Evolution of neural activity in circuits bridging sensory and abstract knowledge

Francesca Mastrogiuseppe[1], Naoki Hiratani[2], Peter Latham[1]

[1] Gatsby Computational Neuroscience Unit, University College London, UK
[2] Center for Brain Science, Harvard University, USA

The ability to associate sensory stimuli with abstract classes is critical for survival. How are these associations implemented in brain circuits? And what governs how neural activity evolves during abstract knowledge acquisition? To investigate these questions, we consider a circuit model that learns to map sensory inputs into abstract classes via gradient descent synaptic plasticity. We focus on typical neuroscience tasks (simple, and context-dependent, categorization), and study how both synaptic connectivity and neural activity evolve during learning. To make contact with the current generation of experiments we focus on the latter, and analyze activity via standard measures such as selectivity, correlations, and tuning symmetry. We find that the model is able to capture experimental observations, including seemingly disparate ones. We determine how, in the model, the behaviour of these activity measures depends on details of the circuit and the task. These dependencies make experimentally-testable predictions about the circuitry supporting abstract knowledge acquisition in the brain.

Introduction

Everyday decisions do not depend on the state of the world alone; they also depend on internal, non-sensory variables that are acquired with experience. For instance, over time we learn that in most situations salads are good for us while burgers are not, while in other contexts (for example, before a long hike in the mountains) the opposite is true. The ability to associate sensory stimuli with abstract variables is critical for survival; how these associations are learned is, however, poorly understood.

Although we do not know how associations are learned, we do have access to a large number of experimental studies addressing how neural activity evolves while animals learn to classify stimuli into abstract categories [1–4]. Such experiments have probed two kinds of associations between stimuli and categories: fixed associations [3, 5, 6] (in which, for example, stimuli are either in category A or in category B), and flexible ones [4, 7–9] (in which, for example, stimuli are in category A in one context and category B in another).

A consistent finding in these experiments is that single neurons in associative cortex develop selectivity to task-relevant abstract variables, such as category [1, 2, 4, 5], context [7, 8, 10], and often both (so-called mixed-selectivity [11]). In contrast, how the population as a whole responds to abstract variables is less consistent across experiments. In one set of experiments, neurons displayed positive signal correlations with respect to category: neurons that learned to respond strongly to category A also responded strongly to category B, and neurons that learned to respond weakly to A also responded weakly to B [12]. However, in another set of experiments, neurons displayed negative correlations: neurons that learned to respond strongly to category A responded weakly to category B, and vice versa [6, 9, 13]. In [12], furthermore, population tuning was strongly asymmetric: almost all neurons displayed the strongest response to the same category. In other experiments, instead, population tuning was almost symmetric: about the same number of neurons responded strongly to category A as to category B [4, 10]. The positively correlated and asymmetric responses reported in [12] are somewhat surprising: they are at odds with classical findings from sensory cortex, which indicate that sensory features (for example, two opposite orientations) are encoded in negatively-correlated [14, 15] and symmetric [16] responses.

To shed light on these experimental results, we hypothesize that synaptic connectivity in neural circuits evolves by implementing gradient descent on an error function [17]. A large body of work has demonstrated that, under gradient-descent plasticity, neural networks can achieve high performance on both simple and complex tasks [18]. Recent studies have furthermore shown that gradient-descent learning can be implemented, at least approximately, in a biologically plausible way [19–23]. Concomitantly, gradient-based learning has been used to construct network models for a variety of brain regions and functions [24–27].

Motivated by this hypothesis, we study a minimal circuit model that learns through gradient descent to associate sensory stimuli with abstract categories, with a focus on tasks inspired by those used in experimental studies. Via mathematical analysis and simulations, we show that the model can capture the experimental findings discussed above. In particular, after learning, neurons in the model become selective to category and, if present, context; this result is robust, and independent of the details of the circuit and the task. On the other hand, whether correlations after learning are positive or negative, and whether population tuning to different categories is asymmetric or not, is not uniquely determined, but depends on details. We determined how, in the model, these measures are modulated by circuit details (gain and sparsity of activity, relative learning rates, initial connectivity) and task features (number of stimuli, context-dependence). These dependencies make experimentally testable predictions about the underlying circuitry. Overall, the model provides a framework for interpreting seemingly disparate experimental findings, and for making novel experimental predictions.

Results

We consider classification into mutually exclusive abstract classes which, as above, we refer to as category A and category B. We consider two tasks: a simple, linearly-separable one [3, 5, 6] and a context-dependent, non linearly-separable, one [4, 7, 9]. We assume that for both, classification is implemented with a two-layer circuit, as shown in Fig. 1A, and that the synaptic weights evolve via gradient descent. Our goal is to determine how the activity in the intermediate layer evolves with learning, and how this evolution depends on the task and the biophysical details of the circuit. We start by describing, in detail, the model. We then consider circuits that learn the simple, linearly-separable, categorization task, and use numerical and mathematical analysis to characterize how learning drives changes in activity. We then extend the analysis to the context-dependent, non linearly-separable, task.

Circuit model

We consider a simple feedforward circuit as in Fig. 1A. A vector $\boldsymbol{x}$, which models the input from sensory areas, is fed into an intermediate layer of neurons which represents a higher-level, associative area. The activity in that area is given by

$$\boldsymbol{y} = \Psi(\boldsymbol{u} \cdot \boldsymbol{x}). \tag{1}$$

For simplicity, we assume that both the input and the intermediate layers consist of N neurons, so that $\boldsymbol{u}$ is an $N \times N$ connectivity matrix, and we assume N to be large. The output of the circuit is given by a readout neuron, which learns, over time, to predict the category associated with each sensory input. The activity of the readout neuron, z, is taken to be

$$z = \Phi(\boldsymbol{w} \cdot \boldsymbol{y}), \tag{2}$$

where $\boldsymbol{w}$ is an N-dimensional readout vector. The activation functions Φ and Ψ are sigmoidals parametrized by a gain and an offset (Fig. 1B; Methods 2.1) that encapsulate the biophysical properties of single neurons.

Over learning, the goal of the circuit is to adjust the synaptic weights, $\boldsymbol{u}$ and $\boldsymbol{w}$, so that the readout neuron fires at rate $z = z^{\mathsf{A}}$ when the sensory input is associated with category A, and at rate $z = z^{\mathsf{B}}$ when the sensory input is associated with B; for definiteness, we assume $z^{\mathsf{A}} > z^{\mathsf{B}}$. In the simple categorization task, the first half of stimuli is associated with category A, and the second half with B; in the context-dependent one, associations are reversed across contexts. We use $\mathcal{E}(\boldsymbol{u}, \boldsymbol{w})$ to denote the average error between z and its target value, and assume that the synaptic weights evolve, via gradient descent, to minimize the error. If the learning rates are small, the weights evolve according to

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = -\eta_u \frac{\partial \mathcal{E}(\boldsymbol{u}, \boldsymbol{w})}{\partial \boldsymbol{u}} \tag{3a}$$

$$\frac{\mathrm{d}\boldsymbol{w}}{\mathrm{d}t} = -\eta_w \frac{\partial \mathcal{E}(\boldsymbol{u}, \boldsymbol{w})}{\partial \boldsymbol{w}}\,, \tag{3b}$$

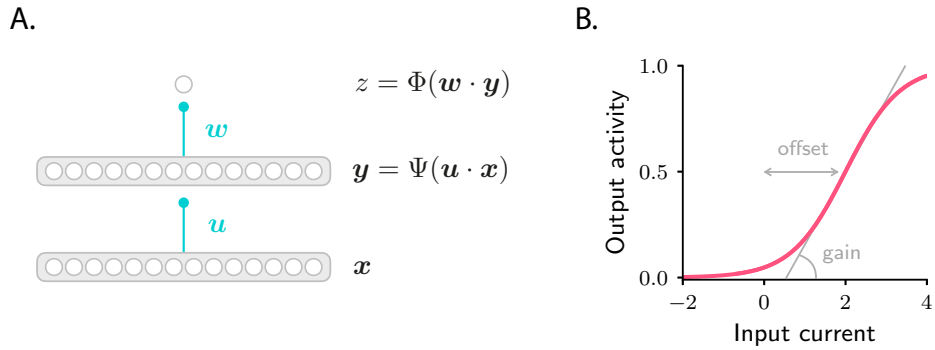

Fig. 1: **Circuit model. A.** The circuit consists of a sensory input layer, $\boldsymbol{x}$, an intermediate layer, $\boldsymbol{y}$, and a readout neuron, z. The intermediate ($\boldsymbol{u}$) and readout weights ($\boldsymbol{w}$) evolve under gradient descent plasticity (Eq. 3). **B.** The activation functions, Ψ and Φ, are taken to be positive sigmoids characterized by a gain and an offset (Methods 2.1). The gain, which controls the sensitivity of activity to inputs, is the slope of the function at its steepest point; the offset, which controls activity sparsity, is the distance from the steepest point to zero.

where t represents learning time and η_u and η_w are learning rates which, for generality, we allow to be different (Fig. 1A; Methods 2.2). Before learning, the weights are random, and chosen so that their amplitude scales as $1/\sqrt{N}$; this ensures that, initially, the activity stays in a reasonable range: it neither vanishes nor saturates when N becomes large. The fact that initial connectivity is random implies that, before learning, activity in the intermediate layer, $\boldsymbol{y}$, is unrelated to abstract categories, and only depends on sensory inputs. As the circuit learns to associate sensory inputs with abstract categories, task-relevant structure emerges in the connectivity matrix, and thus in activity in the intermediate layer as well. Analyzing how activity in the intermediate layer evolves over learning is the focus of this work.

Evolution of activity during the simple categorization task

We first analyze the simple, linearly-separable task, for which we can derive results in a transparent and intuitive form. We then go on to show that similar (although richer) results hold for the context-dependent, non-linearly separable task.

In the simple classification task, each sensory input vector $\boldsymbol{x}^s$ represents a stimulus (for example, an odor, or a visual stimulus), which is associated with one of the two mutually-exclusive categories A and B. Sensory input vectors corresponding to different stimuli are generated at random and assumed to be orthogonal to each other; orthogonality is motivated by the decorrelation performed by sensory brain areas (but this assumption can be dropped without qualitatively changing the main results, see Methods 4.8 and Fig. S5). Since sensory input vectors are orthogonal, they are also linearly separable.

We start our analysis by simulating the circuit numerically, and investigating the properties of neural activity, $\boldsymbol{y}$, in the intermediate layer. To make contact with the current generation of experiments, we focus on activity late in learning, after task acquisition is complete. A common way to characterize the effects of learning on single-neuron activity is through the category selectivity index, a quantity that is positive when activity elicited by within-category stimuli is more similar than activity elicited by across-category stimuli, and negative when the reverse holds. It is defined as [2–4] (Methods 4.3)

$$S_i = \frac{\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff cat}} - \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ same cat}}}{\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff cat}} + \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ same cat}}} \tag{4}$$

where y_i^s represents the activity of neuron i in response to sensory input s, and angle brackets, $\langle \cdot \rangle_{s,s'}$, denote an average over sensory input pairs. The subscript "same cat" refers to averages over the same category (A-A or B-B) and "diff cat" to averages over different categories (A-B).

Before learning, the responses of single neurons to different stimuli are random and unstructured. Thus, responses to stimuli paired with category A are statistically indistinguishable from responses to stimuli paired with category B (Fig. 2A). This makes the category selectivity index zero on average (Fig. 2B). After learning, the responses of single neurons depend on category: within-category responses become more similar than across-category

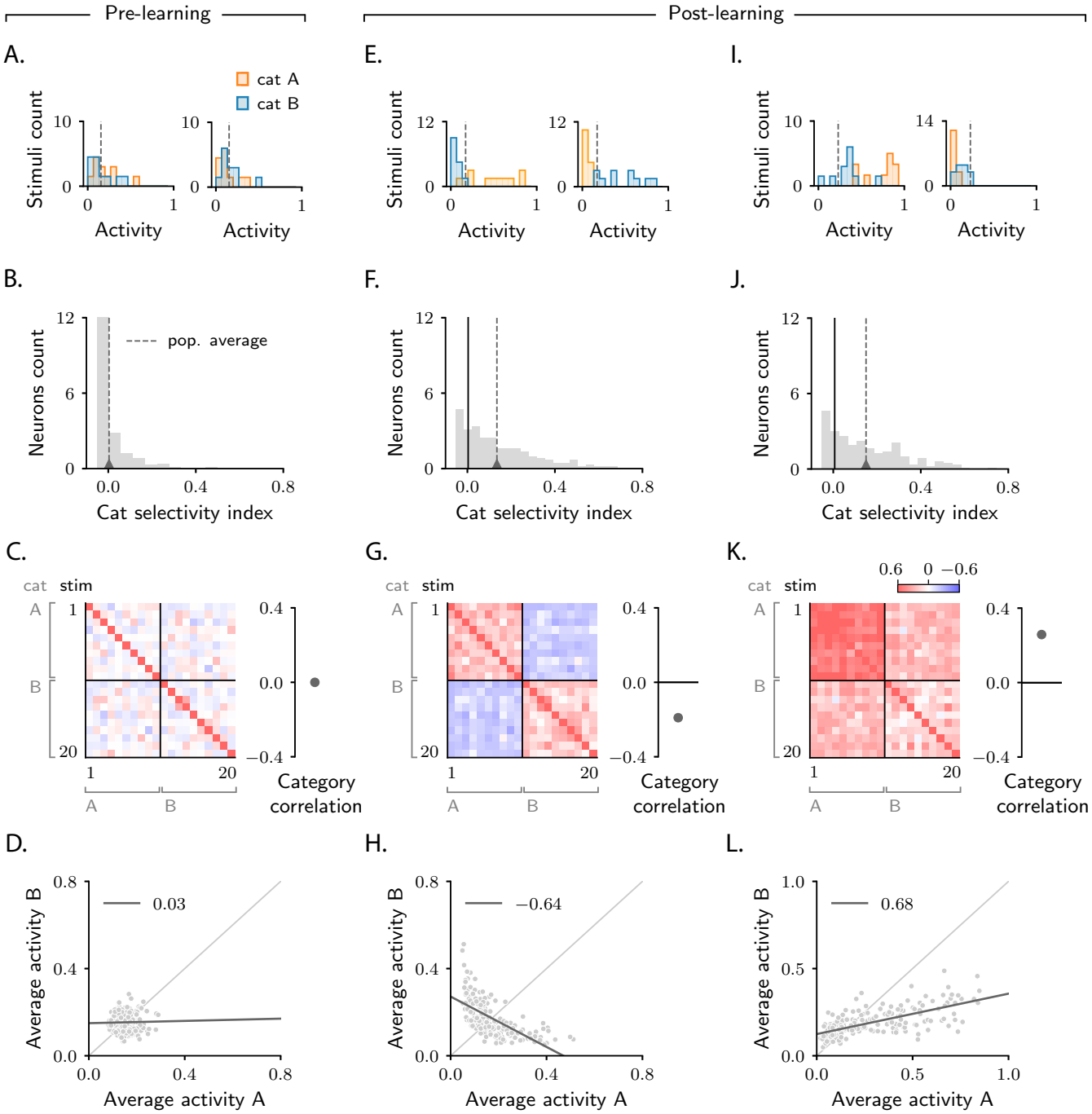
Pre-learning
Post-learning
A.
E.
I.
cat A
cat B
Stimuli count
Activity
B.
F.
J.
Neurons count
pop. average
Cat selectivity index
C.
G.
K.
cat
stim
Category correlation
D.
H.
L.
Average activity B
Average activity A
0.03
−0.64
0.68

Fig. 2 *(previous page)*: **Characterization of activity evolution during the simple categorization task; results from simulations.** The first column (A-D) shows a naive circuit (pre-learning); the second (E-H) and third (I-L) columns show two trained circuits (post-learning), characterized by different sets of parameters (see below). **A**, **E**, **I.** Histograms of single-neuron activity in response to stimuli associated with category A (orange) and category B (blue). Left and right show two sample neurons from the intermediate layer. Grey dashed lines indicate the average activity across the population. **B**, **F**, **J.** Histogram of category selectivity (Eq. 4) across the population of neurons in the intermediate layer. Grey dashed lines indicate the average selectivity across the population. In panels F and J, the black vertical lines indicate the initial value of average selectivity. **C**, **G**, **K.** Signal correlation matrices. Each entry shows the Pearson correlation coefficient between activity elicited by different stimuli; in these examples, we used 20 stimuli. Category correlation (namely, the average of the correlations within the off-diagonal blocks, which contain stimuli in different categories) is shown on the right of the matrices. In panels G and K, the black horizontal lines near zero indicate the initial values of category correlation. **D**, **H**, **L.** Population responses to categories A and B. Each dot represents a neuron in the intermediate layer, with x- and y-axes showing the responses to stimuli associated with categories A and B, respectively, averaged over stimuli. Grey line: linear fit, with Pearson correlation coefficient shown in the figure legend. Parameters are summarized in Table 1 (Methods 6.2).

responses, resulting in two separate distributions (Fig. 2E). As a consequence, the category selectivity index for each neuron increases; correspondingly, average selectivity increases from zero to positive values (Fig. 2F), thus reproducing the behaviour observed in experimental studies [2–4]. To determine whether this effect is robust, we varied the parameters that describe the task (number of stimuli) and the biophysical properties of the circuit (the gain and offset of the activation functions, Fig. 1B, and the learning rates of the two sets of synaptic weights, η_u and η_w). We found that the selectivity increase is a universal property – it is observed in all circuit models that successfully learned the task, independent of the parameters. Activity from a second example circuit is shown in Figs. 2I-J; additional simulations are shown in Fig. S1A.

Category selectivity tells us about the behavior of single neurons. But how does the population as a whole change its activity over learning? To quantify that, we compute signal correlations, defined to be the Pearson correlation coefficient between the activity elicited by two different stimuli [6]. Results are summarized in the correlation matrices displayed in Figs. 2C, G and K. The task involves 20 stimuli, so the correlation matrix is 20×20; stimuli are sorted according to category, with the first half of stimuli being paired with category A, and the second half with category B.

As discussed above, before learning the responses of neurons in the intermediate layer are random and unstructured. In particular, activity in response to different stimuli is uncorrelated; this is illustrated in Fig. 2C, where all non-diagonal entries of the correlation matrix are close to zero. We are interested in particular in the off-diagonal blocks of the matrix, which corresponds to pairs of activity vectors elicited by stimuli in different categories. The average of those correlations, which we refer to as category correlation, is shown on the right panels; before learning, the category correlation is close to zero (Fig. 2C). Over learning, the correlation matrices develop structure. Correlations become different within the two diagonal, and the two off-diagonal blocks, indicating that learning induces category-dependent structure. In Fig. 2G, the average correlation within the off-diagonal blocks is negative; the category correlation is thus negative [6, 9, 13]. The model does not, however, always produce negative correlation: varying model details – either the parameters of the circuit or the number of stimuli – can switch the category correlation from negative to positive [12]; one example is shown in Fig. 2K.

To understand how the population responds when category correlation is negative versus positive, we consider each neuron in the intermediate layer, and plot its average response to stimuli associated with category B (y-axis) versus A (x-axis). Before learning, activity is unstructured, and the dots form a random, uncorrelated cloud (Fig. 2D). After learning, the shape of this cloud depends on category correlation. In Fig. 2H, where the category correlation is negative, the cloud has a negative slope. This is because changes in single-neuron responses to categories A and B have opposite sign: a neuron that increases its activity in response to category A decreases its activity in response to category B (Fig. 2E left), and vice versa (Fig. 2E right). In Fig. 2L, where the category correlation is positive, the cloud has, instead, a positive slope. Here, changes in single-neuron responses to categories A and B have the same sign: a neuron that increases its activity in response to category A also increases its activity in response to category B (Fig. 2I left), and similarly for a decrease (Fig. 2I right).

Negative versus positive slope is not the only difference between Figs. 2H and L: they also differ in symmetry

with respect to the two categories. In Fig. 2H, about the same number of neurons respond more strongly to category A than to category B [4]. In Fig. 2L, however, the number of neurons that respond more strongly to category A is significantly larger than the number of neurons that respond more strongly to B [12]. Furthermore, as observed in [12], the mean population activity in response to category A is larger than to category B, and the range of activity in response to A is larger than to B. The fact that the population response to A is larger than to B is not a trivial consequence of having a larger target for the readout neuron in response to A than to B ($z^{\mathsf{A}} > z^{\mathsf{B}}$): as shown in Figs. S2B and D, example circuits displaying larger responses to B can also be observed. Response asymmetry is discussed in detail in Methods 4.6.

In sum, we simulated activity in circuit models that learn to associate sensory stimuli to abstract categories via gradient descent synaptic plasticity. We observed that single neurons consistently develop selectivity to abstract categories – an observation that is robust to model details. How the population of neurons responds to category depended, however, on model details: we observed both negatively correlated, symmetric responses and positively correlated, asymmetric ones. All of these results are in agreement with experiments. What are the mechanisms that drive these activity changes? And how do the circuit and task details determine how the population responds? In the next section, we address these questions quantitatively, by performing mathematical analysis of the model.

Analysis of the simple categorization task

For the nonlinear circuit model we consider, the evolution of activity during learning can, in general, be arbitrarily complex. However, if the number of neurons in each layer, N, is much larger than the number of sensory inputs to classify – a regime that is relevant to the circuits and tasks we study here – things simplify considerably. In this regime, the number of synaptic weights that the circuit can tune is very large, and so a small change in each weight is sufficient to learn the task. Small changes in the weights make the circuit amenable to mathematical analysis, and allow us to write down approximate expressions for changes in activity during learning – an observation that is in line with a broad set of recent machine-learning studies [28–31]. Full details are reported in Methods 3, 4 and 5; here we discuss the main results.

We start with the simple categorization task illustrated in the previous section, and use the framework developed in Methods to shed light on the simulations described above (Fig. 2). Figure 3A shows, schematically, activity before learning. Each dot represents activity in the intermediate layer in response to sensory input $\boldsymbol{x}^s$ (recall that the superscript s labels sensory inputs). Orange and blue circles indicate activity in response to stimuli associated with categories A and B, respectively. Before learning, activity is determined solely by sensory inputs, which consist of random, orthogonal vectors. Consequently, the initial activity vectors form an unstructured cloud in activity space, with orange and blue circles intermingled (Fig. 3A).

Over learning, activity in the intermediate layer acquires structure. In particular, activity vectors acquire a component that is mainly aligned with a common, stimulus-independent direction. Activity after learning can thus be written

$$\boldsymbol{y}^s \simeq \boldsymbol{y}_0^s + c^s \boldsymbol{d} \tag{5}$$

where $\boldsymbol{y}_0^s$ indicates initial activity, and $\boldsymbol{d}$ indicates the common direction along which activity acquires structure. For reasons that will be clarified in a moment, the coefficients c^s, which measure the strength of the components along the common direction $\boldsymbol{d}$, are mainly determined by category: they are approximately equal to c^{A} if the sensory input $\boldsymbol{x}^s$ is associated with category A, and c^{B} otherwise. Consequently, over learning, activity vectors associated with different categories are pushed apart along $\boldsymbol{d}$; this is illustrated in Figs. 3B and C, which show activity for the two circuits analyzed in the second and third column of Fig. 2. Activity vectors form two distinct clouds along $\boldsymbol{d}$, one for each category; the centers of the two clouds along $\boldsymbol{d}$ are given, approximately, by c^{A} and c^{B}.

The fact that activity clusters by category tells us immediately that the category selectivity index of single neurons increases over learning, as observed in simulations (Figs. 2F and J). To see this quantitatively, note that from the point of view of a single neuron, i, Eq. 5 reads

$$y_i^s \simeq y_{0,i}^s + c^s d_i. \tag{6}$$

Since c^s is category-dependent, while d_i is fixed, the second term in the right-hand side of Eq. 6 separates activity associated with different categories (Figs. 2E and I), and implies an increase in the category selectivity index (Eq. 4; Figs. 2F and J). Importantly, and in agreement with our numerical findings, the generality of Eq. 6 indicates that the increase in selectivity is a robust byproduct of learning, and so can be observed in any circuit that learns the categorization task, regardless of model details.

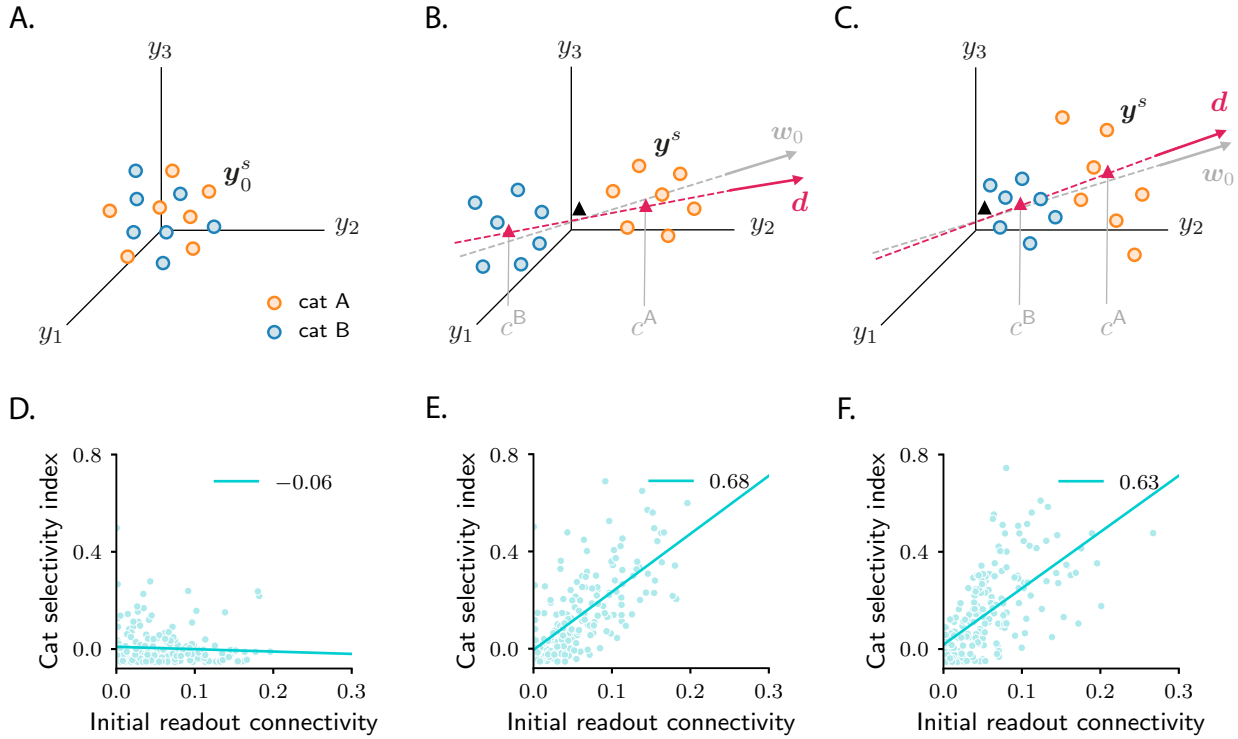

Fig. 3: **Analysis of activity evolution during the simple categorization task.** The three columns are as in Fig. 2: pre-learning (first column) and post-learning for two different circuits (second and third columns). **A**, **B**, **C.** Cartoons illustrating how activity evolves over learning. Circles show activity in the intermediate layer, in response to different stimuli, displayed in a three-dimensional space where axes correspond to the activity of three sample neurons. Orange and blue circles are associated, respectively, with category A and B. Before learning, activity is unstructured (panel A). After learning (panels B and C), the activity vectors develop a component along the common direction $\boldsymbol{d}$ (Eq. 5), shown as a magenta line, and form two clouds, one for each category. The centers of those clouds are indicated by magenta triangles; their positions along $\boldsymbol{d}$ are given, approximately, by c^{A} and c^{B}. The black triangle indicates the center of initial activity. In panel B, c^{A} and c^{B} have opposite sign, so the clouds move in opposite directions with respect to initial activity; in panel C, c^{A} and c^{B} have the same sign, so the clouds move in the same direction. For illustration purposes, we show a smaller number of stimuli (14, instead of 20) than in Fig. 2. Simulated data from the circuits displayed in Fig. 2 are shown in Fig. S1B. **D**, **E**, **F.** Category selectivity as a function of the initial readout connectivity $w_{0,i}$ (in absolute value). Each dot represents a neuron in the intermediate layer. Cyan line: linear fit, with Pearson correlation coefficient shown in the figure legend.

While category selectivity increases for all neurons, the amount it increases depends on d_i, and so can vary markedly across the population. This explains the broad distribution of category selectivity indices in Figs. 2F and J, which reproduces the variability commonly seen in experiments [3–5]. To predict which neurons become the most selective, we need an explicit expression for the direction of category clustering, $\boldsymbol{d}$. We show in Methods 4.2 (Eq. 60) that, for the current task, the category direction $\boldsymbol{d}$ is approximately aligned with the vector that specifies connectivity between the intermediate and the readout neurons, $\boldsymbol{w}$, before learning starts. Denoting this vector $\boldsymbol{w}_0$, we can thus write

$$\boldsymbol{d} \simeq \rho \boldsymbol{w}_0, \tag{7}$$

where ρ is a positive constant. To provide intuition for why Eq. 7 holds, consider for a moment the extreme scenario in which the readout weights, $\boldsymbol{w}$, evolve much more slowly than the intermediate weights, $\boldsymbol{u}$ ($\eta_w \ll \eta_u$), so that the readout weights remain approximately equal to $\boldsymbol{w}_0$ over learning. Under that scenario, we see from Eq. 2 that the output of the network only depends on $\boldsymbol{w}_0 \cdot \boldsymbol{y}^s$, i.e. the projection of activity along the initial readout weights $\boldsymbol{w}_0$. Over learning, activity thus must acquire a component along $\boldsymbol{w}_0$, as implied by Eqs. 5 and 7; the magnitude

of that component is fixed by the task requirement

$$\Phi(\boldsymbol{w}_0 \cdot \boldsymbol{y}^s) = \begin{cases} z^{\mathsf{A}}, & \text{if } \boldsymbol{x}^s \text{ is associated with category } \mathsf{A} \\ z^{\mathsf{B}}, & \text{if } \boldsymbol{x}^s \text{ is associated with category } \mathsf{B}\,. \end{cases} \tag{8}$$

This simple argument provides a qualitative explanation for Eq. 7 when the readout weights do not change. However, Eq. 7 is general, and also holds when the intermediate and the readout weights evolve with similar rates; details are provided in Methods 4.2.

Equation 7 tells us that the component d_i that appears in Eq. 6 is proportional to $w_{0,i}$. Consequently, only neurons that are initially strongly connected to the readout neuron – that is, neurons for which $w_{0,i}$ is large – exhibit a large selectivity index (Figs. 3D-F). In terms of biological circuits, this finding indicates that changes in selectivity are determined by the synaptic connections a neuron makes, before learning, to downstream readout areas. Experiments consistent with this finding have been reported recently: studies in rodents prefrontal cortex [32, 33] found that neurons which were highly selective to a given abstract variable preferentially projected to the same downstream area. These experiments would provide evidence for our model if two conditions were met. First, neurons in the downstream area should behave as readout neurons, meaning after learning their activity should depend only on the abstract variable, and not on the sensory stimuli. Second, the strength of the synaptic connections that neurons make to downstream neurons should correlate with selectivity (Figs. 3E-F). Both could be tested with current experimental technology.

Correlations reflect circuit properties and task difficulty

The above analysis focused on category selectivity, a single-neuron property. We now analyze population responses. We saw in Fig. 2 that, while the behaviour of category selectivity is consistent across all circuit models, the behaviour of population responses is not: over learning, responses can become negatively correlated and symmetric (Figs. 2G-H), or positively correlated and asymmetric (Figs. 2K-L). The reason is illustrated, qualitatively, in Figs 3B and C. In Fig. 3B, the centers of the category clouds along $\boldsymbol{d}$, c^{A} and c^{B}, have, respectively, a positive and a negative sign. As a consequence, the two clouds move in opposite directions relative to the initial cloud of activity (black triangle). The population thus develops, over learning, negative category correlation (Fig. 2G-H): if the activity of a given neuron increases for one category, it decreases for the other, and viceversa. Furthermore, if c^{A} and c^{B} have similar magnitude (which is the case for Fig. 2G-H, and circuits developing strong negative correlations), activity changes for the two categories have similar amplitude, making the response to categories A and B approximately symmetric. In Fig. 3C, on the other hand, c^{A} and c^{B} are both positive; clouds associated with the two categories move in the same direction relative to the initial cloud of activity. This causes the population to develop positive category correlation (Fig. 2K-L): if the activity increases for one category, it also increases for the other, and similarly for a decrease. Because the magnitude of c^{A} is larger than c^{B}, activity changes for category A are larger than for B, making the response to categories A and B asymmetric.

We conclude that whether negative or positive category correlation emerges depends on the relative signs of c^{A} and c^{B}. To understand, qualitatively, why c^{A} and c^{B} can have either opposite or identical sign, consider again the extreme scenario in which the readout weights, $\boldsymbol{w}$, change much more slowly than the intermediate weights, $\boldsymbol{u}$, and thus Eq. 8 holds. To compute $\boldsymbol{w}_0 \cdot \boldsymbol{y}^s$, we combine Eqs. 5 and 7 to yield

$$\boldsymbol{w}_0 \cdot \boldsymbol{y}^s \simeq \boldsymbol{w}_0 \cdot \boldsymbol{y}_0^s + \rho c^s \boldsymbol{w}_0 \cdot \boldsymbol{w}_0 \simeq \rho c^s\,. \tag{9}$$

We neglected the first term, $\boldsymbol{w}_0 \cdot \boldsymbol{y}_0^s$, because initial activity is unstructured, and thus this term yields random contributions; to simplify the second term, we used the fact that $\boldsymbol{w}_0$ has unit norm. Inserting Eq. 9 into Eq. 8, we obtain $\Phi(\rho c^{\mathsf{A}}) \simeq z^{\mathsf{A}}$ and $\Phi(\rho c^{\mathsf{B}}) \simeq z^{\mathsf{B}}$: the coordinates of activity centers, c^{A} and c^{B}, are proportional to the input current needed by the readout neuron to produce the target activity, z^{A} and z^{B}. For fixed target output, the sign and the amplitude of c^{A} and c^{B} are set by the activation function Φ (Fig. 4A). For instance, for the pink activation function in Fig. 4A (which corresponds to the activation function used in Figs. 2G-H), c^{A} and c^{B} have opposite sign. For the grey function, instead (which corresponds to the activation function used in Figs. 2K-L), c^{A} and c^{B} have identical sign. In general, changing Φ via its gain and offset modifies the value of c^{A} and c^{B}, and can change their relative sign.

