Transient recurrent dynamics shape representations in mice

June 17, 2025

Lars Schutzeichel[1,2,3*], Jan Bauer[4,5], Peter Bouss[1,3], Simon Musall[2,6,7], David Dahmen[1,†] and Moritz Helias[1,8,†]

[1]Institute for Advanced Simulation (IAS-6), Jülich Research Centre, Jülich, Germany
[2]Institute of Biological Information Processing (IBI-3), Jülich Research Centre, Jülich, Germany
[3]RWTH Aachen University, Aachen, Germany
[4]Gatsby Unit for Computational Neuroscience, University College London, London, United Kingdom
[5]Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel
[6]Institute of Biology 2, RWTH Aachen University, Aachen, Germany
[7]Faculty of Medicine, Institute of Experimental Epileptology and Cognition Research, University of Bonn, Bonn, Germany
[8]Department of Physics, RWTH Aachen University, Aachen, Germany

*To whom correspondence should be addressed; E-mail: l.schutzeichel@fz-juelich.de.
† shared senior authorship

Abstract

In the brain, different stimuli are represented by responses of a neural population. How is this representation reshaped over time by the dynamics of local recurrent circuits? We investigate this question in Neuropixels recordings of awake behaving mice and recurrent neural network models. We derive a mean-field theory that reduces the dynamics of complex networks to only three relevant dynamical quantities: the mean population activity and two overlaps that reflect the variability of responses within and across stimulus classes, respectively. This theory enables us to quantitatively explain experimental observables and reveals how the three quantities shape the separability of stimulus representations through a dynamic interplay.

We measure the information transmitted from multiple stimuli to the responses with an optimally trained readout on the population signal. This reveals a trade-off between more information conveyed with an increasing number of stimuli, and stimuli becoming less separable due to their larger overlap in the finite-dimensional neuronal space. We find that the experimentally observed small population activity lies within a regime where information increases with the number of stimuli, sharply separated from a second regime in which information asymptotically converges to zero — revealing a crucial advantage of sparse coding.

Introduction

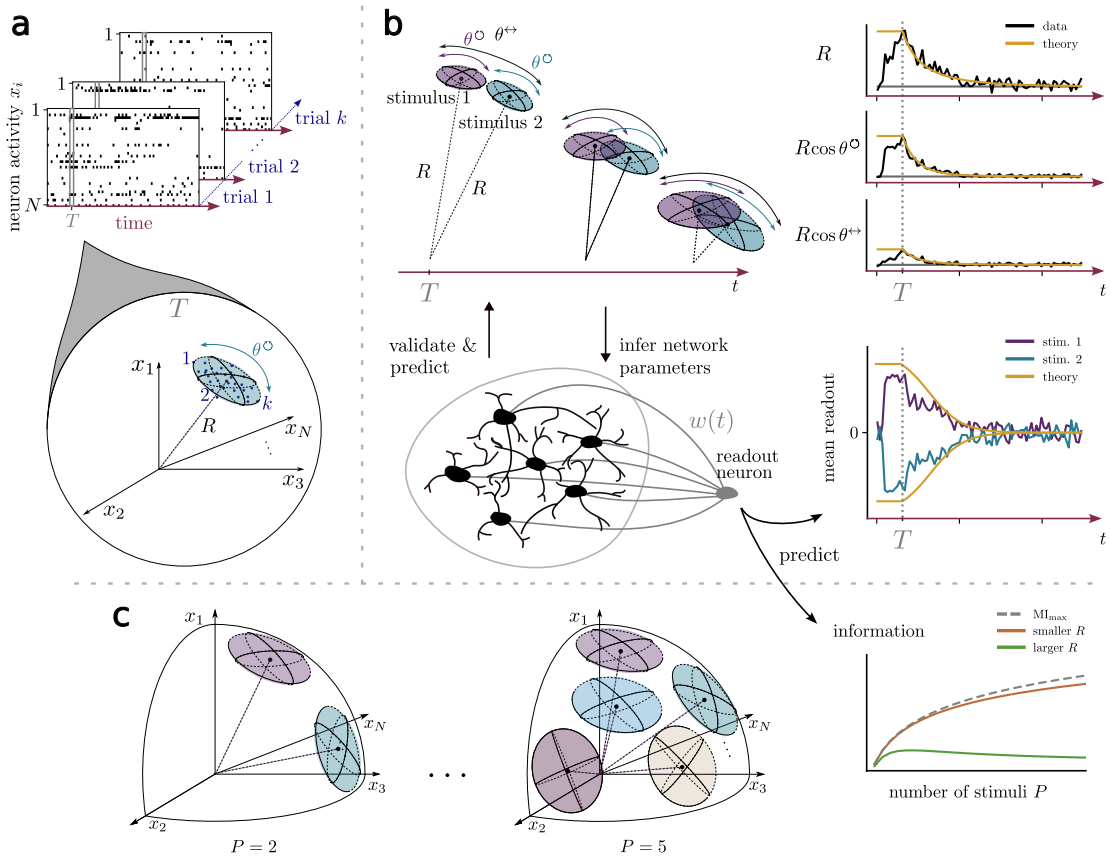

Figure 1: **Overview.** **a**: Raster plots of parallel neural recordings in superior colliculus from awake behaving mice in response to stimulation, e.g. a tactile stimulus (air puff to whisker). For any point in time T and any trial $k = 1, \ldots, N_{\text{trials}}$, the network activity is represented by a neural state vector, depicted as one point in neuronal space. We refer to different types of stimuli (e.g. visual or tactile) as "stimulus classes". Multiple trials of the same stimulus class form a point cloud in neuron space representing this stimulus class. Point clouds in neuron space are characterized by their distance from the origin R and their extent $\theta^{\circlearrowleft}$. These two quantities yield a representative ellipsoid for each cloud. **b**: Temporal evolution of the representation of stimuli from two classes (with relative angle $\theta^{\leftrightarrow}$), for example, tactile (blue) and visual (dark purple), after stimulus offset at time T, shown for data and mean-field theory. An optimally trained linear readout $\boldsymbol{w}(t)$ quantifies the separability between the classes of stimuli for every point in time, which is determined by the time courses of $R(t)$, $Q^{\circlearrowleft}(t) = R(t)\cos\theta^{\circlearrowleft}(t)$, and $Q^{\leftrightarrow}(t) = R(t)\cos\theta^{\leftrightarrow}(t)$. **c**: The separability measure also determines the information content contained in the population signal in the case of $P \geq 2$ stimuli. Encoding more stimuli generally increases information, while crowding of neuron space opposes this effect. The average population rate R controls a sharp transition between expanding (brown, small R) and vanishing (green, larger R) asymptotic information transfer.

Animals perceive external stimuli through an intricate network of neurons. Information about these stimuli is encoded in the activity of neuronal populations. Such population codes have first been found in motor activity and have been associated with motor output [1, 2, 3, 4, 5]. To identify and classify stimuli, neural representations must

preserve similarities between two repetitions of the same stimulus, but they must also be distinguishable between different stimulus classes. In recent years, understanding the geometrical structure underlying neuronal responses has led to insights into representations in both artificial [6, 7, 8] and biological networks [9, 10]. The geometry of neuronal responses in high-level areas has been found to lie on a low-dimensional manifold [11], whose size, shape, and dimensionality determine pattern classification accuracy and capacity [7]. Models have shown that such low-dimensional dynamics emerge, for example, from low-rank connectivity in recurrent networks [12].

In the brain, recurrence plays a crucial role in understanding the deformation of neural representations over time [13]. In the current work, we develop a general theory on how stimulus representations are deformed by recurrent networks and we validate this theory in experiments. In particular, we expose the fundamental mechanisms by which recurrency shapes similarities between stimulus representations and thus separability. We consider separability a minimal requirement for meaningful information processing: The signal propagates to downstream areas, where, along the processing hierarchy, representations of different perceptual objects must become increasingly separable to enable high-level cognition.

In the current work, we begin in Section 1.1 to analyze stimulus responses in the mouse superior colliculus (SC) obtained by massively parallel single unit Neuropixels recordings in awake behaving mice that were passively exposed to different stimuli. The SC is not only important for the processing of visual information in mice [14, 15], but it is also highly multimodal and therefore suitable for the analysis of different stimulus modalities [16, 17, 18].

Some qualitative features of stimulus responses repeat over many brain areas. One such property is neural variability: when presenting identical stimuli, the evoked responses share some characteristics, but they also show trial-to-trial variability across repeated presentations [19]. Neural variability is present also when stimuli are absent and it typically reduces upon stimulus application [20, 21, 22]. Thus, upon stimulus application, the neural activity and its variability form a representation of the encoded information in the N-dimensional activity space of all N recorded neurons, illustrated in Figure 1a. This representation can be visualized as an ellipsoid, where the distance of the ellipsoid's center from the origin is the firing rate averaged over the neuronal population (population mean R). The typical extent of the ellipsoid reflects the diversity of responses, its typical trial-to-trial variability. We denote the angle corresponding to this width as $\theta^{\circlearrowleft}$; the smaller this angle, the more similar the responses to this stimulus. When comparing one neural representation with that of another stimulus, the two are separated by a characteristic angular distance $\theta^{\leftrightarrow}$, which reflects the overlap between neural representations of different stimuli, as shown in Figure 1b.

Using large-scale recurrent network models and tools from statistical physics of disordered systems [23, 24], we show that, in the limit of many neurons, the separability of two neural representations by an optimally trained readout neuron is fully characterized by the three key observables, R, $\theta^{\circlearrowleft}$, and $\theta^{\leftrightarrow}$. Furthermore, we make use of the large-neuron limit to derive an effective mean-field description of the network dynamics. We show that the three observables form a closed set of equations which entirely governs their dynamics. To align the predictions of the recurrent network model to the experimental data, we show that it is sufficient that parameters assume the right statistics; a fine-tuning on the level of individual synaptic weights, for example, is not required. This suggests universal computational properties inherent to such networks. By fitting the model to data, we find parameters that yield a balanced network in which negative feedback stabilizes the dynamics and decorrelates neurons. The intrinsically chaotic network dynamics, in addition, promotes dissimilarities across stimulus responses [25, 26, 27].

As initial stimulus representations are deformed by intrinsic recurrent network dynamics (Figure 1b), the negative feedback decreases the population mean R and thus also the distance of the ellipsoids to the origin. Likewise, the time courses of the angular distances $\theta^{\circlearrowleft}$, $\theta^{\leftrightarrow}$ are driven by the chaotic network dynamics. This results in a dynamic interplay of these three observables. For large times, the angular distances $\theta^{\circlearrowleft}$, $\theta^{\leftrightarrow}$ both converge to the same value, rendering the stimuli indistinguishable.

In Section 1.2, we study how this dynamic interplay shapes stimulus separability. To assess separability, we use a linear readout neuron that is optimally trained using a Bayesian framework to obtain a signal that separates the two stimulus representations (Figure 1b). We show that the three observables also determine the separability achieved by the downstream neuron, establishing a direct link between these variables and the separability of neural representations. We use the network model to extrapolate the results towards realistic numbers of neurons

to estimate the information that would be accessible to a neuron that is part of such a network (Figure 1b).

In Section 1.3, we highlight the importance of the dynamic interplay that shapes separability: We present two settings of the network model which show, counterintuitively, that the population mean is not the main indicator of information retention – different temporal dynamics of R may indeed lead to identical separability. To better understand the mechanism behind this invariance, we derive a time evolution equation for the separability itself. This formulation allows us in Section 1.4 to predict stimulus settings in which a transient increase in separability appears as a result of the recurrent network processing [24]. This setting relies on the chaotic dynamics, which are induced by the discrete signaling of neurons and amplify initially small differences over time [25, 28], thereby facilitating classification [24].

The developed theory allows us to predict the classification of multiple stimuli P in Section 1.5. We consider a snapshot at the time of maximal response, from which we compute the information content of the population signal. To assess the ability to infer the stimulus class from the observed readouts, we evaluate the mutual information between stimulus and response, which represents the information transmitted from the stimuli, through the network, to the responses. For optimal information transmission, the stimulus representations must overlap as little as possible within the neuronal space. This raises the question of how to optimally distribute these representations (Figure 1c), which is linked to the capacity of neural networks [29]. To address this problem, we use spin-glass methods to calculate the distribution of mutual overlaps for P stimuli in a large neural space.

We explore different regimes, shown in Figure 1c, for the mutual information by studying the asymptotic information transmission ($P \to \infty$). We identify two distinct regimes characterized by their respective asymptotic value. They arise from the trade-off between two antagonistic mechanisms: With more stimuli, more information can be transmitted, but at the same time a larger number of stimuli incurs a higher mutual overlap, thus reducing their separability. We observe that the experimental data lie within a regime that is advantageous for information representation. The sparseness of the code is crucial for the separation between these regimes: not only is it beneficial for information representation [30] but, as we show here, a sufficient degree of sparseness is even necessary for non-vanishing asymptotic information transfer.

