Emergent perceptual biases from state-space geometry in spiking recurrent neural networks trained to discriminate time intervals

Luis Serrano-Fernández[a,b], Manuel Beirán[c], and Néstor Parga [*,a,b]

[a]Departamento de Física Teórica, Universidad Autónoma de Madrid, 28049 Madrid, Spain

[b]Centro de Investigación Avanzada en Física Fundamental, Universidad Autónoma de Madrid, 28049 Madrid, Spain

[c]Center for Theoretical Neuroscience, Zuckerman Institute, Columbia University, New York, NY, USA

Abstract

A stimulus held in working memory is perceived as contracted towards the average stimulus. This contraction bias has been extensively studied in psychophysics, but little is known about its origin from neural activity. By training recurrent networks of spiking neurons to discriminate temporal intervals, we explored the causes of this bias and how behavior relates to population firing activity. We found that the networks exhibited animal-like behavior. Various geometric features of neural trajectories in state space encoded warped representations of the durations of the first interval modulated by sensory history. Formulating a novel normative model, we showed that these representations conveyed a Bayesian estimate of the interval durations, thus relating activity and behavior. We conjecture that the neural implementation of the contraction bias in delayed comparison tasks occurs in brain areas where information about the current and preceding stimuli converges.

Keywords— Neural Networks, Decision Making, Dynamical Systems

*Corresponding author at nestor.parga@uam.es

Introduction

Perception is not based purely on the current stimulus but is affected by other factors such as recent stimulation, prior knowledge and context, giving rise to perceptual biases [1]–[9]. Examples of these biases appear in magnitude estimation [10]–[13] and delayed comparison tasks [14]–[17], in which a stimulus parameter is perceived as shifted toward the mean of its distribution. These phenomena, known respectively as the central tendency and the contraction bias, have been extensively investigated in psychophysics [10]–[17] and, more recently and to a lesser extent, with neural activity recordings in temporal reproduction tasks [18]–[24] and delayed comparison tasks [6], [9]. However their causes have not been fully elucidated, particularly in the latter.

The contraction bias is affected by the sensory history [6], [25]–[27]. A remarkable study found that in delayed comparison tasks the activity of posterior parietal cortex (PPC) neurons carries information about past stimuli, but not about the current stimulus [6]. Furthermore, that activity was causally related to the contraction bias. In addition to recent sensory history, neural activity in PPC also reflected longer-term sensory statistics, in the form of the mean of the stimulus distribution.

Within a Bayesian framework, prior experience is integrated with the noisy observation of the current stimulus. This integration would result in a sensory representation biased to the most reliable or likely stimuli, generating the contraction bias in delayed comparison tasks [9], [15], [25], [28], [29]. However, despite their relevance, no study has addressed the fundamental questions of how the integration of those relevant factors takes place in neural circuits and what features of neural population activity relate to behavior. Moreover, it is not known how the neural information about the preceding stimuli relates to Bayesian computations.

New computational techniques enable us to train recurrent neural networks on tasks similar to those used in the experiments [4], [30]–[35]. This approach provides us with a valuable tool to investigate the network mechanisms responsible for perceptual biases. Based on trained networks, we can formulate hypotheses and draw predictions to guide novel experiments. In addition, trained networks allow full access to the state-space dynamics and geometry of the neural trajectories and thus to the study of the relationship between population activity and behavior [20], [21].

In this work, to further the understanding of the causes of the contraction bias and its connection with neural population activity, we trained spiking recurrent neural networks (sRNNs) with the full-FORCE algorithm [36], [37](see also [38]) to compare the durations of two temporal intervals (d_1 and d_2), spaced by a delay interval. The trained networks did exhibit the bias. Examination of the networks' behavior allowed us to identify potential sources of the contraction bias: short- and long-term sensory history and the integration of noisy observations with prior knowledge. Accordingly, an ideal observer would have access to all of them; this led us to propose a novel Bayesian normative model that can be used to guide the analysis of the data and help identify the network computations responsible for the emergence of the contraction bias.