The simplified scenario in which the readout weights do not change with learning provides a qualitative explanation for the observation that category correlation can become either positive or negative over learning. This phenomenon is general, though, and applies also when the intermediate and readout weights evolve with similar

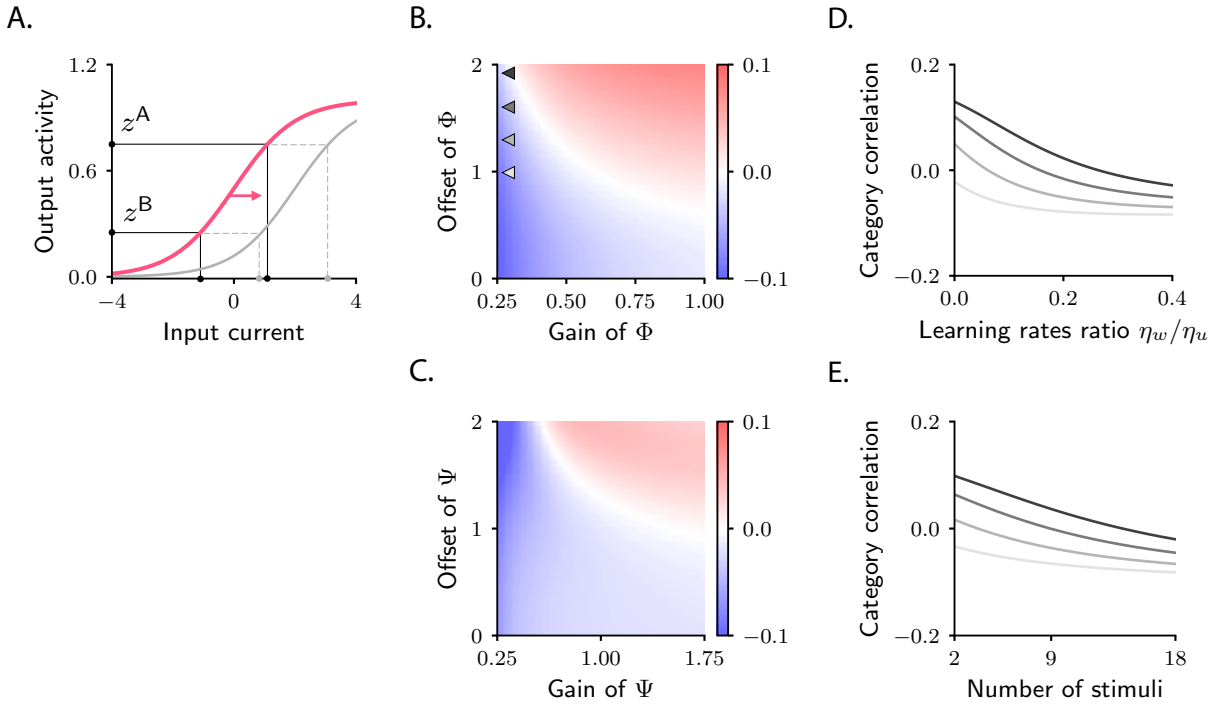

Fig. 4: **Category correlation is modulated by circuit and task details. A.** In the simplified scenario described in the main text ($\eta_w \ll \eta_u$), the coordinates of the activity centers c^{A} and c^{B} are proportional to the input current needed by the readout neuron to to produce the target activity (see text). For fixed targets z^{A} and z^{B}, those depend on the activation function Φ. Increasing the offset of Φ (which is equivalent to changing the drive to the readout neuron) can cause c^{A} and c^{B} to shift from having opposite sign to having the same sign. This in turn can cause a shift from negative to positive category correlation. **B.** Category correlation as a function of the gain and offset of the activation function of the readout neuron, Φ. Grey arrows indicate the offset and gain that are used in panels D and E. The learning rate ratio, η_w/η_u, is set to 0.4 here and in panel C and E. **C.** Category correlation as a function of the gain and offset of the activation function of neurons in the intermediate layer, Ψ; details as in panel B. **D.** Category correlation as a function of the learning rate ratio. The gain and offset of Φ are given by the triangles indicated in panel B, matched by color. **E.** Category correlation as a function of number of stimuli; same color code as in panel D. In all panels, correlations were computed from the approximate theoretical expression given in Methods 4.4 (Eq. 79). Parameters are summarized in Table 1 (Methods 6.2).

rates. In that case, the signs of c^{A} and c^{B} can still be either the opposite or the same (Methods 4.2), and thus both negative and positive correlations can be observed (Methods 4.4, Eq. 79).

In Fig. 4B, we show category correlation as a function of the gain and offset of the readout activation function, Φ. Not surprisingly, given Fig. 4A, we find that the parameters controlling the activation function Φ play an important role in determining the magnitude and sign of correlations. Specifically, for the target values z^{A} and z^{B} we have chosen here, positive correlations are favoured by large gain and offsets. Perhaps less intuitively, category correlation is also affected by the input-output properties of neurons in the intermediate layer, which are determined by the activation function Ψ. This can be seen in Fig. 4C, which shows that larger values of gain and offset tend to favour positive correlation. An equally important role is played by the relative learning rates of the the readout, $\boldsymbol{w}$, and the intermediate weights, $\boldsymbol{u}$. As illustrated in Fig. 4D, increasing the ratio of the learning rates, η_w/η_u, causes the correlation to decrease. Overall, these results indicate that category correlation depends on multiple biophysical aspects of the circuit, which in turn are likely to depend on brain areas. This suggests that correlation can vary across brain areas, which is in agreement with the observation that positive correlations reported in [12] from the lateral intraparietal area of monkeys are robust across experiments, but inconsistent with the correlations observed in different brain regions [6].

Finally, our analysis shows that category correlation also depends on a property of the task; specifically, on the total number of stimuli (Methods 4.4, Eq. 82). This is illustrated in Fig. 4E, which shows that increasing the

number of stimuli causes a systematic decrease in correlation. The model thus predicts that increasing the number of stimuli pushes category correlation from positive to negative values. This is in agreement with the finding that negative correlations are typically observed in sensory cortex, as well as machine-learning models trained on benchmark datasets [34] – that is, in cases where the number of stimuli is much larger than in the current task.

In sum, we analyzed activity in the intermediate layer of circuits that have learned the simple categorization task. We found that activity changes are mainly aligned along a common, stimulus-independent direction (Eq. 5), which is approximately given by the initial readout vector $\boldsymbol{w}_0$ (Eq. 7). Activity vectors associated with different categories develop two distinct clouds along this direction – a fact that explains the increase in category selectivity observed in Figs. 2F and J. We also found that the sign of the category correlation depends on both circuit (activation functions of neurons in the intermediate and readout layers, and relative learning rates) and task (number of stimuli) properties. Modifying any of these can change the direction the clouds of activity move along $\boldsymbol{w}_0$, which in turn changes the sign of category correlation, thus explaining the different behaviours observed in Figs. 2G-H and K-L.

Evolution of activity during the context-dependent categorization task

Motivated by experimental studies [4, 7–9, 35], we now consider a more complex categorization task. Here, stimuli-category associations are not fixed, but context-dependent: stimuli that are associated with category A in context 1 are associated with category B in context 2, and vice versa. The two contexts are signaled by two distinct sets of context cues (for example, two different sets of visual stimuli) [7, 8]. As in the simple task, each stimulus and each context cue are associated with a sensory input vector; those are generated at random and taken to be orthogonal to each other. On every trial, one stimulus and one context cue are presented; the corresponding sensory inputs are combined linearly to yield the total sensory input vector $\boldsymbol{x}^s$ (Methods 5.1). This task is a generalization of the classical XOR task where, rather than just two stimuli and two context cues, there are more than two of each [35]. As in the XOR task, sensory inputs are not linearly separable.

We are again interested in understanding how activity in the intermediate layer, $\boldsymbol{y}$, evolves over learning. We start by investigating activity via simulations (Fig. 5). As in Figs. 2B, F and J, we first measure category selectivity (Eq. 4). Before learning, activity is characterized by small selectivity, which is weakly negative on average (Fig. 5A; the fact that average category selectivity is initially weakly negative is due to the composite nature of inputs for this task, see Methods 5.7). Over learning, the average category selectivity increases (Fig. 5D). We tested the robustness of this behaviour by varying the parameters that control both the circuit (activation functions, learning rates) and task (number of stimuli and context cues). As in the simple task, we found that the average category selectivity increases in all circuit models, regardless of the parameters (Figs. 5G and S6A).

While in the simple task we could only investigate the effect of category on activity, in this task we can also investigate the effect of context. For this we measure context selectivity which, analogously to category selectivity, quantifies the extent to which single-neuron activity is more similar within than across contexts (Methods 5.3, Eq. 127). Context selectivity is shown in Figs. 5B and E. We find, as we did for category selectivity, that average context selectivity increases over learning – a behaviour that is in agreement with experimental findings [7, 8]. The increase in context selectivity is, as for category, highly robust, and does not depend on model details (Figs. 5H and S6A).

Finally, we analyze signal correlations; these are summarized in the correlation matrices displayed in Figs. 5C, F and I. As we used 8 stimuli and 8 context cues, and all stimuli-context cues combinations are permitted, each correlation matrix is 64×64. Trials are sorted according to context cue first and stimulus second; with this ordering, the first half of trials corresponds to context 1 and the second half to context 2, and the off-diagonal blocks are given by pairs of trials from different contexts.

Figure 5C shows the correlation matrix before learning. Here the entries in the correlation matrix are fully specified by sensory inputs, and can take only three values: large (dark red), when both the stimuli and the context cues are identical across the two trials; intermediate (light red), when the stimuli are identical but the context cues are not, or vice versa; and small (white), when both stimulus and context cues are different. Figures 5F and I show correlation matrices after learning for two circuits characterized by different parameters. As in the simple task, the matrices acquire additional structure during learning, and that structure can vary significantly across circuits (Figs. 5F and I). To quantify this, we focus on the off-diagonal blocks (pairs of trials from different contexts) and measure the average of those correlations, which we refer to as context correlation. Context correlation behaves differently in the two circuits displayed in Fig. 5F and I: it decreases over learning in Fig. 5F, whereas it increases in Fig. 5I. Thus, as in the simple task, the behaviour of correlations is variable across circuits. This variability

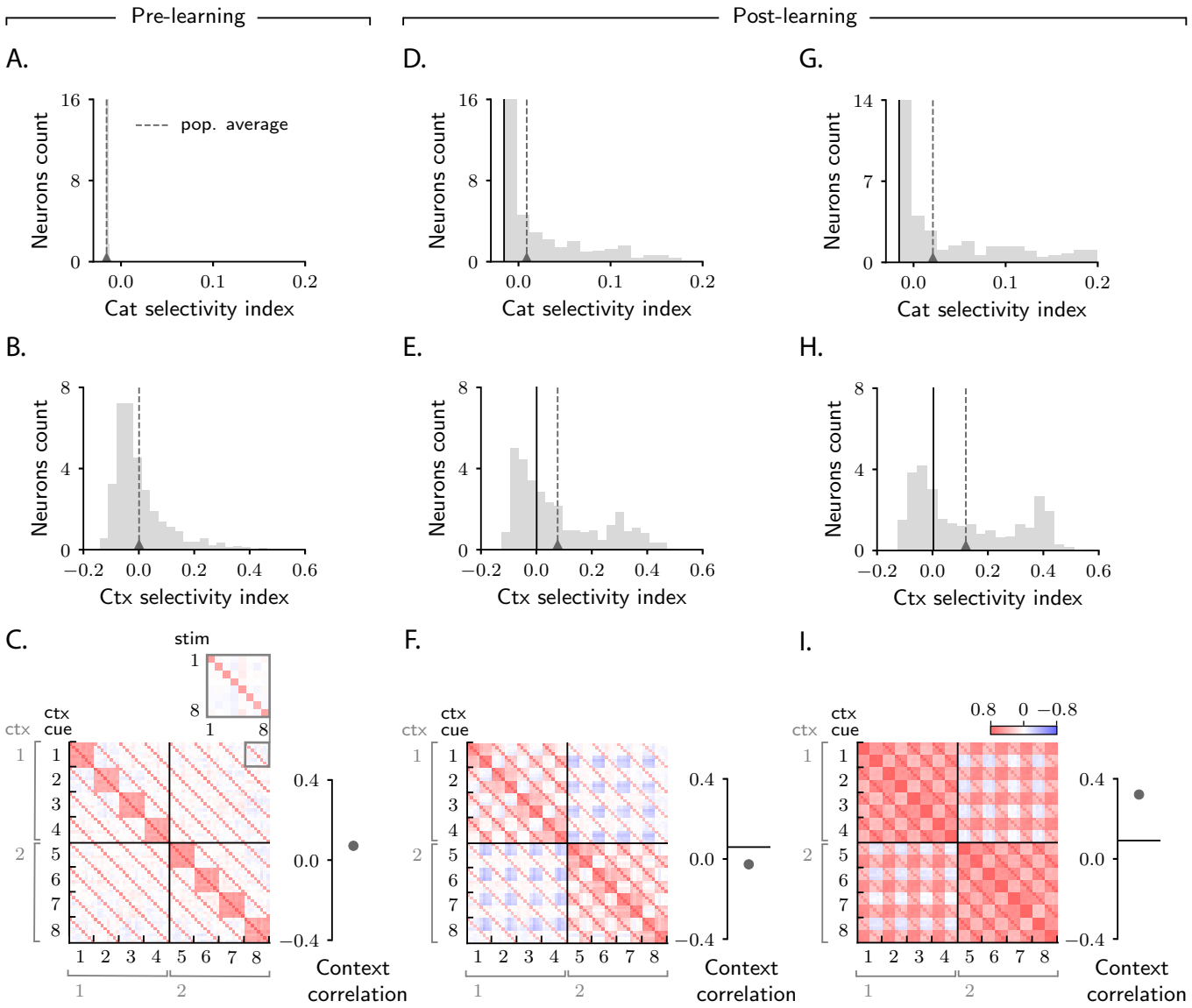

Fig. 5: **Characterization of activity evolution during the context-dependent categorization task; results from simulations**. The first column (A-C) shows a naive circuit (pre-learning); the second (D-F) and third (G-I) columns show two trained circuits (post-learning), characterized by different sets of parameters. **A**, **D**, **G.** Histogram of category selectivity (Eq. 4) across the population of neurons in the intermediate layer. Grey dashed lines indicate the average selectivity across the population. In panels D and G, the black vertical lines indicate the initial value of average selectivity. **B**, **E**, **H.** Histogram of context selectivity (Methods 5.3, Eq. 127), details as in A, D and G. **C**, **F**, **I.** Correlation matrices. Each entry shows the Pearson correlation coefficient between activity from different trials. There are 8 stimuli and 8 context cues, for a total of 64 trials (i.e., 64 stimulus/context cue combinations). The inset on the top of panel C shows, as an example, a magnified view of correlations among trials with context cues 1 and 8, across all stimuli (1 to 8). To the right of the matrices we show the context correlation, defined to be the average of the correlations within the off-diagonal blocks (trials in different contexts). In panels F and I, the black horizontal lines indicate the initial value of context correlation. Parameters are summarized in Table 1 (Methods 6.2).

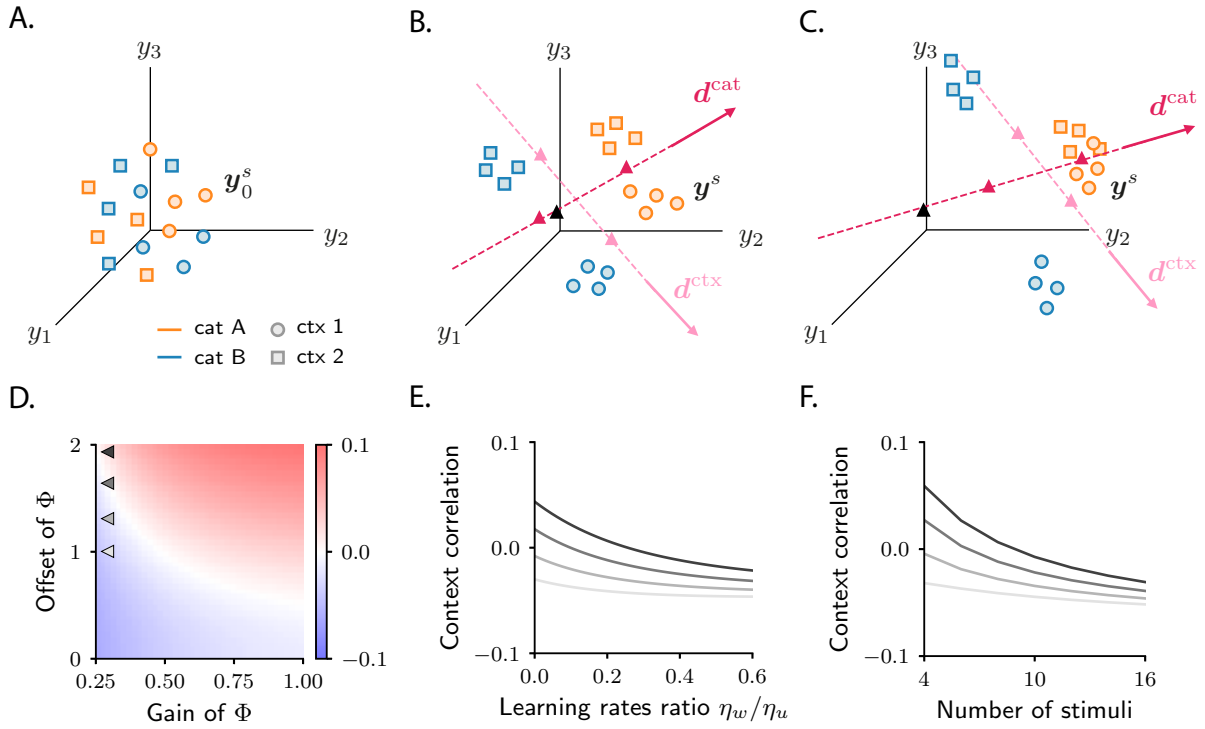

Fig. 6: **Analysis of activity evolution during the context-dependent categorization task. A**, **B**, **C.** Cartoons illustrating how activity evolves over learning. Orange and blue symbols are associated with categories A and B, respectively; circles and squares are associated with contexts 1 and 2. Before learning, activity is mostly unstructured (panel A). After learning, activity forms four clouds, one for each combination of category and context. The center of the activity vectors associated with categories A and B and context 1 and 2 are indicated, respectively, by magenta and pink triangles. The black triangle indicates the center of initial activity. The cartoons in panels A-B-C refer to the three circuits illustrated in the three columns of Fig. 5; for illustration purposes, we show a reduced number of stimuli and context cues (4 instead of 8). Simulated data from the circuits displayed in Fig. 5 are shown in Fig. S6C. **D.** Change in context correlation after learning as a function of the gain and offset of Φ. Grey arrows indicate the offset and gain that are used in panels E and F. **E.** Change in context correlation over learning as a function of the ratio of learning rates in the two layers. **F.** Change in context correlation over learning as a function of the number of stimuli. Correlations in panels D-F were computed from the approximate theoretical expression given in Methods 5.4 (Eq. 132b). Parameters are given in Table 1 (Methods 6).

is not restricted to context correlation: as in the simple task, category correlation is also variable (Fig. S6A); furthermore, the population response to categories A and B can be symmetric or asymmetric depending on model details (Figs. S7A-B).

Analysis of the context-dependent categorization task

To uncover the mechanisms that drive those learning-induced activity changes, we again analyse the circuit mathematically. The addition of context makes the analysis considerably more complicated than for the simple task; most of the details are thus relegated to Methods 5, here we discuss the main results.

Figure 6A shows activity before learning. Each shape (circle or dot) represents activity in a given trial, and is associated with a category (A, orange; B, blue) and a context (1, circles; 2, squares). Before learning, activity is mostly unstructured (Figs. 6A, Methods 5.7); over learning, though, it acquires structure (Figs. 6B and C). As in the simple task (Figs. 3B-C), activity vectors get re-arranged into statistically-distinguishable clouds. While in the simple task clouds were determined by category, here each cloud is associated with a combination of category and context. As a result, four clouds are formed: the cloud of orange circles corresponds to category A and context 1; orange squares to category A and context 2; blue circles to category B and context 1; and blue squares to category

B and context 2.

The transition from unstructured activity (Fig. 6A) to four clouds of activity (Figs. 6B-C) occurs by learning-induced movement along two directions: $\boldsymbol{d}^{\text{cat}}$, which corresponds to category, and $\boldsymbol{d}^{\text{ctx}}$, which corresponds to context. Activity vectors in different categories move by different amounts along $\boldsymbol{d}^{\text{cat}}$; this causes the orange and blue symbols in Figs. 6B-C to move apart, so that activity vectors associated with the same category become closer than vectors associated with opposite categories. As in the simple task, this in turn causes the category selectivity to increase, as shown in Figs. 5D and G (Methods 5.7). Similar learning dynamics occur for context: activity vectors from different contexts move by different amounts along $\boldsymbol{d}^{\text{ctx}}$. This causes the squares and circles in Figs. 6B-C to move apart, so that activity vectors from the same context become closer than vectors from different contexts. Again, not surprisingly, this in turn causes the context selectivity to increase, as shown in Figs. 5E and H (Methods 5.6). Mathematical analysis indicates that the increase in clustering by category and contest is independent of model parameters (Fig. S6B), which explains the robustness of the increase in selectivity observed in simulations.

What is the origin of the structure seen in Figs. 6B and C? In the simple task, category clustering emerged because activity vectors associated with the same category had to produce the same readout: along the direction specified by the readout vector, which was close to $\boldsymbol{w}_0$, activity had to cluster by category (see Eq. 8, and surrounding text). This reasoning does not explain, however, why in Figs. 6B-C activity also clusters by context. In contrast to category, context is not read out. Consequently, the emergence of context-related structure in activity has no direct impact on the circuit output, z, and on the error that the circuit is learning to minimize. The fact that activity clusters by context, and thus context selectivity increases, is therefore a direct signature of the specific learning rule used by the circuit (i.e., gradient descent [36]), rather than a signature of the task. The mechanism through which context selectivity increases is described in detail in Methods 5.6.

While clustering by category and context is robust across circuits, the position of clouds in the activity space is not. As in the simple task, the variability in cloud position explains the variability in context correlation (although the relationship between clouds position and correlations is more complex in this case, see Methods 5). In Figs. 6D-F, we show how context correlation depends on model parameters. This dependence is qualitatively similar to the simple task: context correlation depends on details of the activation functions (compare Figs. 6D and 4B), on the relative learning rate η_w/η_u (compare Figs. 6E and 4D), and on the number of stimuli (compare Figs. 6F and 4E). However, we find that the region of parameters leading to an increase in correlation shrinks substantially compared to the simple task (Fig. S7C, see also Methods 5.2); this is in line with the observation that correlations decrease to negative values when the complexity of the task increases, as shown in Fig. 4E.

Initial activity shapes populations of pure- and mixed-selective neurons

As a final step, we take a closer look at single-neuron selectivity. Analysis from the previous sections indicates that the average selectivity to both category and context increases over learning. And, as in the simple task, the increase is highly variable across neurons (Figs. 5D-E and G-H). To determine which neurons become the most selective to category and context, we analyze the directions along which clustering to category and context occurs, $\boldsymbol{d}^{\text{cat}}$ and $\boldsymbol{d}^{\text{ctx}}$ (Fig. 6B-C). In analogy with the simple task, neurons that strongly increase selectivity to category are characterized by a large component along the category direction $\boldsymbol{d}^{\text{cat}}$; similarly, neurons that strongly increase selectivity to context are characterized by a large component along the context direction $\boldsymbol{d}^{\text{ctx}}$ (Figs. S8A-B).

Analysis in Methods 5.8 shows that both the category and context directions, $\boldsymbol{d}^{\text{cat}}$ and $\boldsymbol{d}^{\text{ctx}}$, are strongly correlated with the initial readout vector $\boldsymbol{w}_0$. As in the simple task, this leads to the prediction that neurons that strongly increase selectivity to either category or context are, before learning, strongly connected to the downstream readout neuron (Fig. 7A).

Although $\boldsymbol{d}^{\text{cat}}$ and $\boldsymbol{d}^{\text{ctx}}$ are both correlated with $\boldsymbol{w}_0$, they are not perfectly aligned (Methods 5.8). In principle, then, for a given neuron (here, neuron i), both d_i^{cat} and d_i^{ctx} could be large (implying mixed selectivity to both abstract variables, category and context), or only one could be large (implying pure selectivity to only one abstract variable, category or context). Evidence for both mixed [11] and pure [33] selectivity has been reported in the brain. While all combinations are possible in principle, in the model they do not all occur. In Fig. 7B, we plot changes in context selectivity as a function of changes in category selectivity. We observe that, among all the neurons that strongly increase their selectivity, some increase selectivity to both category and context (orange sample neuron), and others increase selectivity to category, but not context (yellow sample neuron). In contrast, none increases selectivity to context but not category. This makes the following experimental prediction: among all the neurons that are strongly connected to the readout, neurons with pure selectivity to category and neurons with mixed selectivity to category and context should be observed, but neurons with pure selectivity to context

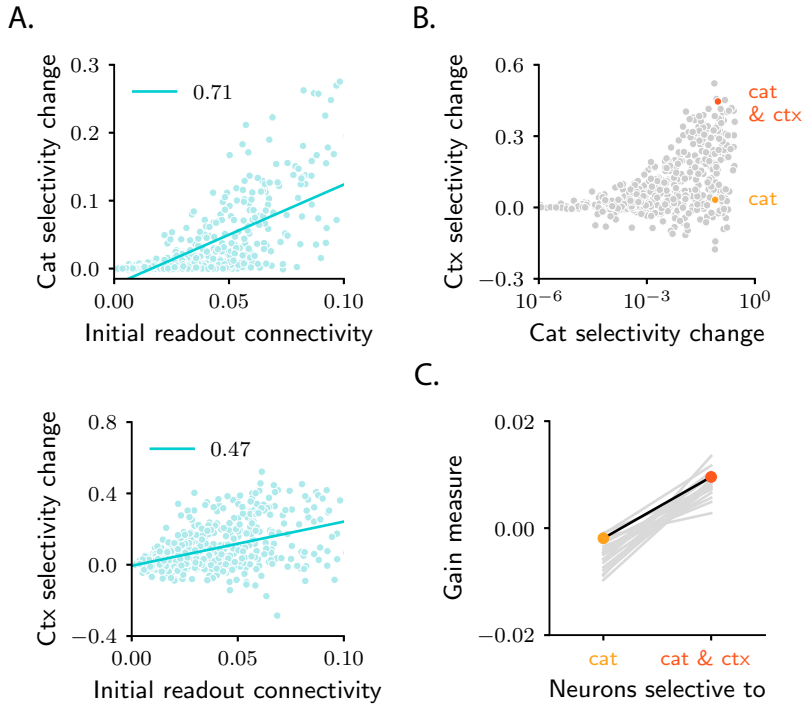

Fig. 7: **Patterns of pure and mixed selectivity to category and context.** **A.** Changes in category selectivity (top) and context selectivity (bottom) as a function of the initial readout connectivity, $w_{0,i}$ (in absolute value). Details as in Figs. 3E-F. **B.** Changes in context selectivity as a function of changes in category selectivity. Note the logarithmic scale on the x-axis; this is required by the heavy-tailed behaviour of category selectivity (Figs. 5D and G). We highlighted two sample neurons: one with strong, pure selectivity to category (yellow) and one with strong, mixed selectivity to category and context (orange). **C.** Neurons that develop pure and mixed selectivity are characterized by different patterns of initial activity. Here, we plot the gain-based measure of activity defined in Eq. 188 for neurons that belong to the former (left), and the latter (right) group. The former group includes neurons for which the change in category selectivity, but not the change in context selectivity, is within the top 15% across the population. The latter group includes neurons for which the change in both category and context selectivity is within the top 15%. Dots show results for the circuit analyzed in panels A and B. Grey lines show results for 20 different circuit realizations; note that the slope is positive for all circuits. All panels in the figure show results for the same circuit displayed in the second column of Fig. 5; the circuit displayed in the third column yields qualitatively similar results (Figs. S8C-D).

should not. The asymmetry between category and context arises because, in the model, the readout neuron learns to read out category, but not context. We show in Figs. S8E-F that if a second readout neuron, which learns to read out context, is included in the circuit, neurons with strong pure selectivity to context are also observed.

What determines whether, over learning, a given neuron develops pure selectivity to category, or mixed selectivity to category and context? Analysis in Methods 5.8 indicates those two populations are characterized by different properties of initial activity. Specifically, before learning, the two populations can be distinguished based on an activity measure, defined in detail in Methods 5.8 (Eq. 188), that is based on the gain of activity across different trials (for a given trial, the gain is given by the slope of the activation function, Ψ, at the response). Specifically: for neurons that develop pure selectivity to category, this quantity is significantly smaller than for neurons that develop mixed selectivity to category and context (Fig. 7C). This result indicates that initial activity, which is mostly unstructured and task-agnostic, plays an important role in learning: it breaks the symmetry among neurons in the intermediate layer, and determines which functional specialization neurons will display over learning.

Discussion

How does the brain learn to link sensory stimuli to abstract variables? Despite decades of experimental [1, 4, 37] and theoretical [38–40] work, the answer to this question remains elusive. Here we hypothesized that learning occurs via gradient descent synaptic plasticity. To explore the implications of this hypothesis, we considered a minimal model: a feedforward circuit with one intermediate layer, assumed to contain a large number of neurons compared to the number of stimuli. This assumption allowed us to thoroughly analyze the model, and thus gain insight into how activity evolves during learning, and how that evolution supports task acquisition.

We focused on two categorization tasks: a simple one (Fig. 2), in which category was determined solely by the stimulus, and a complex one (Fig. 5), in which category was determined by both the stimulus and the context. Through mathematical analysis corroborated with numerical simulations we showed that, over learning, single neurons become selective to abstract variables: category (which is explicitly reinforced) and context (which is not; instead, it embodies the task structure, and is only implicitly cued). From a geometrical perspective, the emergence of selectivity during learning is driven by clustering: activity associated with stimuli in different categories is pushed apart, forming distinct clusters (Fig. 3). In the context-dependent task, additional clustering occurs along a second, context-related axis; this results in activity forming four different clouds, one for each combination of category and context (Fig. 6). While the behaviour of selectivity is highly stereotyped, the behaviour of signal correlations and tuning symmetry is not, but depends on details (Fig. 4). All of these findings are in agreement with experimental observations [3, 4, 6, 7, 12].

Our analysis makes several experimentally testable predictions. First, it makes specific predictions about how category and context correlations should vary with properties of the circuit (shape of the nonlinear mapping from the synaptic input to activity, relative learning rates) and the task (number of stimuli, context-dependence) (Fig. 4). These could be tested with current technology; in particular, testing the dependence on task variables only requires recording neural activity. Second, it predicts that selectivity is shaped by connectivity with downstream areas, a result that is in line with recent experimental observations [32, 33, 41]. More specifically, it predicts that, for a given neuron, selectivity correlates with the strength of the synaptic connection that the neuron makes to the downstream neurons that read out category (Figs. 3E-F and 7A). Moreover, across all neurons that are strongly connected with the readouts, selectivity to category and context is distributed in a highly stereotyped way: during learning, some neurons develop mixed selective to category and context, others develop pure selectivity to category, but none develop pure selectivity to context (Fig. 7B). Initial activity biases neurons towards developing, over learning, one type of selectivity or the other (Fig. 7C).

Related models Previous theoretical studies have investigated how categorization can be implemented in multi-layer circuits, and how activity in these circuits is reshaped during learning [40, 42–44]. These studies considered a reward-dependent Hebbian learning rule, and focused on simple categorization tasks, where stimuli-category associations are fixed. In line with our findings, and with experimental data [1–4], these studies reported an increase in category selectivity over learning.

In our work, we also considered a task where stimuli-category associations are not fixed, but context-dependent; such context-dependent computations are core to a variety of experimental tasks [4, 7, 9, 26, 35, 45]. For that task, we showed that not only category selectivity, but also context selectivity, increases over learning (Figs. 5E-H) – a finding that captures further experimental observations [7, 8]. Whether reward-dependent Hebbian plasticity can be used to learn hard tasks, such as the context-dependent one used here, is currently unclear, and represents an important avenue for future work.

Learning in over-parameterized networks Our mathematical analysis of learning relied on two assumptions: the number of neurons in the circuit, N, is large compared to the number of stimuli to classify, and the initial activity in all layers of the network stays within an intermediate range (i.e., it neither vanishes nor saturates, which is enforced by the $1/\sqrt{N}$ scaling of initial connectivity). These assumptions are reasonable for brain circuits, across time scales ranging from development to animals' lifetimes. The learning dynamics that emerges in circuits characterized by these two assumptions has been named *lazy*, or *kernel*-like, in the machine-learning literature [28–30, 46].

A prominent feature of learning in this regime is that connectivity and activity changes are small in amplitude. This has an important implication: the final configuration of the circuit depends strongly on the initial one. Thus, the selectivity properties that single neurons display at the end of learning are determined by their initial activity and connectivity (Figs. 3D-F and 7A and C). In addition, the initial activity can have non-trivial structure. (For instance, in the context-dependent task, initial selectivity can be negative, Fig. 5A, or have a broad distribution,

Fig. 5B, and the initial correlations are structured, Fig. 5C). This initial structure affects the final activity (Figs. 5D-F and G-I), implying that the effect of learning has to be measured in terms of changes in activity measures, rather than their asymptotic values. These findings stress the importance of recording activity throughout learning: without knowing activity before learning, it can be difficult to correctly interpret neural data [47, 48].

Circuits that do not obey the two assumptions discussed above are characterized by different learning dynamics [46], and could result in different activity evolution. Previous theoretical work characterizing learning dynamics outside of this regime mostly focused on circuits with linear activation functions, which can implement only a restricted set of tasks [49–51]. Developing analytical tools to characterize activity changes in arbitrary circuits (see, e.g., [52, 53]) will prove critical to understand how learning unfolds in the brain [54].

Finally, we have used equations derived in the limit of infinitely wide networks to describe simulated circuits of finite size [31]. This represents, at least in principle, a rough approximation; however, we found that our approximate equations could predict qualitatively (and in some cases quantitatively), the behaviour of the activity measures of interest (Methods 3.4). Including finite-size corrections to the analysis represents an important avenue for future research [55].

Beyond simplified models Our model was based on several simplifying assumptions. First, sensory inputs associated with different stimuli were taken to be orthogonal, and all neurons within the intermediate layer were characterized by the same activation function. However, we showed in Methods 4.8 and Fig. S5 that our main results also hold when these assumptions are dropped. Second, we focused on classification, and did not consider the other cognitive components that are included in classical laboratory tasks, and are relevant for learning. Notably, we did not consider continuous, noisy stimuli or fine discrimination [2, 56, 57], nor did our task include working memory or complex decision-making [3, 7, 27, 58]. Addressing the latter two components would require recurrent connectivity in the model; mathematical tools for analyzing learning dynamics in recurrent networks have recently become available [59, 60], which could allow our analysis to be extended in that direction. Third, the vanilla gradient-descent plasticity rule we considered is not biologically-plausible. However, a number of studies have indicated that gradient-descent learning can be implemented, at least approximately, in biologically-plausible ways [19–23]. Understanding whether these different implementations lead to differences in activity evolution is an important research question that we leave to future work. Fourth, our circuit involves a single readout neuron for both categories A and B; this is sufficient because categories are mutually exclusive. A single readout has been hypothesized in the case of the lateral intraparietal area of monkeys [5, 61]. For other brain regions and tasks, multiple readout neurons might be more appropriate; for example, work in mice prefrontal cortex suggests that two readout circuits are used for valence – one strongly active for positive valence, and one strongly active for negative one [32]. Activity evolution in circuits with two readout neurons can in principle be higher-dimensional, and generate novel patterns of population responses. Constraining model architectures with connectomics, and then using models to interpret neural recordings, represents a promising line of future research (for an example, see the comparison between Fig. 7B, which uses one readout neuron, and Fig. S8E-F, which uses two).

Acknowledgements

F.M. would like to thank Friedrich Schuessler for useful discussions. F.M., N.H. and P.E.L. were supported by the Gatsby Charitable Foundation, and F.M. and P.E.L. were supported by the Wellcome Trust (110114/Z/15/Z).

Competing interests

The authors declare no competing interests.

References

[1] A. Messinger, L. Squire, S. Zola, and T. Albright. Neuronal representations of stimulus associations develop in the temporal lobe during learning. *Proc. Natl. Acad. Sci. USA*, 98(21):12239–12244, 2001.

[2] D. Freedman, M. Riesenhuber, T. Poggio, and E. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316, 2001.

[3] D. Freedman and J. Assad. Experience-dependent representation of visual categories in parietal cortex. *Nature*, 443(7107):85–88, 2006.

[4] S. Reinert, M. Hübener, T. Bonhoeffer, and P. Goltstein. Mouse prefrontal cortex represents learned rules for categorization. *Nature*, 593(7859):411–417, 2021.

[5] J. Fitzgerald, D. Freedman, and J. Assad. Generalized associative representations in parietal cortex. *Nat. Neurosci.*, 14(8):1075–1079, 2011.

[6] J. Cromer, J. Roy, and E. Miller. Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron*, 66(5):796 – 807, 2010.

[7] J. Wallis, K. Anderson, and E. Miller. Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411(6840):953–956, 2001.

[8] G. Stoet and L. Snyder. Single neurons in posterior parietal cortex of monkeys encode cognitive set. *Neuron*, 42(6):1003–1012, 2004.

[9] J. Roy, M. Riesenhuber, T. Poggio, and E. Miller. Prefrontal cortex activity during flexible categorization. *J. Neurosci.*, 30(25):8519–8528, 2010.