1 Results

1.1 Stimulus responses in mouse superior colliculus

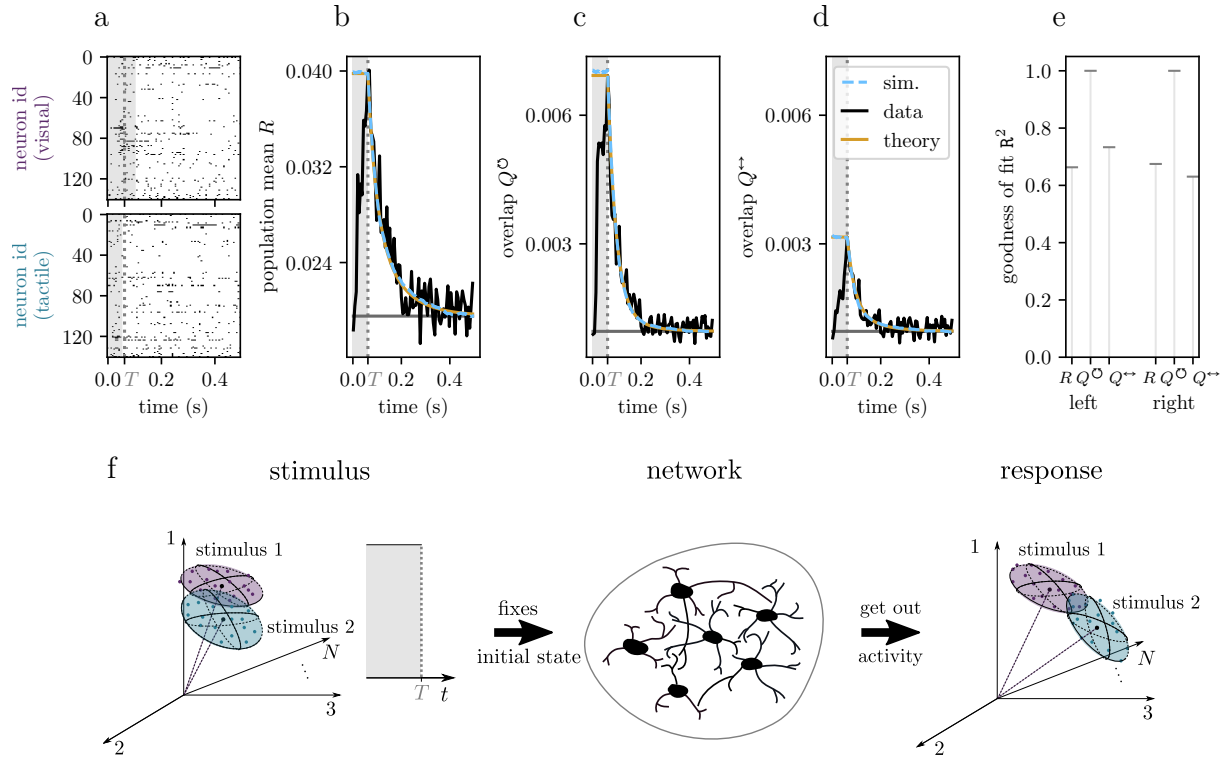

Figure 2: **Experimental data and validation of the network model.** **a**: Exemplary raster plots for visual (top) and tactile stimulus (bottom) in 6 ms bins. Shaded areas mark the duration of the stimulus, either a noise movie or an air puff, respectively. Neurons are sorted according to depth. **b**,**c**,**d:** Experimental observables (black) in counts/bin alongside mean-field theory predictions of the network model (yellow). The overlaps $Q^{\circlearrowleft}$ and $Q^{\leftrightarrow}$ are computed by averaging over all pairs of trials of identical or different stimulus classes, respectively. The decay of the experimental observables, their baselines, and peak heights are the quantities that are used to determine the statistics of the network model parameters (detailed in Methods 3.3). Numerical simulations of network models (light blue) (for network parameters, see Table 1). **e:** Goodness of fit R^2 for the three observables from two sessions for two different mice, in which both stimuli (visual and tactile) were on the left or right side, respectively; see Methods 3.3. **f**: Network theory yields a simplified description of the dynamics: A pair of stimuli is represented by two ellipsoids in stimulus space. After fixing the initial state, the stimulus is switched off at time T and the recurrent dynamics start to deform these initial representations. The time of stimulus application is indicated by a boxcar ending at T. The corresponding gray dotted vertical lines in panels a,b,c,d indicate the stimulus switch-off at T.

The application of different stimuli to mice causes a transient increase of activity in the superior colliculus, an area which shows responses to both visual and tactile stimuli. The mice are exposed to these stimuli over multiple trials. Figure 2a shows two trials where either a visual or a tactile stimulus was presented. In this raster plot, the stimuli evoke different characteristic responses. The firing rate or population mean $R(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$, shown in Figure 2b, measures the population activity at a given point in time averaged over both stimuli. Here, $x_i(t)$ is

the binned and binarized activity of neuron i in a time bin of $6\,\mathrm{ms}$ width at time t.

The population mean increases with stimulus onset and then relaxes back to baseline. Due to the binary nature of the spiking data, the population mean equals the length of the neural state vector.

The responses exhibit variability across trials. We measure the similarity of responses over time between two trials α and β by the overlap of two neural states: $Q^{(\alpha\beta)}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i^{(\alpha)}(t)x_i^{(\beta)}(t)$. Comparing Figure 2c with Figure 2d, as expected, the variability of responses between different stimulus classes is larger than the variability of responses within the same stimulus classes, corresponding to a smaller overlap $Q^{\leftrightarrow}$ in panel d compared to a larger overlap $Q^{\leftrightarrow}$ in panel c. These overlaps are directly related to the angles introduced in Figure 1. As time progresses, both overlaps decrease and ultimately approach the same asymptotic value, indicative of chaotic network dynamics which tend to amplify initial differences of states over time.

To quantitatively understand the dynamics of R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$, we build a recurrent network model of N_{model} neurons. We here employ a binary recurrent neural network model where neurons have a threshold, an average time τ between neuron updates and are connected by a random connectivity matrix with Gaussian distributed weights [24] (see Methods 3.2). The thresholds have an average value δ and a variability γ, while the connectivity has mean $\bar{g}$ and variance $1/N$. For large numbers of neurons N_{model}, we show with mean-field theory (Methods 3.2), that the three quantities R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$ follow a closed set of equations that governs the network dynamics. Moreover, the dynamics of the three quantities, to leading order in model size, do not depend on the precise realizations of the connectivity and the neurons' parameters, but only depend on their statistics, thus exposing a stereotypical collective behavior emerging from the recurrent network. The population mean R follows a self-consistency equation

$$\tau \frac{d}{dt} R = -R + \mathcal{T}(R, R)\,. \tag{1}$$

The concrete form of the function $\mathcal{T}$ and its dependence on the network parameters is given in Methods 3.2. The structure of the equation for R is consistent with previous works [31, 25, 32] and, in particular, is independent of the overlaps.

Mean-field theory for both Qs requires measuring the similarity of responses to different stimuli between two copies of the same network. This is accomplished using a two-replica approach, where each replicon corresponds to one of the two trials, yielding the pair of equations

$$\begin{aligned} \tau \frac{d}{dt} Q^{\circlearrowleft} &= -Q^{\circlearrowleft} + \mathcal{T}(R, Q^{\circlearrowleft})\,, \\ \tau \frac{d}{dt} Q^{\leftrightarrow} &= -Q^{\leftrightarrow} + \mathcal{T}(R, Q^{\leftrightarrow})\,. \end{aligned} \tag{2}$$

The coupling of the differential equation for the population mean (1) and (2) indicates a *dynamic interplay*: The population mean R influences the time evolution of both similarity measures $Q^{\circlearrowleft}$ and $Q^{\leftrightarrow}$ via $\mathcal{T}(R, Q)$. This function is the same as in the time evolution equation for R (1) and can be interpreted as a correlation transmission function, which quantifies the correlation of the output activity of a pair of neurons when driven by a pair of inputs of mean R and of mutual correlation Q.

The mean-field equations are used to determine the statistics of the network model parameters so that R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$ reproduce the experimental observables. We determine the statistics of network parameters by first constraining them by the experimental data and by then performing a least-squares fit (see Methods 3.3).

The mean-field theory, the experiment, and the simulations for the similarities are presented in Figure 2b-d. The goodness of fit of the theory to the experiment is shown in Figure 2e, evaluated here using the coefficient of determination R^2 (see Methods 3.3). The fit yields similar statistics of the network parameters for both mice (see Table 1). In particular, we find that the network connectivity has to be inhibition-dominated ($\bar{g} < 0$) to reproduce the experimentally observed sparse activity.

The inference of the statistics of the network parameters only requires the study of the coupled set of differential equations (1) and (2) for initial conditions on R and Q given by the experimental data. To bring a pair of network models into such an initial state, one needs to apply a pair of artificial stimuli that drive the response into subspaces of the activity space that correspond to this initial state (Figure 2f). In the limit of large networks, only the Gaussian correlations of these artificial stimuli matter. They can thus be represented by ellipsoids in stimulus space. When the networks have settled into their statistically stationary states at time T, the network reproduces the initial activity in the experimental data by the three observables, which corresponds to the ellipsoid abstraction of the point cloud. From this initial time point on, the artificial stimuli are switched off and the ellipsoids resembling the initial representation are continuously deformed by the recurrent dynamics in line with the experimental observations.

1.2 Intrinsic dynamics shape separability of stimulus representations

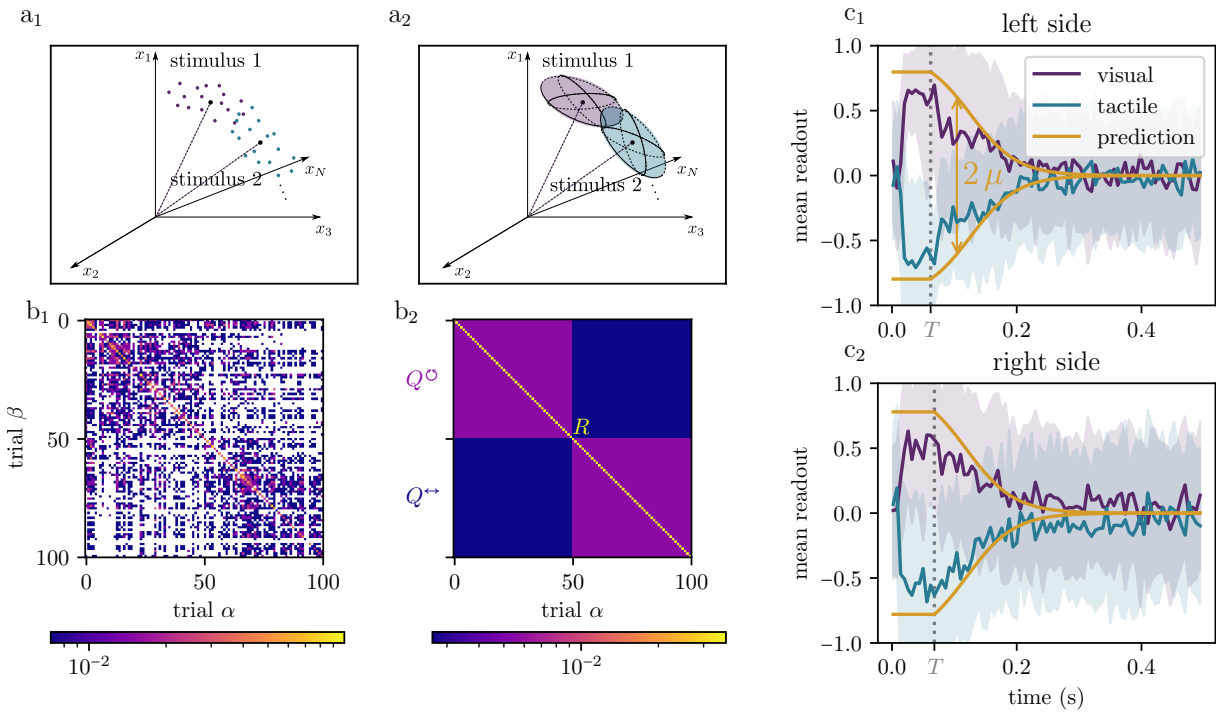

Figure 3: **Separability of neural representations: Experiment and mean-field theory.** Separability of stimuli is shaped by three observables. $\mathbf{a}_1$: Illustration of neural representations for the experimental recordings with a finite number of neurons. $\mathbf{a}_2$: Illustration of many-neuron limit of the mean-field theory, abstracting the experiment. $\mathbf{b}_1$: Experimental kernel matrix corresponding to a_1 (at $t = T$). White spots indicate vanishing entries. The trials are ordered such that the first half belongs to the first stimulus class and the second half to the other class, revealing a block-like structure. $\mathbf{b}_2$: Kernel matrix for the mean-field theory. In the many-neuron limit, the kernel matrix only has three distinct values: R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$, indicated by different colors. $\mathbf{c}_1, \mathbf{c}_2$: Time course of separability between visual (with label $y_\alpha = 1$) and tactile stimulus (with label $y_\alpha = -1$) in mouse superior colliculus (SC), obtained with the kernels from b_1,b_2. $\mathbf{c}_1$: Visual and tactile stimulus on the left side. $\mathbf{c}_2$: Visual and tactile stimulus on the right side. The shaded area represents the standard deviation of predicted labels over test data points and seeds from sampling the test data. For parameters of network model and simulation parameters, see Table 1.

The three observables R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$ provide an intuitive understanding of how different stimuli are represented. Subsequently, we aim to quantitatively assess the separability of these representations to characterize their compu-

tational properties. In experimental recordings, we have access to a limited number of neurons. After measuring the observables, determining the parameters of the model, and validating the model, we can use the network model to assess separability for a larger number of neurons than the recorded subset. This will allow us to make predictions for the separability of stimuli in the superior colliculus.

To decode stimulus classes from neuronal activity, we use a linear readout, which may be interpreted as a readout neuron, given that neurons perform a weighted summation of their inputs. At each time point t, the readout is $y_\alpha(t) = \boldsymbol{w}(t)^{\mathrm{T}}\boldsymbol{x}_\alpha(t) + \epsilon(t)$, with the time-dependent readout vector $\boldsymbol{w}(t) \in \mathbb{R}^N$ and Gaussian noise $\epsilon(t)$, accounting for stochasticity, for example due to afferent signals to the readout neuron from other brain areas (see Methods 3.4). Using Bayesian inference, we determine the optimal readout vector for each time point t to assess what the downstream neuron could optimally capture.

We split the data set into training $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{2N_{\mathrm{train}}}) \in \mathbb{R}^{N\times 2N_{\mathrm{train}}}$ and test data. The binary classification task is mapped to a regression problem by assigning either of two outputs $y_\alpha \in \{-1, 1\}$ depending on the stimulus class. These labels are gathered into a vector $\boldsymbol{y} \in \mathbb{R}^{2\,N_{\mathrm{train}}}$.

