf{r}_t + \eta_r (\mathbf{s}_t - \mathbf{W}_{nr}\mathbf{n}_t - \alpha \mathbf{r}_t)$; `// recurrent neural dynamics`
  end while
  $\mathbf{m} \leftarrow \mathbf{m} + \eta_m(\text{diag}(\mathbf{W}_{rn}\mathbf{W}_{nr}) - \mathbf{1})$; `// weight normalization update`
  $\mathbf{g} \leftarrow [\mathbf{g} + \eta_g (\mathbf{z}_t \circ \mathbf{z}_t - \mathbf{1})]_+$; `// gains update`
  $\mathbf{W}_{rn} \leftarrow \left[\mathbf{W}_{rn} + \eta_w \left(\mathbf{n}_t \mathbf{r}_t^\top - \text{diag}(\mathbf{g} + \mathbf{m})\mathbf{W}_{rn}\right)\right]_+$; `// synaptic weights update`
  $\mathbf{W}_{nr} \leftarrow \left[\mathbf{W}_{nr} + \eta_w \left(\mathbf{r}_t \mathbf{n}_t^\top - \mathbf{W}_{nr}\text{diag}(\mathbf{g} + \mathbf{m})\right)\right]_+$
end for