[10] I. White and S. Wise. Rule-dependent neuronal activity in the prefrontal cortex. *Exp. Brain Res.*, 126(3):315–335, 1999.

[11] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

[12] J. Fitzgerald, D. Freedman, A. Fanini, S. Bennur, J. Gold, and J. Assad. Biased associative representations in parietal cortex. *Neuron*, 77(1):180–191, 2013.

[13] D. Freedman and E. Miller. Neural mechanisms of visual categorization: Insights from neurophysiology. *Neurosci. Biobehav. Rev.*, 32(2):311–329, 2008.

[14] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160(1):106–154, 1962.

[15] B. Olshausen and D. Field. Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.*, 14(4):481–487, 2004.

[16] G. DeAngelis and T. Uka. Coding of horizontal disparity and velocity by mt neurons in the alert macaque. *J. Neurophysiol.*, 89(2):1094–1111, 2003.

[17] B. Richards and al. A deep learning framework for neuroscience. *Nat. Neurosci.*, 22(11):1761–1770, 2019.

[18] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436, 2015.

[19] T. Lillicrap, F. Cownden, D. Tweed, and C. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.*, 7(1):13276, 2016.

[20] J. Whittington and R. Bogacz. An Approximation of the Error Backpropagation Algorithm in a Predictive Coding Network with Local Hebbian Synaptic Plasticity. *Neural Comput.*, 29(5):1229–1262, 2017.

[21] J. Sacramento, R. Ponte Costa, Y. Bengio, and W. Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. *Adv. Neural Inf. Process. Syst.*, 31:8721–8732, 2018.

[22] M. Akrout, C. Wilson, P. Humphreys, T. Lillicrap, and D. Tweed. Deep learning without weight transport. *Adv. Neural Inf. Process. Syst.*, 32, 2019.

[23] A. Payeur, J. Guerguiev, F. Zenke, B. Richards, and R. Naud. Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nat. Neurosci.*, 24(7):1010–1019, 2021.

[24] D. Yamins and J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3):356–365, 2016.

[25] A. Kell, D. Yamins, E. Shook, S. Norman-Haignere, and J. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.e16, 2018.

[26] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, 2013.

[27] W. Chaisangmongkon, S. Swaminathan, D. Freedman, and X.-J. Wang. Computing by robust transience: How the fronto-parietal network performs sequential, category-based decisions. *Neuron*, 93(6):1504 – 1517.e4, 2017.

[28] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Adv. Neural Inf. Process. Syst.*, 31:8571–8580, 2018.

[29] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Adv. Neural Inf. Process. Syst.*, 32:8572–8583, 2019.

[30] C. Liu, L. Zhu, and M. Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Adv. Neural Inf. Process. Syst.*, 33:15954–15964, 2020.

[31] W. Hu, L. Xiao, B. Adlam, and J. Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. *Adv. Neural Inf. Process. Syst.*, 33:17116–17128, 2020.

[32] K. Ye, W. Allen, K. Thompson, Q. Tian, B. Hsueh, C. Ramakrishnan, A.-C. Wang, J. Jennings, A. Adhikari, C. Halpern, I. Witten, A. Barth, L. Luo, J. McNab, and K. Deisseroth. Wiring and molecular features of prefrontal ensembles representing distinct experiences. *Cell*, 165(7):1776–1788, 2016.

[33] J. Hirokawa, A. Vaughan, P. Masset, T. Ott, and A. Kepecs. Frontal cortex neuron types categorically encode single decision variables. *Nature*, 576(7787):446–451, 2019.

[34] V. Papyan, X. Han, and D. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proc. Natl. Acad. Sci. USA*, 117(40):24652–24663, 2020.

[35] S. McKenzie, A. Frank, N. Kinsky, B. Porter, P. Rivière, and H. Eichenbaum. Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron*, 83(1):202–215, 2014.

[36] A. Canatar, B. Bordelon, and C. Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nat. Commun.*, 12(1):2914, 2021.

[37] W. Asaad, G. Rainer, and E. Miller. Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, 21(6):1399–1407, 1998.

[38] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 65(6):65–386, 1958.

[39] O. Barak, M. Rigotti, and S. Fusi. The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. *J. Neurosci.*, 33(9):3844–3856, 2013.

[40] T. Engel, W. Chaisangmongkon, D. Freedman, and X.-J. Wang. Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nat. Commun.*, 6(1):6454, 2015.

[41] O. Gschwend, T. Yang, D. van de Lisdonk, X. Zhang, R. Sharma, and B. Li. Prefrontal top-down projections control context-dependent strategy selection. *biorXiv*, 2021.12.14.472559, 2021.

[42] M. Pannunzi, G. Gigante, M. Mattia, G. Deco, S. Fusi, and P. Del Giudice. Learning selective top-down control enhances performance in a visual categorization task. *J. Neurophys.*, 108(11):3124–3137, 2012.

[43] F. Villagrasa, J. Baladron, J. Vitay, H. Schroll, E. Antzoulatos, E. Miller, and F. Hamker. On the role of cortex-basal ganglia interactions for category learning: A neurocomputational approach. *J. Neurosci.*, 38(44):9551–9562, 2018.

[44] B. Min, D. Bliss, A. Sarma, D. Freedman, and X.-J. Wang. A neural circuit mechanism of categorical perception: top-down signaling in the primate cortex. *bioRxiv*, 2020.06.15.151506, 2020.

[45] S. Brincat, M. Siegel, C. von Nicolai, and E. Miller. Gradual progression from sensory to task-related processing in cerebral cortex. *Proc. Natl. Acad. Sci. USA*, 115(30):e7202–e7211, 2018.

[46] L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. *Adv. Neural Inf. Process. Syst.*, 32, 2019.

[47] N Steinmetz, C Aydin, et al. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539):eabf4588, 2021.

[48] K. Latimer and D. Freedman. Low-dimensional encoding of decisions in parietal cortex reflects long-term training history. *bioRxiv*, 2021.10.07.463576, 2021.

[49] S. Saxe, J. McClelland, and S. Ganguli. A mathematical theory of semantic development in deep neural networks. *Proc. Natl. Acad. Sci. USA*, 116(23):11537–11546, 2019.

[50] Q. Li and H. Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Phys. Rev. X*, 11:031059, 2021.

[51] E. Moroshko, B. Woodworth, S. Gunasekar, J. Lee, N. Srebro, and D. Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *Adv. Neural Inf. Process. Syst.*, 33:22182–22193, 2020.

[52] G. Yang and E. Hu. Feature learning in infinite-width neural networks. *arXiv:2011.14522*, 2021.

[53] S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA*, 115(33):E7665–E7671, 2018.

[54] T. Flesch, K. Juechems, T. Dumbalska, A. Saxe, and C. Summerfield. Rich and lazy learning of task representations in brains and neural networks. *bioRxiv*, 2021.04.23.441128, 2021.

[55] J. Huang and H.-T. Yau. Dynamics of deep neural networks and neural tangent hierarchy. *ICML*, 119:4542–4551, 2020.

[56] H. Shan and H. Sompolinsky. A minimum perturbation theory of deep perceptual learning. *bioRxiv*, 2021.10.05.463260, 2021.

[57] L. Bonnasse-Gahot and J.-P. Nadal. Categorical perception: A groundwork for deep learning. *arXiv:2012.05549*, 2021.

[58] S. Bernardi, M. Benna, M. Rigotti, J. Munuera, S. Fusi, and D. Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967.e21, 2021.

[59] F. Schuessler, F. Mastrogiuseppe, A. Dubreuil, S. Ostojic, and O. Barak. The interplay between randomness and structure during learning in rnns. *Adv. Neural Inf. Process. Syst.*, 33:13352–13362, 2020.

[60] J. Smith, S. Linderman, and D. Sussillo. Reverse engineering recurrent neural networks with jacobian switching linear dynamical systems. *Adv. Neural Inf. Process. Syst.*, 35, 2021.

[61] S. Ganguli, J. Bisley, J. Roitman, M. Shadlen, M. Goldberg, and K. Miller. One-dimensional dynamics of attention and decision making in LIP. *Neuron*, 58(1):15 – 25, 2008.

[62] M. Geiger, S. Spigler, A. Jacot, and M. Wyart. Disentangling feature and lazy training in deep neural networks. *J. Stat. Mech.*, 11:113301, 2020.

[63] T. Albright. Direction and orientation selectivity of neurons in visual area mt of the macaque. *J. Neurophysiol.*, 52(6):1106–1130, 1984.

[64] A. Fanini and J. Assad. Direction selectivity of neurons in the macaque lateral intraparietal area. *J. Neurophysiol.*, 101(1):289–305, 2009.

Methods

1 Overview **20**

2 Model **20**
2.1 Circuit . . . 20
2.2 Gradient descent plasticity . . . 21

3 Evolution of connectivity and activity **21**
3.1 Evolution of connectivity in large circuits . . . 22
3.2 Evolution of activity in large circuits . . . 24
3.3 A low-order Taylor expansion is self-consistent in large circuits . . . 24
3.4 Evolution of activity in finite-size networks . . . 25

4 Simple categorization task **26**
4.1 Task definition . . . 26
4.2 Computing activity . . . 26
4.3 Category selectivity . . . 28
4.4 Category correlation . . . 30
4.5 Computing normalized dot products . . . 32
4.6 Asymmetry in category response . . . 33
4.7 Characterizing variability . . . 33
4.8 Simple categorization task with structured inputs and heterogeneity . . . 35

5 Context-dependent categorization task **36**
5.1 Task definition . . . 36
5.2 Computing activity . . . 36
5.3 Category and context selectivity . . . 38
5.4 Category and context correlation . . . 39
5.5 Computing normalized dot products . . . 39
5.6 Detailed analysis of context selectivity . . . 40
5.7 Detailed analysis of category selectivity . . . 43
5.8 Analysis of context and category selectivity across neurons . . . 44

6 Software **47**
6.1 Circuit simulations . . . 47
6.2 Table of parameters . . . 47
6.3 Evaluation of averages . . . 48

1 Overview

In the main text, we made qualitative arguments about the evolution of activity over learning. Here we make those arguments quantitative. We start with a detailed description of the circuit model (Section 2). We then derive approximate analytical expressions that describe how activity in the circuit evolves over learning (Section 3). To this end, we use an approach that is valid for large circuits. This approach allows us to compute, semi-analytically, the value of selectivity and correlations, first for the simple task (Section 4), then for the context-dependent one (Section 5). Finally, we provide details on the numerical implementation of circuit models and analytical expressions (Section 6).

2 Model

2.1 Circuit

We consider a feedforward circuit with a single intermediate layer (Fig. 1A). For simplicity, we assume that the input and the intermediate layer have identical size N, and we consider N to be large. The sensory input vector is

indicated with $\boldsymbol{x}$. Activity in the intermediate layer reads (Eq. 1)

$$\boldsymbol{y} = \Psi(\boldsymbol{k}) \tag{10a}$$

$$\boldsymbol{k} \equiv \boldsymbol{u} \cdot \boldsymbol{x}\,. \tag{10b}$$

Here, $\boldsymbol{k}$ represents the synaptic drive and $\boldsymbol{u}$ is an $N \times N$ connectivity matrix. Activity in the readout layer is given by (Eq. 2)

$$z = \Phi(h) \tag{11a}$$

$$h \equiv \boldsymbol{w} \cdot \boldsymbol{y} \tag{11b}$$

where h is the synaptic drive and $\boldsymbol{w}$ is an N-dimensional readout vector.

The activation functions Ψ and Φ are non-negative, monotonically increasing functions that model the input-to-output properties of units in the intermediate and readout layer, respectively. In simulations, we use sigmoidal functions,

$$\Psi(x) = \frac{1}{1 + \exp(-\Theta_1(x - \Theta_2))}, \tag{12}$$

and similarly for $\Phi(x)$ (Fig. 1B). The parameters of the activation functions, Θ_1 and Θ_2, determine the gain and offset, respectively, with the gain (defined to be the slope at $x = \Theta_2$) given by $\Theta_1/4$. Their values, which vary across simulations, are given in Section 6.2, Tables 1 and 2.

The synaptic weights, $\boldsymbol{u}$ and $\boldsymbol{w}$, are initialized at random from a zero-mean Gaussian distribution with variance $1/N$. The sensory input vectors $\boldsymbol{x}$ are also drawn from a zero-mean Gaussian distribution (see Sections 4.1 and 5.1), but with variance equal to 1,

$$w_{0,i}, u_{0,ij} \sim \mathcal{N}(0, N^{-1}) \tag{13a}$$

$$x_i \sim \mathcal{N}(0, 1) \tag{13b}$$

where the subscript "0" on the weights indicates that those are evaluated before learning starts. This choice of initialization ensures that, before learning, the amplitude of both the synaptic drive (h, and the components of $\boldsymbol{k}$) and the activity (z, and the components of $\boldsymbol{y}$) are independent of the circuit size (i.e., $O(1)$ in N).

2.2 Gradient descent plasticity

We train the circuit on P sensory input vectors $\boldsymbol{x}^s$ ($s = 1, \ldots, P$), and we work in the regime $P \ll N$. For each input vector, the target activity of the readout neuron, $\tilde{z}^s$, is equal to either z^{A} or z^{B} (Section 4.1 and 5.1), which correspond to high and low activity, respectively. The weights are adjusted to minimize the loss, $\mathcal{E}(\boldsymbol{u}, \boldsymbol{w})$, which is defined to be

$$\mathcal{E}(\boldsymbol{u}, \boldsymbol{w}) \equiv \frac{1}{2P} \sum_{s=1}^{P} (\tilde{z}^s - z^s)^2 \tag{14}$$

where z^s is the activity of the readout neuron (Eq. 11) in response to the sensory input $\boldsymbol{x}^s$. The weights are updated according to full-batch vanilla gradient descent. If the learning rates, η_u and η_w, are sufficiently small, the evolution of the connectivity weights can be described by continuous-time equations (Eq. 3),

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = -\eta_u \frac{\partial \mathcal{E}(\boldsymbol{u}, \boldsymbol{w})}{\partial \boldsymbol{u}} \tag{15a}$$

$$\frac{\mathrm{d}\boldsymbol{w}}{\mathrm{d}t} = -\eta_w \frac{\partial \mathcal{E}(\boldsymbol{u}, \boldsymbol{w})}{\partial \boldsymbol{w}} \tag{15b}$$

where t indicates learning time.

3 Evolution of connectivity and activity

Our goal is to understand how learning affects activity in the intermediate layer, $\boldsymbol{y}$. We do that in two steps. In the first step, we analyze the evolution of the synaptic weights. In particular, we determine the weights after learning is complete – meaning after the loss (Eq. 14) has been minimized (Section 3.1). In the second step, we use the learned weights to determine activity (Section 3.2). We work in the large-N regime, which allows us to make analytic headway [28–30]. We then validate our large-N analysis with finite-N simulations (Section 3.4, Figs. S3, S4, S6, S9).

3.1 Evolution of connectivity in large circuits

It is convenient to make the definitions

$$\boldsymbol{u} \equiv \boldsymbol{u}_0 + \Delta\boldsymbol{u} \tag{16a}$$

$$\boldsymbol{w} \equiv \boldsymbol{w}_0 + \Delta\boldsymbol{w} \tag{16b}$$

where $\boldsymbol{u}_0$ and $\boldsymbol{w}_0$ are the initial weights (Eq. 13), and $\Delta\boldsymbol{u}$ and $\Delta\boldsymbol{w}$ are changes changes in the weights induced by learning (Eq. 15). Using Eq. 15, with the loss given by Eq. 14, we see that $\Delta\boldsymbol{u}$ and $\Delta\boldsymbol{w}$ evolve according to

$$\frac{\mathrm{d}\Delta\boldsymbol{u}(t)}{\mathrm{d}t} = -\eta_u \frac{\partial\mathcal{E}}{\partial\Delta\boldsymbol{u}} = \frac{\eta_u}{P}\sum_{s=1}^{P}\epsilon^s(t)\frac{\partial h^s}{\partial\Delta\boldsymbol{u}} \tag{17a}$$

$$\frac{\mathrm{d}\Delta\boldsymbol{w}(t)}{\mathrm{d}t} = -\eta_w \frac{\partial\mathcal{E}}{\partial\Delta\boldsymbol{w}} = \frac{\eta_w}{P}\sum_{s=1}^{P}\epsilon^s(t)\frac{\partial h^s}{\partial\Delta\boldsymbol{w}} \tag{17b}$$

where ϵ^s is proportional to the error associated with sensory input $\boldsymbol{x}^s$,

$$\epsilon^s \equiv \big(\tilde{z}^s - \Phi(h^s)\big)\Phi'(h^s). \tag{18}$$

To evaluate the partial derivatives on the right-hand side of Eq. 17, we need to express h^s in terms of $\Delta\boldsymbol{u}$ and $\Delta\boldsymbol{w}$. Combining Eq. 11b with 10 and 16, we have

$$h^s = (\boldsymbol{w}_0 + \Delta\boldsymbol{w})\cdot\Psi(\boldsymbol{u}_0\cdot\boldsymbol{x}^s + \Delta\boldsymbol{u}\cdot\boldsymbol{x}^s)\,. \tag{19}$$

To proceed, we assume that changes in the connectivity, $\Delta\boldsymbol{u}$ and $\Delta\boldsymbol{w}$, are small. That holds in the large-N limit (the limit we consider here) because when each neuron receives a large number of inputs, none of them has to change very much to cause a large change in the output (we make this reasoning more quantitative in Section 3.3). Then, Taylor-expanding the nonlinear activation function Ψ in Eq. 19, and keeping only terms that are zeroth and first order in the weight changes $\Delta\boldsymbol{u}$ and $\Delta\boldsymbol{w}$, we have

$$h^s \simeq h_0^s + \boldsymbol{w}_0\cdot\big[\Psi'(\boldsymbol{k}_0^s)\odot(\Delta\boldsymbol{u}\cdot\boldsymbol{x}^s)\big] + \Delta\boldsymbol{w}\cdot\Psi(\boldsymbol{k}_0^s), \tag{20}$$

where $\odot$ indicates element-wise multiplication, and we have defined

$$\boldsymbol{k}_0^s \equiv \boldsymbol{u}_0\cdot\boldsymbol{x}^s \tag{21a}$$

$$h_0^s \equiv \boldsymbol{w}_0\cdot\Psi(\boldsymbol{k}_0^s). \tag{21b}$$

For now, we assume that the three terms in the right-hand side of Eq. 20 are of similar magnitude, and that higher-order terms in $\Delta\boldsymbol{u}$ and $\Delta\boldsymbol{w}$ are smaller, and so can be neglected. We will verify these assumptions post-hoc (Section 3.3). Inserting Eq. 20 into Eq. 17, we arrive at

$$\frac{\mathrm{d}\Delta\boldsymbol{u}(t)}{\mathrm{d}t} = \frac{\eta_u}{P}\sum_{s=1}^{P}\epsilon^s(t)\,\big[\boldsymbol{w}_0\odot\Psi'(\boldsymbol{k}_0^s)\big]\,\boldsymbol{x}^s \tag{22a}$$

$$\frac{\mathrm{d}\Delta\boldsymbol{w}(t)}{\mathrm{d}t} = \frac{\eta_w}{P}\sum_{s=1}^{P}\epsilon^s(t)\,\Psi(\boldsymbol{k}_0^s) \tag{22b}$$

where two adjacent vectors are taken to be an outer product (e.g., $(\boldsymbol{ab})_{ij} = a_i b_j$).

The only quantity on the right-hand side of Eq. 22 that depends on time is ϵ^s. Consequently, we can immediately write down the solution,

$$\Delta\boldsymbol{u}(t) = \frac{1}{N}\sum_{s=1}^{P}c^s(t)\,\big[\boldsymbol{w}_0\odot\Psi'(\boldsymbol{k}_0^s)\big]\,\boldsymbol{x}^s \tag{23a}$$

$$\Delta\boldsymbol{w}(t) = \frac{1}{N}\sum_{s=1}^{P}c^s(t)\,\frac{\eta_w}{\eta_u}\,\Psi(\boldsymbol{k}_0^s) \tag{23b}$$

where the coefficients c^s are found by solving the differential equation

$$\frac{P}{\eta_u N}\frac{\mathrm{d}c^s(t)}{\mathrm{d}t} = \epsilon^s(t) \tag{24}$$

with initial conditions $c^s(t=0)=0$. The right-hand side of Eq. 24 depends on time through the synaptic drive, h^s (Eq. 18), which turn depends on $\Delta\boldsymbol{u}$ and $\Delta\boldsymbol{w}$ through Eq. 20, and thus, via Eq. 23, on the coefficients $c^s(t)$. Consequently, Eq. 24 is a closed differential equation for the coefficients $c^s(t)$.

In the general case, Eq. 24 must be solved numerically. If, however, we are not interested in the full learning dynamics, but care only about connectivity and activity once learning has converged ($t\to\infty$), we can use the fact that dynamics in Eq. 22 are guaranteed to converge to a global minimum of the error function $\mathcal{E}$ [30]. For our loss function and tasks, the minimum occurs at $\mathcal{E}=0$. At that point, $z^s(t\to\infty)=\tilde{z}^s$; equivalently,

$$h^s(t\to\infty) = \Phi^{-1}(\tilde{z}^s), \tag{25}$$

where Φ^{-1} is the inverse of the activation function of the readout neurons (which exists because Φ is a monotonically increasing function).

To find $c^s(t\to\infty)$, we simply express $h^s(t\to\infty)$ in terms of $c^s(t\to\infty)$, and insert that into Eq. 25. To reduce clutter, we define (in a slight abuse of notation) c^s without an argument to be its asymptotic value,

$$c^s \equiv c^s(t\to\infty)\,. \tag{26}$$

Combining Eq. 20 for h^s with Eq. 23 for $\Delta\boldsymbol{w}$ and $\Delta\boldsymbol{u}$, we have

$$h^s(t\to\infty) = h_0^s + \boldsymbol{w}_0\cdot\left(\frac{1}{N}\sum_{q=1}^{P} c^q\,(\boldsymbol{x}^q\cdot\boldsymbol{x}^s)\,\left[\boldsymbol{w}_0\odot\Psi'(\boldsymbol{k}_0^q)\odot\Psi'(\boldsymbol{k}_0^s)\right]\right) + \frac{\eta_w}{\eta_u}\,\frac{1}{N}\sum_{q=1}^{P} c^q\,\left(\Psi(\boldsymbol{k}_0^q)\cdot\Psi(\boldsymbol{k}_0^s)\right). \tag{27}$$

We can simplify the second term in the right-hand side by explicitly evaluating the dot product,

$$\begin{aligned}\boldsymbol{w}_0\cdot\left[\boldsymbol{w}_0\odot\Psi'(\boldsymbol{k}_0^q)\odot\Psi'(\boldsymbol{k}_0^s)\right] &= \sum_{i=1}^{N} w_{0,i}^2\,\Psi'(k_{0,i}^q)\,\Psi'(k_{0,i}^s)\\ &\equiv N\,\langle\, w_{0,i}^2\,\Psi'(k_{0,i}^q)\,\Psi'(k_{0,i}^s)\,\rangle_i\end{aligned} \tag{28}$$

where the notation $\langle\,.\,\rangle_i$ indicates an average over the index i.

Since N is large, we can interpret population averages such as Eq. 28 as expectations over the probability distribution of $w_{0,i}$ and $k_{0,i}^s$. An immediate implication is that Eq. 28 simplifies,

$$\begin{aligned}N\langle\, w_{0,i}^2\,\Psi'(k_{0,i}^q)\,\Psi'(k_{0,i}^s)\,\rangle_i &= N\langle\, w_{0,i}^2\,\rangle_i\,\langle\,\Psi'(k_{0,i}^q)\,\Psi'(k_{0,i}^s)\,\rangle_i\\ &= \langle\,\Psi'(k_{0,i}^q)\,\Psi'(k_{0,i}^s)\,\rangle_i.\end{aligned} \tag{29}$$

For the first equality we used the independence of $w_{0,i}$ and $k_{0,i}$; for the second we used the fact that the elements of $\boldsymbol{w}_0$ are drawn from a zero-mean Gaussian with variance N^{-1} (Eq. 13a). We can thus rewrite Eq. 27 as

$$h^s(t\to\infty) = h_0^s + \sum_{q=1}^{P} c^q\,\langle\, x_i^q x_i^s\,\rangle_i\,\langle\,\Psi'(k_{0,i}^q)\,\Psi'(k_{0,i}^s)\,\rangle_i + \frac{\eta_w}{\eta_u}\sum_{q=1}^{P} c^q\,\langle\,\Psi(k_{0,i}^q)\,\Psi(k_{0,i}^s)\,\rangle_i. \tag{30}$$

Combining this with Eq. 25, we conclude that

$$\Phi^{-1}(\tilde{z}^s) - h_0^s = \sum_{q=1}^{P}\left[\langle\, x_i^q x_i^s\,\rangle_i\,\langle\,\Psi'(k_{0,i}^q)\,\Psi'(k_{0,i}^s)\,\rangle_i + \frac{\eta_w}{\eta_u}\,\langle\,\Psi(k_{0,i}^q)\,\Psi(k_{0,i}^s)\,\rangle_i\right] c^q\,. \tag{31}$$

Equation 31 is a P-dimensional linear system of equations for the coefficients c^s, $s=1,..,P$ (the term in brackets is a $P\times P$ matrix with indices s and q). For the tasks we consider (Sections 4 and 5), this system can be solved analytically, yielding a closed-form expression for the coefficients c^s.

3.2 Evolution of activity in large circuits

It is now straightforward to determine how activity in the intermediate layer, $\boldsymbol{y}^s(t)$, evolves. Inserting Eq. 23 into Eq. 10, and Taylor-expanding the nonlinear activation function Ψ to first order in $\Delta\boldsymbol{u}$, we arrive at

$$\boldsymbol{y}^s(t) \equiv \boldsymbol{y}_0^s + \Delta\boldsymbol{y}^s(t) \simeq \boldsymbol{y}_0^s + \sum_{q=1}^{P} c^q(t)\, \boldsymbol{v}^{qs} \tag{32}$$

where

$$\boldsymbol{v}^{qs} \equiv \langle\, x_i^q x_i^s \,\rangle_i\, \boldsymbol{w}_0 \odot \Psi'(\boldsymbol{k}_0^q) \odot \Psi'(\boldsymbol{k}_0^s). \tag{33}$$

To reduce clutter, we define (following the notation in the previous section) $\boldsymbol{y}^s$ without an argument to be its asymptotic value: $\boldsymbol{y}^s \equiv \boldsymbol{y}^s(t \to \infty)$. Thus, Eq. 32 becomes

$$\boldsymbol{y}^s \simeq \boldsymbol{y}_0^s + \sum_{q=1}^{P} c^q\, \boldsymbol{v}^{qs}. \tag{34}$$

Because of the term $\boldsymbol{w}_0$ on the right-hand side of Eq. 33, the elements of $\boldsymbol{v}^{qs}$ scale as $N^{-1/2}$. Thus, changes in activity are small compared to the initial activity, which is $O(1)$.

In what follows, we refer to $\{\boldsymbol{v}^{qs}\}_{qs}$ as spanning vectors, and to the coefficients c^q as the activity coordinates. We observe that all spanning vectors have a non-zero overlap with the initial readout vector $\boldsymbol{w}_0$, as

$$\boldsymbol{v}^{qs} \cdot \boldsymbol{w}_0 = \langle\, x_i^q x_i^s \,\rangle_i\, \langle\, \Psi'(k_{0,i}^q)\, \Psi'(k_{0,i}^s) \,\rangle_i \equiv \rho^{qs}. \tag{35}$$

This implies that, for every spanning vector, we can write

$$\boldsymbol{v}^{qs} = \rho^{qs}\boldsymbol{w}_0 + \delta\boldsymbol{v}^{qs} \tag{36}$$

where ρ^{qs} is given by Eq. 35 (since $\boldsymbol{w}_0 \cdot \boldsymbol{w}_0 = 1$) and $\delta\boldsymbol{v}^{qs}$ is a residual component due to the nonlinearity of the activation function Ψ:

$$\delta\boldsymbol{v}^{qs} = \langle\, x_i^q x_i^s \,\rangle_i\, \boldsymbol{w}_0 \odot \Big(\Psi'(\boldsymbol{k}_0^q) \odot \Psi'(\boldsymbol{k}_0^s) - \langle\, \Psi'(k_{0,i}^q)\, \Psi'(k_{0,i}^s) \,\rangle_i\, \mathbf{1}\Big). \tag{37}$$

The notation $\mathbf{1}$ indicates a vector whose components are all equal to 1: $\mathbf{1} \equiv (1, 1, ..., 1)$.

3.3 A low-order Taylor expansion is self-consistent in large circuits

To conclude our theoretical derivation, we verify that the approximations we made in Section 3.1 are valid in large circuits. Specifically, we show that the approximate expression for h^s, Eq. 20 (which was derived by Taylor-expanding the nonlinear activation function Ψ), is self-consistent when N is large. As a first step, we compute the size of $\Delta\boldsymbol{u}$ and $\Delta\boldsymbol{w}$, and show that in the large-N limit they are small compared to $\boldsymbol{u}_0$ and $\boldsymbol{w}_0$, respectively. We then Taylor-expand Ψ in Eq. 19 to all orders, and show that the terms that were included in Eq. 20 (zeroth- and first-order terms in connectivity changes) are indeed the dominant ones.

Assuming that the term in brackets in Eq. 31, when viewed as a $P \times P$ matrix, is invertible (which is generically the case when $P \ll N$), it follows that, with respect to N

$$c^s \sim O(1). \tag{38}$$

This result applies to the asymptotic ($t \to \infty$) value of c^s (Eq. 26). We assume, though, that the learning process is smooth enough that $c^s(t)$ remains at most $O(1)$ for all t. Under this assumption, the results we derive in this section are valid at any point during learning.

Using Eq. 38, along with the fact that $w_{0,i} \sim O(N^{-1/2})$ while all other variables are $O(1)$, we see from Eq. 23 that

$$\Delta u_{ij} \sim O(N^{-3/2}) \tag{39a}$$

$$\Delta w_i \sim O(N^{-1}). \tag{39b}$$

When N is large, both are small compared to the initial weights $\boldsymbol{u}_0$ and $\boldsymbol{w}_0$ whose elements are $O(N^{-1/2})$ (Eq. 13).

Equation 39 suggests that a low-order Taylor expansion is self-consistent, but it is not proof. We thus turn directly to Eq. 19. The three terms in the right-hand side of Eq. 20 are re-written in Eq. 30, and it is clear from that expression that they are all $O(1)$. To determine the size of the higher order terms, we need the complete Taylor expansion of Eq. 19. That is given by

$$h^s = \sum_{n=0}^{\infty} \frac{1}{n!} (\boldsymbol{w}_0 + \Delta \boldsymbol{w}) \cdot \left(\Psi^{(n)}(\boldsymbol{k}_0^s) \odot (\Delta \boldsymbol{u} \cdot \boldsymbol{x}^s)^n \right) \tag{40}$$

where $\Psi^{(n)}$ is the nth derivative of Ψ, and the exponentiation in $(\Delta \boldsymbol{u} \cdot \boldsymbol{x}^s)^n$ is taken element-wise. The higher order terms (i.e., the terms not included in Eq. 20) are

$$h^s_{\text{higher order}} = \sum_{n=2}^{\infty} \frac{N}{n!} \langle\, w_{0,i}\, \Psi^{(n)}(k^s_{0,i}) (\Delta \boldsymbol{u} \cdot \boldsymbol{x}^s)^n_i \,\rangle_i + \sum_{n=1}^{\infty} \frac{N}{n!} \langle\, \Delta w_i\, \Psi^{(n)}(k^s_{0,i}) (\Delta \boldsymbol{u} \cdot \boldsymbol{x}^s)^n_i \,\rangle_i \tag{41}$$

where we have replaced dot products with averages over indices. Using Eq. 23a, and taking into account the fact that $w_{0,i}$ and $k_{0,i}$ are independent, we observe that

$$\frac{N}{n!} \langle\, w_{0,i}\, \Psi^{(n)}(k^s_{0,i})\, (\Delta \boldsymbol{u} \cdot \boldsymbol{x}^s)^n_i \,\rangle_i \sim N \langle\, w^{n+1}_{0,i} \,\rangle_i \times O(1). \tag{42}$$

Similarly, this time using both Eq. 23a and b, we have

$$\frac{N}{n!} \langle\, \Delta w_i\, \Psi^{(n)}(k^s_{0,i})\, (\Delta \boldsymbol{u} \cdot \boldsymbol{x}^s)^n_i \,\rangle_i \sim \langle\, w^n_0 \,\rangle_i \times O(1). \tag{43}$$

Inserting these into Eq. 41 then gives us

$$h^s_{\text{higher order}} \sim \sum_{n=2}^{\infty} \frac{1}{n!} N \langle\, w^{n+1}_{0i} \,\rangle_i \times O(1) + \sum_{n=1}^{\infty} \frac{1}{n!} \langle\, w^n_{0i} \,\rangle_i \times O(1). \tag{44}$$

Finally, using the fact that the $w_{0,i}$ are drawn independently from a zero-mean Gaussian with variance N^{-1} (Eq. 13), we see that $\langle\, w^n_{0,i} \,\rangle_i$ is proportional to $N^{-n/2}$ when n is even and $N^{-(n+1)/2}$ when n is odd. Consequently, the largest term in the expression for $h^s_{\text{higher order}}$ is proportional to N^{-1}. The higher order terms can, therefore, be neglected in the large N limit.

3.4 Evolution of activity in finite-size networks

The equations that describe the evolution of connectivity and activity that were derived in Sections 3.1 and 3.2 are accurate if two assumptions are satisfied: (i) the circuit is very large ($N \gg 1$), and (ii) the synaptic weights are initialized to be $O(N^{-1/2})$ (Eq. 13a), which guarantees that synaptic drives and activity neither vanish nor explode at initialization. Both assumptions are reasonable for brain circuits, and correspond to rather standard modeling choices in theoretical neuroscience.

In this work, we use the analytical expressions derived for N large to capture activity evolution in finite-size networks. This is a crude approximation, as dealing with finite N would require, in principle, integrating corrective terms into our equations [55]. How accurate is this approximation? Ongoing work in the machine-learning literature is committed to address this question across tasks, architectures, and loss functions [28, 31, 46, 52, 62]. Because of the Taylor expansions used in Sections 3.1 and 3.2, for fixed N, good accuracy is expected when the amplitude of activity changes is small. Via Eq. 34, we see that the latter increases with the number of sensory input vectors P, implying that good accuracy is expected when P is small. For fixed P, furthermore, the amplitude of activity changes increases with correlations among sensory inputs (Eq. 33), implying that good accuracy is expected when sensory input correlations are small. As detailed in Sections 4.1 and 5.1, sensory input correlations are smaller in the simple than in the context-dependent task, which implies that accuracy in the former task is expected to be higher than in the latter. The amplitude of activity changes also depends on the amplitude of activity coordinates c^s (Eq. 34). We show in Sections 4.2 and 5.2 that activity coordinates are usually smaller in the simple than in the complex task, which again implies that accuracy in the former task is expected to be higher than in the latter. Overall, those arguments suggest that good accuracy is expected when the task is easy, and thus the training loss

converges to zero very quickly [31]. Finally, because of the Taylor expansion (Eq. 20), we also expect accuracy to depend on properties of the activation function Ψ, with accuracy increasing as Ψ becomes more linear in its effective activation range.

In Figs. S3, S4, S6 and S9, we evaluate accuracy by performing a systematic comparison between analytical expressions and circuit simulations with finite N. We find good agreement for the full range of parameters considered in the study. Specifically, the theory correctly predicts qualitatively, and in some cases also quantitatively, the behaviour of all activity measures discussed in the main text. As expected, the agreement is stronger in the simple (Figs. S3 and S4) than in the context-dependent task (Figs. S6 and S9).

4 Simple categorization task

4.1 Task definition

We first consider a simple categorization task. Each stimulus is represented by an input pattern $\boldsymbol{\mu}^S$, with $S = 1, \ldots, Q$ where Q is the total number of stimuli. The $\boldsymbol{\mu}^S$ are random vectors whose entries are drawn independently from a zero-mean, unit-variance Gaussian distribution. Every sensory input vector $\boldsymbol{x}^s$ corresponds a stimulus,

$$\boldsymbol{x}^s = \boldsymbol{\mu}^S; \tag{45}$$

consequently, the number of sensory input vectors, P, is equal to Q (the upper-case notation S is used for consistency with the context-dependent task; see Sec. 5.1). To leading order in N, sensory input vectors are thus orthonormal,

$$\frac{\boldsymbol{x}^s \cdot \boldsymbol{x}^{s'}}{N} = \langle\, x_i^s x_i^{s'} \,\rangle_i \simeq \delta_{ss'} \tag{46}$$

where $\delta_{ss'}$ is the Kronecker delta.