In the limit of many neurons, the expected label of a test point follows a Gaussian process posterior with mean $\mu(\boldsymbol{x}^*_\alpha) \coloneqq \langle y^*(\boldsymbol{x}^*_\alpha)\rangle = K(\boldsymbol{x}^*_\alpha, \boldsymbol{X})\left(K(\boldsymbol{X}, \boldsymbol{X}) + \kappa\mathbb{1}\right)^{-1}\boldsymbol{y}$, where $K(\circ, \circ)$ is the kernel matrix of the process [33, 34]. The entries of this kernel are pairwise similarities of network responses, $K(\boldsymbol{x}_\alpha(t), \boldsymbol{x}_\beta(t)) \coloneqq 1/N\,\boldsymbol{x}_\alpha(t)^{\mathrm{T}}\boldsymbol{x}_\beta(t) = 1/N\,\sum_{i=1}^N x_{\alpha i}(t)x_{\beta i}(t)$. Importantly, in this limit the kernel concentrates around its mean value and can hence be expressed in terms of the key observables introduced earlier, i.e. $1/N\,\boldsymbol{x}_\alpha(t)^{\mathrm{T}}\boldsymbol{x}_\alpha(t) \to R(t)$, $1/N\,\boldsymbol{x}_\alpha(t)^{\mathrm{T}}\boldsymbol{x}_\beta(t) \to Q^{\circlearrowleft}(t)$ for trials α and β belonging to the same class, and $1/N\,\boldsymbol{x}_\alpha(t)^{\mathrm{T}}\boldsymbol{x}_\beta(t) \to Q^{\leftrightarrow}(t)$ for trials α and β belonging to different classes. This limit corresponds to the abstraction step from a point cloud in Figure 3a$_1$ to the ellipsoid picture in panel a$_2$. Both the experimental and the mean-field kernel show a block-like structure in Figure 3b$_1$,b$_2$. The mean-field kernel has entries R on the diagonal, $Q^{\circlearrowleft}$ for entries corresponding to the same class (diagonal blocks), and $Q^{\leftrightarrow}$ for entries corresponding to different stimulus classes (off-diagonal blocks).

The scalar separability $\mu \coloneqq |\mu(\boldsymbol{x}^*_\alpha)|$ can be derived analytically

$$\mu = \left(1 + \frac{1}{N_{\mathrm{train}}}\frac{\kappa + R - Q^{\circlearrowleft}}{Q^{\circlearrowleft} - Q^{\leftrightarrow}}\right)^{-1}, \tag{3}$$

where κ is the variance of the noise (derived in Methods 3.4).

Therefore, we can understand separability assessed by a downstream neuron by the interplay of the three observables R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$. Specifically, the separability depends on a ratio of differences: It increases with the difference between the population mean and the within-class overlap (total variability $\kappa + R - Q^{\circlearrowleft}$), and decreases with the difference between the within-class and the between-class overlap (signal $Q^{\circlearrowleft} - Q^{\leftrightarrow}$). The former is a measure for the variability within one class compared to the maximal overlap R, while the latter can be understood as the signal, such that the ratio of differences in equation (3) resembles a ratio of the total variability to the signal. Notably, separability vanishes if stimuli within one class are as similar as stimuli between the two different classes, $Q^{\circlearrowleft} = Q^{\leftrightarrow}$. Furthermore, separability increases with the amount of training data per class, N_{train}, and decreases with noise κ. Adding noise has the same effect as increasing the population mean, while keeping the overlaps the same.

The experimentally observed separability and the extrapolation towards larger numbers of neurons than those recorded, obtained by equation (3), are shown in Figure 3c$_1$ and c$_2$. As expected, the extrapolated separability from mean-field theory exceeds the one obtained from the experimentally subsampled neurons; this is because the kernel matrix in the theory is homogeneous within each block (Figure 3b$_2$), while the experimental one has some heterogeneity (Figure 3b$_1$). Separability is highest immediately after stimulus application, and the time courses of the experimental and the predicted separability match well.

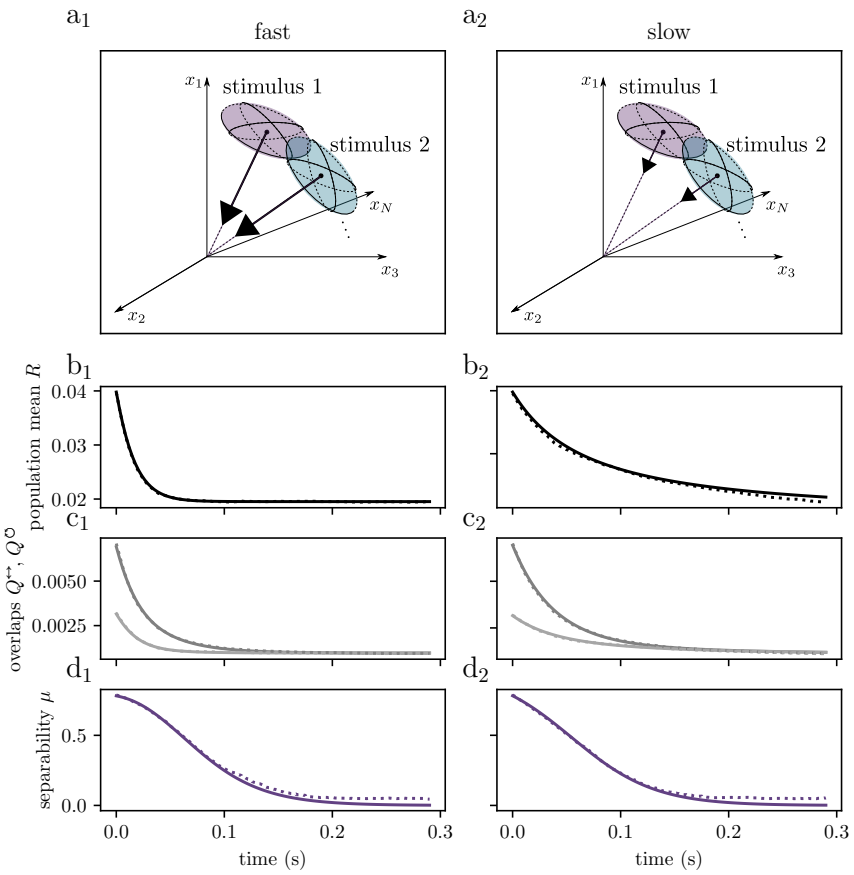

Figure 4: **Two scenarios of fast and slow decay of population mean with similar decay of separability.** $\mathbf{a}_1, \mathbf{a}_2$: Initial representations in neural state space. We compare two scenarios with different parameters that affect the decay of the population mean. The arrows indicate the different velocities of the population mean decay. $\mathbf{a}_1$: Fast. $\mathbf{a}_2$: Slow. $\mathbf{b}_1, \mathbf{b}_2$: Decay of the population mean that alters the decay of the within-class overlaps (dark gray) and between-class overlaps (light gray) in $\mathbf{c}_1$ and $\mathbf{c}_2$. Solid curves are theoretical predictions, dotted curves are simulations. $\mathbf{d}_1, \mathbf{d}_2$: Decay of the predictive mean for the different velocities of the decays of the population mean. The network parameters are those from the network model (see "left side" in Table 1), except for the ratio between inhibition $\bar{g}$ and threshold δ, respectively. The decay of the population mean is modulated by changing the slope of $\mathcal{T}(R, R)$ (Methods 3.2). Resulting parameters: Fast: $\bar{g}_{\text{fast}} = -10.03$, $\delta_{\text{fast}} = -0.11$. Slow: $\bar{g}_{\text{slow}} = -3.70$, $\delta_{\text{slow}} = -0.24$.

1.3 Firing rate decay alone does not reflect stimulus information retention

Having obtained an expression for the separability μ, we now ask how it is affected by the quantities R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$. Traditionally, a persistent firing rate or population mean R is considered indicative of information retention. In the model, the time course of R is controlled by inhibitory balancing, whereas $Q^{\circlearrowleft}$ and $Q^{\leftrightarrow}$ are driven by chaotic dynamics. Specifically, the rate of decay of R can be adjusted by modifying the ratio between inhibition and threshold (see Methods 3.2). All other parameters are obtained from the network adapted to the dataset of mouse 1 (stimuli on left side). In this particular case, despite the differences in the decay of R (Figure 4b$_1$, b$_2$), the decay of the overlaps in Figure 4c$_1$, c$_2$ changes in such a way that the separability in Figure 4d$_1$, d$_2$ evolves similarly. The three important observables R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$ enter the mean-field equations (2) and the equation for the predictive mean (3) in a dynamic fashion, but remarkably, the alterations to R cancel each other out in this particular case and yield a similar decay of the separability. Why does such a cancellation happen?

We gain insight into the decay of the separability by deriving its time evolution analytically (additional details in Section 3.4)

$$\tau \frac{d}{dt}\mu = -\mu\left(1-\mu\right)\left[s_1 - s_2\right], \tag{4}$$

which shows that the separability evolves slowly whenever $\mu \approx 0$ or $\mu \approx 1$. The right hand side is, however, not expressible in terms of μ alone, which means that the time evolution of the separability is not completely determined by the separability at an initial time – rather, the time course of the three observables R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$ explicitly influence the separability via the two terms s_1, s_2.

The first term, s_1, measures the relative change of the total variability $\kappa + R - Q^{\circlearrowleft}$ while the second term s_2 measures the relative change of the signal $Q^{\circlearrowleft} - Q^{\leftrightarrow}$. The difference between these two terms, by equation (4), modulates the speed by which μ evolves. Although the total variability decays faster with a rapidly decreasing population mean, the separability can exhibit a similar decay in both cases if the signal decay adjusts accordingly. The time evolution of the separability may also be understood geometrically, see Supplement 4.1.1.

This analysis highlights that firing rate alone is insufficient to infer information retention. Instead, our model suggests that the dynamic interplay of the three observables R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$ is crucial. This interplay can lead to a similar decay in separability even when the firing rates decay on different time scales. In particular, even a quickly decaying firing rate may be accompanied by correlations Q that are informative about the stimulus for a considerably longer time. Thus, stimuli can still be separable after the transient in the population mean has significantly subsided.

1.4 Transient increase in separability for low response variability and hard tasks

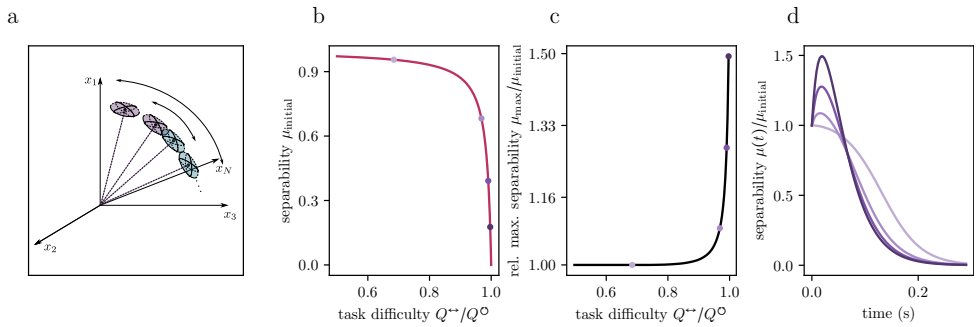

Figure 5: **Separability transiently increases for stimuli that are initially hard to separate.** **a**: Illustration of the initial state that is modified compared to the experiment. The overlap within one stimulus class $Q^{\circlearrowleft}$ is increased, reflected in narrower ellipsoids compared to Figure 4. The hardness of the task $Q^{\leftrightarrow}/Q^{\circlearrowleft}$ is varied as indicated by the black arrows. **b**: Initial value of separability μ_{initial} decreases with task difficulty. **c**: The maximum value of separability over time, $\mu_{\max}$ as a function of the task difficulty in relation to the initial value μ_{initial}. **d**: Separability over time normalized by the initial value. Different shadings correspond to different task difficulties, as indicated by markers in b and c. Parameter values: $Q^{\circlearrowleft}_{\max} = 0.99\,R_{\max}$, ratio of similarities: $Q^{\leftrightarrow}_{\max}/Q^{\circlearrowleft}_{\max} = \{0.72, 0.973, 0.992, 0.997\}$, variance of noise $\kappa = 0.5R_{\max}$, amount of training data per class $N_{\text{train}} = 35$.

The analytical form of the time evolution of separability (equation (4)) allows us to explore how the decay changes for different values of $Q^{\leftrightarrow}$ and $Q^{\circlearrowleft}$.

In the experimental data, tactile and visual stimuli can be clearly distinguished as shown by the separability in Figure 3c$_1$,c$_2$. To increase the difficulty of distinguishing stimuli, we decrease the distance between stimulus classes in the model, thus increasing their overlap $Q^{\leftrightarrow}$: the centers of the ellipsoids in Figure 5a are closer together. Simultaneously, we increase the overlap within classes to ensure that it remains larger than the overlap between classes. This corresponds to the ellipsoids becoming narrower in Figure 5a compared to Figure 4. The ratio between the distance of the ellipsoids and their extent $Q^{\leftrightarrow}/Q^{\circlearrowleft}$ reflects the task difficulty.

In the regime of larger task difficulty, we observe a transient increase in separability over time (Figure 5d). The reason is that chaos drives smaller overlaps (or larger distances) faster apart than larger overlaps (or smaller distances), particularly in biologically-inspired neuronal networks that have threshold-like activation functions [24]. This is due to the correlation transmission function $\mathcal{T}(R, Q)$ becoming infinitely steep at the maximum correlation $Q = R$, inducing a transient increase in separability.

According to equation (4), a transient expansion occurs when the signal grows faster than the total variability, i.e., when the second term s_2 is larger than the first term s_1 in equation (4). Since the slope of the correlation transmission function increases monotonically with Q, s_1 is always greater than s_2 in the absence of noise $\kappa = 0$ (for details see Supplement 4.1.1). So for a transient increase the signal has to increase ($s_2 > 1$) and the noise κ has to be sufficiently large to suppress the relative increase in total variability (see Methods 3.4 for more details).

While separability decreases with task difficulty in Figure 5b, the expansion is more pronounced for harder tasks (Figure 5c,d), where the ellipsoids are initially closer together. The transient increase in separability occurs at lower response variability than seen in the current set of experimental data. We expect that more similar pairs of stimuli will lead to representations that are harder to distinguish and may therefore enter the regime of transient increase.