The activity of the neural population described trajectories in a low-dimensional state space. The geometry of the trajectories during the task encoded the durations of the first interval but did so by compressing them with respect to their true values. These compressions were hypothesized to convey a Bayesian estimate of the interval, thus relating population activity and behavior. Indeed, the Bayesian estimator of d_1 showed similar compressions as those of the state-space trajectories, thereby supporting

a Bayesian explanation of the contraction bias.

Our work leads to testable predictions about the causes of the contraction bias and how they are interrelated. It predicts a relationship between geometric properties of state-space trajectories and behavior. Moreover, the normative model predicts that the sensory history contributes not only to the mean of the observations of the first interval but also modulates their variance; this modulation has behavioral and physiological consequences, since an increase in observation variance implies a larger contraction bias and affects the geometric properties of the trajectories. These findings are a consequence of the recurrence of the network and the existence of a delay period between the two presented stimuli. More generally, then, we expect the contraction bias be explained in a similar way in other two-interval forced-choice tasks.

Results

Network training and numerical experiments. We trained sRNNs to decide which of two time intervals, d_1 and d_2, presented sequentially and separated by a delay interval, was longer (Fig. 1a). Networks were trained with a set of duration pairs (d_1, d_2), also named classes (Fig. 1b), selected randomly and independently in each trial, using the full-FORCE algorithm for sRNNs [36]–[38] (Supplementary Fig. S1). Trials were presented continuously, without altering the network activity at the end of each trial. Once a network was trained, a large number of test trials were obtained for further analyses of task performance and population activity (Methods). We considered three different trial blocks: a block in which only diagonal classes (classes labeled from 1 to 10 in Fig. 1b) were tested, a block where only horizontal classes (d_2=370 ms) were tested and a block where all classes were tested (Methods). These will be referred to as the diagonal, horizontal and full blocks (Fig. 1b). Each block defined a context with a distinct prior distribution of d_1 (Fig. 1c); note that some classes appeared in more than one context.

In the main numerical experiment we trained a total of 35 networks (obtained from random initializations of the network parameters), with a fixed delay interval of 1,500 ms. Unless otherwise specified, the analyses were performed on this experiment. In a separate numerical experiment, to examine the effect of the delay period duration on the contraction bias, we trained 15 sRNNs by selecting its duration on each trial randomly from a set of four possible values, respectively (Methods). In another experiment, we trained 7 networks resetting the population activity at the end of every trial (Methods).

The trained networks exhibited the contraction bias. The psychometric curves were computed using test trials from the horizontal classes in the full block. Fig. 1d shows this curve for their mean over all trained networks (left) and for example network 1 (right) (see Extended Data Fig. 1a for example network 2). Psychometric curves were heterogeneous across the 35 trained networks (Supplementary Fig. S2). Next we analyzed the discrimination performance using test trials from the diagonal block. Note that classes in this block have a fixed duration difference $\Delta d = |d_1 - d_2|$ (60 ms; Fig. 1b). However, contrary to what would be expected for classes with the same duration differences, accuracy was not uniform, as seen in its average over networks (Fig. 1e, left) and in the two example networks, Fig. 1e (right) and Extended Data Fig. 1b, respectively. The contraction bias can explain this disparity. This bias is usually described as a shift of the encoding of duration of the first stimulus

towards the mean of its distribution (i.e., $\langle d_1 \rangle$=370 ms, Fig. 1f). This means that a short duration (i.e., d_1=250 ms) is encoded as longer, while a long duration is encoded as shorter. Therefore, correct decisions for classes with $d_1 < d_2$ are increasingly impaired by the bias as d_1 decreases (upper diagonal, Fig. 1f), while classes with $d_1 > d_2$ are increasingly facilitated by the bias (lower diagonal, Fig. 1f). To make this bias structure explicit, we numbered the classes such that those with favorable bias effects are near the extremes, while classes with disfavorable effects are around the center (Fig. 1f). With this arrangement of classes the decision accuracy was higher for classes near the two extremes (Fig. 1e). Note that since the networks were trained with sequences of temporally uncorrelated d_1 values, this result implies that correlations in the training stimuli are not crucial for generating the bias.