Each stimulus is associated with one among the two mutually-exclusive categories A and B: the first half of stimuli is associated with A, the second half with B. The target value $\tilde{z}^s$ for the readout neuron is thus equal to z^{A} for the first half of sensory inputs and z^{B} for the second half. Since sensory input vectors are approximately orthogonal to each other, they are also linearly separable.

Our goal is to derive explicit expressions for the quantities analyzed in the main text: category selectivity (defined in Eq. 61 below), and category correlation (defined in Eq. 76 below). Both quantities depend on activity in the intermediate layer, $\boldsymbol{y}^s$, after learning, which is given in Eq. 34. In the next section, we then write down an explicit expression for $\boldsymbol{y}^s$; after that, we compute category selectivity (Section 4.3) and category correlation (Section 4.4). Further mathematical details are discussed in Sections 4.5, 4.6 and 4.7; a generalization of the current task is discussed and analyzed in Section 4.8.

4.2 Computing activity

Examining Eq. 34, we see that to compute the activity in the intermediate layer, $\boldsymbol{y}$, we need the asymptotic activity coordinates, c^s, and the spanning vectors, $\boldsymbol{v}^{qs}$. We start with the coordinates. To compute them, we solve the linear system of equations given in Eq. 31. Using Eq. 46, that system of equations becomes

$$\Phi^{-1}(\tilde{z}^s) - h_0^s = \langle\, \Psi'(k_{0,i}^s)^2 \,\rangle_i\, c^s + \frac{\eta_w}{\eta_u} \sum_{q=1}^{P} \langle\, \Psi(k_{0,i}^s)\, \Psi(k_{0,i}^q) \,\rangle_i\, c^q\,. \tag{47}$$

As a first step, we simplify the averages in the right-hand side. The law of large number guarantees that, when N is large, the elements of the synaptic drive, $k_{0,i}^s$, are independently drawn from a Gaussian distribution. The statistics of this distribution are given by

$$\langle\, k_{0,i}^s \,\rangle_i = \sum_{j=1}^{N} \langle\, u_{0,ij} \,\rangle_i\, x_j^s = 0 \tag{48a}$$

$$\langle\, k_{0,i}^q\, k_{0,i}^s \,\rangle_i = \sum_{j=1}^{N} \sum_{j'=1}^{N} \langle\, u_{0,ij}\, u_{0,ij'} \,\rangle_i\, x_j^q x_{j'}^s = \langle\, x_j^q x_j^s \,\rangle_j \tag{48b}$$

where we have used the fact that, because of Eq. 13a, $\langle\, u_{0,ij}\, u_{0,ij'}\,\rangle_i = \delta_{jj'}/N$. Equation 48, combined with Eq. 46, implies that the $k^s_{0,i}$ have zero mean and unit variance, and are uncorrelated across stimuli. In addition, because the statistics of $k^s_{0,i}$ are independent of s, averages over i of any function of $k^s_{0,i}$ are independent of s.

Using these observations, Eq. 47 can be written as

$$\Phi^{-1}(\tilde{z}^s) - h^s_0 = \left(\langle \Psi'^2\rangle\, c^s + \frac{\eta_w}{\eta_u}\,\left(\langle\Psi^2\rangle - \langle\Psi\rangle^2\right)\right) + \frac{\eta_w}{\eta_u}\,\langle\Psi\rangle^2 \sum_{q=1}^{P} c^q \tag{49}$$

here we used the short-hand notation $\langle F\rangle$ to indicate the average of a function F whose argument is drawn from a zero-mean, unit-variance Gaussian distribution. I.e.,

$$\langle F\rangle \equiv \langle\, F(a)\,\rangle_a \tag{50}$$

where a is a zero-mean, unit-variance Gaussian variable. This average can be computed via numerical integration, as detailed in Section 6.3 (Eq. 189).

The left-hand side of Eq. 49 consists of two terms: the target $\Phi^{-1}(\tilde{z}^s)$, which is fixed by the task, and h^s_0 (representing the synaptic drive of the readout neuron at initialization), which fluctuates across model realizations. The presence of the latter term indicates that connectivity and activity changes are not fully self-averaging; they are rather tuned to compensate for the initial state of the readout neuron. Here we seek to analyze the *average* behaviour of the model, and so we drop the second, variable term. This approximation is discussed in detail in Section 4.7.

With the variable terms neglected, the left-hand side of Eq. 49 can take only two values: $\Phi^{-1}(z^{\mathsf{A}})$ and $\Phi^{-1}(z^{\mathsf{B}})$. Combined with the symmetry of the right-hand side, this implies that the coordinates c^s themselves can take only two values. Specifically, we have

$$c^s = \begin{cases} c^{\mathsf{A}} & s \text{ in category } \mathsf{A} \\ c^{\mathsf{B}} & s \text{ in category } \mathsf{B}\,. \end{cases} \tag{51}$$

The category-dependent coordinates, c^{A} and c^{B}, are determined by the two-dimensional linear system of equations

$$\Phi^{-1}(z^{\mathsf{A}}) = \alpha\, c^{\mathsf{A}} + \beta\,(c^{\mathsf{A}} + c^{\mathsf{B}}) \tag{52a}$$

$$\Phi^{-1}(z^{\mathsf{B}}) = \alpha\, c^{\mathsf{B}} + \beta\,(c^{\mathsf{A}} + c^{\mathsf{B}}) \tag{52b}$$

where the scalars α and β are defined as

$$\alpha = \langle \Psi'^2\rangle + \frac{\eta_w}{\eta_u}\,(\langle\Psi^2\rangle - \langle\Psi\rangle^2) \tag{53a}$$

$$\beta = \frac{\eta_w}{\eta_u}\frac{Q}{2}\langle\Psi\rangle^2. \tag{53b}$$

This system is easily solved, yielding

$$c^{\mathsf{A}} = \frac{1}{\alpha + 2\beta}\big(\Phi^{-1}(z^{\mathsf{A}}) + \gamma\big) \tag{54a}$$

$$c^{\mathsf{B}} = \frac{1}{\alpha + 2\beta}\big(\Phi^{-1}(z^{\mathsf{B}}) - \gamma\big) \tag{54b}$$

where we have defined the shift

$$\gamma = \frac{\beta}{\alpha}\big(\Phi^{-1}(z^{\mathsf{A}}) - \Phi^{-1}(z^{\mathsf{B}})\big). \tag{55}$$

Note that γ is positive, as α, $\beta > 0$ and $\Phi^{-1}(z^{\mathsf{A}}) > \Phi^{-1}(z^{\mathsf{B}})$, which in turn indicates that $c^{\mathsf{A}} > c^{\mathsf{B}}$.

To conclude the derivation of activity, we evaluate the spanning vectors, $\boldsymbol{v}^{qs}$ (Eq. 33). Because the sensory inputs $\boldsymbol{x}^s$ are orthogonal (Eq. 46), spanning vectors with $q \neq s$ vanish. Consequently, the activity, $\boldsymbol{y}^s$ (Eq. 34), reads

$$\boldsymbol{y}^s = \begin{cases} \boldsymbol{y}^s_0 + c^{\mathsf{A}}\boldsymbol{v}^{ss} & s \text{ in category } \mathsf{A} \\ \boldsymbol{y}^s_0 + c^{\mathsf{B}}\boldsymbol{v}^{ss} & s \text{ in category } \mathsf{B}\,. \end{cases} \tag{56}$$

Using Eq. 36, we can rewrite this as

$$\boldsymbol{y}^s = \begin{cases} \boldsymbol{y}_0^s + c^{\mathsf{A}}\rho\boldsymbol{w}_0 + c^{\mathsf{A}}\delta\boldsymbol{v}^{ss} & s \text{ in category } \mathsf{A} \\ \boldsymbol{y}_0^s + c^{\mathsf{B}}\rho\boldsymbol{w}_0 + c^{\mathsf{B}}\delta\boldsymbol{v}^{ss} & s \text{ in category } \mathsf{B} \end{cases} \tag{57}$$

where we used Eq. 35 to define

$$\rho \equiv \rho^{ss} = \langle \Psi'^2 \rangle \,. \tag{58}$$

Equation 57 indicates that activity consists of three components. The first one coincide with initial activity, $\boldsymbol{y}_0^s$, which for this task is fully unstructured. The second one is a common component along $\boldsymbol{w}_0$ (whose strength is category-dependent, as it is given by c^{A} or c^{B}). The third one is a non-shared component along the residuals $\delta\boldsymbol{v}^{ss}$, which represent the components of the spanning vectors that are perpendicular to the initial readout $\boldsymbol{w}_0$. For the current task, the latter component is orthogonal across activity vectors, implying that activity vectors only overlap along $\boldsymbol{w}_0$. To leading order in N, in fact

$$\delta\boldsymbol{v}^{ss} \cdot \delta\boldsymbol{v}^{s's'} = \langle\, x_i^s x_i^s \,\rangle_i \,\langle\, x_i^{s'} x_i^{s'} \,\rangle_i \left[\langle\, \Psi'(k_{0,i}^s)\, \Psi'(k_{0,i}^s)\, \Psi'(k_{0,i}^{s'})\, \Psi'(k_{0,i}^{s'}) \,\rangle_i - \langle\, \Psi'(k_{0,i}^s)^2 \,\rangle_i^2 \right] \simeq 0 \,, \tag{59}$$

which follows because $k_{0,i}^s$ and $k_{0,i}^{s'}$ are uncorrelated.

We observe that Eq. 57 is similar, but not identical to the expressions that we used in the main text to describe activity evolution. By combining Eq. 5 with 7, those read

$$\boldsymbol{y}^s = \begin{cases} \boldsymbol{y}_0^s + c^{\mathsf{A}}\rho\boldsymbol{w}_0 & s \text{ in category } \mathsf{A} \\ \boldsymbol{y}_0^s + c^{\mathsf{B}}\rho\boldsymbol{w}_0 & s \text{ in category } \mathsf{B}\,. \end{cases} \tag{60}$$

Comparing Eq. 57 with 60, we see that the residuals $\delta\boldsymbol{v}^{ss}$ were neglected in the main text. This could be done because, for the current task (but not for the context-dependent one, see Section 5.2), residuals are all orthogonal to each other (Eq. 59). As such, they do not add novel structure to activity, and do not significantly contribute to activity measures. This is showed and justified, in detail, in the next sections.

4.3 Category selectivity

In this section, we evaluate the category selectivity of neurons in the intermediate layer (Figs. 2B, F and J). For each neuron i, we evaluate the standard selectivity index [3], defined in Eq. 4. We repeat that definition here for convenience,

$$S_i = \frac{\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff cat}} - \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s\neq s' \text{ same cat}}}{\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff cat}} + \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s\neq s' \text{ same cat}}} \tag{61}$$

where the notation $\langle\, \cdot \,\rangle_{s,s'}$ denotes an average over sensory input pairs associated either with different, or the same, category. To evaluate this expression, we assume that the number of stimuli, $Q = P$, is moderately large ($1 \ll Q \ll N$). We show that, under this assumption, the category selectivity index for each neuron, which is approximately zero at $t = 0$, becomes positive over learning.

We start with

$$y_i^s = y_{0i}^s + c^s(\rho w_{0i} + \delta v_i^{ss}) \,, \tag{62}$$

which follows from Eq. 57. The first term of the right-hand side is $O(1)$, while both terms in parentheses are $O(N^{-1/2})$. Thus, when evaluating the denominator in Eq. 61, to lowest non-vanishing order in N we can replace y_i with y_{0i}. Doing that, and expanding the square, we have

$$\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff cat}} + \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s\neq s' \text{ same cat}} \simeq 4\langle\, (y_{0i}^s)^2 \,\rangle_s - 2\langle\, y_{0i}^s y_{0i}^{s'} \,\rangle_{s,\, s' \text{ diff cat}} - 2\langle\, y_{0i}^s y_{0i}^{s'} \,\rangle_{s\neq s' \text{ same cat}} \,. \tag{63}$$

Noting that $y_{0i}^s = \Psi(k_{0i}^s)$, and using Eq. 48, we see that the second two averages in the above equation are both equal to $\langle\, y_{0i}^s \,\rangle_s^2$. Consequently,

$$\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff cat}} + \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s\neq s' \text{ same cat}} \simeq 4\langle\, (y_{0i}^s)^2 \,\rangle_s - 4\langle\, y_{0i}^s \,\rangle_s^2 \,. \tag{64}$$

Strictly speaking, this step is accurate only in the large-Q limit, but is a good approximation even for moderate Q. Since Q is moderately large, we can further approximate this as

$$\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff cat}} + \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s\neq s' \text{ same cat}} \simeq 4\langle \Psi^2 \rangle - 4\langle \Psi \,\rangle^2 \,, \tag{65}$$

where averages can be computed as described in Section 6.3.

For the numerator of Eq. 61, the minus sign causes the $(y_{0i}^s)^2$ terms to cancel, so we have

$$\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff cat}} - \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s \neq s' \text{ same cat}} = 2\langle\, y_i^s y_i^{s'} \,\rangle_{s \neq s' \text{ same cat}} - 2\langle\, y_i^s y_i^{s'} \,\rangle_{s,\, s' \text{ diff cat}} \,. \tag{66}$$

Using Eq. 62, we have (for $s \neq s'$)

$$y_i^s y_i^{s'} = y_{i0}^s y_{i0}^{s'} + c^{s'} y_{i0}^s (\rho w_{0i} + \delta v_i^{s's}) + c^s y_{i0}^{s'} (\rho w_{0i} + \delta v_i^{ss}) + c^s c^{s'} (\rho w_{0i} + \delta v_i^{s's})(\rho w_{0i} + \delta v_i^{ss}) \,. \tag{67}$$

Apart from the first term, and the term proportional to w_{0i}^2, all terms in the right-hand side have essentially random signs. Neglecting those for a moment, we obtain

$$y_i^s y_i^{s'} \simeq y_{i0}^s y_{i0}^{s'} + c^s c^{s'} \rho^2 w_{0i}^2. \tag{68}$$

Inserting this into Eq. 66, using the fact that $\langle\, y_{0i}^s y_{0i}^{s'} \,\rangle$ is independent of s and s', and performing a small amount of algebra, we arrive at

$$\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff cat}} - \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s \neq s' \text{ same cat}} \simeq w_{0i}^2 (c^{\mathsf{A}} - c^{\mathsf{B}})^2 \rho^2. \tag{69}$$

Combining this with Eq. 65, and using Eq. 58 for ρ, we arrive at

$$S_i \simeq \frac{w_{0,i}^2\, (c^{\mathsf{A}} - c^{\mathsf{B}})^2 \langle \Psi'^2 \rangle^2}{4\, (\langle \Psi^2 \rangle - \langle \Psi \rangle^2)}. \tag{70}$$

We conclude that single-neuron selectivity vanishes at $t = 0$ (when $c^{\mathsf{A}} = c^{\mathsf{B}} = 0$), and is positive at the end of learning. Furthermore, for each neuron, selectivity is determined by the value of $w_{0,i}$, which measures initial connectivity with the readout neuron. As a result, neurons with large initial connectivity develop large selectivity values (Fig. 3D-F).

Because of the factor w_{0i}^2, the right-hand side of Eq. 70 is $O(N^{-1})$. To derive Eq. 70, we neglected terms in the numerator that have random sign and thus contribute as noise. The dominant random terms are $O(1)$ in N, but $O(Q^{-1})$ in Q. This implies that, in simulated circuits with finite Q, random deviations from Eq. 70 occur. For example, Fig. 2B shows that selectivity values at $t = 0$ are small but non-zero; Figs. 3D-F, instead, shows that the values of S_i and $w_{0,i}^2$ are not perfectly correlated across the population.

We can, finally, average Eq. 70 over neurons, yielding

$$S = \langle\, S_i \,\rangle_i \simeq \frac{(c^{\mathsf{A}} - c^{\mathsf{B}})^2 \langle \Psi'^2 \rangle^2}{4N\, (\langle \Psi^2 \rangle - \langle \Psi \rangle^2)} \tag{71}$$

where neglected random terms are now $O(N^{-1/2} Q^{-1})$. In Figs. S3 and S4, we compare this approximate analytical expression for average category selectivity with values measured in finite-size circuits, and find good agreement between the two.

Category clustering Our derivation of the average category selectivity, Eq. 71, was based on several assumptions: we assumed that the number of stimuli, Q, was large, and that terms with random signs could be neglected. A different, but related, activity measure is given by category clustering [40, 58]. That is defined as

$$\tilde{S} = \frac{\langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s,\, s' \text{ diff cat}} - \langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s \neq s' \text{ same cat}}}{\langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s,\, s' \text{ diff cat}} + \langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s \neq s' \text{ same cat}}}. \tag{72}$$

This measure is positive if activity vectors elicited by within-category stimuli are more similar, in norm, than activity vectors elicited by across-category stimuli – and negative otherwise. In contrast to average category selectivity, category clustering can be evaluated straightforwardly, and for any value of Q. We show this in the following.

By using the statistical homogeneity of activity vectors, we can rewrite

$$\tilde{S} = \frac{-\langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s,\, s' \text{ diff cat}} + \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s \neq s' \text{ same cat}}}{2\langle\, \langle\, (y_i^s)^2 \,\rangle_i \,\rangle_s - \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s,\, s' \text{ diff cat}} - \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s \neq s' \text{ same cat}}}. \tag{73}$$

Expressions in the form of $\langle y_i^s y_i^{s'} \rangle_i$ are evaluated in Section 4.5; the derivation involves lengthy, but straightforward algebra. Using those results (Eqs. 95 and 100), we have:

$$\tilde{S} = \frac{(c^{\mathsf{A}} - c^{\mathsf{B}})^2 \, \langle \Psi'^2 \rangle^2}{4N(\langle \Psi^2 \rangle - \langle \Psi \rangle^2) + 2 \, [(c^{\mathsf{A}})^2 + (c^{\mathsf{B}})^2] \, \langle \Psi'^4 \rangle - \, (c^{\mathsf{A}} + c^{\mathsf{B}})^2 \, \langle \Psi'^2 \rangle^2}. \tag{74}$$

To the leading order in N, we obtain

$$\tilde{S} = \frac{(c^{\mathsf{A}} - c^{\mathsf{B}})^2 \, \langle \Psi'^2 \rangle^2}{4N \, (\langle \Psi^2 \rangle - \langle \Psi \rangle^2)} \tag{75}$$

which is identical to the expression obtained for average category selectivity evaluated with Q large (Eq. 71).

To better understand the relationship between selectivity and clustering, we observe that clustering coincide with the average selectivity, $S = \langle \, S_i \, \rangle_i$, if the average over the numerator and the denominator of S_i (Eq. 61) is factorized. In general, the numerator and the denominator of S_i are correlated, and the average cannot be factorized. We have however shown that, in the limit where both Q and N are large, S_i can be approximated by an expression where the denominator is independent of i (Eq. 70). In that regime, the average can be factorized; average category selectivity S and category clustering $\tilde{S}$ thus take very similar values, as quantified by Eqs. 71 and 75. A detailed comparison between average selectivity and clustering within data from simulated circuits is provided in Figs. S3 and S4.

4.4 Category correlation

To quantify how the population as a whole responds to the two categories, we evaluate category correlation. This quantity, denoted C, is given by the average Pearson correlation coefficient of activity in response to stimuli associated with different categories. We have:

$$C = \langle \, C^{s_{\mathsf{A}} s_{\mathsf{B}}} \, \rangle_{s_{\mathsf{A}} s_{\mathsf{B}}} \tag{76}$$

where s_{A} and s_{B} are indices that denote sensory inputs associated, respectively, with category A and B. The Pearson correlation $C^{s_{\mathsf{A}} s_{\mathsf{B}}}$ is given by

$$\begin{aligned} C^{s_{\mathsf{A}} s_{\mathsf{B}}} &= \frac{\langle \, y_i^{s_{\mathsf{A}}} y_i^{s_{\mathsf{B}}} \, \rangle_i - \langle \, y_i^{s_{\mathsf{A}}} \, \rangle_i \langle \, y_i^{s_{\mathsf{B}}} \, \rangle_i}{\sqrt{\langle \, y_i^{s_{\mathsf{A}}} y_i^{s_{\mathsf{A}}} \, \rangle_i - \langle \, y_i^{s_{\mathsf{A}}} \, \rangle_i^2} \sqrt{\langle \, y_i^{s_{\mathsf{B}}} y_i^{s_{\mathsf{B}}} \, \rangle_i - \langle \, y_i^{s_{\mathsf{B}}} \, \rangle_i^2}} \\ &= \frac{\langle \, y_i^{s_{\mathsf{A}}} y_i^{s_{\mathsf{B}}} \, \rangle_i - \langle \Psi \rangle^2}{\sqrt{\langle \, y_i^{s_{\mathsf{A}}} y_i^{s_{\mathsf{A}}} \, \rangle_i - \langle \Psi \rangle^2} \sqrt{\langle \, y_i^{s_{\mathsf{B}}} y_i^{s_{\mathsf{B}}} \, \rangle_i - \langle \Psi \rangle^2}}. \end{aligned} \tag{77}$$

To go from the first to the second line, we used the fact that, for each sensory input,

$$\langle \, y_i^s \, \rangle_i = \langle \, y_{0,i}^s \, \rangle_i + c^s \langle \, v_i^{ss} \, \rangle_i = \langle \Psi \rangle \tag{78}$$

where the second equality follows from $\langle \, w_{0,i} \, \rangle_i = 0$ (Eq. 13a), which in turns implies that $\langle \, v_i^{ss} \, \rangle_i = 0$ (Eq. 33). Pearson correlation coefficients are displayed in the correlation matrices of Figs. 2C, G and K.

As we show in Section 4.5, in the large-N limit, $\langle \, y_i^s y_i^{s'} \, \rangle_i$ only depends on the category s and s' are in. This makes the average over s_{A} and s_{B} in Eq. 76 trivial. Using Eqs. 95 and 100, we arrive at

$$C = \frac{c^{\mathsf{A}} c^{\mathsf{B}} \, \langle \Psi'^2 \rangle^2}{\sqrt{N\big(\langle \Psi^2 \rangle - \langle \Psi \rangle^2\big) + (c^{\mathsf{A}})^2 \langle \Psi'^4 \rangle} \sqrt{N\big(\langle \Psi^2 \rangle - \langle \Psi \rangle^2\big) + (c^{\mathsf{B}})^2 \langle \Psi'^4 \rangle}}. \tag{79}$$

In Figs. S3 and S4, we compare this approximate analytical expression with values measured in finite-size circuits, and find good agreement between the two. We can further simplify Eq. 79 by Taylor-expanding in N. To leading order, we obtain

$$C = \frac{1}{N} \frac{c^{\mathsf{A}} c^{\mathsf{B}} \, \langle \Psi'^2 \rangle^2}{\langle \Psi^2 \rangle - \langle \Psi \rangle^2}. \tag{80}$$

Before learning, $c^{\mathsf{A}} = c^{\mathsf{B}} = 0$, and so correlation vanishes. After learning, C is non-zero, and its sign is given by the sign of the product $c^{\mathsf{A}} c^{\mathsf{B}}$. This has a simple geometric explanation: after mean subtraction, activity vectors associated with opposite categories only overlap along the direction spanned by the initial readout vector $\boldsymbol{w}_0$. The

coordinates of vectors associated with category A and B along this direction are proportional, respectively, to c^{A} and c^{B} (Eq. 60). When c^{A} and c^{B} have opposite sign, activity vectors acquire opposite components along $\boldsymbol{w}_0$, which generates negative category correlation. When c^{A} and c^{B} have identical sign, instead, activity vectors acquire aligned components, which generates positive category correlation.

To determine how the product $c^{\mathsf{A}}c^{\mathsf{B}}$ depends on parameters, we use Eq. 54 for c^{A} and c^{B} to write

$$c^{\mathsf{A}}c^{\mathsf{B}} = \frac{1}{(\alpha+2\beta)^2}\left(\Phi^{-1}(z^{\mathsf{A}})+\gamma\right)\left(\Phi^{-1}(z^{\mathsf{B}})-\gamma\right) \tag{81}$$

where the (positive) scalars α and β are defined in Eq. 53, and γ in Eq. 55. Consequently, the sign of $c^{\mathsf{A}}c^{\mathsf{B}}$, and thus, the sign of the category correlation C, depends on the value of the target synaptic drives $\Phi^{-1}(z^{\mathsf{A}})$ and $\Phi^{-1}(z^{\mathsf{B}})$ (Fig. 4A), as well as on γ.

In particular, when $\Phi^{-1}(z^{\mathsf{A}})$ and $\Phi^{-1}(z^{\mathsf{B}})$ have opposite sign, Eq. 81 can only be negative, and thus category correlation can only be negative. When $\Phi^{-1}(z^{\mathsf{A}})$ and $\Phi^{-1}(z^{\mathsf{B}})$ have identical sign, Eq. 81 can be either negative or positive, depending on the value of the shift γ, and thus category correlation can be either negative or positive.

What controls the value of the shift γ, and thus, the sign of correlation, in the latter case? Combining Eq. 55 for γ with Eq. 53 for α and β, we have

$$\gamma = \frac{Q\langle\Psi\rangle^2}{2\left[(\eta_w/\eta_u)^{-1}\,\langle\Psi'^2\rangle + (\langle\Psi^2\rangle - \langle\Psi\rangle^2)\right]}\left(\Phi^{-1}(z^{\mathsf{A}}) - \Phi^{-1}(z^{\mathsf{B}})\right). \tag{82}$$

Recall that $\Phi^{-1}(z^{\mathsf{A}}) - \Phi^{-1}(z^{\mathsf{B}})$ is always positive; see Fig. 4A. We observe that γ depends on the learning rate ratio η_w/η_u: increasing this ratio increases the value of γ and thus, via Eq. 81, favors negative correlation (Fig. 4D). It also depends on the number of stimuli, Q: increasing Q increases the value of γ, and thus also favors negative correlation (Fig. 4E). Finally, γ depends on the activation function of neurons in the intermediate layer, Ψ, through nonlinear population averages; by computing those averages, we find that decreasing the gain and offset of Ψ favors negative correlation (Fig. 4C).

Alternative definition For completeness, we observe that an alternative way of quantifying category correlation consists of averaging activity over stimuli first (Fig. 2D, H and L), and then computing the Person correlation coefficient between averaged responses. The correlation values obtained via this procedure are displayed in the legend of Figs. 2D, H and L. This alternative definition yields qualitatively identical results to Eq. 76; we show this below.

We start by defining the category-averaged activity

$$\boldsymbol{y}^{\mathsf{A}} = \langle\, \boldsymbol{y}^{s_{\mathsf{A}}} \,\rangle_{s_{\mathsf{A}}} \tag{83a}$$

$$\boldsymbol{y}^{\mathsf{B}} = \langle\, \boldsymbol{y}^{s_{\mathsf{B}}} \,\rangle_{s_{\mathsf{B}}}. \tag{83b}$$

We then define category correlation as

$$\begin{aligned} C &= \frac{\langle\, y_i^{\mathsf{A}} y_i^{\mathsf{B}} \,\rangle_i - \langle\, y_i^{\mathsf{A}} \,\rangle_i \langle\, y_i^{\mathsf{B}} \,\rangle_i}{\sqrt{\langle\, y_i^{\mathsf{A}} y_i^{\mathsf{A}} \,\rangle_i - \langle\, y_i^{\mathsf{A}} \,\rangle_i^2}\sqrt{\langle\, y_i^{\mathsf{B}} y_i^{\mathsf{B}} \,\rangle_i - \langle\, y_i^{\mathsf{B}} \,\rangle_i^2}} \\ &= \frac{\langle\, y_i^{\mathsf{A}} y_i^{\mathsf{B}} \,\rangle_i - \langle\Psi\rangle^2}{\sqrt{\langle\, y_i^{\mathsf{A}} y_i^{\mathsf{A}} \,\rangle_i - \langle\Psi\rangle^2}\sqrt{\langle\, y_i^{\mathsf{B}} y_i^{\mathsf{B}} \,\rangle_i - \langle\Psi\rangle^2}}. \end{aligned} \tag{84}$$

Thsn, using Eqs. 95 and 100 from Section 4.5, we have

$$\langle\, y_i^{\mathsf{A}} y_i^{\mathsf{B}} \,\rangle_i = \left(\frac{2}{Q}\right)^2 \sum_{s_{\mathsf{A}}=1}^{Q/2}\sum_{s_{\mathsf{B}}=1}^{Q/2} \langle\, y_i^{s_{\mathsf{A}}} y_i^{s_{\mathsf{B}}} \,\rangle_i = \langle\Psi\rangle^2 + N^{-1}\, c^{\mathsf{A}}c^{\mathsf{B}}\langle\Psi'^2\rangle^2\,, \tag{85}$$

while

$$\begin{aligned} \langle\, y_i^{\mathsf{A}} y_i^{\mathsf{A}} \,\rangle_i &= \left(\frac{2}{Q}\right)^2 \sum_{s_{\mathsf{A}}=1}^{Q/2}\left[\langle\, y_i^{s_{\mathsf{A}}} y_i^{s_{\mathsf{A}}} \,\rangle_i + \sum_{s'_{\mathsf{A}}\neq s_{\mathsf{A}}} \langle\, y_i^{s_{\mathsf{A}}} y_i^{s'_{\mathsf{A}}} \,\rangle_i\right] \\ &= \langle\Psi\rangle^2 + \frac{2}{Q}\left(\langle\Psi^2\rangle - \langle\Psi\rangle^2\right) + N^{-1}\,(c^{\mathsf{A}})^2\left[\langle\Psi'^2\rangle^2 + \frac{2}{Q}\left(\langle\Psi'^4\rangle - \langle\Psi'^2\rangle^2\right)\right] \end{aligned} \tag{86}$$

and similarly for $\langle\, y_i^{\mathsf{B}} y_i^{\mathsf{B}} \,\rangle_i$, by replacing c^{A} with c^{B}. Inserting this into Eq. 84, we arrive at

$$C = \frac{c^{\mathsf{A}} c^{\mathsf{B}} \langle \Psi'^2 \rangle^2}{\sqrt{\frac{2}{Q} N \big(\langle \Psi^2 \rangle - \langle \Psi \rangle^2\big) + (c^{\mathsf{A}})^2 \left[\langle \Psi'^2 \rangle^2 + \frac{2}{Q}\big(\langle \Psi'^4 \rangle - \langle \Psi'^2 \rangle^2\big)\right]} \sqrt{\frac{2}{Q} N \big(\langle \Psi^2 \rangle - \langle \Psi \rangle^2\big) + (c^{\mathsf{B}})^2 \left[\langle \Psi'^2 \rangle^2 + \frac{2}{Q}\big(\langle \Psi'^4 \rangle - \langle \Psi'^2 \rangle^2\big)\right]}}. \tag{87}$$

Although the denominator of this expression is different from Eq. 79, the numerator is identical. As the denominators in both expressions are positive, the qualitative behaviour of Eq. 87 is identical to Eq. 79. Furthermore, to leading order in N, we obtain

$$C = \frac{Q}{2N} \frac{c^{\mathsf{A}} c^{\mathsf{B}} \, \langle \Psi'^2 \rangle^2}{\langle \Psi^2 \rangle - \langle \Psi \rangle^2} \tag{88}$$

which is proportional to Eq. 80, with constant of proportionality equal to $Q/2$.

4.5 Computing normalized dot products

We now compute the normalized dot products among pairs of activity vectors; namely

$$\langle\, y_i^s y_i^{s'} \,\rangle_i = \frac{\boldsymbol{y}^s \cdot \boldsymbol{y}^{s'}}{N} \,. \tag{89}$$

Those were used above to derive the behaviour of category clustering (Section 4.3) and correlations (Section 4.4).

The dot product takes different values depending on whether or not sensory inputs s and s' coincide. We start with the former,

$$\langle\, (y_i^s)^2 \,\rangle_i = \langle\, (y_{0,i}^s)^2 \,\rangle_i + \langle\, (\Delta y_i^s)^2 \,\rangle_i. \tag{90}$$

We used the fact that the cross-term $\langle\, y_{0,i}^s \, \Delta y_i^s \,\rangle_i$ vanishes on average,

$$\langle\, y_{0,i}^s \, \Delta y_i^s \,\rangle_i = c^s \langle\, y_{0,i}^s \, v_i^{ss} \,\rangle_i = c^s \,\langle\, w_{0,i} \,\rangle_i \,\langle\, y_{0,i}^s \, \Psi'(k_{0,i}^s)^2 \,\rangle_i = 0 \tag{91}$$

where we used Eq. 56 for the first equality, Eq. 33 for the second, and Eq. 13a for the third. By definition,

$$\langle\, (y_{0,i}^s)^2 \,\rangle_i = \langle \Psi^2 \rangle \tag{92}$$

while

$$\langle\, (\Delta y_i^s)^2 \,\rangle_i = (c^s)^2 \,\langle\, (v_i^{ss})^2 \,\rangle_i = N^{-1} (c^s)^2 \langle \Psi'^4 \rangle \tag{93}$$

where we have used the fact that, from Eq. 33

$$\langle\, (v_i^{ss})^2 \,\rangle_i = \langle\, x_i^s x_i^s \,\rangle_i^2 \,\langle\, w_{0,i}^2 \, \Psi'(k_{0,i}^s)^4 \,\rangle_i = N^{-1} \langle \Psi'^4 \rangle. \tag{94}$$

Putting these results together, we have

$$\langle\, (y_i^s)^2 \,\rangle_i = \begin{cases} \langle \Psi^2 \rangle + N^{-1} \, (c^{\mathsf{A}})^2 \, \langle \Psi'^4 \rangle & s \text{ in category } \mathsf{A} \\ \langle \Psi^2 \rangle + N^{-1} \, (c^{\mathsf{B}})^2 \, \langle \Psi'^4 \rangle & s \text{ in category } \mathsf{B}\,. \end{cases} \tag{95}$$

Note that activity vectors associated with different categories are characterized by different norms (unless coordinates are fine-tuned to be symmetric: $c^{\mathsf{A}} = -c^{\mathsf{B}}$, which occurs when $\Phi^{-1}(z^{\mathsf{A}}) = -\Phi^{-1}(z^{\mathsf{B}})$, as in Figs. 2E-H). Asymmetry of activity in response to different categories is discussed in detail in Section 4.6.

For dot products among different activity vectors, we have

$$\langle\, y_i^s \, y_i^{s'} \,\rangle_i = \langle\, y_{0,i}^s \, y_{0,i}^{s'} \,\rangle_i + \langle\, \Delta y_i^s \, \Delta y_i^{s'} \,\rangle_i. \tag{96}$$

with $s \neq s'$. In this case,

$$\langle\, y_{0,i}^s \, y_{0,i}^{s'} \,\rangle_i = \langle \Psi \rangle^2 \tag{97}$$

while

$$\langle\, \Delta y_i^s \, \Delta y_i^{s'} \,\rangle_i = c^s c^{s'} \,\langle\, v_i^{ss} \, v_i^{s's'} \,\rangle_i = N^{-1} \, c^s c^{s'} \langle \Psi'^2 \rangle^2 \,, \tag{98}$$

which comes from

$$\langle\, v_i^{ss} \, v_i^{s's'} \,\rangle_i = \langle\, x_i^s x_i^s \,\rangle_i \,\langle\, x_i^{s'} x_i^{s'} \,\rangle_i \,\langle\, w_{0,i}^2 \, \Psi'(k_{0,i}^s)^2 \, \Psi'(k_{0,i}^{s'})^2 \,\rangle_i = N^{-1} \langle \Psi'^2 \rangle^2. \tag{99}$$

Putting this together, we arrive at

$$\langle\, y_i^s \, y_i^{s'} \,\rangle_i = \begin{cases} \langle \Psi \rangle^2 + N^{-1} \, (c^{\mathsf{A}})^2 \, \langle \Psi'^2 \rangle^2 & s,\, s' \text{ in category } \mathsf{A} \\ \langle \Psi \rangle^2 + N^{-1} \, (c^{\mathsf{B}})^2 \, \langle \Psi'^2 \rangle^2 & s,\, s' \text{ in category } \mathsf{B} \\ \langle \Psi \rangle^2 + N^{-1} \, c^{\mathsf{A}} c^{\mathsf{B}} \, \langle \Psi'^2 \rangle^2 & s,\, s' \text{ in diff. categories.} \end{cases} \tag{100}$$

Equation 100 has a simple geometric interpretation. The first term in the right-hand side, $\langle \Psi \rangle^2$, is generated by the overlap between the activity vectors along the direction spanned by the unit vector $\mathbf{1}$. This component is due to the activation function Ψ being positive, and is approximately constant over learning. The second term on the right-hand side emerges over learning. This arises because activity vectors become aligned, via the spanning vectors (Eq. 35), along the direction spanned by the initial readout vector $\boldsymbol{w}_0$. Note that the components of activity that are aligned with the residual directions $\delta \boldsymbol{v}^{ss}$ (Eq. 57) do not contribute to the dot product. This can be verified by computing the dot product directly from Eq. 60, where residuals are neglected, and observing that the same result is obtained. This was expected, as we have showed in Eq. 59 that, for the current task task, residuals are orthogonal to each other.