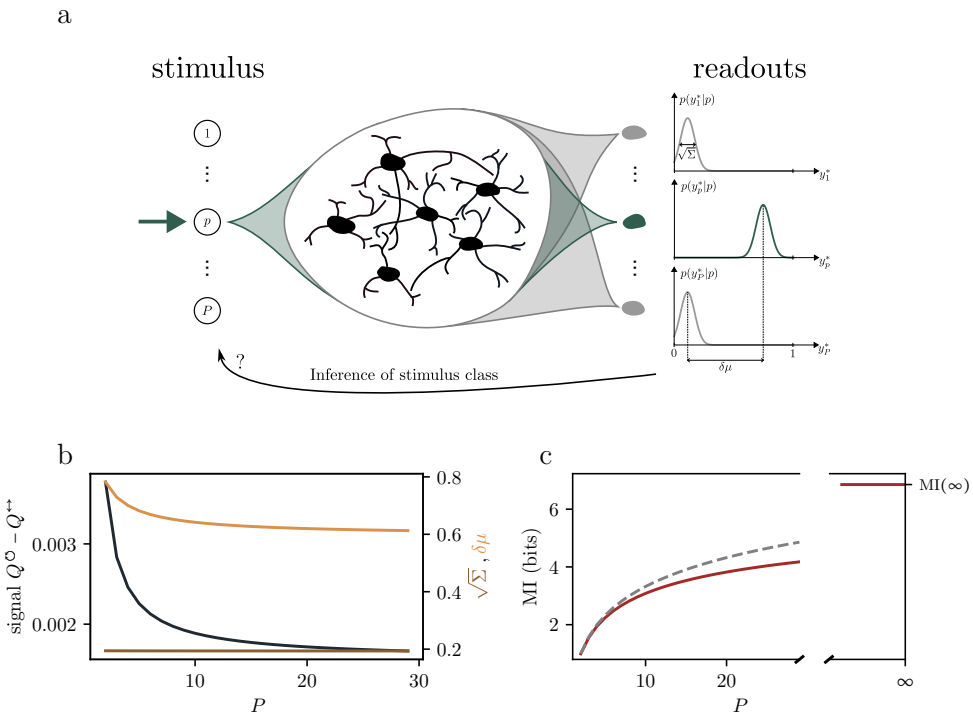

Figure 6: **Separability for more than two stimuli: Information transmitted from stimuli through the network to readouts. a**: Schematic drawing: The network is subject to one of P stimuli. The elicited network state is read out by P linear readout neurons, one for each stimulus: The p-th linear readout is trained such that its response is 1 for the p-th stimulus and zero otherwise (one-hot encoding). Through the overlap with the other stimuli, the other linear readouts also show a response for the p-th stimulus. **b**: The effect of crowding in neuron space. Left axis: The signal (the difference between within- and between-class overlap) decreases with increasing P (black). Right axis: Separability $\delta\mu$ decreases accordingly and posterior variance approximately stays constant as a function of P. **c**: Left: Mutual information $\mathrm{MI}(\{y_p^*\}_{p=1,\ldots,P}; C)$ between readouts y_p^* and stimulus classes C as a function of number of stimuli P for the experimental observables (red curve). The dashed curve indicates the maximally possible mutual information $\log_2 P$. The asymptotic information transmission (the saturation value of the mutual information for $P \to \infty$) is indicated in the broken axis panel. Parameter values are as in Table 1.

1.5 Information in population signal

Having elucidated the mechanisms by which recurrent dynamics shape stimulus representations, we next investigate how much information about the stimulus class is encoded in the neural activity. As the brain must routinely distinguish among a far greater number of sensory inputs, we leverage the model to extrapolate beyond the case of binary classification studied so far. The aim is to examine how information encoding scales with the number of stimulus classes. This allows us to assess the model's capacity for stimulus discrimination in more realistic settings.

In addition, we systematically vary the parameters of the model to explore whether different regimes emerge with respect to information transmission. To obtain the maximal information capacity across time, we choose the time point of peak population response, providing a snapshot of maximal representational separation. We then analyze the asymptotic behavior as the number of stimulus classes increases, which serves as a proxy for the model's information capacity. Finally, by locating the parameters in the phase space of the model, we characterize the regime in which the superior colliculus (SC) operates.

We consider P stimuli, each eliciting a characteristic network state, see Figure 6a. Rather than focusing on the distance between two stimulus representations, we must now consider the joint embedding of P stimuli, which inevitably generates overlap between pairs of representations. This overlap $Q^{\leftrightarrow}(P)$ in the limit of large numbers of neurons, can be derived by help of spin-glass methods (more details in Methods 3.5) as

$$Q^{\leftrightarrow}(P) = \frac{1}{N}\sum_{i=1}^{N} R_i^2 - 2\frac{\mathcal{E}}{\sqrt{N}}\sqrt{\frac{\sum_{i=1}^{N} R_i^2(1-R_i)^2}{P(P-1)}}. \tag{5}$$

The elicited network states in response to stimuli are binary vectors in neuron space, where each neuron i has a firing rate R_i, measured from trial averages in the experiment. The first term $\frac{1}{N}\sum_{i=1}^{N} R_i^2$ in equation (5) is the typical overlap of representations if neurons were firing independently across stimuli. An optimal system will use the space more efficiently, which is captured by the second term reducing the overlap $Q^{\leftrightarrow}(P)$. The degree to which the brain is close to this optimum is controlled by the efficiency $\mathcal{E}/\sqrt{N}$, which is determined using the experimentally measured overlap of two stimuli $Q^{\leftrightarrow}(2) \overset{!}{=} Q^{\leftrightarrow,\mathrm{exp}}$. From equation (5), we observe that as P increases, distributing stimuli becomes more difficult because of their finite extent: The overlap increases and asymptotically approaches the average overlap given by the first term. This leads to a reduction in the signal, as shown in Figure 6b. Consequently, the separability in the limit of many neurons also decreases with P, as seen in Figure 6b. In the limit of many neurons, the separability is again given by equation (3), denoted as $\delta\mu(P)$ here, with $Q^{\leftrightarrow} = Q^{\leftrightarrow}(P)$ depending on P (for more details see Supplement 5).

Balancing information and overlap

How much information is contained in the population signal for P stimuli? In general, more stimuli provide more information; however, the neuronal space becomes increasingly crowded, thereby increasing $Q^{\leftrightarrow}(P)$, which makes it more difficult to infer the stimulus class from the responses in Figure 6a. To quantify the trade-off between the information gain and the decreasing separability, we use the Kullback-Leibler divergence $D_{\mathrm{KL}}(\circ\|\circ)$ to assess the mutual information between stimuli and responses

$$\mathrm{MI} = D_{\mathrm{KL}}\left(p(C,\boldsymbol{y}^*)\|p(C)\,p(\boldsymbol{y}^*)\right). \tag{6}$$

It captures the information that is transmitted from the stimuli through the network to the responses. We consider the case of all P stimuli being equally likely $p(C) = \frac{1}{P}$. The responses are given by the readouts. Each stimulus evokes a distinct network state, which is subsequently decoded by the set of P linear readout neurons. Here, the p-th readout neuron is trained so that its response is largest for the p-th stimulus and is supposed to be zero else. However, non-matching readout neurons also show non-vanishing responses because of the mutual overlap

between different stimulus representations.. In this context, separability refers to the mean difference between the target readout's output and that of all others, see Figure 6a. We denote the variability of the outputs with Σ. In comparison to the separability, this variability $\Sigma(P) \approx \Sigma$ approximately stays constant, as shown in Figure 6c.

The ratio between these two quantities

$$\Delta(P) = \frac{\delta\mu(P)}{\sqrt{\Sigma(P)}} \approx \frac{\delta\mu(P)}{\sqrt{\Sigma}} \tag{7}$$

is the effective separability, which, in addition to P, fully determines the mutual information in equation (6) as

$$\mathrm{MI}(P, \Delta) \approx \log_2(e)\,(\Delta(P))^2 - \log_2(e)\, m(\Delta(P), P). \tag{8}$$

Here, $m(\Delta(P), P)$ is a parameter resulting from an approximation of a sum of random variables (see Methods 3.5 for more details).

Using the experimentally measured values, we find that the mutual information increases as the number of stimuli grows, following a trend similar to its theoretical maximum, $\log_2 P$, as shown in Figure 6c.

For an infinite number of stimuli, the mutual information, however, saturates

$$\mathrm{MI}(P) \overset{P\to\infty}{\to} \frac{1}{2}\log_2(e)\,(\Delta(\infty))^2\,. \tag{9}$$

This limit is formally derived in Supplement 5.3, and it admits an intuitive interpretation: Each new readout provides an additional projection of the network state. However, as the number of readouts increases, the incremental benefit of an additional projection diminishes, as most of the network has already been effectively accessed.

The effect that limits the growth of the mutual information is the crowding of stimulus representations in neural space. The saturation value of the mutual information arises from the balance between the crowding and the information gained from more stimuli in the population signal. How does this balance depend on the population mean R? To operationalize this question, we scale each individual neurons' rate R_i by the same factor, which preserves the shape of the experimentally measured firing rate distribution $\{R_i\}_i$. This scaling argument is based on the typically observed exponential dependence of the firing rate on the mean input to the neuron [35] (for more details, see Methods 3.6 and Supplement 5.5). From this consideration it follows how the overlap between classes changes with the population mean in equation (5).

In Figure 7a, we observe two factors that enhance mutual information: (1) low response variability within classes and (2) a small population mean. A cross-section of mutual information along the line preserving the experimentally measured Pearson correlation coefficient within classes is shown in Figure 7b. This Pearson correlation coefficient ρ_i quantifies the selectivity of every neuron to stimulus classes. For a very small population mean, the mutual information decreases due to the finite noise κ on the readout, as seen in Figure 7b. Two phases emerge in Figure 7a: In the expanding regime (for small population mean), the saturation value of mutual information is positive, and mutual information increases with P in a similar fashion to its theoretical maximum, $\log_2 P$. In the vanishing regime (for higher population mean), the mutual information initially follows the $\log_2 P$ trend, reaches a peak, but then declines when P increases further, eventually approaching zero. Examples for the two cases in comparison to the case for the experimentally observed population mean are shown in Figure 7c. The transition between the regimes is continuous, with the point of vanishing mutual information corresponding to where the overlap within classes equals the overlap between classes $Q^{\circlearrowleft} = Q^{\leftrightarrow}$. Assuming that neuron class selectivity remains approximately constant $\rho(R) \approx \rho(R^{\mathrm{exp}})$ as the firing rate increases from the experimentally measured R^{exp}, the overlap between classes $Q^{\leftrightarrow}$ grows more rapidly with the firing rate than the overlap within classes $Q^{\circlearrowleft}$, leading to such a transition point. This can be seen from Figure 7a, where the phase boundary between the expanding and the vanishing regime corresponds to $Q^{\circlearrowleft} = Q^{\leftrightarrow}$. As the population mean R increases, the line indicating this boundary has a steeper slope than the black line indicating the overlap within classes $Q^{\circlearrowleft}$. The transition point between the two phases - where these two lines intersect (near the green point) - remains the same when scaling class selectivities $\rho \to \rho^{\mathrm{small}}, \rho^{\mathrm{large}}$ (dashed gray lines).

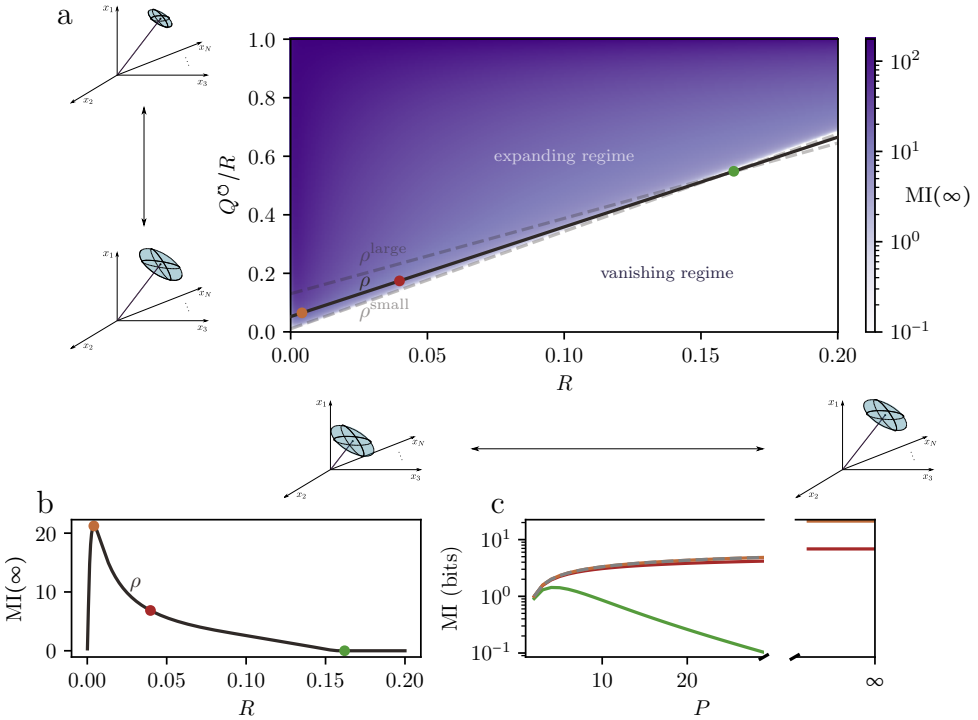

Figure 7: **Regimes for asymptotic information transmission.** **a**: Asymptotic value $\mathrm{MI}(\infty)$ for different population means R and overlap within classes $Q^{\circlearrowleft}/R$. $Q^{\circlearrowleft}/R$ can be understood as the (cosine of the) angle $\theta^{\circlearrowleft}$ that measures the extent (variability) of the neural representation ellipsoid for one stimulus. The colored area represents the regime where $\mathrm{MI}(\infty) \neq 0$. White space represents the vanishing regime, $\mathrm{MI}(\infty) = 0$. The vanishing regime (white) is separated from the expanding regime (purple) by a line indicating the phase boundary. Along this line, $Q^{\circlearrowleft} = Q^{\leftrightarrow}$. Values smaller than $0.1\,\mathrm{bits}$ are colored white. The black curve in panel a represents values of $Q^{\circlearrowleft}$ as a function of R for constant Pearson correlation coefficient ρ_i for every neuron i (the gray dashed line indicates smaller (ρ^{small}) or larger (ρ^{large}) Pearson correlation coefficients than in the experiment; see Methods 3.6). Dots indicate different population means. The intermediate population mean (red) is identical to the experiment with $R^{\text{exp}} = 0.04$ (same as in Figure 6c). The small population mean (brown dot) indicates optimal sparseness. The large population mean (green dot) is an example of vanishing asymptotic information transfer in the vanishing regime. **b**: Asymptotic mutual information $\mathrm{MI}(\infty)$ along the black line of panel a. **c**: Asymptotic mutual information $\mathrm{MI}(\infty)$ as a function of P. Colors correspond to the different population means that are indicated as dots (green, brown) in panel b. The curve for the optimal asymptotic (brown) value starts very close to its theoretical maximum $\mathrm{MI}_{\max} = \log_2 P$ (dashed, gray). The mutual information curve decreases and ultimately vanishes if the crowding effect is too strong (green). The saturation value of the MI is indicated by the broken axis. Parameter values are as in Table 1.

In conclusion, a lower population mean R expands the effective encoding space, leading to better information representation. Sparse coding, therefore, is not only advantageous for information representation but is also a necessary condition for non-vanishing asymptotic information transfer.

2 Discussion

In this study, we investigate the representation of stimuli in neuronal activity recorded from Neuropixels probes in the superior colliculus (SC) in mice. The neural activity of each trial can be thought of as describing a neural trajectory. Recent years have seen growing interest in the functional implications of such neural trajectories traversing neural state space – particularly within low-dimensional manifolds of high-level areas – through both theoretical and experimental work [2, 36, 37, 11, 12].

We here find that upon stimulus presentation, these neural trajectories align and subsequently decorrelate again. Eventually, they reach a state similar to the one before stimulus application. For large networks, we demonstrate that a downstream neuron's ability to separate the neural representations is given by three key observables: the *population firing rate* (R) and two types of *correlations* ($Q^{\circlearrowleft}$ and $Q^{\leftrightarrow}$) between pairs of neural trajectories.

Our analysis concerns the superior colliculus, a midbrain structure that integrates high-dimensional input from the retina, the primary visual cortex, and primary somatosensory cortex. It distills this multisensory information and initiates motor commands [15]. As neural representations of different stimuli are transmitted from the superior colliculus to higher-level areas, they need to become separable.

In this work, we demonstrate that this separability is shaped by the dynamic interplay of three key observables R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$ in recurrent networks: We extend the two-replica mean-field theory from [24] by incorporating transient firing rate dynamics and temporally sparse activity, and thereby reveal the critical role of these three observables in the characterization of neural representations. Moreover, this theory provides analytically tractable time-evolution equations that describe the dynamical interaction of these observables in recurrent neural networks in a self-consistent manner. Specifically, it uncovers how inhibitory balancing controls the time course of the population mean R, while chaotic dynamics drive the overlaps $Q^{\circlearrowleft}$ and $Q^{\leftrightarrow}$. Together, these three observables govern stimulus separation in a downstream readout neuron, uncovering the mechanisms that underlie separability.

We test this theory by applying it to experimental data. The theory shows that network parameters tuning inhibition-excitation balance, thresholds, and the neuronal time constant enter only at the level of their statistics. To explain the time courses observed in experimental data, it is thus sufficient to fit these statistics – there is no fine-tuning needed on the level of individual neurons or synapses. This suggests that the network model contains the crucial building blocks for network-level processing. Furthermore, the agreement between the theoretical predictions of separability dynamics and experimental results supports the sufficiency of this theoretical framework.

With this network model, we can explore further implications of the theory. We highlight that information retention crucially depends on the mutual interplay between the three observables rather than on the firing rate alone: We demonstrate two cases where separability decays similarly, despite differences in firing rate decay. Previous studies have shown that a transient change in firing rate causes long lasting trial-to-trial decorrelation in disordered balanced networks [38, 28]. Our findings suggest that the information is contained within this decorrelation. We predict a transient increase in separability for stimulus representations that are hard to separate and exhibit low within-class variability. This transient increase occurs because the signal grows faster than the total variability, proposing a hypothesis on how chaotic dynamics on the network level contribute to computation.

By considering multiple stimuli, we identify two regimes for information transfer in the limit of large numbers of stimuli, that are separated by a sharp phase boundary: the expanding regime at low firing rate and a vanishing regime at higher rates. The experimental data with its sparse activity and corresponding low rate lies within the expanding regime. We conclude that sparse coding is not only beneficial, but also necessary for information representation and information transmission.

Our network model has only few parameters, which, surprisingly, are sufficient to reproduce and explain the main features of the data. In this sense, the theory is minimal. Specifically, we assume random Gaussian connectivity, whose sufficiency lies in the fact that the resulting model tracks the key observables relevant for a downstream neuron reading out from a large network. In particular this connectivity gives rise to the overlap Q, which captures the variability homogeneously across neurons, leading to the depiction of the neural activity over multiple trials as ellipsoids in Figure 1a. This formulation is advantageous because it remains agnostic to the specific direction of variability.

In reality, the superior colliculus is a midbrain structure composed of different laminae, which in turn are composed of different types of neurons that differ in their responses [15]. For neurons reading out from specific parts of a brain area or along some specific axis, it may become important to incorporate structure in the connectivity. Accounting for structured connectivity [12, 39, 40] would skew the ellipsoids, altering the directions of variability they measure and thereby affecting their subsequent decorrelation. Measuring the variability along a specific axis may also be advantageous when neural activity is confined to a manifold with a specific directionality [7, 41]. This may be essential for more complex task setups (beyond the passive stimulation examined here) and recordings in higher-level areas. In our theoretical formalism, including structure in the connectivity can be achieved by relaxing the assumption of statistical homogeneity across all neurons, thereby introducing heterogeneity.

The prediction of separability in the mean-field approach corresponds to an extrapolation to large networks, allowing a readout neuron to achieve a higher signal-to-noise ratio than that directly estimated from the observed experimental data. This is consistent with previous findings showing that finite-size corrections reduce separability [42]. The expression for separability is derived in the limit of large network size [33, 34], while keeping the amount of data finite. In machine learning, this corresponds to the lazy-learning regime [43], where training the network is equivalent to training only the readout [44, 45]. In other words, in this limit, training the whole network does not substantially change the recurrent weights relative to their initial values. Consequently, the simple and analytically tractable expression for separability remains valid.

For networks that learn features, one could adopt a Bayesian approach similar to that used to study artificial networks, which is agnostic to the learning rule [46, 47], or implement plasticity into the field-theoretic model [48, 49]. This would enable modeling of learning and task adaptation, with a focus on how the interaction between neuronal and plasticity timescales influences neural representations and their separability.

In the experimental data, a transient increase of separability has not been observed. However, we predict a transient increase in separability that peaks at a characteristic time for stimuli that are more difficult to separate (i.e., higher task difficulty) and exhibit lower variability within stimulus classes. Such a transient increase would correspondingly enhance the amount of information conveyed in the population signal, with a maximum occurring later than immediately after stimulus application.

In addition to stimulus-related sources, variability can also stem from ongoing neural activity, presumably reflecting varying brain states [19] or corresponding to unobserved behavior [50, 51, 52]. It would be interesting to measure how trial-to-trial variability changes with learning [53, 50], particularly between and within highly similar stimuli, and to test the theory under conditions of lower response variability.

To quantify mutual information, we consider a set of P stimuli. Rather than considering the full network state, we project it onto a number of readout neurons equal to the number of stimuli, which then defines the response. This presents a dimensionality reduction, where every readout neuron is optimal in a Bayesian sense to maximize the mutual information between inputs and outputs [54, 55, 56, 57]. We again leverage the fact that a downstream neuron separates network states according to the geometry in neural activity space given by the three observables R, $Q^{\circlearrowleft}$, and $Q^{\leftrightarrow}$. This allows us to relate mutual information directly to separability.

The amount of mutual information depends on the choice of the prior variance for the readout weights (g_w), which sets the scale for the noise in the readout neurons. We choose this variance such that the correlation between readout neurons matches that of the experimentally recorded neurons.

The asymptotic amount of information increases with decreasing firing rate, supporting the sparse coding strategy of the brain [30]. This occurs because the total variability, $\kappa + R - Q^{\circlearrowleft}$, decreases faster with decreasing population mean than the signal, $Q^{\circlearrowleft} - Q^{\leftrightarrow}$, leading to an enhanced signal-to-noise ratio that determines the separability $\delta\mu$.

Sparse coding enables efficient storage [29, 58], communication, and computation despite drastic compression [59]. In addition to these computational advantages, it also reduces energetic costs - another benefit of sparse coding [60]. Our findings do not rely on these previous arguments and thus offer a novel perspective on the

classical advantages provided by sparse coding.

For sufficiently high rates - at low levels of sparseness - we observe a regime in which the asymptotic information transfer vanishes and a maximum amount of information appears at a certain number of stimuli. This optimal number of separable stimuli arises from a trade-off between transmitting more information per stimulus and the neuron space becoming too crowded, which increases the mutual overlap of responses to unrelated stimuli. The vanishing regime arises for a larger population mean than that observed in the experiment. To extrapolate to such higher population means, we scaled the firing rate of each neuron by a common factor, which is valid for small population means, where the neuronal activation function behaves approximately like an exponential [35]. As the activity increases, the activation function becomes more linear. Reassuringly, however, the vanishing regime remains prevalent and is robust to these changes, as discussed in Methods 3.6. In fact, there is a third regime, distinct from the expanding and vanishing regimes, with negligible width, where the mutual information curve is non-monotonous, as in the vanishing regime, but saturates to a non-zero value (more details in Supplement 5.6).

It would be interesting to experimentally test whether neural representations of stimuli in a finite neuronal space are organized in a way that optimally transmits information according to the mutual information considered here. Concretely, the behavioral accuracy of mice as a function of task complexity could be investigated to test whether the data falls into the expanding or vanishing regime.

3 Methods

3.1 Experimental design and statistical analysis

We analyze parallel neuronal in vivo electrophysiology recordings from mouse superior colliculus. All experiments were carried out in accordance with the German animal protection law and local ethics committee (LANUV, NRW)

Animals:

Animals were 16-week old male mice. Surgical details for electrophysiological recordings are described in [61].

Recordings:

To record from superior colliculus (SC), mice were head-fixed on a rubber coated wheel and Neuropixels probes [62, 63] were inserted into the left hemisphere of the brain. The bottom 384 channels of the Neuropixels probe were then used to record a high pass filtered action potential signal ($30\,\mathrm{kHz}$). We simultaneously recorded $N = 141$ (mouse 1) and $N = 65$ (mouse 2) neurons in the SC.

Stimulation:

Visual stimulation was performed using Gaussian noise movies displayed at $60\,\mathrm{fps}$ for $100\,\mathrm{ms}$ at a $17\,\mathrm{cm}$ distance from the right eye on a gamma-corrected LED-backlit LCD monitor (Viewsonic VX3276-2K-MHD-2, 32”, $60\,\mathrm{Hz}$ refresh rate). Monitor brightness was calibrated to $60\,\mathrm{lux}$. Tactile stimulation consisted of puffs of air that were given in short bursts with a duration of $50\,\mathrm{ms}$ and a pressure of $0.03\,\mathrm{bar}$. Spouts delivering the stimulation were targeted to hit the whiskers of the mouse and aimed so that they did only elicit whisker deflections but no additional startle reactions. Stimuli were delivered for 50 (mouse 1) and 40 (mouse 2) trials each. The time between individual trials was 3-$5\,\mathrm{s}$.

Pre-processing:

The Neuropixel recordings provide voltage traces for each channel, from which spike waveforms stemming from different neurons have to be identified and classified. For spike sorting, Kilosort 2 was utilized for the data set of mouse 1 and Kilosort 2.5 for mouse 2 [64].

We bin the data in $6\,\mathrm{ms}$ bins, leaving us with a discretized time axis and an entry $x_{\alpha i}(t) \in \mathbb{N}_0$ for every neuron i, trial α, and time t. For this bin size only 1.2% (mouse 1) and 1.0% (mouse 2) of all non-zero entries are greater than 1. In the following we set these entries to 1 to ensure Boolean values $x_{\alpha i}(t) \in \{0, 1\}$. Additional details can be found in Supplement 1.

3.2 Recurrent neural network model

We use a network of binary neurons, where the neurons can be in either of two states: actively firing $x_i(t) = 1$ or quiescent $x_i(t) = 0$ [65, 66, 25, 67, 68, 69]. The state of each neuron i is dynamically updated according to the synaptic input $h_i(t) = \sum_{j=1}^{N} J_{ij} x_j(t)$, which is the weighted sum of the neurons' activities $x_j(t)$ with the connectivity matrix J. If a neuron's input h_i exceeds its threshold, it transitions to the active state; otherwise, it will be in the quiescent state $x_i(t) = 0$. Neuronal activation is thus described by the Heaviside function $H(\circ)$. This constitutes a well established simple model for brain cortical activity [70, 25, 26]. For propagating the state of the neurons, we use Glauber dynamics [71], selecting neurons for update according to a Poisson process with exponential interval distribution of mean τ.

Network parameters

The network model is minimal in the sense that it is described by only four network parameters: The process of updating neurons is governed by a Poisson process with time constant τ. The connectivity matrix is drawn randomly $J_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\bar{g}, 1/N)$, where the average strength of neuronal couplings is $\bar{g}$, the sign of which determines whether the network is excitation- or inhibition-dominated. Every neuron has an individual threshold, which the synaptic input has to surpass for the neuron to fire. Thresholds are distributed with mean δ and variability γ across neurons. The variance of the connectivity is fixed to $1/N$ without loss of generality; a different numerator would merely rescale the fields and could be absorbed by the other parameters. This is due to the Heaviside activation function having no inherent scale $H(ax) = H(x) \quad \forall\, a > 0$ [69].