4.6 Asymmetry in category response

In Fig. 2L in the main text, activity in response to category A and B is asymmetric: the number of neurons that respond more strongly to category A is significantly larger than the number that respond more strongly to B. Furthermore, the mean and variance of activity across the population is larger in response to A than to B. Such asymmetry is not present at $t = 0$ (Fig. 2D), and is thus a consequence of learning. Asymmetry has been reported in experimental data as well [12], where it was referred to as *biased category representations*. Here we discuss in detail why and how response asymmetry arises in the model. We show that asymmetry is controlled by the value of the target readout activity, z^{A} and z^{B}, and also by the shape of the activation functions of the intermediate and readout layer, Ψ and Φ.

Figure 2L displays activity in response to category A and B averaged over stimuli; those are denoted, respectively, by $\boldsymbol{y}^{\mathsf{A}}$ and $\boldsymbol{y}^{\mathsf{B}}$ (Eq. 83). We start deriving an explicit expression for $\boldsymbol{y}^{\mathsf{A}}$, from which the mean and variance across the population can be computed. Since initial activity is symmetric, we focus on the part of activity that is induced by learning. Combining Eq. 56 with 33, we have

$$\Delta y_i^{\mathsf{A}} \equiv \langle \, \Delta y_i^{s_{\mathsf{A}}} \, \rangle_{s_{\mathsf{A}}} \simeq c^{\mathsf{A}} \, \langle \Psi'^2 \rangle \, w_{0,i} \tag{101}$$

where the last approximate equality follows if Q is sufficiently large. The variance across the population is, therefore, given by

$$\langle \, (\Delta y_i^{\mathsf{A}})^2 \, \rangle_i = N^{-1} \, (c^{\mathsf{A}})^2 \, \langle \Psi'^2 \rangle^2. \tag{102}$$

For the variance across the population in response to category B, we simply replace c^{A} with c^{B}.

Consequently, the variances in response to categories A and B are identical only if $(c^{\mathsf{A}})^2 = (c^{\mathsf{B}})^2$. From Eq. 54, we see that this happens only if $\Phi^{-1}(z^{\mathsf{A}}) = -\Phi^{-1}(z^{\mathsf{B}})$, which yields $c^{\mathsf{A}} = -c^{\mathsf{B}}$. Figure 2H shows a circuit where the activation function of the readout neuron, Φ, was chosen to satisfy this relationship. In general, however, the two variances differ, and can have either $(c^{\mathsf{A}})^2 > (c^{\mathsf{B}})^2$ (the variance in response to A is lather than to B), or $(c^{\mathsf{A}})^2 < (c^{\mathsf{B}})^2$ (the opposite). Figure 2L corresponds to the first scenario, $(c^{\mathsf{A}})^2 > (c^{\mathsf{B}})^2$; this was achieved by setting $\Phi^{-1}(z^{\mathsf{A}}) > \Phi^{-1}(z^{\mathsf{B}}) > 0$, which yielded $c^{\mathsf{A}} > c^{\mathsf{B}} > 0$ (Fig. 4A). Figures S2A-B correspond to the second scenario, $(c^{\mathsf{A}})^2 < (c^{\mathsf{B}})^2$; this was achieved by setting $\Phi^{-1}(z^{\mathsf{B}}) < \Phi^{-1}(z^{\mathsf{A}}) < 0$, which yielded $c^{\mathsf{B}} < c^{\mathsf{A}} < 0$. Note that in both cases, $c^{\mathsf{A}} > c^{\mathsf{B}}$, as it must be (Eq. 54).

In Fig. 2L, activity in response to category A is not only characterized by larger variance, but also larger mean. This observation does not emerge immediately from our analysis, since our equations predict that the mean of activity changes vanishes both in response to A and B: from Eq. 13, we see that in response to category A,

$$\langle \, \Delta y_i^{\mathsf{A}} \, \rangle_i = c^{\mathsf{A}} \, \langle \Psi'^2 \rangle \langle \, w_{0,i} \, \rangle_i = 0 \tag{103}$$

and similarly for category B. To understand how Eq. 103 can be reconciled with Fig. 2L, recall that the equations we use for activity changes (Eq. 34) provide a linearized estimate of activity changes, which is strictly valid only in infinitely wide networks. In finite width networks, a non-zero mean response can emerge from higher-order terms in the expansion of Eq. 32. The leading higher-order terms of this expansion are quadratic, implying that the behaviour of the mean is controlled by the second-order derivative of the activation function of neurons in the intermediate layer, Ψ''. When the offset of Ψ is positive (so that activity is initialized close to the lower bound of Ψ), the second-order derivative Ψ'' is positive on average. Combined with $(c^{\mathsf{A}})^2 > (c^{\mathsf{B}})^2$, this implies that the mean of activity in response to category A is larger than to B; this case is illustrated in Fig. 2L. When the offset of Ψ is negative (so that activity is initialized close to the upper bound of Ψ), the second-order derivative Ψ'' is negative on average. Combined with $(c^{\mathsf{A}})^2 > (c^{\mathsf{B}})^2$, this implies that the mean of activity in response to category A is smaller than to B; this case is illustrated in Figs. S2C-D.

Finally, Eq. 103 suggest that non-vanishing mean activity could also be obtained if the initial readout weights $w_{0,i}$ have a non-zero mean. This is likely to be verified in the brain, where intra-area connectivity is mainly excitatory. We leave the incorporation of non-zero mean connectivity, along with Dale's law, to future investigations.

4.7 Characterizing variability

In Section 4.2, when computing the value of activity coordinates c^s, we neglected the second terms within the left-hand side of Eq. 47; because of this, the coordinates took on only two values, namely c^{A} and c^{B} (Eq. 51). The neglected terms do not

self-average, and thus fluctuate at random across model realizations. Had we included these variable terms, Eq. 51 would have read

$$c^s = \begin{cases} c^{\mathsf{A}} + \delta^s & s \text{ in category } \mathsf{A} \\ c^{\mathsf{B}} + \delta^s & s \text{ in category } \mathsf{B} \end{cases} \tag{104}$$

where the δ^s obey the linear system of equations

$$-h_0^s = \delta^s \left(\langle \Psi'^2 \rangle + \frac{\eta_w}{\eta_u} \left(\langle \Psi^2 \rangle - \langle \Psi \rangle^2 \right) \right) + \sum_{q=1}^{P} \delta^q \, \frac{\eta_w}{\eta_u} \, \langle \Psi \rangle^2. \tag{105}$$

Here, we further characterize the behaviour of the neglected terms δ^s. For simplicity, we consider the case in which plasticity in the readout weights is much slower than plasticity in the input connectivity ($\eta_w \ll \eta_u$). In that regime, Eq. 105 greatly simplifies, and we obtain

$$\delta^s = -\frac{h_0^s}{\langle \Psi'^2 \rangle}. \tag{106}$$

There are two sources of random fluctuations in h_0^s: different realizations of the circuit (via different initializations of the intermediate and readout connectivity, $\boldsymbol{u}$ and $\boldsymbol{w}$), and different sensory inputs. In the following, we show that these two sources of variability can be decomposed, and one can write

$$h_0^s = \langle \Psi \rangle \kappa + \sqrt{\langle \Psi^2 \rangle - \langle \Psi \rangle^2} \, \varepsilon^s \tag{107}$$

where κ and ε^s are zero-mean, unit-variance Gaussian variables. For a given circuit realization, the value of κ is fixed, while the value of ε^s fluctuates across different sensory inputs. Combining Eq. 107 with 104, we conclude that two different forms of variability (one that is frozen for a given circuit realization, represented by κ, and one that is not, represented by ε^s) impact activity coordinates c^{A} and c^{B}; the absolute and relative amplitude of the two contributions is controlled by the shape of the activation function Ψ. Such factorization of variability is illustrated, for an example simulated circuit, in Fig. S2E.

To derive Eq. 107, we consider a given circuit realization, and assume that the number of stimuli Q is sufficiently large, so that averages over stimuli approximately self-average. We start from Eq. 11b, and compute the mean of of h_0^s over sensory inputs, which yields

$$\langle \, h_0^s \, \rangle_s = \sum_{i=1}^{N} w_{0,i} \, \langle \, \Psi(k_{0,i}^s) \, \rangle_s = \langle \Psi \rangle \sum_{i=1}^{N} w_{0,i}. \tag{108}$$

By defining $\kappa \equiv \sum_{i=1}^{N} w_{0,i}$, the first term in the right-hand side of Eq. 107 follows. We then compute the variance of of h_0^s over sensory inputs. By using:

$$\langle \, (h_0^s)^2 \, \rangle_s = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{0,i} \, w_{0,j} \, \langle \, \Psi(k_{0,i}^s) \Psi(k_{0,j}^s) \, \rangle_s = \langle \Psi^2 \rangle + \langle \Psi \rangle^2 \sum_{i=1}^{N} \sum_{j \neq i} w_{0,i} \, w_{0,j} \tag{109}$$

and, from Eq. 108

$$\langle \, h_0^s \, \rangle_s^2 = \langle \Psi \rangle^2 \sum_{i=1}^{N} \sum_{j=1}^{N} w_{0,i} w_{0,j} = \langle \Psi \rangle^2 + \langle \Psi \rangle^2 \sum_{i=1}^{N} \sum_{j \neq i} w_{0,i} w_{0,j} \tag{110}$$

we conclude that:

$$\langle \, h_0^{s\,2} \, \rangle_s - \langle \, h_0^s \, \rangle_s^2 = \langle \Psi^2 \rangle - \langle \Psi \rangle^2 \, , \tag{111}$$

from which the second term in the right-hand side of Eq. 107 follows.

Equations 71 and 79 indicate that activity measures such as category selectivity and correlation depend on the value of activity coordinates c^{A} and c^{B}. As coordinates are variable (Eq. 104), activity measures are variable as well. Importantly, activity measures involve averages over sensory inputs (see Eqs. 61 and 76). This implies that the two forms of variability described by Eq. 107 are expected to contribute in different ways: variability originating from the second term (which fluctuates across stimuli, and thus can be averaged out) is expected to be small, while variability originating from the first term (which is fixed for each circuit realization) is expected to be large.

Variability in simulated circuits is quantified in Figs. S3-S4, where it is represented as error bars. Figures S3A and S4A show that variability in c^{A} and c^{B} is modulated by properties of the activation function Ψ (third column); this is in agreement with Eq. 107, which indicates that the magnitude of variability is Ψ-dependent. Figures S3B-C and S4B-C show, furthermore, that variability in correlation is typically much larger than in average selectivity. This can be explained by observing that average selectivity (Eq. 71) only depends on the difference between c^{A} and c^{B}, so variability originating from the first, frozen term of Eq. 107 is expected to cancel; this is not the case for correlation (Eq. 79), for which the cancellation does not occur.

4.8 Simple categorization task with structured inputs and heterogeneity

The circuit and task we considered so far are characterized by several simplifying modelling assumptions, which allowed us to analyze activity evolution in great detail and develop useful analytical intuition. One important assumption is that sensory input vectors corresponding to different stimuli are orthogonal to each other. This choice was motivated by two observations: first, in many tasks from the experimental literature, sensory stimuli are taken to be very different from each other, and thus sensory inputs are expected to be uncorrelated [1, 5, 7]; second, in tasks where sensory stimuli obey a continuous statistical structure [3], pre-processing from sensory brain regions [63] is expected to decorrelate, at least partially, inputs to higher-level associative areas. A second important assumption is that neurons in the intermediate layer are statistically homogeneous, as they receive statistically identical inputs and are characterized by the same nonlinearity Ψ.

For some tasks and brain regions, those two assumptions might be inaccurate. For example, data collected during passive conditions [64] indicate that some LIP neurons [3, 5, 12] display weak, but significant direction tuning, which might be due to structured sensory inputs. Furthermore, activity profiles are heterogeneous, with different neurons characterized by different baseline activity levels. To investigate whether our findings extrapolate beyond our two simplifying hypothesis, here we construct a more biologically-grounded model, and use simulations to systematically investigate activity evolution in the resulting circuit.

To begin with, we use sensory input vectors characterized by a continuous statistical structure, which implies continuous tuning in the intermediate layer activity prior to learning. We set

$$\boldsymbol{x}^s = \sqrt{1-\Sigma^2}\,\boldsymbol{\mu}^s + \Sigma\left[\boldsymbol{\xi}^1\,\cos(\theta^s) + \boldsymbol{\xi}^2\,\sin(\theta^s)\right] \tag{112}$$

where Σ is a scalar that measures the fraction of inputs variance that is continuous. We fixed $\Sigma = 1/3$. Like $\boldsymbol{\mu}^s$, entries of the vectors $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ are generated at random from a zero-mean, unit-variance Gaussian distribution. We furthermore set

$$\theta^s = s\frac{2\pi}{Q}\,. \tag{113}$$

With this choice, when $s \neq s'$, we have $\langle\, x_i^s x_i^{s'}\,\rangle = \Sigma^2\cos(2\pi(s-s')/Q)$, so stimuli with similar values of s are more strongly correlated than stimuli with very different values of s. As in [3], we take $Q = 12$. Similar to the standard task we analyzed so far, sensory inputs with $s = 1, \ldots, Q/2$ are associated with category A, while $s = Q/2, \ldots, Q$ are associated with category B. Note that, as in the simple categorization task we analyzed so far, sensory input vectors are linearly separable for every value of Σ.

To introduce heterogeneity in the intermediate layer, we add an offset, so Eq. 10 becomes

$$\boldsymbol{y} = \Psi(\boldsymbol{k}) \tag{114a}$$

$$\boldsymbol{k} \equiv \boldsymbol{u}\cdot\boldsymbol{x} + \boldsymbol{b}\,. \tag{114b}$$

The entries of $\boldsymbol{b}$ are fixed bias terms that control the value of baseline activity for each neuron. We generate those entries from a zero-mean Gaussian distribution with standard deviation 0.2.

In contrast to the model we analyzed so far, initial activity is characterized by non-trivial activity measures. Specifically, initial population tuning is characterized by non-vanishing category correlation; the latter is modulated both by heterogeneity (which tends to increase signal correlations) and the continuous inputs structure (which tends to decrease them). For our choice of parameters, these two effects roughly balance each other, so that initial activity is characterized by initial correlation that is small in magnitude (Figs. S5A).

We investigated numerically the evolution of activity with learning for this model. Two sample circuits are shown in Figs. S5B-C; extensive analysis is presented in Figs. S5D-E. We find that the behaviour of both category selectivity and correlation is qualitatively consistent with the behaviour of the simpler model analyzed so far. Specifically, we find that average category selectivity increases over learning (Figs. S5D); this behaviour is robust, and does not depend on circuit details. For completeness, we tested two definitions of category selectivity. The first one is identical to Eq. 61; as initial activity is structured, this gives slightly positive initial values; the second one (which is used in related experimental work [2, 3]) is again identical to Eq. 61 – but pairs of stimuli ss' are subsampled in a way that is tailored to inputs structure to yield vanishing initial selectivity. We show in Fig. S5D that both selectivity definitions give qualitatively similar results. Whether category correlation increases or decreases over learning depends, on the other hand, on parameters (Figs. S5B-C and E). Correlation depends on parameters in a way that is consistent with the simple task: it is strongly modulated by properties of the readout activation function Φ (Fig. S5E, different shades of gray). It also depends on the activation function of neurons in the intermediate layer Ψ (Fig. S5E left). Finally, it decreases with the learning ratio η_w/η_u (Fig. S5E center) and with the number of stimuli Q (Fig. S5E right).

5 Context-dependent categorization task

5.1 Task definition

The second task we consider is a context-dependent categorization task. On each trial, both a stimulus, and a context cue, are presented to the network. For simplicity, we assume that the number of stimuli and context cues is identical, and is equal to Q. As in the simple task, each stimulus is represented by an input vector $\boldsymbol{\mu}^S$, with $S = 1, \dots, Q$; each context cue is also represented by an input vector, denoted $\boldsymbol{\nu}^C$, with $C = 1, \dots, Q$. The entries of both vectors, $\boldsymbol{\mu}^S$ and $\boldsymbol{\nu}^C$, are generated independently from a zero-mean, unit-variance Gaussian distribution. The total sensory input on each trial, $\boldsymbol{x}^s$, is given by the linear combination of the stimulus and context cue inputs,

$$\boldsymbol{x}^s = \frac{1}{\sqrt{2}} \left(\boldsymbol{\mu}^{S_s} + \boldsymbol{\nu}^{C_s} \right). \tag{115}$$

All combinations of stimuli and context cues are permitted; the total number of trials and sensory inputs is thus $P = Q^2$. Each trial s is thus specified by a stimulus and context index: $s = (S_s C_s)$. In contrast to the simple task, sensory input vectors are not orthogonal among each other; using Eq. 115, we see that to the leading order in N,

$$\langle\, x_i^s x_i^{s'} \,\rangle_i \simeq \begin{cases} 1 & S_s = S'_s \text{ and } C_s = C'_s \ (s = s') \\ 1/2 & S_s = S'_s \text{ or } C_s = C'_s \\ 0 & \text{otherwise.} \end{cases} \tag{116}$$

The task is defined as follows. When the context cue C ranges between 1 and $Q/2$, context takes value 1. In context 1, the first half of the Q stimuli is associated with category A ($\tilde{z} = z^{\mathsf{A}}$), and the second half with B ($\tilde{z} = z^{\mathsf{B}}$). When the context cue C ranges between $Q/2$ and Q, context takes value 2. In context 2, stimuli-category associations are reversed: the first half of the Q stimuli is associated with category B ($\tilde{z} = z^{\mathsf{B}}$), and the second half with A ($\tilde{z} = z^{\mathsf{A}}$).

Correlations in the sensory inputs (Eq. 116) are such that, for every value of Q, inputs are not linearly-separable [39]. For $Q = 2$, the task is equivalent to a classical XOR computation. We focus however on $Q > 2$, for which each context is signaled by more than one context cue. As in experimental work [7, 8, 45], this allows to dissociate the activity dependence on the *abstract* variable context from the *sensory* variable context cue (see Eqs. 127 and 128 in Section 5.3).

We start by writing down explicit expressions for the activity (Eq. 34) in the current task (Sections 5.2). We then derive the expressions that quantify how activity measures, such as selectivity and correlations, evolve over learning (Sections 5.3, 5.4 and 5.5). These expressions are rather complex, and require numerical evaluation. To gain further mathematical insight, in Sections 5.6, 5.7 and 5.8 we consider specific cases and quantities, and derive their behaviour analytically.

5.2 Computing activity

We start by computing the value of coordinates c^s, which are solution to the linear system in Eq. 31. As in Section 4.2 (see also Section 4.7), we neglect the variable term h_0^s in the left-hand side of that equation and, after a small amount of algebra, we find that it can be rewritten as

$$\Phi^{-1}(\tilde{z}^s) = c^s \left(\langle \Psi'^2 \rangle + \frac{\eta_w}{\eta_u} \left(\langle \Psi^2 \rangle - \langle \Psi \rangle^2 \right) \right) + \sum_{q \in N(s)} c^q \left(\frac{1}{2} \langle \Psi' \Psi' \rangle + \frac{\eta_w}{\eta_u} \left(\langle \Psi \Psi \rangle - \langle \Psi \rangle^2 \right) \right) + \sum_{q=1}^{P} c^q \, \frac{\eta_w}{\eta_u} \, \langle \Psi \rangle^2 \tag{117}$$

where we used the short-hand notation $N(s)$ to indicate the set of trials that are *neighbours* to s (i.e., trials that have either the same stimulus or the same context cue of s). We have used the notation $\langle FF \rangle$ to indicate the average over the product of two nonlinear functions, F, whose arguments are given by two zero-mean and unit-variance Gaussian variables with covariance 1/2. I.e.,

$$\langle FF \rangle \equiv \langle\, F(a) F(b) \,\rangle_{a,b} \tag{118}$$

where both a and b are zero-mean, unit-variance Gaussian random variables with covariance 1/2. Detail on how these averages are computed numerically is given in Section 6.3 (Eq. 191).

As in the simple task (Eq. 51), because the left-hand side can take on only two values, the coordinates c^s can take on only two values,

$$c^s = \begin{cases} c^{\mathsf{A}} & s \text{ in category } \mathsf{A} \\ c^{\mathsf{B}} & s \text{ in category } \mathsf{B}\,. \end{cases} \tag{119}$$

The values of c^{A} and c^{B} are determined by the same linear system as in Eq. 52, except now α and β are given by

$$\alpha = \langle \Psi'^2 \rangle + \frac{\eta_w}{\eta_u} \left(\langle \Psi^2 \rangle - \langle \Psi \rangle^2 \right) - 2 \left[\frac{1}{2} \langle \Psi' \Psi' \rangle + \frac{\eta_w}{\eta_u} \left(\langle \Psi \Psi \rangle - \langle \Psi \rangle^2 \right) \right] \tag{120a}$$

$$\beta = Q \left[\frac{1}{2} \langle \Psi' \Psi' \rangle + \frac{\eta_w}{\eta_u} \left(\langle \Psi \Psi \rangle - \langle \Psi \rangle^2 \right) \right] + \frac{\eta_w}{\eta_u} \frac{Q^2}{2} \langle \Psi \rangle^2. \tag{120b}$$

To derive the expression above, we used the fact that every sensory input has $2Q-2$ neighbours, of which $Q-2$ are associated with the same category, and Q with the opposite one. The final expression for c^{A} and c^{B} is thus given by Eq. 54; that expression depends on γ, which is given in Eq. 55.

By comparing Eq. 120 with 53 we see that, with respect to the simple task, the expressions for α and β include extra terms (shown in square brackets in the right-hand side of Eq. 120). These arise because, unlike in the simple task, different inputs can be correlated (Eq. 116). The extra term in the expression for β (Eq. 120b) scales with Q, while the extra term for α (Eq. 120a) does not; this indicates the typical value of γ (Eq. 55), which is proportional to β/α, is larger in this task than in the simple one. This in turn implies that the parameter region where one has approximately $c^{\mathsf{A}} \simeq -c^{\mathsf{B}}$ is larger in the current task than in the simple one; this approximation will later be used in Section 5.8. In the simple task, the parameter region where $c^{\mathsf{A}} \simeq -c^{\mathsf{B}}$ coincided with the region where category correlation were negative (Eq. 79, Section 4.4). This suggests that the parameter region where correlations are negative, also, is larger in this task than in the simple one. As it will be shown in Section 5.4, however, the expressions for correlations are much more complex in the current task than Eq. 79; this hypothesis thus needs to be carefully verified – which is done, using numerical integration, in Fig. S7C.

Since this task is an extension of the XOR task, sensory inputs are not linearly-separable. This shows up as a singularity when the intermediate layer is linear (e.g., $\Psi(\boldsymbol{x}) = \boldsymbol{x}$). Indeed, in that case, the value of γ (Eq. 55) diverges, which in turn means both c^{A} and c^{B} diverge (Eq. 54). That's because γ is proportional to the ratio β/α, and α vanishes, while β does not. To see that α vanishes, we use Eqs. 120a to write

$$\begin{aligned}\alpha &= 1 + \frac{\eta_w}{\eta_u}\left(\langle\Psi^2\rangle - \langle\Psi\rangle^2\right) - 2\left[\frac{1}{2} + \frac{\eta_w}{\eta_u}\left(\langle\Psi\Psi\rangle - \langle\Psi\rangle^2\right)\right] \\ &= \frac{\eta_w}{\eta_u}\left[\langle\Psi^2\rangle - 2\langle\Psi\Psi\rangle + \langle\Psi\rangle^2\right] \\ &= \frac{\eta_w}{\eta_u}\left[1 - 2\cdot\frac{1}{2} + 0\right] = 0.\end{aligned} \tag{121}$$

When the activation function Ψ is nonlinear, instead, the values of c^{A} and c^{B} are finite; their magnitude depends on how close to linear Ψ is in its effective activation range.

To conclude our characterization of activity, we evaluate spanning vectors, $\boldsymbol{v}^{qs}$, by combining Eqs. 33 and 116. Unlike in the simple task, for each activity vector, $\boldsymbol{y}^s$, there exists more than one spanning vector; those are given by $\boldsymbol{v}^{ss}$, and all vectors $\boldsymbol{v}^{qs}$ for which $q \in N(s)$. Eq. 34 thus reads

$$\begin{aligned}\boldsymbol{y}^s &= \boldsymbol{y}_0^s + c^s\boldsymbol{v}^{ss} + \sum_{q\in N(s)} c^q\boldsymbol{v}^{qs} \\ &= \boldsymbol{y}_0^s + c^s\,\boldsymbol{w}_0\odot\Psi'(\boldsymbol{k}_0^s)\odot\Psi'(\boldsymbol{k}_0^s) + \frac{1}{2}\sum_{q\in N(s)} c^q\,\boldsymbol{w}_0\odot\Psi'(\boldsymbol{k}_0^q)\odot\Psi'(\boldsymbol{k}_0^s)\end{aligned} \tag{122}$$

where the second line follows from Eq. 33 and the coordinates c^q take values c^{A} or c^{B} depending on the category $\boldsymbol{x}^q$ is associated with (Eq. 119). Using the notation $s = (S_sC_s)$, Eq. 122 can also be written in the compact form

$$\boldsymbol{y}^{S_sC_s} = \boldsymbol{y}_0^{S_sC_s} + \frac{1}{2}\sum_{S=1}^{Q} c^{SC_s}\,\boldsymbol{w}_0\odot{\Psi'}^{SC_s}\odot{\Psi'}^{S_sC_s} + \frac{1}{2}\sum_{C=1}^{Q} c^{S_sC}\,\boldsymbol{w}_0\odot{\Psi'}^{S_sC}\odot{\Psi'}^{S_sC_s} \tag{123}$$

where we used the short-hand notation ${\Psi'}^{SC} \equiv \Psi'(\boldsymbol{k}_0^{SC})$.

To isolate the effect of the nonlinearity Ψ, it will be instructive (see Sections 5.6 and 5.7) to also compute the synaptic drive, $\boldsymbol{k}^s$, after learning. Using Eq. 10b and 23a, it is easy to see that

$$\boldsymbol{k}^s \equiv \boldsymbol{k}_0^s + \Delta\boldsymbol{k}^s = \boldsymbol{k}_0^s + c^s\,\boldsymbol{w}_0\odot\Psi'(\boldsymbol{k}_0^s) + \frac{1}{2}\sum_{q\in N(s)} c^q\,\boldsymbol{w}_0\odot\Psi'(\boldsymbol{k}_0^q), \tag{124}$$

or, equivalently,

$$\boldsymbol{k}^{S_sC_s} = \boldsymbol{k}_0^{S_sC_s} + \frac{1}{2}\sum_{S=1}^{Q} c^{SC_s}\,\boldsymbol{w}_0\odot{\Psi'}^{SC_s} + \frac{1}{2}\sum_{C=1}^{Q} c^{S_sC}\,\boldsymbol{w}_0\odot{\Psi'}^{S_sC}. \tag{125}$$

We conclude with a remark on the geometry of the spanning vectors, $\boldsymbol{v}^{qs}$. As in the simple task, those include a component that is aligned with the initial readout vector, $\boldsymbol{w}_0$, and a residual component that is perpendicular to it, $\delta\boldsymbol{v}^{qs}$ (Eq. 36). In the simple task, residual components could be neglected (Eq. 60) because they were orthogonal to each other, and did not contribute to novel activity structure. In this task, residual components are not, in general, orthogonal to each other, and thus cannot be neglected. In fact, we have

$$\delta\boldsymbol{v}^{qs}\cdot\delta\boldsymbol{v}^{q's'} = \langle\, x_i^q x_i^s\,\rangle_i\,\langle\, x_i^{q'} x_i^{s'}\,\rangle_i\,\left[\langle\,\Psi'(k_{0,i}^q)\,\Psi'(k_{0,i}^s)\,\Psi'(k_{0,i}^{q'})\,\Psi'(k_{0,i}^{s'})\,\rangle_i - \langle\,\Psi'(k_{0,i}^q)\Psi'(k_{0,i}^s)\,\rangle_i\langle\,\Psi'(k_{0,i}^{q'})\Psi'(k_{0,i}^{s'})\,\rangle_i\right]. \tag{126}$$

The term in the right-hand side can be non-zero even when sq are different from $s'q'$; this is due to Eqs. 48b and 116, which imply that $k_{0,i}$ variables can be correlated among each other. The fact that residuals $\delta \boldsymbol{v}^{qs}$ cannot be neglected implies that activity evolution is not effectively one-dimensional, as it was the simple task, but higher-dimensional (this is evident in the PC plots in Fig. S6C-D). All the directions along which activity evolve are, however, correlated with the initial readout vector $\boldsymbol{w}_0$ (Eq. 35).

5.3 Category and context selectivity

In the present task, we can compute category, as well as context selectivity. In analogy with category selectivity, Eq. 61, context selectivity is defined as

$$S_i^{\text{ctx}} = \frac{\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff ctx}} - \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s \neq s' \text{ same ctx, diff ctx cue}}}{\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff ctx}} + \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s \neq s' \text{ same ctx, diff ctx cue}}}. \tag{127}$$

Note that, in the average over pairs of trials from the same context, we excluded pairs of trials with the same context cue. This was done to exclude the possibility that context selectivity increases simply because activity in response to the same context cue become more similar over learning. For completeness, we also compute

$$S_i^{\text{ctx},2} = \frac{\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff ctx}} - \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s \neq s' \text{ same ctx}}}{\langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s,\, s' \text{ diff ctx}} + \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_{s \neq s' \text{ same ctx}}}, \tag{128}$$

which we plot in Figs. S6A-B. Those plots show that the behaviour under this definition is similar to that of Eq. 127.

We are interested in deriving theoretical expressions for average category and context selectivity, obtained by averaging Eqs. 61 and 127 (or 128) over i. For the present task, that is hard. Consequently, we use results from the simple task (Section 4.3) which indicated that, in the limit $N \gg Q \gg 1$, average category selectivity can be approximated with the category clustering measure, Eq. 72; the latter is equivalent to separately averaging the numerator and denominator of selectivity over neurons.

For category, clustering is the same as in the simple task, Eqs. 72 and 73, which we repeat here for convenience,

$$\begin{aligned} \tilde{S}^{\text{cat}} &= \frac{\langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s,\, s' \text{ diff cat}} - \langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s \neq s' \text{ same cat}}}{\langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s,\, s' \text{ diff cat}} + \langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s \neq s' \text{ same cat}}} \\ &= \frac{-\langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s,\, s' \text{ diff cat}} + \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s \neq s' \text{ same cat}}}{2\langle\, \langle\, (y_i^s)^2 \,\rangle_i \,\rangle_s - \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s,\, s' \text{ diff cat}} - \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s \neq s' \text{ same cat}}} \end{aligned} \tag{129}$$

where we used the statistical homogeneity of activity vectors. Similarly, for context selectivity, we may write

$$\begin{aligned} \tilde{S}^{\text{ctx}} &= \frac{\langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s,\, s' \text{ diff ctx}} - \langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s \neq s' \text{ same ctx, diff ctx cue}}}{\langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s,\, s' \text{ diff ctx}} + \langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s,\, s' \text{ same ctx, diff ctx cue}}} \\ &= \frac{-\langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s,\, s' \text{ diff ctx}} + \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s \neq s' \text{ same ctx, diff ctx cue}}}{2\langle\, \langle\, (y_i^s)^2 \,\rangle_i \,\rangle_s - \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s,\, s' \text{ diff ctx}} - \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s,\, s' \text{ same ctx, diff ctx cue}}} \end{aligned} \tag{130}$$

and

$$\begin{aligned} \tilde{S}^{\text{ctx},2} &= \frac{\langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s,\, s' \text{ diff ctx}} - \langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s \neq s' \text{ same ctx}}}{\langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s,\, s' \text{ diff ctx}} + \langle\, \langle\, (y_i^s - y_i^{s'})^2 \,\rangle_i \,\rangle_{s \neq s' \text{ same ctx}}} \\ &= \frac{-\langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s,\, s' \text{ diff ctx}} + \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s \neq s' \text{ same ctx}}}{2\langle\, \langle\, (y_i^s)^2 \,\rangle_i \,\rangle_s - \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s,\, s' \text{ diff ctx}} - \langle\, \langle\, y_i^s\, y_i^{s'} \,\rangle_i \,\rangle_{s \neq s' \text{ same ctx}}}. \end{aligned} \tag{131}$$

.

To evaluate those expressions, we need the normalized dot products over activity, $\langle\, y_i^s\, y_i^{s'} \,\rangle_i$. These are computed in Section 5.5. Finally, averages over trials are performed numerically. The resulting theoretical estimates for $\tilde{S}^{\text{cat}}$ and $\tilde{S}^{\text{ctx}}$ are shown in Figs. S6B and S9.

In Figs. S9A-C, we compare theoretical estimates with simulations. Agreement is relatively good, although it is worse than for the simple task; as argued in Section 3.4, that is expected. Note that the values of average selectivity and clustering are not close (this is only verified in the $N \gg Q \gg 1$ limit, and would require values of N larger than those used in simulations); the qualitative behaviour of the two quantities is, however, identical. In Fig. S6B, we plot the theoretical estimates across a broad range of task and circuit parameters. These theoretical estimates indicate that, in all cases, category (Eq. 129) and context (Eqs. 130, 131) selectivity increase. This is in agreement with simulations, which are reported in Fig. S6A.

5.4 Category and context correlation

To quantify how the population as a whole encodes category and context, we evaluate category and context correlations. Those quantities, denoted C^{cat} and C^{ctx}, are given by the average Pearson correlation coefficient for trials in different categories and contexts. C^{cat} is defined as in Eq. 76. Similarly, C^{ctx} is defined as

$$C^{\text{ctx}} = \langle\, C^{s_1 s_2} \,\rangle_{s_1 s_2} \tag{132}$$

where s_1 and s_2 are indices that denote, respectively, trials from contexts 1 and 2. Similarly to Eq. 77, the Pearson correlation coefficient $C^{s_1 s_2}$ is given by

$$C^{s_1 s_2} = \frac{\langle\, y_i^{s_1} y_i^{s_2} \,\rangle_i - \langle \Psi \rangle^2}{\sqrt{\langle\, y_i^{s_1} y_i^{s_1} \,\rangle_i - \langle \Psi \rangle^2}\sqrt{\langle\, y_i^{s_2} y_i^{s_2} \,\rangle_i - \langle \Psi \rangle^2}}. \tag{133}$$

To evaluate these expressions, we use the normalized dot products $\langle\, y_i^s\, y_i^{s'} \,\rangle_i$ that are computed in Section 5.5. Averaging over trials is, finally, done numerically.