Dynamical mean-field theory

A recurrent network model with binary units has been shown to be analytically tractable using dynamical mean-field theory [25, 24, 23]. We here extend this theory by a time-varying mean activity to account for stimulus-driven network transients. The problem can be described in a path-integral formalism using a probability distribution approach. We formulate a joint probability distribution for the states of neurons $\boldsymbol{x}$ and synaptic fields $\boldsymbol{h}$

$$\rho[\boldsymbol{x}, \boldsymbol{h}] = \rho[\boldsymbol{x}|\boldsymbol{h}]\, \rho[\boldsymbol{h}], \quad \rho[\boldsymbol{x}|\boldsymbol{h}] := \prod_{i=1}^{N} \rho[x_i|h_i], \tag{10}$$

where the latter factorization shows that only the single component h_i of the input field influences the state of the i-th neuron. By providing the relation from the input field back to the neuronal state $\rho[\boldsymbol{h}(t)|\boldsymbol{x}(t)] = \delta\,[\boldsymbol{h}(t) - J\boldsymbol{x}(t)]$ with the connectivity matrix J of elements $J_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\bar{g}/N, 1/N)$, these equations can be closed (see Supplement 2 for details of the derivation)

$$\begin{aligned}\rho[\boldsymbol{h}] &= \int \mathcal{D}\boldsymbol{x}\, \rho[\boldsymbol{x}|\boldsymbol{h}] \delta[\boldsymbol{h} - \boldsymbol{J}\boldsymbol{x}] \\ &= \int \mathcal{D}\boldsymbol{x} \int \mathcal{D}\hat{\boldsymbol{h}}\, \rho[\boldsymbol{x}|\boldsymbol{h}] \exp\big(\hat{\boldsymbol{h}}^{\mathrm{T}} \boldsymbol{h}\big) \exp\big(-\hat{\boldsymbol{h}}^{\mathrm{T}} \boldsymbol{J}\, \boldsymbol{x}\big).\end{aligned}$$

In the limit of many neurons we can average over all possible connectivity matrices J (disorder average), which amounts to mean-field theory. This step provides a description in terms of the macroscopic variables

$$\mathcal{R}(t) := \frac{1}{N}\sum_{j=1}^{N} x_j(t) \quad \text{and} \quad \mathcal{Q}(t,s) := \frac{1}{N}\sum_{j=1}^{N} x_j(t)x_j(s). \tag{11}$$

Enforcing the definitions in equation (11) by delta constraints give rise to (imaginary) auxiliary fields $\hat{\mathcal{R}}$ and $\hat{\mathcal{Q}}$ [23, Chapter 12], leaving us with the disorder averaged probability distribution of network states

$$\begin{aligned}&\int \mathcal{D}\boldsymbol{h}\,\big\langle \rho[\boldsymbol{h}]\big\rangle_{\boldsymbol{J}}\\ &= \int \mathcal{D}\{\mathcal{Q},\mathcal{R},\hat{\mathcal{Q}},\hat{\mathcal{R}}\}\exp\big(N\,\Omega[\mathcal{R},\mathcal{Q},\hat{\mathcal{R}},\hat{\mathcal{Q}}]\big),\end{aligned}$$

with the exponent (action)

$$\Omega[\mathcal{R},\mathcal{Q},\hat{\mathcal{R}},\hat{\mathcal{Q}}] := -\mathcal{R}^{\mathrm{T}}\hat{\mathcal{R}} - \mathcal{Q}^{\mathrm{T}}\hat{\mathcal{Q}} + \ln\int \mathcal{D}\{x,h,\hat{h}\}\,\rho[x|h]\exp\Big(\hat{h}^{\mathrm{T}}h - \bar{g}\hat{h}^{\mathrm{T}}\mathcal{R} + \frac{1}{2}\,\hat{h}^{\mathrm{T}}\mathcal{Q}\hat{h} + \hat{\mathcal{R}}^{\mathrm{T}}x + x^{\mathrm{T}}\hat{\mathcal{Q}}x\Big). \tag{12}$$

By a saddle-point approximation we get the most likely values $R, Q, \hat{R}, \hat{Q}$ of the fields $\mathcal{R}, \mathcal{Q}, \hat{\mathcal{R}}, \hat{\mathcal{Q}}$, which for the auxiliary fields are zero, $\hat{R} = \hat{Q} = 0$. In addition, we obtain self consistent equations for the fields R, Q. For more details, see Supplement 2.1.

Two-replica approach

To study how a network responds to two different stimuli, we consider two instances of the network (replicas) that share their connectivity, resembling two trials from the experiment. Reformulating the probability distribution for these two networks $\rho[\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)}|\boldsymbol{h}^{(1)},\boldsymbol{h}^{(2)}] = \prod_{i=1}^{N}\rho[x_i^{(1)},x_i^{(2)}|h_i^{(1)},h_i^{(2)}]$ and in expectation over the connectivity (disorder average) gives rise to a term that couples the replicas

$$\begin{aligned}&\Big\langle \exp\big(-\hat{\boldsymbol{h}}^{(1)\mathrm{T}}\boldsymbol{J}\boldsymbol{x}^{(1)} - \hat{\boldsymbol{h}}^{(2)\mathrm{T}}\boldsymbol{J}\boldsymbol{x}^{(2)}\big)\Big\rangle_{\boldsymbol{J}\overset{\text{i.i.d.}}{\sim}\mathcal{N}(\frac{\bar{g}}{N},\frac{1}{N})}\\ &= \prod_{i=1}^{N}\prod_{\alpha=1}^{2}\exp\Big(-\frac{\bar{g}}{N}\,\hat{h}_i^{(\alpha)\mathrm{T}}\sum_{j=1}^{N}x_j^{(\alpha)} + \frac{1}{2}\sum_{j=1}^{N}\big(\hat{h}_i^{(\alpha)\mathrm{T}}x_j^{(\alpha)}\big)^2\Big)\\ &\qquad\times\exp\Big(\frac{1}{N}\sum_{j=1}^{N}\hat{h}_i^{(1)\mathrm{T}}x_j^{(1)}\,\hat{h}_i^{(2)\mathrm{T}}x_j^{(2)}\Big).\end{aligned} \tag{13}$$

The coupling term requires the introduction of an additional field $\mathcal{Q}^{(12)}$ and this definition is again enforced by a delta constraint yielding one additional auxiliary field. This motivates the notation $\mathcal{Q}^{(\alpha\beta)}(t,s) := \frac{1}{N}\sum_{j=1}^{N}x_j^{(\alpha)}(t)\,x_j^{(\beta)}(s)$. We again take the saddle-point approximation giving rise to self consistent equations for these fields, see Supplement 2.2.

Addition of thresholds

Thresholds are accounted for by a static contribution of the neuronal field $\boldsymbol{h}^{(\alpha)} \to \boldsymbol{h}^{(\alpha)} + \boldsymbol{h}_{\text{stat.}}$, where $h_{\text{stat},i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\delta,\gamma)$. This still allows for the factorization in equation (12) after averaging over the disorder in terms of connectivity and $\boldsymbol{h}_{\text{stat.}}$.

Time dynamics

The update times follow a Poisson process, so that the intervals Δt between update times follow an exponential distribution $p(\Delta t = t) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$ with $t \geq 0$. If a neuron is selected, it can update its state dependent on the input fields $\boldsymbol{h}(t)$. The probability for a neuron i to be active at some time t only depends on the field $h_i^{(\alpha)}$ of the replicon α it belongs to and is influenced by the probability to be active at a prior time point t'

$$\rho[x_i^{(\alpha)}(t) = 1 | h_i^{(\alpha)}] = \int_{-\infty}^{t} \frac{dt'}{\tau} e^{-\frac{t-t'}{\tau}} H\left(h_i^{(\alpha)}(t')\right). \tag{14}$$

The Markov property renders the time evolution equations local in time. Inserting the expression in equation (14) for the neuronal input-to-output relationship into the self-consistency equations for $\mathcal{R}, \mathcal{Q}$, we arrive at equations (1), (2) with the correlation transmission function $\mathcal{T}(\circ, \circ)$ (see Supplement 2.3 for details). The time evolution equations for the solutions $R(t)$ and $Q(t,t)$ form a closed set of coupled equations. The Heaviside activation function H enters in the calculation of $\mathcal{T}(\circ, \circ)$ as

$$\mathcal{T}(R, Q) = \left\langle H(h^{(1)}) H(h^{(2)})\}\right\rangle_{(h^{(1)}, h^{(2)}) \sim \mathcal{N}(\{R, Q\})}$$

with

$$\mathcal{N}(\{R, Q\}) = \mathcal{N}\left(\begin{pmatrix} \bar{g}R + \delta \\ \bar{g}R + \delta \end{pmatrix}, \begin{pmatrix} R + \gamma & Q + \gamma \\ Q + \gamma & R + \gamma \end{pmatrix}\right)$$

and specifically

$$\mathcal{T}(R, R) = \left\langle H(h)\right\rangle_{h \sim \mathcal{N}(\bar{g}R + \delta,\, R + \gamma)}.$$

The derivative of $\mathcal{T}(R, R)$ determines the speed of the decay of R up to linear order. To achieve a slowly or quickly decaying population mean in Section 1.3 we choose parameters such that the derivative at the baseline R_0 is $\frac{d}{dR}\mathcal{T}(R, R)\big|_{R=R_0} = 1$ or -1, respectively. Additional details are given in Supplement 3.5. For a list of all network parameters, see Table 1.

3.3 Determining network parameters from experimental data

The right-hand sides of the time evolution equations (1),(2) only depend on certain constellations of observables and parameters. By inferring the experimental values of these constellations and constraining the parameters $\bar{g}, \delta, \gamma$ accordingly, we can ensure that the mean-field theory matches the experimental baselines. This leaves us with parameters τ, δ, with which we can fit the decay by minimizing a weighted mean-squared error

$$\mathcal{L}(\tau, \delta) = \sum_{t \in \{t_{\text{decay}}\}} \left[\frac{(R(t) - R(t))^2}{\sigma_R^2(t)} + \frac{\left(Q^{\circlearrowleft}(t) - Q^{\circlearrowleft,\text{exp}}(t)\right)^2}{\sigma^2_{Q^{\circlearrowleft},\text{exp}}(t)} + \frac{(Q^{\leftrightarrow}(t) - Q^{\leftrightarrow,\text{exp}}(t))^2}{\sigma^2_{Q^{\leftrightarrow},\text{exp}}(t)} \right].$$

The individual error terms are weighted inversely proportional to the accuracy of the experimental measurement, estimated by the variance over trials. More details on the fitting procedure in Supplement 3.

Goodness of fit

For the experimental observables (expm) and the model fit (model_fit) of $R, Q^{\circlearrowleft}, Q^{\leftrightarrow}$, we perform a smoothing-window approach with a window size of 3 bins, i.e. $t_{\text{window}} = 3 \cdot 6\text{ms} = 18\text{ms}$. We calculate the sliding window average

$$\bar{\text{expm}}(t) = \langle \text{expm}(t') \rangle_{t'=t\pm t_{\text{window}}}$$

and variance

$$\text{Var}(t) = \left\langle \left(\text{expm}(t') - \bar{\text{expm}}(t')\right)^2 \right\rangle_{t'=t\pm t_{\text{window}}} .$$

The goodness of fit is then evaluated using the coefficient of determination

$$\text{R}^2 = 1 - \left\langle \frac{\left(\bar{\text{expm}}(t) - \text{model_fit}(t)\right)^2}{\text{Var}(t)} \right\rangle_t .$$

This leaves us with three coefficients of determination, evaluating the goodness of fit for each of the three observables separately (see Figure 2e).

3.4 Separability

We quantify separability using a linear readout model $y(t) = \boldsymbol{w}^{\text{T}}\boldsymbol{x}(t) + \epsilon$ with $\boldsymbol{w} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, g_w^2/N\right)$ and $\epsilon \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_n^2\right)$. After conditioning on training data $\boldsymbol{X} \in \mathbb{R}^{N\times 2N_{\text{train}}}$ with labels $\boldsymbol{y} \in \{-1, 1\}^{2N_{\text{train}}}$ in a Bayesian manner, we can obtain the predictive statistics for the test data $\boldsymbol{X}^* \in \mathbb{R}^{N\times 2N_{\text{test}}}$. We obtain the posterior distribution for the readout weight $\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}$ through

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})\, p(\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{X})}$$

and the posterior distribution $y^*|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}$ for some test vector $\boldsymbol{x}^* \in \mathbb{R}^N$ by using $p(y^*|\boldsymbol{x}^*, \boldsymbol{y}, \boldsymbol{X}) = \int d\boldsymbol{w}\, p(y^*|\boldsymbol{x}^*, \boldsymbol{w})\, p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})$, which follows a Gaussian process [33]. The two cumulants of this Gaussian process can be calculated using the kernel: The kernel matrix directly follows as the dot-product between two neural state vectors $K(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) = \frac{g_w}{N}\boldsymbol{x}_\alpha^{\text{T}}\boldsymbol{x}_\beta$. In the limit of large numbers of neurons the entries of this kernel concentrate around the three distinct values

$$\begin{aligned}
&K(\boldsymbol{x}_\alpha, \boldsymbol{x}_\alpha) \to R\,, \\
&K(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) \to Q^{\circlearrowleft} \quad \text{for} \quad C(\alpha) = C(\beta)\,, \\
&K(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) \to Q^{\leftrightarrow} \quad \text{for} \quad C(\alpha) \neq C(\beta)\,,
\end{aligned}$$

which can be expressed by the three observables. Here, $C(\circ)$ denotes the class membership of the sample. When sorting the trials by class-membership the block structure becomes evident. This block-structure of the kernel allows for the analytical evaluation of the predictive mean in equation (3) (more details in Supplement 4). It depends on the three observables $R, Q^{\circlearrowleft}, Q^{\leftrightarrow}$ in addition to the noise $\kappa = \sigma_n^2/g_w^2$ and the number of training points N_{train}.