For completeness, we also consider the alternative definition of correlations, where activity is averaged over trials first, and then the Pearson correlation is computed. The alternative definition for category correlation is identical to Eq. 84. The alternative definition for context correlation is given by

$$\begin{aligned} C^{\text{ctx}} &= \frac{\langle\, y_i^1 y_i^2 \,\rangle_i - \langle\, y_i^1 \,\rangle_i \langle\, y_i^2 \,\rangle_i}{\sqrt{\langle\, y_i^1 y_i^1 \,\rangle_i - \langle\, y_i^1 \,\rangle_i^2}\sqrt{\langle\, y_i^2 y_i^2 \,\rangle_i - \langle\, y_i^2 \,\rangle_i^2}} \\ &= \frac{\langle\, y_i^1 y_i^2 \,\rangle_i - \langle \Psi \rangle^2}{\sqrt{\langle\, y_i^1 y_i^1 \,\rangle_i - \langle \Psi \rangle^2}\sqrt{\langle\, y_i^2 y_i^2 \,\rangle_i - \langle \Psi \rangle^2}} \end{aligned} \tag{134}$$

where we have defined

$$\begin{aligned} \boldsymbol{y}^1 &= \langle\, \boldsymbol{y}^{s_1} \,\rangle_{s_1} \\ \boldsymbol{y}^2 &= \langle\, \boldsymbol{y}^{s_2} \,\rangle_{s_2}. \end{aligned} \tag{135}$$

For the current task, there exists no simple mathematical relationship between correlations obtained from the standard, and the alternative definition. We thus checked numerically the behaviour of both quantities; results are reported in Fig. S6B. As in the simple task, we found that the qualitative behaviour of both quantities is not fixed, but depends on task and circuit parameters. This is in agreement with simulations, which are illustrated in Fig. S6A.

5.5 Computing normalized dot products

To conclude, we illustrate how normalized dot products, Eq. 89, are computed for the current task. We start from Eq. 96, which we repeat here for completeness,

$$\langle\, y_i^s y_i^{s'} \,\rangle_i = \langle\, y_{0,i}^s\, y_{0,i}^{s'} \,\rangle_i + \langle\, \Delta y_i^s \Delta y_i^{s'} \,\rangle_i. \tag{136}$$

The first term of the right-hand side reads

$$\langle\, y_{0,i}^s\, y_{0,i}^{s'} \,\rangle_i = \begin{cases} \langle \Psi^2 \rangle & S_s = S_s' \text{ and } C_s = C_s' \ (s = s') \\ \langle \Psi\Psi \rangle & S_s = S_s' \text{ or } C_s = C_s' \\ \langle \Psi \rangle^2 & \text{otherwise,} \end{cases} \tag{137}$$

where we used Eq. 10 together with Eqs. 48b and 116. Using Eq. 34 together with 33, the second term of the right-hand side of Eq. 136 reads

$$\langle\, \Delta y_i^s \Delta y_i^{s'} \,\rangle_i = \frac{1}{N} \sum_{q=1}^{P} \sum_{q'=1}^{P} c^q c^{q'} \,\langle\, x_i^q x_i^s \,\rangle_i \,\langle\, x_i^{q'} x_i^{s'} \,\rangle_i \,\langle\, \Psi'(k_{0,i}^q)\Psi'(k_{0,i}^s)\Psi'(k_{0,i}^{q'})\Psi'(k_{0,i}^{s'}) \,\rangle_i \tag{138}$$

where sensory input correlations, $\langle\, x_i^q x_i^s \,\rangle_i$, are given in Eq. 116.

Because $\langle\, x_i^q x_i^s \,\rangle_i$ can be non-zero even when $s \neq q$ (Eq. 116), the number of non-zero terms in the sum in Eq. 138 is, in general, large. Each term contains an average, $\langle\, \Psi'(k_{0,i}^q)\Psi'(k_{0,i}^s)\Psi'(k_{0,i}^{q'})\Psi'(k_{0,i}^{s'}) \,\rangle_i$, that includes four nonlinear functions. The value of those averages is specified by the correlations among the arguments, $k_{0,i}$, which in turn depend on the values of s, q, s' and q' (Eq. 116, via Eq. 48b). Averages are evaluated numerically; detail on how this is done is given in Section 6.3.

This procedure yields a set of normalized dot products that can be used to evaluate, numerically, the expressions for activity selectivity and correlation derived in Sections 5.3 and 5.4. As we rely on numerics, the results we obtain in this way are hard to interpret. For this reason, in the next sections we focus on specific cases were results can be obtained analytically; this allows us to extract a more intuitive understanding of how activity measures evolve over learning.

5.6 Detailed analysis of context selectivity

We start clarifying how context selectivity increases over learning. Results from simulations, and numerical integration of Eq. 130, indicate that context selectivity increases for the synaptic drive, $\boldsymbol{k}^s$; this increase is then reflected in the activity, $\boldsymbol{y}^s$ (Figs. S6A-B and S9B). In this section, we analyze the behaviour of context selectivity for the synaptic drive. Focusing on the synaptic drive, instead of activity, allows us to derive results analytically. In the following, we start from Eq. 130 and show that, for the synaptic drive $\boldsymbol{k}^s$, the value of $\tilde{S}^{\text{ctx}}$ increases over learning. At the end of this section, we comment on the insights provided by such derivation.

We start by simplifying the sums over trials contained in Eq. 130, which involve pairs of trials ss' from the same, or different context. To this end we observe that, because of task symmetries, these sums involve a large number of identical terms; for example, the term with $s = (11)$ and $s' = (12)$ is identical to $s = (21)$ and $s' = (22)$ (both pairs of trials are neighbours, and are associated with the same category). We thus perform averages over a reduced, and less redundant subset of pairs of trials. First, we consider only two values of s: for concreteness, we take $s = (11)$ and $s = ((\hat{Q}+1)1)$, where we defined

$$\hat{Q} \equiv \frac{Q}{2}. \tag{139}$$

These s trials are associated, respectively, with category A and B. Second, for each value of s, we consider s' trials with context cue equal to $C = 2$ and $C = \hat{Q}+1$; these are associated, respectively, with context 1 and 2 (note that $C = 1$ must be avoided, as trials with the same context cue must be excluded, see Eq. 130). This allows us to rewrite the averages contained in Eq. 130 as

$$\langle\, \langle\, (k_i^s - k_i^{s'})^2 \,\rangle_i \,\rangle_{s,\, s' \text{ same ctx, diff ctx cue}} = \frac{1}{2Q}\left[\sum_{\bar{S}=1}^{Q} \langle\, (k_i^{11} - k_i^{\bar{S}2})^2 \,\rangle_i + \sum_{\bar{S}=1}^{Q} \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{\bar{S}2})^2 \,\rangle_i\right] \tag{140a}$$

$$\langle\, \langle\, (k_i^s - k_i^{s'})^2 \,\rangle_i \,\rangle_{s,\, s' \text{ diff ctx}} = \frac{1}{2Q}\left[\sum_{\bar{S}=1}^{Q} \langle\, (k_i^{11} - k_i^{\bar{S}(\hat{Q}+1)})^2 \,\rangle_i + \sum_{\bar{S}=1}^{Q} \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{\bar{S}(\hat{Q}+1)})^2 \,\rangle_i\right]. \tag{140b}$$

The sums over $\bar{S}$ can further be simplified. By using again symmetries, we have:

$$\sum_{\bar{S}=1}^{Q} \langle\, (k_i^{11} - k_i^{\bar{S}2})^2 \,\rangle_i = \langle\, (k_i^{11} - k_i^{12})^2 \,\rangle_i + (\hat{Q}-1)\langle\, (k_i^{11} - k_i^{22})^2 \,\rangle_i + \hat{Q}\langle\, (k_i^{11} - k_i^{(\hat{Q}+1)2})^2 \,\rangle_i. \tag{141}$$

We can do the same for the other sums, yielding:

$$\sum_{\bar{S}=1}^{Q} \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{\bar{S}2})^2 \,\rangle_i = \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+1)2})^2 \,\rangle_i + (\hat{Q}-1)\langle\, (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+2)2})^2 \,\rangle_i + \hat{Q}\langle\, (k_i^{(\hat{Q}+1)1} - k_i^{12})^2 \,\rangle_i \tag{142a}$$

$$\sum_{\bar{S}=1}^{Q} \langle\, (k_i^{11} - k_i^{\bar{S}(\hat{Q}+1)})^2 \,\rangle_i = \langle\, (k_i^{11} - k_i^{1(\hat{Q}+1)})^2 \,\rangle_i + (\hat{Q}-1)\langle\, (k_i^{11} - k_i^{2(\hat{Q}+1)})^2 \,\rangle_i + \hat{Q}\langle\, (k_i^{11} - k_i^{(\hat{Q}+1)(\hat{Q}+1)})^2 \,\rangle_i \tag{142b}$$

$$\sum_{\bar{S}=1}^{Q} \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{\bar{S}(\hat{Q}+1)})^2 \,\rangle_i = \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+1)(\hat{Q}+1)})^2 \,\rangle_i + (\hat{Q}-1)\langle\, (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+2)(\hat{Q}+1)})^2 \,\rangle_i + \hat{Q}\langle\, (k_i^{(\hat{Q}+1)1} - k_i^{1(\hat{Q}+1)})^2 \,\rangle_i. \tag{142c}$$

It is easy to verify that, before learning starts, the right-hand sides of Eqs. 140a and b are identical. This implies that the initial value of context selectivity, Eq. 130, vanishes (Fig. S9B). To show that context selectivity increases over learning, we thus need to show that the numerator of Eq. 130 becomes positive over learning. This is equivalent to show that Eq. 140a is smaller than Eq. 140b. Using Eq. 141 and 142, this condition can be rewritten as

$$\begin{aligned}
&\hat{Q}\left[\langle\, (k_i^{11} - k_i^{22})^2 \,\rangle_i + \langle\, (k_i^{11} - k_i^{(\hat{Q}+1)2})^2 \,\rangle_i + \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+2)2})^2 \,\rangle_i + \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{12})^2 \,\rangle_i\right] + \\
&\langle\, (k_i^{11} - k_i^{12})^2 \,\rangle_i + \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+1)2})^2 \,\rangle_i - \langle\, (k_i^{11} - k_i^{22})^2 \,\rangle_i - \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+2)2})^2 \,\rangle_i < \\
&\hat{Q}\left[\langle\, (k_i^{11} - k_i^{2(\hat{Q}+1)})^2 \,\rangle_i + \langle\, (k_i^{11} - k_i^{(\hat{Q}+1)(\hat{Q}+1)})^2 \,\rangle_i + \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+2)(\hat{Q}+1)})^2 \,\rangle_i + \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{1(\hat{Q}+1)})^2 \,\rangle_i\right] + \\
&\langle\, (k_i^{11} - k_i^{1(\hat{Q}+1)})^2 \,\rangle_i + \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+1)(\hat{Q}+1)})^2 \,\rangle_i - \langle\, (k_i^{11} - k_i^{2(\hat{Q}+1)})^2 \,\rangle_i - \langle\, (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+2)(\hat{Q}+1)})^2 \,\rangle_i.
\end{aligned} \tag{143}$$

We now use Eq. 124 to write

$$\langle\, (k_i^s - k_i^{s'})^2 \,\rangle_i = \langle\, (k_{0,i}^s)^2 \,\rangle_i + \langle\, (k_{0,i}^{s'})^2 \,\rangle_i - 2\langle\, k_{0,i}^s k_{0,i}^{s'} \,\rangle_i + \langle\, (\Delta k_i^s)^2 \,\rangle_i + \langle\, (\Delta k_i^{s'})^2 \,\rangle_i - 2\langle\, \Delta k_i^s \Delta k_i^{s'} \,\rangle_i \tag{144}$$

where the terms containing the cross-products between k_i^s and Δk_i^s vanish on average because of Eq. 13a. By using the statistical homogeneity of activity across contexts, we can rewrite Eq. 143 as

$$\begin{aligned}
&\hat{Q}\left[\langle\, \Delta k_i^{11}\Delta k_i^{22}\,\rangle_i + \langle\, \Delta k_i^{11}\Delta k_i^{(\hat{Q}+1)2}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+2)2}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{12}\,\rangle_i\right] + \\
&\langle\, \Delta k_i^{11}\Delta k_i^{12}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+1)2}\,\rangle_i - \langle\, \Delta k_i^{11}\Delta k_i^{22}\,\rangle_i - \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+2)2}\,\rangle_i > \\
&\hat{Q}\left[\langle\, \Delta k_i^{11}\Delta k_i^{2(\hat{Q}+1)}\,\rangle_i + \langle\, \Delta k_i^{11}\Delta k_i^{(\hat{Q}+1)(\hat{Q}+1)}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+2)(\hat{Q}+1)}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{1(\hat{Q}+1)}\,\rangle_i\right] + \\
&\langle\, \Delta k_i^{11}\Delta k_i^{1(\hat{Q}+1)}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+1)(\hat{Q}+1)}\,\rangle_i - \langle\, \Delta k_i^{11}\Delta k_i^{2(\hat{Q}+1)}\,\rangle_i - \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+2)(\hat{Q}+1)}\,\rangle_i
\end{aligned} \tag{145}$$

or, re-arranging terms,

$$\begin{aligned}
&\hat{Q}\left\{\left[\langle\, \Delta k_i^{11}\Delta k_i^{22}\,\rangle_i + \langle\, \Delta k_i^{11}\Delta k_i^{(\hat{Q}+1)2}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+2)2}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{12}\,\rangle_i\right] - \right. \\
&\left.\left[\langle\, \Delta k_i^{11}\Delta k_i^{2(\hat{Q}+1)}\,\rangle_i + \langle\, \Delta k_i^{11}\Delta k_i^{(\hat{Q}+1)(\hat{Q}+1)}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+2)(\hat{Q}+1)}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{1(\hat{Q}+1)}\,\rangle_i\right]\right\} + \\
&\left\{\left[\langle\, \Delta k_i^{11}\Delta k_i^{12}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+1)2}\,\rangle_i - \langle\, \Delta k_i^{11}\Delta k_i^{22}\,\rangle_i - \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+2)2}\,\rangle_i\right] - \right. \\
&\left.\left[\langle\, \Delta k_i^{11}\Delta k_i^{1(\hat{Q}+1)}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+1)(\hat{Q}+1)}\,\rangle_i - \langle\, \Delta k_i^{11}\Delta k_i^{2(\hat{Q}+1)}\,\rangle_i - \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+2)(\hat{Q}+1)}\,\rangle_i\right]\right\} > 0.
\end{aligned} \tag{146}$$

To show that context selectivity increases over learning, we need to verify that the equation above holds. To this end, we evaluate analytically the normalized dot products $\langle\, \Delta k_i^{SC}\Delta k_i^{S'C'}\,\rangle_i$ for each pair of trials involved. This is done in the next paragraph; here we simply use those results (Eqs. 161, 162, 163 and 164).

We start evaluating the difference within the first set of curly parenthesis of Eq. 146, which correspond to the dominant contribution in Q. By using Eq. 164, we see that this can be rewritten as

$$\left[\langle\, \Delta k_i^{11}\Delta k_i^{22}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{(\hat{Q}+2)2}\,\rangle_i\right] - \left[\langle\, \Delta k_i^{11}\Delta k_i^{(\hat{Q}+1)(\hat{Q}+1)}\,\rangle_i + \langle\, \Delta k_i^{(\hat{Q}+1)1}\Delta k_i^{1(\hat{Q}+1)}\,\rangle_i\right]. \tag{147}$$

Using Eq. 163, this becomes

$$\frac{(c^{\mathsf{A}} - c^{\mathsf{B}})^2}{N}[\mathcal{A}_1 - \mathcal{A}_2]. \tag{148}$$

We then evaluate the difference within the second set of curly parenthesis. Using Eqs. 161, 162, 163 and 164 it is straightforward to see that that difference vanishes. Putting results together, our condition to verify (Eq. 146) becomes simply:

$$(c^{\mathsf{A}} - c^{\mathsf{B}})^2[\mathcal{A}_1 - \mathcal{A}_2] > 0, \tag{149}$$

which is satisfied whenever $\mathcal{A}_1 - \mathcal{A}_2 > 0$. This is always verified, as from Eqs. 155 and 156 we have

$$\mathcal{A}_1 = \sum_{S+,S'_+} \langle S_+1, S'_+2\rangle = \frac{Q}{2}\langle \Psi'\Psi'\rangle + \frac{Q}{2}\left(\frac{Q}{2} - 1\right)\langle \Psi'\rangle^2 \tag{150}$$

while

$$\mathcal{A}_2 = \sum_{S+,S'_-} \langle S_+1, S'_-2\rangle = \left(\frac{Q}{2}\right)^2 \langle \Psi'\rangle^2 \tag{151}$$

so that

$$\mathcal{A}_1 - \mathcal{A}_2 = \frac{Q}{2}(\langle \Psi'\Psi'\rangle - \langle \Psi'\rangle^2) > 0 \tag{152}$$

which concludes our derivation. We remark that Eq. 152 vanishes when Ψ is linear. This indicates that, even if context selectivity also increases for synaptic drives (which are a linear transformation of the sensory inputs), this phenomenon is due to the nonlinearity of activation functions.

Computing normalized dot products We now compute the normalized dot product expressions, $\langle\, \Delta k_i^{SC}\Delta k_i^{S'C'}\,\rangle_i$, for each pair of trials involved in Eq. 146. We illustrate in detail how one example dot product, $\langle\, \Delta k_i^{11}\Delta k_i^{12}\,\rangle_i$, is computed. Other expressions are computed in a similar way; results are given below (Eqs. 162, 163 and 164).

We start from:

$$\begin{aligned}
\langle\, \Delta k_i^{11}\Delta k_i^{12}\,\rangle_i = \frac{1}{4N}\Bigg[&\sum_{S,S'} c^{S1}c^{S'2}\langle\, {\Psi'}_i^{S1}{\Psi'}_i^{S'2}\,\rangle_i + \sum_{C,C'} c^{1C}c^{1C'}\langle\, {\Psi'}_i^{1C}{\Psi'}_i^{1C'}\,\rangle_i \\
&+ \sum_{S,C'} c^{S1}c^{1C'}\langle\, {\Psi'}_i^{S1}{\Psi'}_i^{1C'}\,\rangle_i + \sum_{C,S'} c^{1C}c^{S'2}\langle\, {\Psi'}_i^{1C}{\Psi'}_i^{S'2}\,\rangle_i\Bigg]
\end{aligned} \tag{153}$$

which was derived from Eq. 125 together with Eq. 13a. We then rewrite the sums in the right-hand side by expanding each index in two set of indices: one running from 1 to $Q/2$ (denoted by the subscript $+$), and one running from $Q/2+1$ to Q (denoted by the subscript $-$). The first sum in Eq. 153 becomes:

$$\sum_{S,S'} c^{S1} c^{S'2} \langle \Psi_i'^{S1} \Psi_i'^{S'2} \rangle_i = c^{\mathsf{A}2} \sum_{S_+,S'_+} \langle S_+1, S'_+2 \rangle + c^{\mathsf{A}} c^{\mathsf{B}} \sum_{S_+,S'_-} \langle S_+1, S'_-2 \rangle + c^{\mathsf{A}} c^{\mathsf{B}} \sum_{S_-,S'_+} \langle S_-1, S'_+2 \rangle + c^{\mathsf{B}2} \sum_{S_-,S'_-} \langle S_-1, S'_-2 \rangle \quad (154)$$

where we have used the short-hand notation $\langle SC, S'C' \rangle \equiv \langle \Psi_i'^{SC} \Psi_i'^{S'C} \rangle_i$. We now observe that

$$\sum_{S_+,S'_+} \langle S_+1, S'_+2 \rangle = \sum_{S_-,S'_-} \langle S_-1, S'_-2 \rangle \equiv \mathcal{A}_1 \quad (155)$$

while

$$\sum_{S_+,S'_-} \langle S_+1, S'_-2 \rangle = \sum_{S_-,S'_+} \langle S_-1, S'_+2 \rangle \equiv \mathcal{A}_2 \quad (156)$$

so that

$$\sum_{S,S'} c^{S1} c^{S'2} \langle \Psi_i'^{S1} \Psi_i'^{S'2} \rangle_i = (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_1 + 2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_2. \quad (157)$$

The second sum in Eq. 153 gives:

$$\begin{aligned} \sum_{C,C'} c^{1C} c^{1C'} \langle \Psi_i'^{1C} \Psi_i'^{1C'} \rangle_i &= c^{\mathsf{A}2} \sum_{C_+,C'_+} \langle 1C_+, 1C'_+ \rangle + c^{\mathsf{A}} c^{\mathsf{B}} \sum_{C_+,C'_-} \langle 1C_+, 1C'_- \rangle + c^{\mathsf{A}} c^{\mathsf{B}} \sum_{C_-,C'_+} \langle 1C_-, 1C'_+ \rangle + c^{\mathsf{B}2} \sum_{C_-,C'_-} \langle 1C_-, 1C'_- \rangle \\ &\equiv (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_3 + 2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_4 \end{aligned} \quad (158)$$

by appropriately defining $\mathcal{A}_3$ and $\mathcal{A}_4$. The third sum gives:

$$\begin{aligned} \sum_{S,C'} c^{S1} c^{1C'} \langle \Psi_i'^{S1} \Psi_i'^{1C'} \rangle_i &= c^{\mathsf{A}2} \sum_{S_+,C'_+} \langle S_+1, 1C_+ \rangle + c^{\mathsf{A}} c^{\mathsf{B}} \sum_{S_+,C'_-} \langle S_+1, 1C_- \rangle + c^{\mathsf{A}} c^{\mathsf{B}} \sum_{S_-,C'_+} \langle S_-1, 1C_+ \rangle + c^{\mathsf{B}2} \sum_{S_-,C'_-} \langle S_-1, 1C_- \rangle \\ &\equiv c^{\mathsf{A}2} \mathcal{A}_5 + c^{\mathsf{B}2} \mathcal{A}_6 + 2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_7 \end{aligned} \quad (159)$$

and, similarly, the fourth one:

$$\sum_{C,S'} c^{1C} c^{S'2} \langle \Psi_i'^{1C} \Psi_i'^{S2} \rangle_i = c^{\mathsf{A}2} \mathcal{A}_5 + c^{\mathsf{B}2} \mathcal{A}_6 + 2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_7. \quad (160)$$

By putting those results together, we conclude that

$$\langle \Delta k_i^{11} \Delta k_i^{12} \rangle_i = \frac{1}{4N} \left[(c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_1 + 2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_2 + (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_3 + 2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_4 + 2c^{\mathsf{A}2} \mathcal{A}_5 + 2c^{\mathsf{B}2} \mathcal{A}_6 + 4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_7 \right]. \quad (161)$$

We can use the same procedure to evaluate dot products for all the remaining pairs of trials. This gives:

$$\langle \Delta k_i^{(\hat{Q}+1)1} \Delta k_i^{(\hat{Q}+1)2} \rangle_i = \frac{1}{4N} \left[(c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_1 + 2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_2 + (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_3 + 2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_4 + 2c^{\mathsf{B}2} \mathcal{A}_5 + 2c^{\mathsf{A}2} \mathcal{A}_6 + 4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_7 \right]$$

$$\langle \Delta k_i^{11} \Delta k_i^{1(\hat{Q}+1)} \rangle_i = \frac{1}{4N} \left[2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_1 + (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_2 + (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_3 + 2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_4 + (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_5 + (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_6 + 4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_7 \right]$$

$$\langle \Delta k_i^{(\hat{Q}+1)1} \Delta k_i^{(\hat{Q}+1)(\hat{Q}+1)} \rangle_i = \frac{1}{4N} \left[2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_1 + (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_2 + (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_3 + 2c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_4 + (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_5 + (c^{\mathsf{A}2} + c^{\mathsf{B}2}) \mathcal{A}_6 + 4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_7 \right]$$

while

$$\langle \Delta k_i^{11} \Delta k_i^{22} \rangle_i = \frac{1}{4N} \left[(2c^{\mathsf{A}2} + 2c^{\mathsf{B}2}) \mathcal{A}_1 + 4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_2 + 2c^{\mathsf{A}2} \mathcal{A}_5 + 2c^{\mathsf{B}2} \mathcal{A}_6 + 4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_7 \right] \quad (163a)$$

$$\langle \Delta k_i^{(\hat{Q}+1)1} \Delta k_i^{(\hat{Q}+2)2} \rangle_i = \frac{1}{4N} \left[(2c^{\mathsf{A}2} + 2c^{\mathsf{B}2}) \mathcal{A}_1 + 4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_2 + 2c^{\mathsf{B}2} \mathcal{A}_5 + 2c^{\mathsf{A}2} \mathcal{A}_6 + 4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_7 \right] \quad (163b)$$

$$\langle \Delta k_i^{11} \Delta k_i^{(\hat{Q}+1)(\hat{Q}+1)} \rangle_i = \frac{1}{4N} \left[4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_1 + (2c^{\mathsf{A}2} + 2c^{\mathsf{B}2}) \mathcal{A}_2 + 2c^{\mathsf{B}2} \mathcal{A}_5 + 2c^{\mathsf{A}2} \mathcal{A}_6 + 4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_7 \right] \quad (163c)$$

$$\langle \Delta k_i^{(\hat{Q}+1)1} \Delta k_i^{1(\hat{Q}+1)} \rangle_i = \frac{1}{4N} \left[4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_1 + (2c^{\mathsf{A}2} + 2c^{\mathsf{B}2}) \mathcal{A}_2 + 2c^{\mathsf{A}2} \mathcal{A}_5 + 2c^{\mathsf{B}2} \mathcal{A}_6 + 4c^{\mathsf{A}} c^{\mathsf{B}} \mathcal{A}_7 \right] \quad (163d)$$

(163e)

and

$$\langle\, \Delta k_i^{11} \Delta k_i^{(\hat{Q}+1)2} \,\rangle_i = \langle\, \Delta k_i^{11} \Delta k_i^{2(\hat{Q}+1)} \,\rangle_i = \langle\, \Delta k_i^{(\hat{Q}+1)1} \Delta k_i^{12} \,\rangle_i = \langle\, (\Delta k_i^{(\hat{Q}+1)1} \Delta k_i^{(\hat{Q}+2)(\hat{Q}+1)} \,\rangle_i$$
$$= \frac{1}{4N} \left[({c^{\mathsf{A}}}^2 + {c^{\mathsf{B}}}^2 + 2c^{\mathsf{A}}c^{\mathsf{B}})\mathcal{A}_1 + ({c^{\mathsf{A}}}^2 + {c^{\mathsf{B}}}^2 + 2c^{\mathsf{A}}c^{\mathsf{B}})\mathcal{A}_2 + ({c^{\mathsf{A}}}^2 + {c^{\mathsf{B}}}^2)\mathcal{A}_5 + ({c^{\mathsf{A}}}^2 + {c^{\mathsf{B}}}^2)\mathcal{A}_6 + 4c^{\mathsf{A}}c^{\mathsf{B}}\mathcal{A}_7 \right]. \tag{164}$$

All the $\mathcal{A}$ coefficients can easily be evaluated analytically. However, we have shown in the previous paragraph that the only coefficients that do not cancel in Eq. 146 are $\mathcal{A}_1$ and $\mathcal{A}_2$; these two are evaluated analytically in Eqs. 150 and 151.

Extracting intuition Can we derive a more intuitive picture of why and how context selectivity increases over learning? We have seen in the previous paragraphs that context selectivity increases because the difference within the first set of curly parenthesis of Eq. 146 is positive (while the difference within the second set of curly parenthesis vanishes). To simplify the math, we assume that $c^{\mathsf{A}} = -c^{\mathsf{B}}$; this condition thus reads:

$$\langle\, \Delta k_i^{11} \Delta k_i^{22} \,\rangle_i - \langle\, \Delta k_i^{11} \Delta k_i^{(\hat{Q}+1)(\hat{Q}+1)} \,\rangle_i > 0. \tag{165}$$

(With respect to Eq. 147, we could get rid of pairs of trials with $s = ((\hat{Q}+1)1)$ because, when $c^{\mathsf{A}} = -c^{\mathsf{B}}$, they give identical results to $s = (11)$.)

Eq. 165 indicates that, over learning, activity from trial $s = (11)$ becomes closer (i.e., more correlated) to activity from trials with the same category and context, such as $s' = (22)$, than trials with the same category but different context, such as $s' = ((\hat{Q}+1)(\hat{Q}+1))$. On the contrary, activity from trial $s = (11)$ becomes equally close to activity from trials with different category and same context, such as $s' = ((\hat{Q}+1)2)$, and trials with different category and different context, such as $s' = (2(\hat{Q}+1))$. This can be seen from Eq. 164, from which

$$\langle\, \Delta k_i^{11} \Delta k_i^{(\hat{Q}+1)2} \,\rangle_i - \langle\, \Delta k_i^{11} \Delta k_i^{2(\hat{Q}+1)} \,\rangle_i = 0 \tag{166}$$

follows.

The geometrical relationships implied by both Eqs. 165 and 166 can be easily verified in Fig. S6C, which shows the synaptic drive from simulated circuits; the middle panel shows a circuit for which we have exactly $c^{\mathsf{A}} = -c^{\mathsf{B}}$. Taken together, Eqs. 165 and 166 indicate that the increase in context selectivity comes from activity clustering by context over learning; such clustering is, however, category-dependent. This leads to the emergence of four statistically-distinguishable clouds, one for each combination of category and context. This is visible in simulated activity from Fig. S6C, and is illustrated in Figs. 6A-C.

5.7 Detailed analysis of category selectivity

We now provide extra detail on the behaviour of category selectivity. We start explaining why, as observed in Figs. 5A and S9A, initial selectivity does not vanish, but is weakly negative. This phenomenon is observed both for the synaptic drive $\boldsymbol{k}^s$ and the activity $\boldsymbol{y}^s$; for the sake of simplicity, we focus on the former.

Consider for a moment the case $Q = 2$ (XOR computation). The geometry of the initial synaptic drive is in that case particularly simple, and is illustrated in Fig. S7D. As can be easily verified by using Eqs. 10b and 115, each synaptic drive is given by the linear superposition of two vectors: a vector among $\tilde{\boldsymbol{\mu}}^1$ and $\tilde{\boldsymbol{\mu}}^2$, and a vector among $\tilde{\boldsymbol{\nu}}^1$ and $\tilde{\boldsymbol{\nu}}^2$. Vectors $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\nu}}$ are obtained by applying the initial connectivity $\boldsymbol{u}_0$ to vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ (Eq. 115); for example, $\tilde{\boldsymbol{\mu}}^1 = \boldsymbol{u}_0 \cdot \boldsymbol{\mu}^1$. In the plane spanned by vectors $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\nu}}$, the geometry of synaptic drives is square-like (Fig. S7D). To verify that, observe that the squared distance between consecutive vertices is identical – for example,

$$\langle\, (k_{0,i}^{11} - k_{0,i}^{12})^2 \,\rangle_i = 2\langle\, (k_{0,i}^{11})^2 \,\rangle_i - 2\langle\, k_{0,i}^{11} k_{0,i}^{12} \,\rangle_i = 2 - 1 = 1 \tag{167}$$

where we used Eq. 48 together with 116. Opposite vertices have instead double squared distance – for example,

$$\langle\, (k_{0,i}^{11} - k_{0,i}^{22})^2 \,\rangle_i = 2\langle\, (k_{0,i}^{11})^2 \,\rangle_i - 2\langle\, k_{0,i}^{11} k_{0,i}^{22} \,\rangle_i = 2 - 0 = 2 \tag{168}$$

as expected for a square. Importantly, consecutive vertices are associated with different categories, while opposite vertices are associated with the same category; this implies that initial category selectivity is negative. In fact, using Eqs. 167 and 168 into Eq. 129 yields:

$$\tilde{S}^{\text{cat}} = \frac{1-2}{1+2} = -\frac{1}{3} < 0. \tag{169}$$

It is easy to see that initial category selectivity is negative also when $Q > 2$. However, its magnitude converges to zero as the number of stimuli and context cues, Q, increases (Fig. S9A). This is due to the fact that, as Q becomes large, both the within-category and the across-category averages in Eq. 129 become dominated by pairs of trials with different stimulus and context cue; activity from those pairs of trials are characterized by identical initial distances (=2, as in Eq. 168), and thus the two averages become similar.

We now shed light on a second phenomenon: the fact that category selectivity increases over learning for the activity $\boldsymbol{y}^s$, but remains identical for the synaptic drive $\boldsymbol{k}^s$. This is observed both in simulations (Figs. S6B and S9A), and in numerical integration of theoretical expressions (Figs. S6A and S9A). To see why this happens, we assume that the number of stimuli and context cues, Q, is fairly large ($1 \ll Q \ll N$). As discussed above, in this limit, initial category selectivity is approximately close to zero. To compute selectivity after learning, we use Eq. 129, and evaluate the within-category and the across-category averages. We compute averages to the dominant terms in Q, which correspond to pairs of trials with different stimulus and context cue. Using the same s and s' trials as in Section 5.6, we obtain

$$\begin{aligned}
&\langle \langle (k_i^s - k_i^{s'})^2 \rangle_i \rangle_{s,\, s' \text{ same cat}} \simeq \langle (k_i^{11} - k_i^{22})^2 \rangle_i + \langle (k_i^{11} - k_i^{(\hat{Q}+1)(\hat{Q}+1)})^2 \rangle_i + \langle (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+2)2})^2 \rangle_i + \langle (k_i^{(\hat{Q}+1)1} - k_i^{1(\hat{Q}+1)})^2 \rangle_i \\
&\langle \langle (k_i^s - k_i^{s'})^2 \rangle_i \rangle_{s,\, s' \text{ diff cat}} \simeq \langle (k_i^{11} - k_i^{(\hat{Q}+1)2})^2 \rangle_i + \langle (k_i^{11} - k_i^{2(\hat{Q}+1)})^2 \rangle_i + \langle (k_i^{(\hat{Q}+1)1} - k_i^{12})^2 \rangle_i + \langle (k_i^{(\hat{Q}+1)1} - k_i^{(\hat{Q}+2)(\hat{Q}+1)})^2 \rangle_i .
\end{aligned} \tag{170}$$

To show that category selectivity does not change over learning, we need to show that the two lines above are identical. Using Eq. 144, this condition can be written as:

$$\begin{aligned}
&\langle \Delta k_i^{11} \Delta k_i^{22} \rangle_i + \langle \Delta k_i^{11} \Delta k_i^{(\hat{Q}+1)(\hat{Q}+1)} \rangle_i + \langle \Delta k_i^{(\hat{Q}+1)1} \Delta k_i^{(\hat{Q}+2)2} \rangle_i + \langle \Delta k_i^{(\hat{Q}+1)1} \Delta k_i^{1(\hat{Q}+1)} \rangle_i = \\
&\langle \Delta k_i^{11} \Delta k_i^{(\hat{Q}+1)2} \rangle_i + \langle \Delta k_i^{11} \Delta k_i^{2(\hat{Q}+1)} \rangle_i + \langle \Delta k_i^{(\hat{Q}+1)1} \Delta k_i^{12} \rangle_i + \langle \Delta k_i^{(\hat{Q}+1)1} \Delta k_i^{(\hat{Q}+2)(\hat{Q}+1)} \rangle_i ,
\end{aligned} \tag{171}$$

which can now be easily verified by using Eqs. 163 and 164.

Equation 171 indicates that, on average across contexts, synaptic drives from trials with the same category are as close as trials with different category. This geometrical relationship can be easily verified in Fig. S6C, which shows the synaptic drive from simulated circuits. We focus on the middle panel, where we have $c^{\mathsf{A}} = -c^{\mathsf{B}}$. The four activity clouds corresponding to different combinations of category and context values are approximately arranged on the vertices of a square; consecutive vertices are associated with different categories, while opposite vertices are associated with the same category. To see why Eq. 171 holds, note that squared distances among synaptic drives associated with different category are approximately identical, while squared distances among synaptic drives associated within the same category are either 0 (approximately, half of the times), or twice the across-category distance (the other half). It is interesting to observe that this square-like configuration, which emerges over learning from an almost unstructured one (Fig. S6C left), strongly resembles the initial configuration of the XOR task (Fig. S7D).

A fundamental feature of this configuration is that synaptic drives are not linearly-separable by category. The activity vectors $\boldsymbol{y}^s$, on the other hand, are linearly-separable. Before learning, linear separability is guaranteed by the nonlinearity Ψ, which makes activity vectors linearly-separable along random directions [39]. After learning, activity vectors become linearly-separable also along task-relevant directions. In the simplified scenario where $\eta_w \ll \eta_u$, the activity vectors become linearly-separable along $\boldsymbol{w}_0$; in the general case, they become linearly-separable along a direction that is correlated with $\boldsymbol{w}_0$. This is shown in Fig. S6D: the configuration of activity is very similar to synaptic drives, but activity vectors associated with different categories clusters, and thus become linearly separable, along an emerging, orthogonal direction. This drives the increase in category selectivity that was observed both in equations and simulations (Figs. S6A-B and S9A). A further insight on the relationship between selectivity and activity geometry is given in the next section.