Time evolution of separability

For the time evolution of the separability, the noise $\kappa + R - Q^{\circlearrowleft}$ and the signal $Q^{\circlearrowleft} - Q^{\leftrightarrow}$ entering the formula (3) are of particular importance. The relative change of the total variability and the signal are governed by the terms s_1 and s_2, respectively. This can be seen by rewriting equations (1) and (2) as

$$\begin{aligned}
\tau \frac{d}{dt} \ln\left(\kappa + R - Q^{\circlearrowleft}\right) &= -1 + s_1\,, \\
\tau \frac{d}{dt} \ln\left(Q^{\circlearrowleft} - Q^{\leftrightarrow}\right) &= -1 + s_2\,.
\end{aligned} \tag{15}$$

A large readout noise attenuates changes of the total noise $\tau \frac{d}{dt} \ln\left(\kappa + R - Q^{\circlearrowleft}\right) \xrightarrow{\kappa \to \infty} 0$, facilitating a transient increase in separability for $s_2 > 1$. For more details, see Supplement 4.1.1.

3.5 Information transmission

We take P linear readouts $y_p = \boldsymbol{w}_p^{\mathrm{T}} \boldsymbol{x} + \epsilon$ with one weight vector $\boldsymbol{w}_p$ for every stimulus class $p = 1, \ldots, P$. The p-th readout is obtained through Bayesian inference to produce the highest output when applying the p-th stimulus. The labels are one-hot encoded. Prior weights are as before chosen according to $w_{pi} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g_w^2/N)$. Because prior and likelihood factorize over p [72, Chapter 11, (11.2)], the posterior covariance is also diagonal between readouts p. The Bayesian posterior quantifies the conditional probability distribution of responses $p(\boldsymbol{y}^* | C)$, where $\boldsymbol{y}^* = (y_1^*, \ldots, y_P^*)^T$, which is necessary for evaluating the mutual information. Together with the probability distribution across stimuli $p(C) = 1/P$, the mutual information is fully determined. Since the mutual information between the stimuli and responses depends only on the ratio $\Delta = \frac{\delta\mu}{\sqrt{\Sigma}}$ it suffices to consider the difference of the mean readouts. The latter is given by equation (3). The overlap within classes $Q^{\circlearrowleft}$ is kept constant across the different numbers of stimuli, i.e. the extent of each stimulus is kept the same. This assumption is based on experimental observations.

The mutual information of the set of P readouts on the class label of the stimulus can then be written as (see Supplement 5.2)

$$\mathrm{MI}(P, \Delta) = \log_2(e)\Delta^2 - \int dy_1^* \cdots \int dy_P^* \, \mathcal{N}(y_1^* | \Delta, 1) \left[\prod_{C=2}^{P} \mathcal{N}(y_C^* | 0, 1) \right] \log_2 \left(\frac{1}{P} \sum_{C=1}^{P} \exp\left(\Delta \, y_C^*\right) \right), \tag{16}$$

where $\Delta = \Delta(P)$ is still a function of P through equation (7). We have to evaluate one additional integral for every additional readout. This is very expensive to compute numerically using Monte Carlo techniques, even after exploiting the symmetries of equation (16), and may even be intractable. We therefore rewrite the formula as

$$\mathrm{MI}(P, \Delta) \approx \log_2(e) \left(\Delta(P)\right)^2 - \log_2(e) \, m(\Delta(P), P)$$

with the two-dimensional Gaussian integral

$$m(\Delta(P), P) = \int dy_1^* \int dY \, \mathcal{N}(y_1^* | \Delta, 1) \, \mathcal{N}(Y | \mu_S, \Sigma_S) \log_2 \left(\frac{1}{P} \exp\left(\Delta \, y_1^*\right) + \frac{P-1}{P} \exp\left(Y\right) \right).$$

We here approximated the sum of identically and independently distributed lognormal variables appearing in equation (16) as a lognormally distributed variable

$$\exp\left(Y\right) = \frac{1}{P-1} \sum_{C=2}^{P} \exp\left(\Delta \, y_C^*\right), \tag{17}$$

with Y drawn from the Gaussian distribution $\mathcal{N}\left(\mu_Y, \Sigma_Y\right)$. The parameters μ_Y, Σ_Y for the distribution of Y are obtained using the Yeh-Schwartz approximation [73]. We empirically validate this approximation in Supplement 5.2.

Saturation of mutual information

For large P, the mean of the term in the logarithm of equation (16) evaluates to

$$\left\langle \frac{1}{P} \sum_{C=1}^{P} \exp\left(\Delta \, y_C^*\right) \right\rangle_{y_1^*, \ldots, y_P^*} \xrightarrow{P \to \infty} \exp\left(\frac{1}{2} \left(\Delta(\infty)\right)^2 \right),$$

which leads to equation (9). Because this term has the particular form of $\frac{1}{P}\sum_{C=1}^{P} z_C$ with z_C drawn i.i.d from regular probability distributions, the k-th cumulant goes as $\propto P^{-k+1}$, such that the distribution concentrates around this mean for large P. Note that lognormal variables with heavy tails (large Δ) only concentrate slowly as P increases.

Optimal overlap between classes

We here calculate the optimal overlap $Q^{\leftrightarrow}$ between different stimulus representations in equation (5) for the case that stimuli are distributed efficiently in the available neuronal space. The assumptions entering the derivation of this overlap are informed by both the data and network model. We observe that stimuli evoke similarly high responses with similar population firing rates and similar extents. Therefore, we keep R and $Q^{\circlearrowleft}$ the same for all stimulus classes. Network responses are captured well by the mean-field theory, in which cross-correlations between neurons vanish. Therefore, we assume that neurons i fire independently from each other as $x_{\alpha i} \overset{\text{i.i.d over i}}{\sim}$ Bernoulli (R_i), but coordinated across trials.

In the following, we illustrate the derivation for the special case $R_i = R$, the general case is detailed in Supplement 5.4. We would like to know the probability $p\,(Q|R)$ for an overlap Q, given all stimuli have the same mean R. For binary vectors $x_{\alpha i} \overset{\text{i.i.d}}{\sim}$ Bernoulli (R) the probability distribution concentrates around the random overlap $\frac{1}{N}\sum_{i=1}^{N} R_i^2$. To quantify "more efficient than random" we need to estimate the width of the probability distribution $p\,(Q|R)$. The problem comes down to calculating the overlaps of binary vectors with the same length.

We would like to determine the number of vectors $\boldsymbol{x}_\alpha \in \{0,1\}^N$ with the following properties:

C1. The (normalized) pairwise overlap between any pair of vectors should be $Q^{\leftrightarrow} = N^{-1}\sum_{i=1}^{N} x_{\alpha i}\, x_{\beta i} \quad \forall \alpha \neq \beta$.

C2. The (normalized) number of ones for each neuron should be $R = \frac{1}{N}\sum_{i=1}^{N} x_{\alpha i} \quad \forall \alpha$.

We construct the counting function

$$Z(C,j) := \sum_{\{x_{\alpha i}\}\in\{0,1\}^{P\times N}} \left\{\prod_{\alpha=1}^{P} \delta\left[j_\alpha; N^{-1}\sum_{i=1}^{N} x_{\alpha i}\right]\right\}\left\{\prod_{\alpha\neq\beta=1}^{P} \delta\left[C_{\alpha\beta}; N^{-1}\sum_{i=1}^{N} x_{\alpha i}x_{\beta i}\right]\right\}, \tag{18}$$

where $\delta[\circ;\circ]$ is the Kronecker δ. The j_α tunes the firing rates of the experimentally observed neurons and the matrix C tunes the mutual overlaps between vectors. The first term in the curly braces enforces condition C1 and the second term condition C2, respectively. In Supplement 5.4 it is detailed how the form of the counting function changes when additionally enforcing the firing rate distribution $\{R_i\}_i$. By expressing the Kronecker delta with help of the Dirac delta $\delta[a;b] = \lim_{\epsilon\searrow 0}\int_{-\epsilon}^{\epsilon} \delta(a-b-\iota)\, d\iota$ and expressing the latter by its Fourier representation $\delta(x) = \int_{-i\infty}^{i\infty} \frac{d\tilde{x}}{2\pi i} e^{\tilde{x}x}$, one has

$$\begin{aligned} Z(C,j) &= \int d^{P\times P}\iota \int \mathcal{D}\tilde{j} \int \mathcal{D}\tilde{C} \exp\Big(-\sum_{\alpha=1}^{P} \tilde{j}_\alpha(j_\alpha+\iota_\alpha) - \sum_{\alpha\neq\beta=1}^{P} \tilde{C}_{\alpha\beta}(C_{\alpha\beta}+\iota_{\alpha\beta}) + \mathcal{W}(\tilde{C})\Big), \\ \mathcal{W}(\tilde{C},\tilde{j}) &= N \ln \sum_{\{x_\alpha\}\in\{0,1\}^P} \exp\Big(\sum_{\alpha=1}^{P} \frac{\tilde{j}_\alpha}{N} x_\alpha + \sum_{\alpha\neq\beta=1}^{P} \frac{\tilde{C}_{\alpha\beta}}{N} x_\alpha x_\beta\Big) \end{aligned} \tag{19}$$

with $\int \mathcal{D}\tilde{j} = \prod_{\alpha=1}^{P}\int \frac{d\tilde{j}_\alpha}{2\pi i}$, $\int \mathcal{D}\tilde{C} = \prod_{\alpha\neq\beta=1}^{P}\int \frac{d\tilde{C}_{\alpha\beta}}{2\pi i}$ and the cumulant-generating function $\mathcal{W}$. The form of $\mathcal{W}$ is that of the cumulant-generating function (free energy) of the Ising model [74].

For large numbers of neurons $N \to \infty$, we can use the Gaertner-Ellis theorem [75] (a large-deviation principle, l.d.p.) to approximate

$$Z(C,j) \overset{\text{l.d.p}}{\simeq} \int d^{P\times P}\iota \exp\big(-\Gamma(C+\iota,\, j+\iota)\big),$$
$$\Gamma(C,j) := \sup_{\tilde{C},\tilde{j}} \left\{ \sum_{\alpha=1}^{P} \tilde{j}_\alpha j_\alpha + \sum_{\alpha\neq\beta=1}^{P} \tilde{C}_{\alpha\beta} C_{\alpha\beta} - \mathcal{W}(\tilde{C},\tilde{j}) \right\}.$$

In this sense $Z(C,j)$ becomes proportional to the normalized probability distribution $p(C,j)$. Using the two appearing Legendre transformations from the cumulant-generating function $\mathcal{W}$ to the effective action Γ, we can enforce conditions C1 and C2, obtaining $p(Q,R) \propto \exp\left(-\Gamma(Q,R)\right)$. For the Legendre transformation in $\tilde{C}_{\alpha\beta}$ we expand up to second order (TAP-approximation) around this non-Gaussian theory [76]. The validity of this approximation is confirmed in Supplement 5.4. Lastly, in splitting $p(Q,R) = p(R)p(Q|R)$ we obtain the Gaussian distribution $p(Q|R) = \mathcal{N}\left(Q|R^2, \frac{2}{N\,P(P-1)}R(1-R)\right)$. Demanding $E^{-1} \overset{!}{=} \int_0^{Q^{\leftrightarrow}} dQ\, p\left(Q|R\right)$ for some efficiency E to realize an overlap for a certain number of possible solutions leads to the equation (5) (for the case of additionally preserving the firing rate distribution). In the main text, $\mathcal{E} = -\mathrm{erf}^{-1}\left(2E^{-1}-1\right)$.

3.6 Robustness of the vanishing regime

In Section 1.5, we encountered the *vanishing regime*, where the mutual information asymptotically approaches zero as the number of stimuli increases. One assumption is that increasing the population mean preserves the shape of the firing rate distribution, scaling it by some factor to achieve the higher population mean. Here, we show that the vanishing regime is robust to modifications of the firing rate distribution *other than scaling*. These modifications arise from different regimes of the transfer function $\phi(h)$, which defines the relationship between input current h and firing rate $R = \phi(h)$. More details can be found in Supplement 5.5.

When examining the effect of changing the population mean R, a relationship between the overlaps $Q^{\circlearrowleft}$ and $Q^{\leftrightarrow}$ must be specified. While the between-classes overlap $Q^{\leftrightarrow}$ scales according to equation (5), for the within-class overlaps $Q^{\circlearrowleft}$, we assume that the selectivity of neurons across different classes ρ_i remains constant as a function of the population mean R. This class selectivity for each neuron i is captured by the Pearson correlation within classes, $C(\alpha) = C(\beta)$

$$\rho_i = \frac{\mathrm{cov}(x_{\alpha i} x_{\beta i})_{(\alpha\beta)}}{\sqrt{\mathrm{var}(x_{\alpha i})_\alpha\, \mathrm{var}(x_{\alpha i})_\beta}}.$$

We determine ρ_i from the experiment. To calculate the population mean R^*, for which the mutual information vanishes, we can equate within-class overlap to between-classes overlap,

$$\begin{aligned} & Q^{\circlearrowleft} = Q^{\leftrightarrow} \\ \Rightarrow \quad & \frac{1}{N}\sum_{i=1}^{N} \rho_i R_i = \frac{1}{N}\sum_{i=1}^{N} \rho_i R_i^2. \end{aligned} \tag{20}$$

Notably, the resulting population mean marking the transition given is invariant under the rescaling $\rho_i \to \mathrm{const}\,\rho_i$, such that the lines for constant ρ_i, $\rho_i^{\mathrm{small}} = 0.2\,\rho_i$, and $\rho_i^{\mathrm{large}} = 2.5\,\rho_i$ (black and dashed lines) show the same point of transition R^* in Figure 7a.

Exponential transfer function

If the population mean is low, the neural transfer function $R = \phi(h)$ has an exponential shape [35], leading to the scaling of the firing rates by a common factor when the inputs obtain an additive term $h \to h + \mathrm{const.} \Rightarrow \exp(h) \to \exp(\mathrm{const.})\exp(h)$.

If the degree of class-selectivity is similar for all neurons, $\rho_i \approx \rho$, the firing rate R^* beyond which the vanishing regime sets in takes the form

$$R^* = \frac{\left(\frac{1}{N}\sum_{i=1}^{N} R_i^{\exp}\right)^2}{\frac{1}{N}\sum_{i=1}^{N}\left(R_i^{\exp}\right)^2} \,. \tag{21}$$

It decreases the larger the random overlap $\frac{1}{N}\sum_{i=1}^{N}\left(R_i^{\exp}\right)^2$ (denominator) is compared to the squared population mean (numerator). Thus, the larger the heterogeneity in firing rates, the smaller the population mean has to be for the vanishing regime.

Linear transfer function

Neurons at sufficiently high firing rates typically show an affine linear dependence $R = \phi(h)$ [35, 77]. An additive term to their input then results in an additive term to every firing rate $h \to h + \mathrm{const.} \Rightarrow \phi(h) \to \phi(h) + \mathrm{const.}$. This transformation does not affect the variance $\langle\!\langle R_i^2 \rangle\!\rangle_i$. By rewriting equation (20) as $R(1-R) = \langle\!\langle R_i^2 \rangle\!\rangle_i$ we obtain a solution for the rate R^* to transition into the vanishing regime that is larger than the R^* in (21). In conclusion, for a transfer function with an exponential shape at low rates and a linear part at higher rates, the transition into the vanishing regime lies between the two values of R^* calculated for the exponential and linear case, respectively. Thus the transition to the vanishing regime persists.

Experimental observables	left side	right side	
$R_{\max}$	6.63	12.05	Initial population mean $[\mathrm{Hz}]$
R_0	3.25	8.86	Population mean baseline $[\mathrm{Hz}]$
$Q^{\circlearrowleft}_{\max}$	3.08	2.53	Initial variability within stimulus class $[\mathrm{Hz}^2]$
$Q^{\leftrightarrow}_{\max}$	0.52	1.22	Initial variability across stimulus classes $[\mathrm{Hz}^2]$
Q_0	0.16	0.98	Variability baseline $[\mathrm{Hz}^2]$
N	141	65	number of recorded neurons

Inferred network parameters			
τ	29.29	25.65	time-constant $[\mathrm{ms}]$
$\bar{g}$	-4.47	-1.69	average coupling strength (inhibition)
δ	-0.22	-0.30	average threshold
γ	0.0031	0.0062	variability of thresholds
T	62	68	stimulus switch-off $[\mathrm{ms}]$

Bayesian inference parameters			
N_{trials}	50	40	
N_{train}	35	28	number of training samples/trials
N_{test}	15	12	number of test samples/trials
κ	$0.1\,R_{\max}$	$0.1\,R_{\max}$	readout noise
g_w/N	$1/N$	$1/N$	variance of prior
	100	100	number of seeds for sampling training, test data

Simulation Parameters		
N_{trials}	50	
$\mathrm{d}t$	0.6	Simulation step size $[\mathrm{ms}]$
N_{model}	35250	Number of neurons

Table 1: **Network parameters.** Left/right side refers to the side of stimulus application for the visual and tactile stimuli on the mouse. The experimental observable R is averaged over all trials, and the overlaps Q are averaged over pairs of trials. The overlap within classes $Q^{\circlearrowleft}$ is computed by averaging over all pairs of trials with the same stimulus class (both visual or both tactile).

Code and data availability

The datasets generated during and analyzed during the current study are available from the corresponding author on reasonable request.

References

[1] Georgopoulos, A. P., Schwartz, A. B. & Kettner, R. E. Neuronal population coding of movement direction. *Science* **233**, 1416–1419 (1986).

[2] Buonomano, D. V. & Maass, W. State-dependent computations: spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.* **10**, 113–125 (2009).

[3] Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17**, 401 (2013).

[4] Saxena, S. & Cunningham, J. P. Towards the neural population doctrine. *Curr. Opin. Neurobiol.* **55**, 103–111 (2019).

[5] Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation through neural population dynamics. *Annu. Rev. Neurosci.* **43**, 249–275 (2020).

[6] Chung, S., Lee, D. D. & Sompolinsky, H. Linear readout of object manifolds. *Phys. Rev. E* **93**, 060301 (2016).

[7] Chung, S., Lee, D. D. & Sompolinsky, H. Classification and geometry of general perceptual manifolds. *Phys. Rev. X* **8**, 031003 (2018).

[8] Cohen, U., Chung, S., Lee, D. D. & Sompolinsky, H. Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* **11**, 746 (2020).

[9] Bernardi, S. *et al.* The geometry of abstraction in hippocampus and prefrontal cortex. *Cell* **183**, 954 (2020).

[10] Fascianelli, V. *et al.* Neural representational geometries reflect behavioral differences in monkeys and recurrent neural networks. *Nat. Commun.* **15**, 6479 (2024).

[11] Gallego, J. A., Perich, M. G., Miller, L. E. & Solla, S. A. Neural manifolds for the control of movement. *Neuron* **94**, 978–984 (2017).

[12] Mastrogiuseppe, F. & Ostojic, S. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron* **99**, 609–623 (2018).

[13] Kietzmann, T. C. *et al.* Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. USA* **116**, 21854–21863 (2019).

[14] Wang, L., McAlonan, K., Goldstein, S., Gerfen, C. R. & Krauzlis, R. J. A causal role for mouse superior colliculus in visual perceptual decision-making. *J. Neurosci.* **40**, 3768–3782 (2020).

[15] Ito, S. & Feldheim, D. A. The mouse superior colliculus: An emerging model for studying circuit formation and function. *Front. Neural Circuits* **12** (2018).

[16] Meredith, M. A. & Stein, B. E. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J. Neurophysiol.* **56**, 640–662 (1986).

[17] Benavidez, N. L. *et al.* Organization of the inputs and outputs of the mouse superior colliculus. *Nat. Commun.* **12**, 4004 (2021).

[18] White, B. J. & Munoz, D. P. The superior colliculus. In *The Oxford Handbook of Eye Movements* (Oxford University Press, 2011).

[19] Arieli, A., Sterkin, A., Grinvald, A. & Aertsen, A. Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science* **273**, 1868–1871 (1996).

[20] Churchland, M. M. *et al.* Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat. Neurosci.* **13**, 369–378 (2010).

[21] Litwin-Kumar, A. & Doiron, B. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nat. Neurosci.* **15**, 1498–1505 (2012).

[22] Mazzucato, L., Fontanini, A. & La Camera, G. Stimuli reduce the dimensionality of cortical activity. *Front. Syst. Neurosci.* **10**, 11 (2016).

[23] Helias, M. & Dahmen, D. *Statistical Field Theory for Neural Networks* (Springer International Publishing, 2020).

[24] Keup, C., Kühn, T., Dahmen, D. & Helias, M. Transient chaotic dimensionality expansion by recurrent networks. *Phys. Rev. X* **11**, 021064 (2021).

[25] van Vreeswijk, C. & Sompolinsky, H. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* **274**, 1724–1726 (1996).

[26] Renart, A. *et al.* The asynchronous state in cortical circuits. *Science* **327**, 587–590 (2010).

[27] Tetzlaff, T., Helias, M., Einevoll, G. T. & Diesmann, M. Decorrelation of neural-network activity by inhibitory feedback. *PLOS Comput. Biol.* **8**, e1002596 (2012).

[28] Monteforte, M. & Wolf, F. Dynamic flux tubes form reservoirs of stability in neuronal circuits. *Phys. Rev. X* **2**, 041007 (2012).

[29] Gardner, E. The space of interactions in neural network models. *J. Phys. A Math. Gen.* **21**, 257 (1988).

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

[31] van Vreeswijk, C. & Sompolinsky, H. Chaotic balanced state in a model of cortical circuits. *Neural Comput.* **10**, 1321–1371 (1998).

[32] Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J. & Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 29* (2016).

[33] Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning (MIT Press, Cambridge, Mass, 2006).

[34] Lee, J. *et al.* Deep neural networks as Gaussian processes. In *International Conference on Learning Representations* (OpenReview.net, Vancouver, 2018).

[35] Roxin, A., Brunel, N., Hansel, D., Mongillo, G. & van Vreeswijk, C. On the distribution of firing rates in networks of cortical neurons. *J. Neurosci.* **31**, 16217–16226 (2011).

[36] Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51 (2012).

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

[38] London, M., Roth, A., Beeren, L., Häusser, M. & Latham, P. E. Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature* **466**, 123–128 (2010).

[39] Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F. & Ostojic, S. The role of population structure in computations through neural dynamics. *Nat. Neurosci.* **25**, 783–794 (2022).

[40] Clark, D. G., Marschall, O., van Meegen, A. & Litwin-Kumar, A. Connectivity structure and dynamics of nonlinear recurrent neural networks (2024).

[41] Schuessler, F., Mastrogiuseppe, F., Ostojic, S. & Barak, O. Aligned and oblique dynamics in recurrent neural networks. *eLife* **13**, RP93060 (2024).

[42] Lindner, J., Dahmen, D., Krämer, M. & Helias, M. A theory of data variability in neural network Bayesian inference. *arXiv preprint arXiv:2307.16695* (2023).

[43] Chizat, L., Oyallon, E. & Bach, F. On lazy training in differentiable programming. In *Adv. Neural Inf. Process. Syst.*, vol. 32 (2019).

[44] Yang, G. Wide feedforward or recurrent neural networks of any architecture are Gaussian processes. In Wallach, H. *et al.* (eds.) *Adv. Neural Inf. Process. Syst.*, vol. 32 (Curran Associates, Inc., 2019).

[45] Lee, J. *et al.* Wide neural networks of any depth evolve as linear models under gradient descent. *J. Stat. Mech. Theory Exp.* **2020**, 124002 (2020).

[46] Seroussi, I., Naveh, G. & Ringel, Z. Separation of scales and a thermodynamic description of feature learning in some cnns. *Nat. Commun.* **14**, 908 (2023).

[47] Fischer, K. *et al.* Critical feature learning in deep neural networks. *arXiv preprint arXiv:2405.10761* (2024).

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

[49] Clark, D. G. & Abbott, L. F. Theory of coupled neuronal-synaptic dynamics. *Phys. Rev. X* **14**, 021001 (2024).

[50] Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nat. Neurosci.* **22**, 1677–1686 (2019).

[51] Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, eaav7893 (2019).

[52] Wang, Z. A. *et al.* Not everything, not everywhere, not all at once: a study of brain-wide encoding of movement. *bioRxiv* (2023).

[53] Mandelblat-Cerf, Y., Paz, R. & Vaadia, E. Trial-to-trial variability of single cells in motor cortices is dynamically modified during visuomotor adaptation. *J. Neurosci.* **29**, 15053–15062 (2009).

[54] Barlow, H. B. *Possible principles underlying the transformations of sensory messages*, 217–234 (MIT Press, 1961).

[55] Linsker, R. Perceptual neural organization: some approaches based on network models and information theory. *Annu. Rev. Neurosci.* **13**, 257–281 (1990).

[56] Atick, J. J. & Redlich, A. N. What does the retina know about natural scenes? *Neural Comput.* **4**, 196–210 (1992).

[57] Barlow, H. Redundancy reduction revisited. *Netw. Comput. Neural Syst.* **12**, 241–253 (2001).

[58] Palm, G. Neural associative memories and sparse coding. *Neural Netw.* **37**, 165–171 (2013).

[59] Ganguli, S. & Sompolinsky, H. Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annu. Rev. Neurosci.* **35**, 485–508 (2012).

[60] Lennie, P. The cost of cortical computation. *Curr. Biol.* **13**, 493–497 (2003).

[61] Balla, E. *et al.* Broadband visual stimuli improve neuronal representation and sensory perception. *Nat. Commun.* **16**, 2957 (2025).

[62] Putzeys, J. *et al.* Neuropixels data-acquisition system: A scalable platform for parallel recording of 10 000+ electrophysiological signals. *IEEE Trans. Biomed. Circuits Syst.* **13**, 1635–1644 (2019).

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

[64] Pachitariu, M., Sridhar, S., Pennington, J. & Stringer, C. Spike sorting with Kilosort4. *Nat. Methods* **21**, 914–921 (2024).

[65] McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in neural nets. *Bull. Math. Biol.* **5**, 115–137 (1943).

[66] Hebb, D. O. *The organization of behavior: A neuropsychological theory* (John Wiley & Sons, New York, 1949).

[67] Amit, D. J. & Brunel, N. Dynamics of a recurrent network of spiking neurons before and following learning. *Netw. Comput. Neural Syst.* **8**, 373–404 (1997).

[68] Helias, M., Tetzlaff, T. & Diesmann, M. The correlation structure of local cortical networks intrinsically results from recurrent dynamics. *PLOS Comput. Biol.* **10**, e1003428 (2014).

[69] Dahmen, D., Bos, H. & Helias, M. Correlated fluctuations in strongly coupled binary networks beyond equilibrium. *Phys. Rev. X* **6**, 031024 (2016).

[70] Ginzburg, I. & Sompolinsky, H. Theory of correlations in stochastic neural networks. *Phys. Rev. E* **50**, 3171–3191 (1994).

[71] Glauber, R. Time-dependent statistics of the Ising model. *J. Math. Phys.* **4**, 294–307 (1963).

[72] Murphy, K. P. *Probabilistic Machine Learning: An introduction* (MIT Press, 2022).

[73] Schwartz, S. C. & Yeh, Y. S. On the distribution function and moments of power sums with log-normal components. *Bell Syst. Tech. J.* **61**, 1441–1462 (1982).

[74] Goldenfeld, N. *Lectures on phase transitions and the renormalization group* (Perseus books, Reading, Massachusetts, 1992).

[75] Touchette, H. The large deviation approach to statistical mechanics. *Phys. Rep.* **478**, 1–69 (2009).

[76] Kühn, T. & Helias, M. Expansion of the effective action around non-Gaussian theories. *J. Phys. A Math. Theor.* **51**, 375004 (2018).

[77] Tuckwell, H. C. *Introduction to Theoretical Neurobiology*, vol. 2 (Cambridge University Press, Cambridge, 1988).

Acknowledgments

We thank Tobias Kühn, Inés Samengo, Severin Graff and Thomas Rüland for helpful discussions. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the SPP 2205 – 533396241 and the Gatsby Charitable Foundation. Open access publication funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 491111487. The authors gratefully acknowledge the computing time granted by the JARA Vergabegremium and provided on the JARA Partition part of the supercomputer JURECA at Forschungszentrum Jülich (computation grant JINB33).

Author contributions

MH, DD, SM designed the study.
SM designed the experiments.
LS, JB, PB, DD, MH developed the analysis methods and mathematical models.
LS analyzed the data and performed model simulations and the analytical calculations for the network model.
All authors discussed the findings and wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.