We conclude with a remark. Although for activity variables category selectivity robustly increases, the fact that selectivity is weakly negative before learning implies that asymptotic values can be small, or even negative. This is compatible with findings in [58], where very small values of category clustering (Eq. 129) were observed. This observation stresses the importance of measuring, in experimental preparations, neural activity across multiple stages of learning.

5.8 Analysis of context and category selectivity across neurons

In this section, we investigate how changes in context and category selectivity are distributed across neurons.

In the simple task, we found that the magnitude of selectivity changes for a given neuron, i, was correlated with the magnitude of the i-th entry of the initial readout vector $\boldsymbol{w}_0$ (Eq. 70, Figs. 3D-F). This vector defines the direction along which clustering by category takes place. In fact, if one draws the vector joining the centers of the activity clouds associated with different categories, $\boldsymbol{y}^{\mathsf{A}}$ and $\boldsymbol{y}^{\mathsf{B}}$ (Eqs. 83), the resulting direction is correlated with $\boldsymbol{w}_0$ (Eq. 101). This direction is indicated with $\boldsymbol{d}$ in the main text; cloud centers $\boldsymbol{y}^{\mathsf{A}}$ and $\boldsymbol{y}^{\mathsf{B}}$ are plotted, in Figs. 3B-C and S1B, as magenta triangles.

In analogy with the simple task, we now hypothesize that the magnitude of changes in context and category selectivity for a given neuron, i, is related to the magnitude of the i-th entry of the context and category directions, $\boldsymbol{d}^{\text{ctx}}$ and $\boldsymbol{d}^{\text{cat}}$. Those coincide with the directions along which clustering to context and category emerges (Figs. 6B-C), and are given by the vectors joining the centers of the activity clouds associated with different contexts (Eq. 135) and categories (Eq. 83). The cloud centers for category and context are plotted, in Figs. 6B-C and S6C-D, as magenta and pink triangles. This assumption is verified in Figs. S8A-B, which shows that selectivity changes and context and category directions are highly correlated. Our reasoning implies that, in order to understand how selectivity changes are distributed across neurons, we need to evaluate the entries of the context and category directions; this is done, analytically, in the rest of this section.

As we are interested in selectivity changes, we focus on activity changes, and approximate

$$\boldsymbol{d}^{\text{ctx}} \simeq \Delta \boldsymbol{y}^1 - \Delta \boldsymbol{y}^2 \tag{172a}$$

$$\boldsymbol{d}^{\text{cat}} \simeq \Delta \boldsymbol{y}^{\mathsf{A}} - \Delta \boldsymbol{y}^{\mathsf{B}} \tag{172b}$$

where, similarly to Eqs. 83 and 135, we have taken

$$\Delta \boldsymbol{y}^1 = \langle\, \Delta \boldsymbol{y}^{s_1} \,\rangle_{s_1} \tag{173a}$$

$$\Delta \boldsymbol{y}^2 = \langle\, \Delta \boldsymbol{y}^{s_2} \,\rangle_{s_2} \tag{173b}$$

and

$$\Delta \boldsymbol{y}^{\mathsf{A}} = \langle\, \Delta \boldsymbol{y}^{s_{\mathsf{A}}} \,\rangle_{s_{\mathsf{A}}} \tag{174a}$$

$$\Delta \boldsymbol{y}^{\mathsf{B}} = \langle\, \Delta \boldsymbol{y}^{s_{\mathsf{B}}} \,\rangle_{s_{\mathsf{B}}}. \tag{174b}$$

We start with context. We have seen in Section 5.6 that context selectivity can also be studied at the level of the synaptic drive $\boldsymbol{k}^s$, which greatly simplifies the analysis. Starting from Eq. 125, we thus compute

$$\begin{aligned}
\Delta k_i^1 &= \frac{w_{0,i}}{2} \left[\sum_S \langle\, c^{SC_{s_1}}\, \Psi_i'^{SC_{s_1}} \,\rangle_{s_1} + \sum_C \langle\, c^{S_{s_1}C}\, \Psi_i'^{S_{s_1}C} \,\rangle_{s_1} \right] \\
&= \frac{w_{0,i}}{2} \left[\sum_S \langle\, c^{SC}\, \Psi_i'^{SC_+} \,\rangle_{C_+} + \sum_C \langle\, c^{SC}\, \Psi_i'^{SC} \,\rangle_S \right] \\
&= \frac{w_{0,i}\, Q}{4} \left[c^{\mathsf{A}} \langle\, \Psi_i'^{S_+C_+} \,\rangle_{S_+,\, C_+} + c^{\mathsf{B}} \langle\, \Psi_i'^{S_-C_-} \,\rangle_{S_-,\, C_+,} + 2 \langle\, c^{SC}\, \Psi_i'^{SC} \,\rangle_{S,\, C} \right].
\end{aligned} \tag{175}$$

As in Section 5.6, indices S_+ and S_- (and, similarly, C_+ and C_-) run, respectively, from 1 to $Q/2$ and from $Q/2+1$ to Q. Similarly,

$$\Delta k_i^2 = \frac{w_{0,i}\, Q}{4} \left[c^{\mathsf{B}} \langle\, \Psi_i'^{S_+C_-} \,\rangle_{S_+,\, C_-} + c^{\mathsf{A}} \langle\, \Psi_i'^{S_-C_-} \,\rangle_{S_-,\, C_-} + 2 \langle\, c^{SC}\, \Psi_i'^{SC} \,\rangle_{S,\, C} \right]. \tag{176}$$

Note that, because of the first two terms in the right-hand sides, Eqs. 175 and 176 are not identical.

To further simplify the analysis, we assume that $c^{\mathsf{B}} \simeq -c^{\mathsf{A}}$. As discussed in Section 5.2, in the current task, this represents a good approximation for a large space of parameters; we verified with simulations that our main results also hold, qualitatively, in circuits where this approximation fails (notably, in the circuit illustrated in the third column of Fig. 5, see Figs. S8C-D). Combining Eq. 172a with Eqs. 175 and 176, we then obtain

$$\begin{aligned}
d_i^{\text{ctx}} = \Delta k_i^1 - \Delta k_i^2 &= \frac{w_{0,i}\, c^{\mathsf{A}}\, Q}{4} \langle\, \langle\, \Psi_i'^{S_+C} \,\rangle_{S_+} - \langle\, \Psi_i'^{S_-C} \,\rangle_{S_-} \,\rangle_C \\
&\equiv \frac{w_{0,i}\, c^{\mathsf{A}}\, Q}{4} \langle\, D_i^C \,\rangle_C
\end{aligned} \tag{177}$$

where we have defined

$$D_i^C \equiv \langle\, \Psi_i'^{S_+C} \,\rangle_{S_+} - \langle\, \Psi_i'^{S_-C} \,\rangle_{S_-}. \tag{178}$$

Equation 177 indicates that neurons exhibiting a strong increase in context selectivity are characterized by: (i) strong readout connectivity, before learning, as quantified by $w_{0,i}$, and (ii) a large value of D_i^C, averaged over context cues. D_i^C is a function of the response gain function, Ψ', evaluated before learning; specifically, D_i^C measures the difference in the initial gain in response to the two classes of stimuli (the first half, $S_+ = 1, \ldots, Q/2$, and the second half, $S_- = Q/2, \ldots, Q$). These predictions, which were derived for the synaptic drive $\boldsymbol{k}^s$, also hold, qualitatively, for the activity $\boldsymbol{y}^s$ (Fig. 7).

We next compute the category direction $\boldsymbol{d}^{cat}$; we focus again on the synaptic drive $\boldsymbol{k}^s$ rather than activity $\boldsymbol{y}^s$. We observe that, before learning, the centers of synaptic drive vectors associated with category A and B are perfectly identical. In fact,

$$\begin{aligned}
k_{0,i}^{\mathsf{A}} &= \frac{\langle\, \tilde{\mu}_i^{S_{s_{\mathsf{A}}}} \,\rangle_{s_{\mathsf{A}}} + \langle\, \tilde{\nu}_i^{C_{s_{\mathsf{A}}}} \,\rangle_{s_{\mathsf{A}}}}{\sqrt{2}} \\
&= \frac{\langle\, \tilde{\mu}_i^{S} \,\rangle_S + \langle\, \tilde{\nu}_i^{C} \,\rangle_C}{\sqrt{2}},
\end{aligned} \tag{179}$$

and an identical expression is obtained for $k_{0,i}^{\mathsf{B}}$. The fact that the centers are identical is due to the fact that sensory inputs for the two categories are collinear, and perfectly intermingled (Fig. S7D). We now consider the synaptic drive changes over

learning. Starting from Eq. 125, we have

$$\begin{aligned}\Delta k_i^{\mathsf{A}} &= \frac{w_{0,i}}{2}\left[\sum_S \langle\, c^{SC_{s_{\mathsf{A}}}}\, {\Psi'}_i^{SC_{s_{\mathsf{A}}}}\,\rangle_{s_{\mathsf{A}}} + \sum_C \langle\, c^{S_{s_{\mathsf{A}}}C}\, {\Psi'}_i^{S_{s_{\mathsf{A}}}C}\,\rangle_{s_{\mathsf{A}}}\right] \\ &= \frac{w_{0,i}}{2}\left[\sum_S \langle\, c^{SC}\, {\Psi'}_i^{SC}\,\rangle_C + \sum_C \langle\, c^{SC}\, {\Psi'}_i^{SC}\,\rangle_S\right] \\ &= w_{0,i}\, Q\, \langle\, c^{SC}\, {\Psi'}_i^{SC}\,\rangle_{S,\,C}.\end{aligned} \tag{180}$$

It is easy to show that Δk_i^{B} yields the same result, implying that the centers for synaptic drive vectors associated with categories A and B remain identical over learning (Fig. S6C, magenta triangles). This happens because the synaptic drive vectors associated with category A and B remain intermingled, and non linearly-separable, over learning. We conclude that the category axis $\boldsymbol{d}^{\text{cat}}$ (Eq. 172b) vanishes, which is in agreement with the observation that category selectivity does not change for synaptic drives (Section 5.7).

To compute $\boldsymbol{d}^{\text{cat}}$, we thus turn to activity $\boldsymbol{y}$. We start from Eq. 123, and write

$$\Delta y_i^{\mathsf{A}} = \frac{w_{0,i}}{2}\left[\sum_S \langle\, c^{SC_{s_{\mathsf{A}}}}\, {\Psi'}_i^{S_{s_{\mathsf{A}}}C_{s_{\mathsf{A}}}}\, {\Psi'}_i^{SC_{s_{\mathsf{A}}}}\,\rangle_{s_{\mathsf{A}}} + \sum_C \langle\, c^{S_{s_{\mathsf{A}}}C}\, {\Psi'}_i^{S_{s_{\mathsf{A}}}C_{s_{\mathsf{A}}}}\, {\Psi'}_i^{S_{s_{\mathsf{A}}}C}\,\rangle_{s_{\mathsf{A}}}\right] \tag{181a}$$

$$\Delta y_i^{\mathsf{B}} = \frac{w_{0,i}}{2}\left[\sum_S \langle\, c^{SC_{s_{\mathsf{B}}}}\, {\Psi'}_i^{S_{s_{\mathsf{B}}}C_{s_{\mathsf{B}}}}\, {\Psi'}_i^{SC_{s_{\mathsf{B}}}}\,\rangle_{s_{\mathsf{B}}} + \sum_C \langle\, c^{S_{s_{\mathsf{B}}}C}\, {\Psi'}_i^{S_{s_{\mathsf{B}}}C_{s_{\mathsf{B}}}}\, {\Psi'}_i^{S_{s_{\mathsf{B}}}C}\,\rangle_{s_{\mathsf{B}}}\right]. \tag{181b}$$

We then expand indices over stimuli and context cues, which yields

$$\begin{aligned}\Delta y_i^{\mathsf{A}} = \frac{w_{0,i}\, Q}{2}\Big[&c^{\mathsf{A}}\langle\, \Psi_i^{S_+C_+}\Psi_i^{\bar{S}_+C_+}\,\rangle + c^{\mathsf{B}}\langle\, \Psi_i^{S_-C_-}\Psi_i^{\bar{S}_+C_-}\,\rangle + c^{\mathsf{B}}\langle\, \Psi_i^{S_+C_+}\Psi_i^{\bar{S}_-C_+}\,\rangle + c^{\mathsf{A}}\langle\, \Psi_i^{S_-C_-}\Psi_i^{\bar{S}_-C_-}\,\rangle + \\ &c^{\mathsf{A}}\langle\, \Psi_i^{S_+C_+}\Psi_i^{S_+\bar{C}_+}\,\rangle + c^{\mathsf{B}}\langle\, \Psi_i^{S_-C_-}\Psi_i^{S_-\bar{C}_+}\,\rangle + c^{\mathsf{B}}\langle\, \Psi_i^{S_+C_+}\Psi_i^{S_+\bar{C}_-}\,\rangle + c^{\mathsf{A}}\langle\, \Psi_i^{S_-C_-}\Psi_i^{S_-\bar{C}_-}\,\rangle\Big]\end{aligned} \tag{182}$$

and

$$\begin{aligned}\Delta y_i^{\mathsf{B}} = \frac{w_{0,i}\, Q}{2}\Big[&c^{\mathsf{A}}\langle\, \Psi_i^{S_-C_+}\Psi_i^{\bar{S}_+C_+}\,\rangle + c^{\mathsf{B}}\langle\, \Psi_i^{S_+C_-}\Psi_i^{\bar{S}_+C_-}\,\rangle + c^{\mathsf{B}}\langle\, \Psi_i^{S_-C_+}\Psi_i^{\bar{S}_-C_+}\,\rangle + c^{\mathsf{A}}\langle\, \Psi_i^{S_+C_-}\Psi_i^{\bar{S}_-C_-}\,\rangle + \\ &c^{\mathsf{A}}\langle\, \Psi_i^{S_+C_-}\Psi_i^{S_+\bar{C}_+}\,\rangle + c^{\mathsf{B}}\langle\, \Psi_i^{S_-C_+}\Psi_i^{S_-\bar{C}_+}\,\rangle + c^{\mathsf{B}}\langle\, \Psi_i^{S_+C_-}\Psi_i^{S_+\bar{C}_-}\,\rangle + c^{\mathsf{A}}\langle\, \Psi_i^{S_-C_+}\Psi_i^{S_-\bar{C}_-}\,\rangle\Big].\end{aligned} \tag{183}$$

To reduce the clutter, we have removed subscripts after brackets $\langle\, .\,\rangle$; those indicate an average taken over all the S and C indices contained within.

As will become clear shortly, the two now centers differ (Fig. S6D, magenta triangles). To simplify those expressions, we again assume that $c^{\mathsf{B}} \simeq -c^{\mathsf{A}}$; this allows us to write

$$\begin{aligned}d_i^{\text{cat}} = \Delta y_i^{\mathsf{A}} - \Delta y_i^{\mathsf{B}} = \frac{w_{0,i}\, c^{\mathsf{A}}}{2}\left(\frac{Q}{2}\right)^2\Bigg[&\sum_{C_+}\sum_{S_+}\Psi_i^{S_+C_+}\left(\sum_{\bar{S}_+}\Psi_i^{\bar{S}_+C_+} - \sum_{\bar{S}_-}\Psi_i^{\bar{S}_-C_+}\right) - \sum_{C_+}\sum_{S_-}\Psi_i^{S_-C_+}\left(\sum_{\bar{S}_+}\Psi_i^{\bar{S}_+C_+} - \sum_{\bar{S}_-}\Psi_i^{\bar{S}_-C_+}\right) \\ &- \sum_{C_-}\sum_{S_-}\Psi_i^{S_-C_-}\left(\sum_{\bar{S}_+}\Psi_i^{\bar{S}_+C_-} - \sum_{\bar{S}_-}\Psi_i^{\bar{S}_-C_-}\right) + \sum_{C_-}\sum_{S_+}\Psi_i^{S_+C_-}\left(\sum_{\bar{S}_+}\Psi_i^{\bar{S}_+C_-} - \sum_{\bar{S}_-}\Psi_i^{\bar{S}_-C_-}\right) \\ &+ \sum_{S_+}\sum_{C_+}\Psi_i^{S_+C_+}\left(\sum_{\bar{C}_+}\Psi_i^{S_+\bar{C}_+} - \sum_{\bar{C}_-}\Psi_i^{S_+\bar{C}_-}\right) - \sum_{S_+}\sum_{C_-}\Psi_i^{S_+C_-}\left(\sum_{\bar{C}_+}\Psi_i^{S_+\bar{C}_+} - \sum_{\bar{C}_-}\Psi_i^{S_+\bar{C}_-}\right) \\ &- \sum_{S_-}\sum_{C_-}\Psi_i^{S_-C_-}\left(\sum_{\bar{C}_+}\Psi_i^{S_-\bar{C}_+} - \sum_{\bar{C}_-}\Psi_i^{S_-\bar{C}_-}\right) + \sum_{S_-}\sum_{C_+}\Psi_i^{S_-C_+}\left(\sum_{\bar{C}_+}\Psi_i^{S_-\bar{C}_+} - \sum_{\bar{C}_-}\Psi_i^{S_-\bar{C}_-}\right)\Bigg].\end{aligned} \tag{184}$$

With a little algebra, we can see that

$$\begin{aligned}d_i^{\text{cat}} = \frac{w_{0,i}\, c^{\mathsf{A}}}{2}\left(\frac{Q}{2}\right)^2\Bigg[&\sum_{C_+}\left(\sum_{\bar{S}_+}\Psi_i^{\bar{S}_+C_+} - \sum_{\bar{S}_-}\Psi_i^{\bar{S}_-C_+}\right)^2 + \sum_{C_-}\left(\sum_{\bar{S}_+}\Psi_i^{\bar{S}_+C_-} - \sum_{\bar{S}_-}\Psi_i^{\bar{S}_-C_-}\right)^2 \\ &+ \sum_{S_+}\left(\sum_{\bar{C}_+}\Psi_i^{S_+\bar{C}_+} - \sum_{\bar{C}_-}\Psi_i^{S_+\bar{C}_-}\right)^2 + \sum_{S_-}\left(\sum_{\bar{C}_+}\Psi_i^{S_-\bar{C}_+} - \sum_{\bar{C}_-}\Psi_i^{S_-\bar{C}_-}\right)^2\Bigg]\end{aligned} \tag{185}$$

or, equivalently

$$
\begin{aligned}
d_i^{\text{cat}} &= \frac{w_{0,i}\, c^{\mathsf{A}}\, Q}{2} \left[\langle \left(\langle \, {\Psi_i}^{S_+ C} \, \rangle_{S_+} - \langle \, {\Psi_i}^{S_- C} \, \rangle_{S_-} \right)^2 \rangle_C + \langle \left(\langle \, {\Psi_i}^{SC_+} \, \rangle_{C_+} - \langle \, {\Psi_i}^{SC_-} \, \rangle_{C_-} \right)^2 \rangle_S \right] \\
&= \frac{w_{0,i}\, c^{\mathsf{A}}\, Q}{2} \left[\langle \, (D_i^C)^2 \, \rangle_C + \langle \, (D_i^S)^2 \, \rangle_S \right]
\end{aligned}
\tag{186}
$$

where we have defined

$$
D_i^S \equiv \langle \, {\Psi'_i}^{SC_+} \, \rangle_{C_+} - \langle \, {\Psi'_i}^{SC_-} \, \rangle_{C_-} \, . \tag{187}
$$

Equation 186 indicates that neurons characterized by a strong increase in category selectivity are characterized by: (i) strong readout connectivity, before learning, as quantified by $w_{0,i}$, and (ii) large values of D_i^C and/or D_i^S, averaged, respectively, over context cues and stimuli.

Note that neurons that are characterized by a strong increase in context selectivity (Eq. 177), which are have large $w_{0,i}$ and D_i^C values, are also characterized by a strong increase in category selectivity (Eq. 186). On the other hand, neurons with large $w_{0,i}$ and D_i^S values are characterized by a strong increase in category selectivity (Eq. 186), but not context (Eq. 177). Overall, strongly selective neurons can thus be classified in two groups: one displaying mixed selectivity to category and context, and one displaying pure selectivity to category. By defining the quantity:

$$
G_i = \langle \, |D_i^C| \, \rangle_C - \langle \, |D_i^S| \, \rangle_S \tag{188}
$$

we see that the former group is characterized by larger values of G_i with respect to the latter. This is verified and illustrated in Figs. 7B-C.

6 Software

6.1 Circuit simulations

Simulations were implemented with the Python programming language. Gradient-descent learning was implemented with the PyTorch package. We used the *SGD* optimization function, with loss *MSELoss*. On every learning epoch, the batch included all sensory input vectors. Training stopped when the loss function dropped below 10^{-5}. Learning rates were taken to be $\eta = 0.1$ for input connectivity $\boldsymbol{u}$, and $\eta \cdot \eta_w/\eta_u$ (with values of η_u and η_w as indicated in Section 6.2) for readout connectivity $\boldsymbol{w}$.

Code will be made available in a public repository upon publication.

6.2 Table of parameters

We summarize below the parameters chosen for the simulations reported in Figures and Supplementary Figures. For figures not included in the tables below, parameters have been detailed in figures captions.

We have taken everywhere $z^{\mathsf{A}} = 0.75$, $z^{\mathsf{B}} = 0.25$ (note that activity variables range between 0 and 1).

Figure	N	Q	η_w/η_u	Θ_1 of Ψ	Θ_2 of Ψ	Θ_1 of Φ	Θ_2 of Φ
Fig. 2-3, first and second column	200	20	0.0	1.0	2.0	1.0	0.0
Fig. 2-3, third column	200	20	0.0	1.0	2.0	1.0	2.0
Fig. 4B	200	20	0.4	2.0	2.0	varies	varies
Fig. 4C	200	20	0.4	varies	varies	1.0	2.0
Fig. 4D	200	20	varies	2.0	2.0	1.0	varies
Fig. 4E	200	varies	0.4	2.0	2.0	1.0	varies
Fig. 5, first and second column	600	8 ($P = 64$)	0.0	1.0	0.0	1.0	0.0
Fig. 5, third column	600	8 ($P = 64$)	0.0	1.0	0.0	1.0	4.0
Fig. 6D	600	8 ($P = 64$)	0.2	2.5	2.0	varies	varies
Fig. 6E	600	8 ($P = 64$)	varies	2.5	2.0	1.0	varies
Fig. 6F	600	varies	0.2	2.5	2.0	1.0	varies

Table 1: Table of parameters for Figures in the main text.

Supplementary Figure	N	Q	η_w/η_u	Θ_1 of Ψ	Θ_2 of Ψ	Θ_1 of Φ	Θ_2 of Φ
Fig. S1A	200	varies	varies	varies	varies	varies	varies
Fig. S2E, first column	200	20	0.0	1.0	2.0	1.0	0.0
Fig. S3, first column	varies	20	0.0	1.0	0.0	1.0	0.0
Fig. S3, second column	200	20	varies	1.0	0.0	1.0	0.0
Fig. S3, third column	200	20	0.0	1.0	varies	1.0	0.0
Fig. S4, first column	varies	20	0.0	1.0	0.0	1.0	2.0
Fig. S4, second column	200	20	varies	1.0	0.0	1.0	2.0
Fig. S4, third column	200	20	0.0	1.0	varies	1.0	2.0
Fig. S5A-B	200	12	0.0	1.0	2.0	1.0	0.0
Fig. S5C	200	12	0.0	1.0	2.0	1.0	2.0
Fig. S5D-E, first column	200	12	0.1	2.0	varies	1.0	varies
Fig. S5D-E, second column	200	12	varies	2.0	2.0	1.0	varies
Fig. S5D-E, third column	200	varies	0.1	2.0	2.0	1.0	varies
Fig. S6A-B	600	varies	varies	varies	varies	varies	varies
Fig. S7A	600	8 ($Q = 64$)	0.0	1.0	3.0	1.0	0.0
Fig. S7B	600	8 ($Q = 64$)	0.0	1.0	3.0	1.0	4.0
Fig. S7C	600	varies	varies	varies	varies	varies	varies
Fig. S9A-B-C	600	varies	0.0	1.0	0.0	1.0	0.0
Fig. S9D-E-F	600	varies	0.0	1.0	0.0	1.0	4.0

Table 2: Table of parameters for Supplementary Figures.

6.3 Evaluation of averages

Evaluating the approximate theoretical expressions for activity measures given in Sections 4 and 5 requires computing a number of Gaussian integrals over nonlinear functions. We compute those averages numerically; technical details are given below.

The simplest average, which only involves one nonlinear function, was denoted by $\langle F \rangle$ (Eq. 50). We rewrite Eq. 50 in an integral form, yielding

$$\langle F \rangle \equiv \int \mathcal{D}a \, F(a) \tag{189}$$

where we have used the short-hand notation

$$\int \mathcal{D}a \equiv \int_{-\infty}^{\infty} \mathrm{d}a \, \frac{\exp(-a^2/2\pi)}{\sqrt{2\pi}} \, . \tag{190}$$

This integral was computed numerically via Hermite-Gaussian quadrature.

Averages involving two nonlinear functions were denoted by $\langle FF \rangle$ (Eq. 118). We rewrite Eq. 118 in an integral form, yielding

$$\langle FF \rangle \equiv \int \mathcal{D}a \int \mathcal{D}b_1 \int \mathcal{D}b_2 \, F\left(\frac{1}{\sqrt{2}}(a+b_1)\right) F\left(\frac{1}{\sqrt{2}}(a+b_2)\right) = \int \mathcal{D}a \left[\int \mathcal{D}b \, F\left(\frac{1}{\sqrt{2}}(a+b)\right)\right]^2 . \tag{191}$$

This integral was computed again via Hermite-Gaussian quadrature.

Averages involving four nonlinear functions, such as $\langle \, \Psi'(k_{0,i}^q)\Psi'(k_{0,i}^s)\Psi'(k_{0,i}^{q'})\Psi'(k_{0,i}^{s'}) \, \rangle_i$ from Eq. 138 (Section 5.5) were computed instead via the function *nquad* from the Python *scipy.integrate* package. We start by rewriting the argument of the average as:

$$\Psi'(k_{0,i}^q)\Psi'(k_{0,i}^s)\Psi'(k_{0,i}^{q'})\Psi'(k_{0,i}^{s'}) = \Psi'\left(\frac{1}{\sqrt{2}}(a_{S_q}+b_{C_q})\right)\Psi'\left(\frac{1}{\sqrt{2}}(a_{S_s}+b_{C_s})\right)\Psi'\left(\frac{1}{\sqrt{2}}(a_{S_{q'}}+b_{C_{q'}})\right)\Psi'\left(\frac{1}{\sqrt{2}}(a_{S_{s'}}+b_{C_{s'}})\right). \tag{192}$$

For each value of the stimulus index S and the context cue index C, a_S and b_C are two independent, zero-mean and unit-variance Gaussian variables. If the values of S and C are different across the four trials q, s, q' and s', then all a and b

variables involved in Eq. 192 are different, and the average reads

$$\langle\, \Psi'(k^q_{0,i})\Psi'(k^s_{0,i})\Psi'(k^{q'}_{0,i})\Psi'(k^{s'}_{0,i})\,\rangle_i = \int \mathcal{D}a_{S_q} \int \mathcal{D}a_{S_s} \int \mathcal{D}a_{S_{q'}} \int \mathcal{D}a_{S_{s'}} \int \mathcal{D}b_{C_q} \int \mathcal{D}b_{C_s} \int \mathcal{D}b_{C_{q'}} \int \mathcal{D}b_{C_{s'}} \times \\ \Psi'\left(\frac{1}{\sqrt{2}}(a_{S_q}+b_{C_q})\right)\Psi'\left(\frac{1}{\sqrt{2}}(a_{S_s}+b_{C_s})\right)\Psi'\left(\frac{1}{\sqrt{2}}(a_{S_{q'}}+b_{C_{q'}})\right)\Psi'\left(\frac{1}{\sqrt{2}}(a_{S_{s'}}+b_{C_{s'}})\right) \tag{193}$$

which simplifies into

$$\langle\, \Psi'(k^q_{0,i})\Psi'(k^s_{0,i})\Psi'(k^{q'}_{0,i})\Psi'(k^{s'}_{0,i})\,\rangle_i = \left[\int \mathcal{D}a_{S_q} \int \mathcal{D}b_{C_q} \Psi'\left(\frac{1}{\sqrt{2}}(a_{S_q}+b_{C_q})\right)\right]^4 = \langle\Psi'\rangle^4. \tag{194}$$

If the stimulus S or the context cue C are, instead, identical across two o more trials (q, s, q' and s'), then some of the a and b variables in Eq. 192 are shared across nonlinear functions. This generates correlations, which determine the final value of the average. For example, assume $S_q = S_s$, while all other S and C values are different among each other. Then the average reads

$$\langle\, \Psi'(k^q_{0,i})\Psi'(k^s_{0,i})\Psi'(k^{q'}_{0,i})\Psi'(k^{s'}_{0,i})\,\rangle_i = \int \mathcal{D}a_{S_q} \int \mathcal{D}a_{S_{q'}} \int \mathcal{D}a_{S_{s'}} \int \mathcal{D}b_{C_q} \int \mathcal{D}b_{C_s} \int \mathcal{D}b_{C_{q'}} \int \mathcal{D}b_{C_{s'}} \times \\ \Psi'\left(\frac{1}{\sqrt{2}}(a_{S_q}+b_{C_q})\right)\Psi'\left(\frac{1}{\sqrt{2}}(a_{S_q}+b_{C_s})\right)\Psi'\left(\frac{1}{\sqrt{2}}(a_{S_{q'}}+b_{C_{q'}})\right)\Psi'\left(\frac{1}{\sqrt{2}}(a_{S_{s'}}+b_{C_{s'}})\right) \tag{195}$$

which simplifies into

$$\langle\, \Psi'(k^q_{0,i})\Psi'(k^s_{0,i})\Psi'(k^{q'}_{0,i})\Psi'(k^{s'}_{0,i})\,\rangle_i = \langle\Psi'\Psi'\rangle\,\langle\Psi'\rangle^2. \tag{196}$$

We considered all the possible configurations of S and C indices that can occur in the context-dependent task, and all the resulting correlation patterns. Then, we used analytics to simplify integrals when possible (as in the cases described above). We finally used numerics to evaluate the remaining integral expressions.

Supplementary Figures

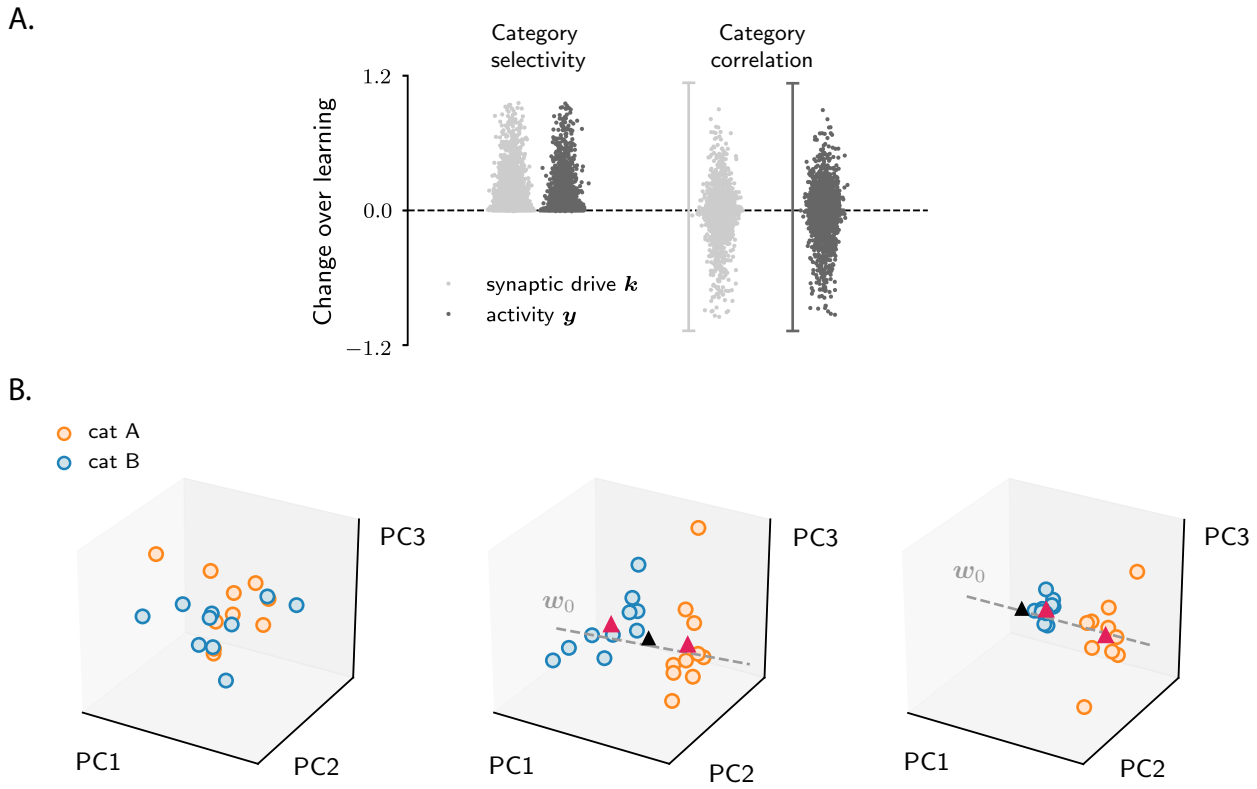

Fig. S1: **Characterization of activity evolution during the simple categorization task; additional results, part I. A.** Extensive results from simulations. Each point represents a model. We simulated 5000 different models, with parameters drawn randomly and uniformly from the following ranges: the number of stimuli, Q, between 10 and 30; the learning rates ratio, η_w/η_u, between 0 and 1; the gain of Ψ and Φ between 0.05 and 1; the offset of Ψ and Φ between -2 and 2. Light points refer to the synaptic drive, $\boldsymbol{k}$, and dark ones to the activity, $\boldsymbol{y}$. To quantify category selectivity, we used Eq. 61, and then averaged over neurons. To quantify category correlation, we used Eq. 76. The vertical bars in the category correlation plots indicate the range of results obtained when quantifying correlation with the alternative definition in Eq. 84. Note that selectivity changes are positive, while correlation changes take both positive and negative values. For a few parameter sets, learning convergence was pathological (i.e., the loss displayed strong oscillations over epochs, or did not converge); those cases have been excluded from the analysis. **B.** Activity from circuits in Figs. 2 A-D (left), E-H (center) and I-L (right) projected on the first three principal components. Before computing the principal components, we subtracted from activity the mean across trials and neurons. Over learning, activity develops two clouds, one for each category, as schematically illustrated in Figs. 3A-C. The clustering direction is approximately given by $\boldsymbol{w}_0$ (dashed grey line). The center of the two clouds is indicated by magenta triangles; the center of initial activity is indicated by the black triangle.

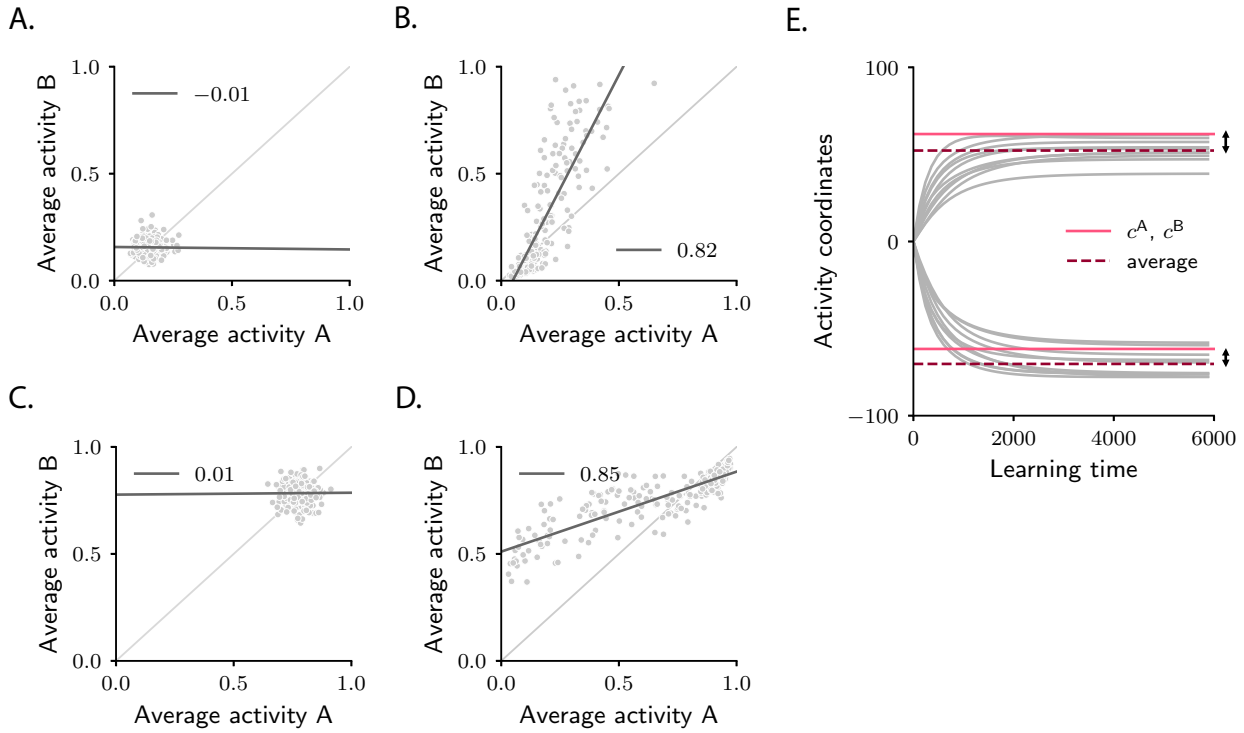

Fig. S2: **Characterization of activity evolution during the simple categorization task; additional results, part II. A-D.** Population response to categories A and B, averaged over stimuli. Each dot represents a neuron. Panels A-B and C-D display two sample circuits, different from those displayed in Fig. 2. Panels A and C and B and D display, respectively, pre- and post-learning activity. All details and parameters are as in Fig. 2I-L, except: in A and B, we set the offset of Φ to -2; in C and D, we set the offset of Ψ to -1.5. In panel B, in contrast to Fig. 2L, the variance of activity in response to B is larger than the variance in response to A. In panel D, as in Fig. 2L, the variance of activity in response to A is larger than the variance in response to B. In contrast to Fig. 2L, however, the mean activity in response to B is larger than the mean activity in response to A (Methods 4.6). Note that, in all cases, category correlation remains positive. **E.** Variability across circuits realizations (Methods 4.7): illustration from a sample circuit. Grey lines illustrate activity coordinates, c^s, as a function of learning time, for all sensory inputs. Those were estimated by taking the dot product $\Delta \boldsymbol{y}^s \cdot \boldsymbol{w}_0$, and then dividing by $\langle \Psi'^2 \rangle$ (Eqs. 60 and 58). Note that c^s values corresponding to sensory inputs in the same category do not saturate to identical values; their average is indicated by dashed lines. Pink continuous lines indicate the values of c^A and c^B computed through the theory by neglecting variability (Eq. 54). As predicted by Eq. 104, the dashed and continuous pink lines do not coincide (Methods 4.7).

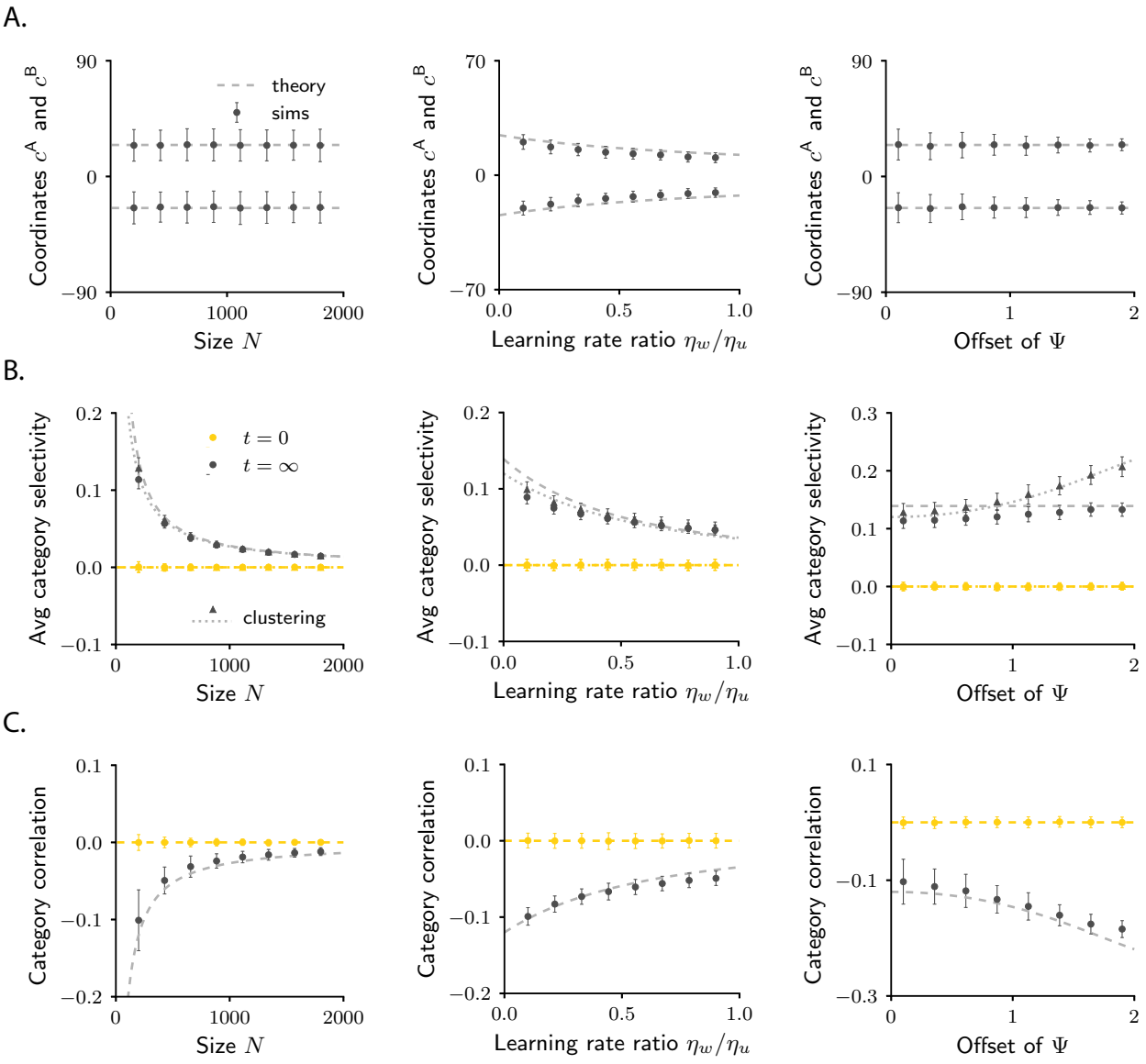

Fig. S3: **Comparison between finite-size networks and approximate mathematical description for the simple categorization task, part I.** Dashed lines show the theoretical predictions; dots show the average over 400 simulations where both the initial connectivity and the sensory inputs were drawn at random. Error bars show the standard deviation across simulations; they thus quantify variability across initializations (Methods 4.7). Yellow: pre-learning; gray: post-learning. **A.** Activity coordinates c^{A} and c^{B}. Theory was computed from Eq. 54. Simulations results were computed by taking the dot product $\Delta \boldsymbol{y}^s \cdot \boldsymbol{w}_0$ for two activity vectors associated with different categories, and then dividing by $\langle \Psi'^2 \rangle$ (Eqs. 60 and 58). **B.** Average category selectivity. Theory was computed from Eq. 71. Simulations results were computed by using Eq. 61, and then averaging across neurons. We also plot results for category clustering. Theory (dotted lines) was computed from Eq. 74. Simulations results (triangles) were computed by using Eq. 72. Note that average selectivity and clustering are not identical for all parameters (although they become close in the limit $1 \ll Q \ll N$, as discussed in Methods 4.3). **C.** Category correlation. Theory was computed from Eq. 79. Simulations results were obtained by applying Eq. 76 to simulated data.

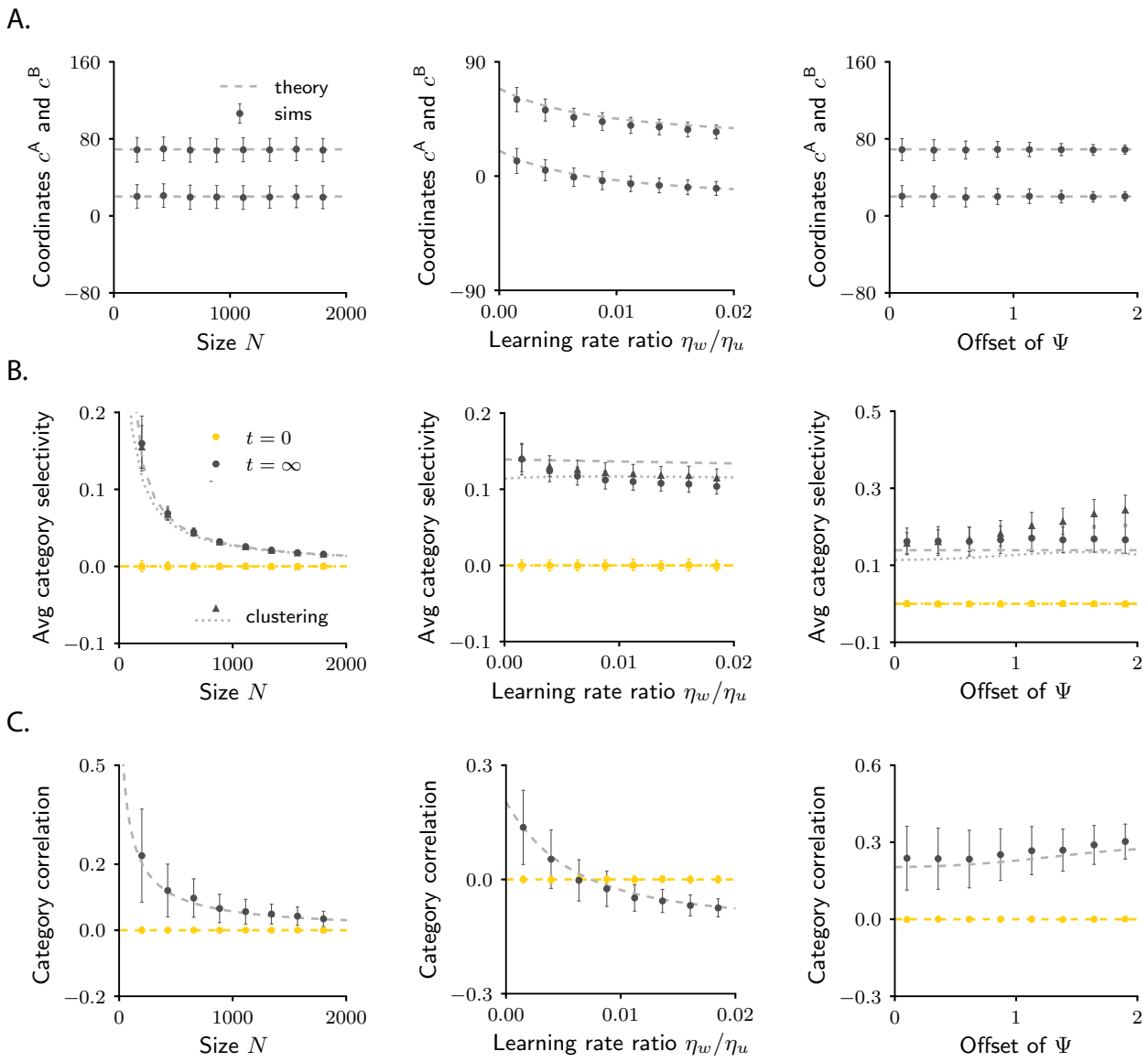

Fig. S4: **Comparison between finite-size networks and approximate mathematical description for the simple categorization task, part II.** Details as in Fig. S3. We used different parameters (see Table 2), which lead to positive category correlation.

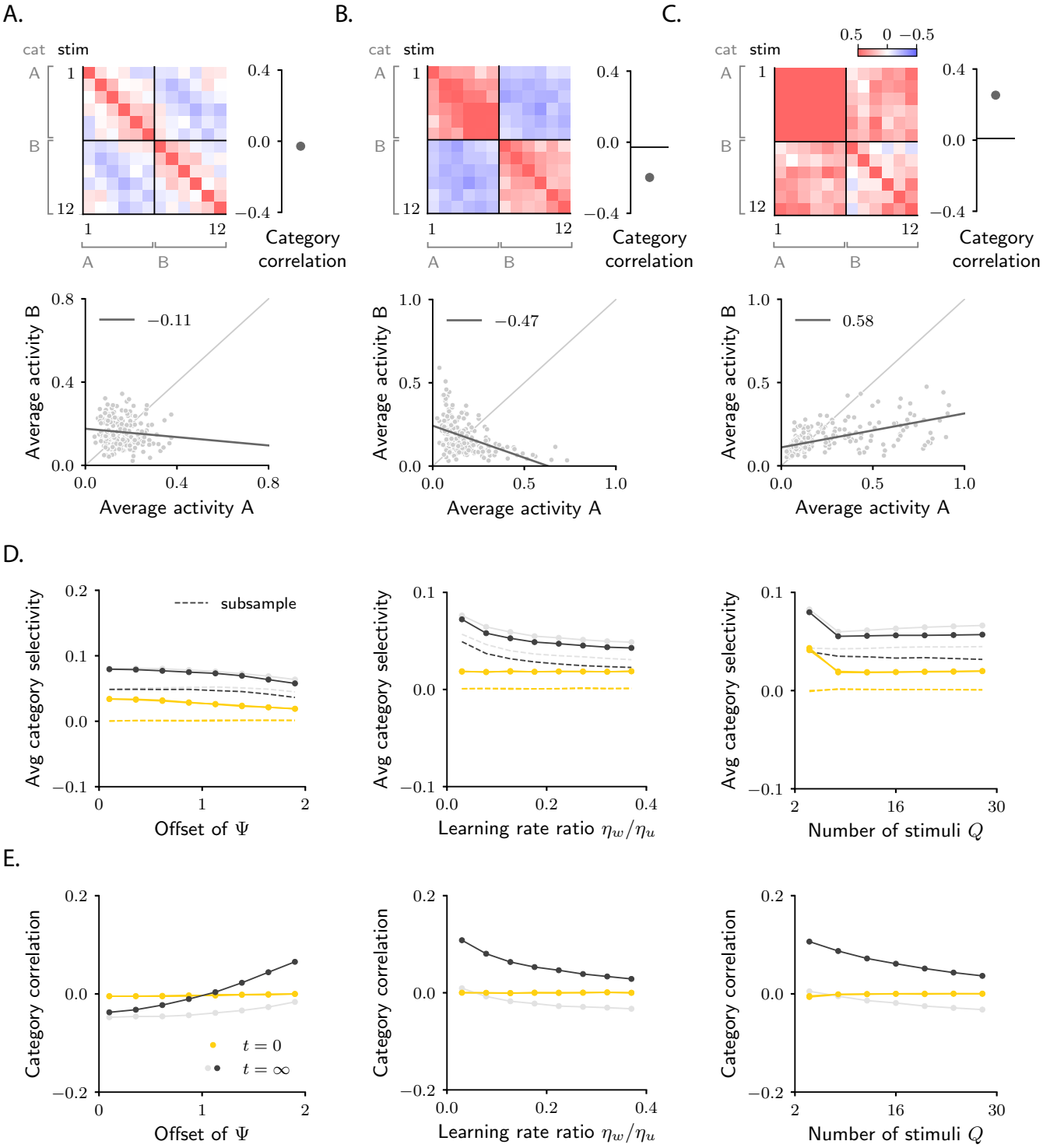

Fig. S5: **Simple categorization task with structured inputs and heterogeneity (Methods 4.8). A-C.** Analysis of category correlation for a naive circuit (A), and two trained ones (B-C). Details are, respectively, as in Figs. 2C-D, G-H and K-L. **D.** Extensive analysis of average category selectivity across circuits and task parameters. Details are as in Fig. S3 (except that no theory, only simulations, are shown). As in Figs. 4D-E, dark and light grey indicate different values for the offset of the activation function Φ. Dashed lines show results obtained by subsampling pairs of sensory inputs in the category selectivity definition (Eq. 61) so that initial selectivity vanishes (see Methods 4.8 for details). **E.** Extensive analysis of category correlation across circuit and task parameters. Details as in panel D.

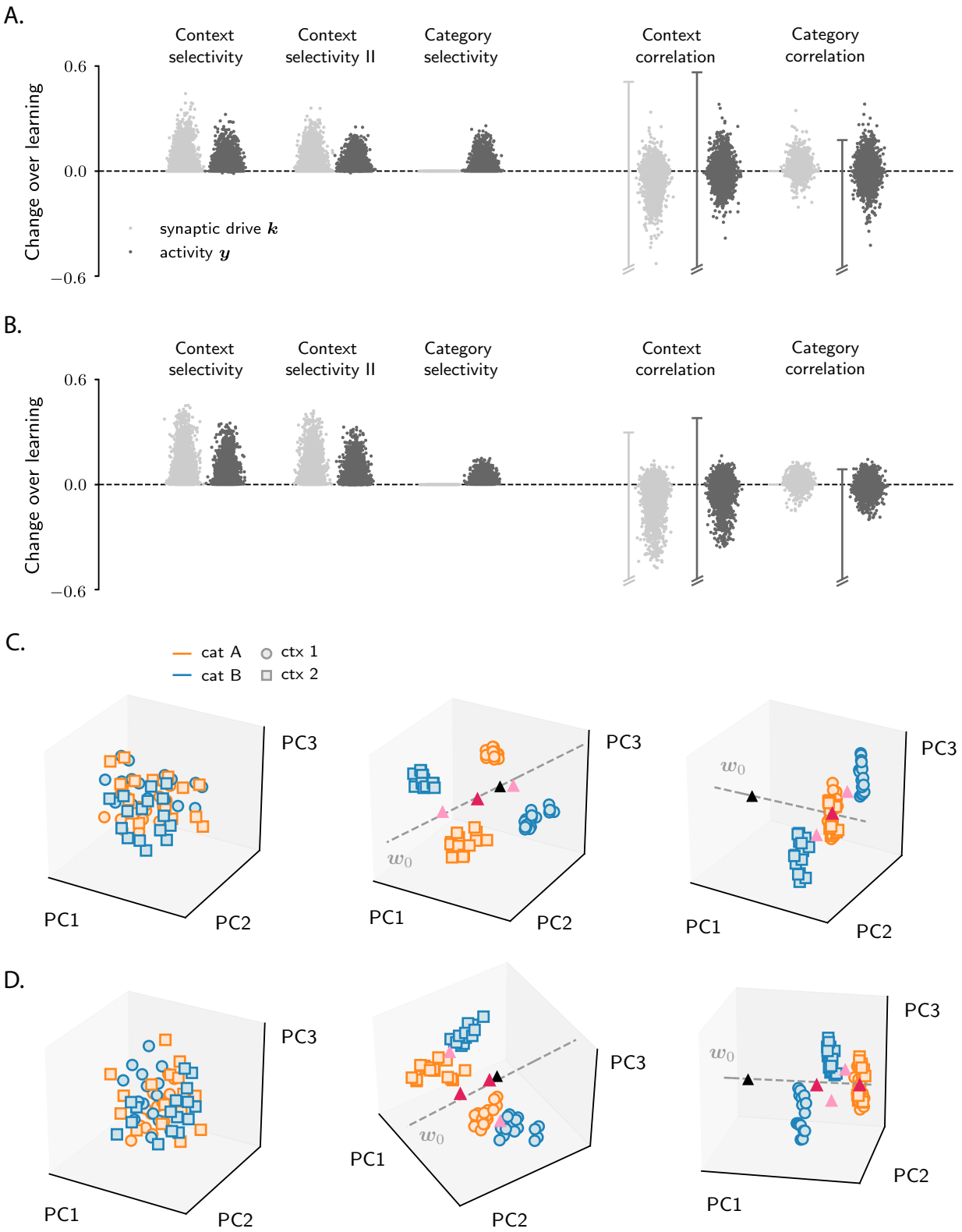
A.
Context selectivity
Context selectivity II
Category selectivity
Context correlation
Category correlation
Change over learning
0.6
0.0
−0.6
synaptic drive $\boldsymbol{k}$
activity $\boldsymbol{y}$
B.
Context selectivity
Context selectivity II
Category selectivity
Context correlation
Category correlation
Change over learning
0.6
0.0
−0.6
C.
cat A
cat B
ctx 1
ctx 2
PC1
PC2
PC3
$\boldsymbol{w}_0$
D.
PC1
PC2
PC3
$\boldsymbol{w}_0$

Fig. S6 *(previous page)*: **Characterization of activity evolution during the context-dependent categorization task; additional results, part I. A.** Extensive results from simulations. Each point represents a model. We simulated 5000 different models, with parameters drawn randomly. Details as in Fig. S1A, except that the number of stimuli and context cues, Q, ranges between 4 and 20. To quantify context selectivity, we used Eqs. 127 and 128, and then averaged over neurons. To quantify category selectivity, we used Eq. 61, and then averaged over neurons. To quantify context and category correlation, we used respectively Eqs. 132 and 76. The vertical bars in correlation plots indicate the range of results obtained when quantifying correlation with the alternative definitions in Eqs. 134 and 84; for illustration purposes, the range was cut on the negative semi-axis. For a few parameter sets, learning convergence was pathological (i.e., the loss displayed strong oscillations over epochs, or did not converge); those cases have been excluded from the analysis. Note that, as in Fig. S1A, selectivity changes are positive, while correlation changes take both positive and negative values. **B.** Same as in A, except that changes were computed from theoretical expressions. For context selectivity, we used Eqs. 130 and 131. For category selectivity, we used Eq. 129. For context and category correlation, we used the same equations as in panel A, but evaluated them via the theoretical framework. **C-D.** Activity from circuits in Figs. 5A-C (left), D-F (center) and G-I (right) projected on the first three principal components. Details as in Fig. S1B. Panels C and D display, respectively, results for the synaptic drive $\boldsymbol{k}$ and the activity $\boldsymbol{y}$. The center of activity clouds for category A and B (resp. context 1 and 2) is indicated by magenta (resp. pink) triangles; the center of initial activity is indicated by the black triangle.

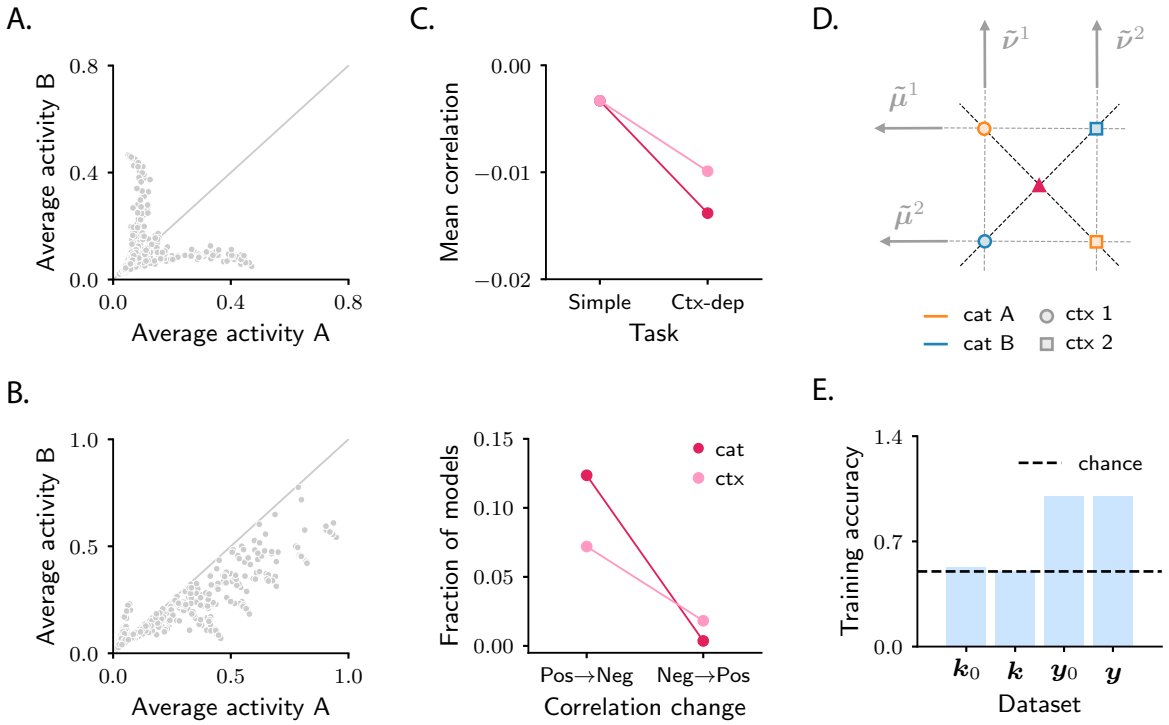

Fig. S7: **Characterization of activity evolution during the context-dependent categorization task; additional results, part II. A-B.** Population response to categories A and B, averaged over trials. Each dot represents a neuron. Panels A and B correspond to two sample circuits, characterized by different values of the offset of the activation function Φ. Note that responses in panels A are approximately symmetric, while those in panel B are strongly asymmetric. **C.** Context dependence makes categorization more complex and causes a drop in signal correlations: extensive quantification. We computed changes in correlations for the simple and the context-dependent task across a broad range of models; details as in Fig. S6B. The top panel displays the mean of correlation changes across models. Note that correlation changes are more negative in the context-dependent task. For the context-dependent task, we computed both the category and context correlation; results are displayed, respectively, in magenta and pink. In the bottom panel, we computed the fraction of models where correlation changes were positive in the simple task, and became negative in the context-dependent one (left), and viceversa (right). Note that the former substantially outnumbered the latter, which confirms that context-dependency shrinks the parameter space where changes in correlations are positive. **D.** Schematic representation of the initial synaptic drive, $\boldsymbol{k}_0^s$, in the context-dependent task with $Q = 2$ (XOR). The centers of activity vectors associated with category A and B are indicated by magenta triangles; note that the two triangles coincide exactly (Methods 5.8). **E.** Accuracy of a linear classifier trained to decode category from initial and final synaptic drive, $\boldsymbol{k}_0$ and $\boldsymbol{k}$, and initial and final activity, $\boldsymbol{y}_0$ and $\boldsymbol{y}$. Data are from the model displayed in Fig. 5D-F. To train the linear classifier, we used the function svm.SVC with a linear kernel from the package sklearn; the regularization parameter C was fixed through the GridSearchCV routine.

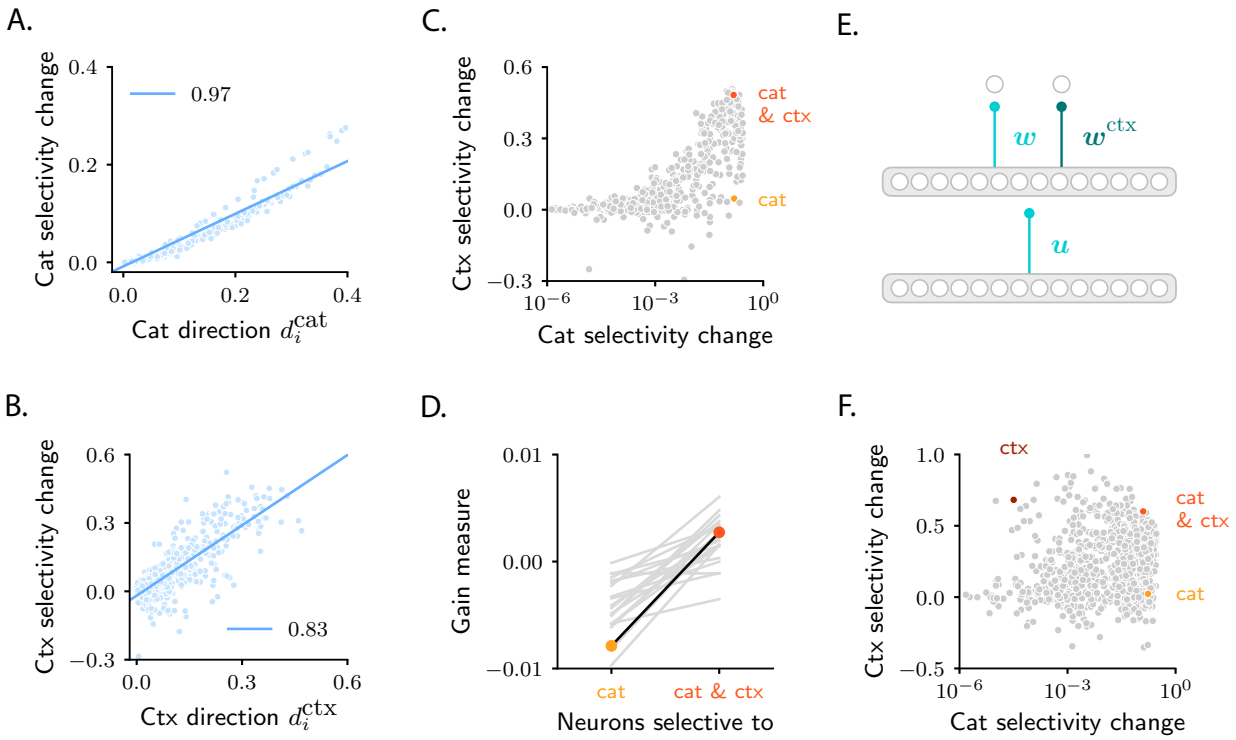

Fig. S8: **Characterization of activity evolution during the context-dependent categorization task; additional results, part III. A-B.** Changes in category (panel A) and context (panel B) selectivity as a function of the components on the category and context directions, d_i^{cat} and d_i^{ctx} (in absolute value). The latter are defined in Eq. 172, Methods 5.8. Each dot represents a neuron. Data are from the same model displayed in Fig. 7. Blue line: linear fit, with Pearson correlation coefficient shown in the figure legend. Note that correlations for both panels are high. **C-D.** Same analysis as in Figs. 7B-C, but for the sample circuit displayed in the third column of Fig. 5. **E-F.** Analysis of a circuit that includes a second readout neuron, trained to report context. Panel E: circuit architecture; this is identical to the standard model (Fig. 1A), except for the presence of the extra readout. All details and parameters are as in Fig. 7. To make sure that the learning process associated with the two readouts drives activity changes of similar magnitude in the intermediate layer, the target values for the context readout were taken to be $z^{\mathsf{A}} = 0.95$, $z^{\mathsf{B}} = 0.05$. Panel F: analysis as in Fig. 7B. Note that here, in contrast to Fig. 7B, some neurons display a strong increase in selectivity to context, but not category (brown sample neuron).

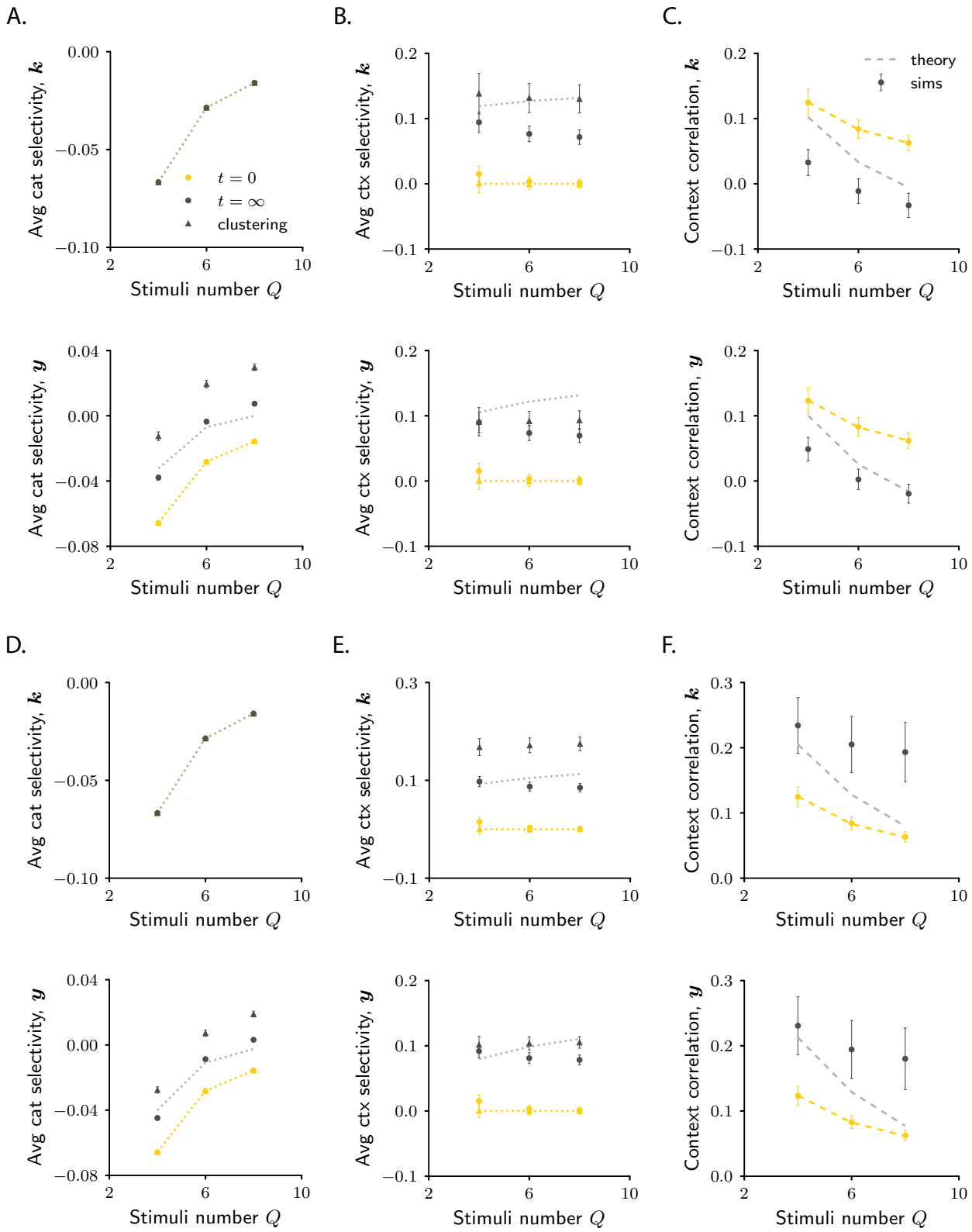
A.
B.
C.
D.
E.
F.
Avg cat selectivity, $\boldsymbol{k}$
Avg cat selectivity, $\boldsymbol{y}$
Avg ctx selectivity, $\boldsymbol{k}$
Avg ctx selectivity, $\boldsymbol{y}$
Context correlation, $\boldsymbol{k}$
Context correlation, $\boldsymbol{y}$
Stimuli number Q
$t = 0$
$t = \infty$
clustering
theory
sims

Fig. S9 *(previous page)*: **Comparison between finite-size networks and approximate mathematical description for the context-dependent categorization task.** Details are as in Fig. S3. In all panels, the top and bottom plots show, respectively, results for the synaptic drive $\boldsymbol{k}$ and the activity $\boldsymbol{y}$. **A.** Average category selectivity. Theory was computed from Eq. 129. Simulations results were computed by using Eq. 61, and then averaging across neurons. We also plot results for category clustering (triangles); clustering was computed by applying Eq. 129 to simulated data. **B.** Average context selectivity. Theory was computed from Eq. 130. Simulations results were computed by using Eq. 127, and then averaging across neurons. Context clustering was computed from Eq. 130. **C.** Context correlation. Theory was computed from Eq. 132. Simulations results were obtained applying Eq. 132 to simulated data. **D-E-F.** Details as in A-B-C. We used different parameters (see Table 2), which lead to positive changes in context correlation.