The contraction bias of a network was quantified as the root-mean-square deviation of the performance, relative to the diagonal classes with $d_1 = 370$ ms, a quantity denoted as B (Methods). The distribution of this bias metric spanned approximately one order of magnitude across networks (Fig. 1g).

Behavior and possible causes of the contraction bias. We postulate three different sources for the contraction bias. One possible explanation is inter-trial effects, through the persistence of the neural representations of stimuli applied in previous trials. This would produce a contraction driven by the mean of the stimulus parameter or by the most recently presented stimuli [6], [25]. In addition, it has been proposed that the bias can be generated by a Bayesian mechanism that occurs if, at the time of comparison of the two stimuli, the uncertainty in the observation of the first stimulus is greater than the uncertainty in the observation of the second one -which creates a tendency to rely more on the *a priori* knowledge on the first stimulus [15]. In support of this idea, behavioral data from monkeys discriminating tactile frequencies were well fit with a Bayesian model [9]. Finally, the stimulus statistics could be stored in the network synapses during the learning stage, then affecting behavior with a contraction of the intervals toward the learned mean [28]. We describe below the results of numerical experiments performed to examine these three possible causes of contraction bias.

To check that the class in the previous trial can modulate behavior, we sorted a large number of test trials by the preceding class and calculated the choice accuracy for each case (Methods). The analysis confirmed that the preceding trial modulates performance (see Fig. 2a and Extended Data Fig. 1c for example networks 1 and 2, respectively). In addition, we observed the presence of a choice bias, which depended on whether the preceding choice had been rewarded (Methods).

If the contraction bias were due solely to the preceding stimulation then it should remain constant or weaken as the duration of the delay period is increased [29]. However, if a Bayesian mechanism were operating in our networks, we would expect that as the duration of the delay epoch increases, the noise of the representation of the first interval would also increase, resulting in a larger contraction bias [29]. To compare the bias for different durations of the delay period, we trained 15 networks to perform time interval discrimination for four delay durations, randomly interleaving these values during training (Methods). After training, blocks of test trials (diagonal classes) with fixed delay period duration were collected and analyzed. We found that the bias B increased with the delay duration (Fig. 2b).

To gain further insight into the mechanisms that could account for the contraction bias we trained networks by resetting the network activity at the end of each trial, rather than maintaining continuity

of activity during the inter-trial epoch (Methods). If the bias only depended on the information about past trials, then the bias would disappear in the absence of this information. In case the bias persisted, then we would have to look for some other mechanism capable of producing it. To answer this question we investigated the seven networks trained under this condition. The accuracy of six of them showed a clear bias, as can be observed in the network mean of the psychometric function (Fig. 2c, left) and the accuracy (Fig. 2c, middle). It is noteworthy that the bias B turns out to be weaker than in the networks trained with no reset (Fig. 2c, right).

So far we have analyzed behavioral data from trained networks with an eye to potential sources of the contraction bias: sensory history, noisy observations and prior knowledge and memorized stationary statistics. We will return to these results later (section "Ideal observer model"), to examine them in light of a normative model that incorporates the presumed causes of contraction bias. Before that we examined whether these factors were present in the neuronal activity.

State-space dynamics and geometry. We adopted the view that neural circuits represent and process information through the state-space dynamics of neural populations [4], [30], [32], [33], [39]–[43]. Most analyses on neural firing activity were performed at the population level, the population being all those neurons encoding d_1 (Methods). We projected the population activity (z-score) onto the directions in state space responsible for at least 90% of the data variance (Methods). As expected, during the presentation of the first interval the population z-score followed a common path until the end of the duration of the respective d_1, defining a curved manifold (Fig. 3a). This feature also appeared in recordings of monkeys performing time reproduction tasks, during the presentation of the sample interval [19]–[21]. The similarity is not surprising, since both the first interval in the discrimination task and the sample interval in the reproduction task must be represented in the network activity to complete the corresponding task. The trajectories evolved at different speeds during the delay period, so that the longer the duration of their interval d_1, the higher the speed (Fig. 3b). As the trajectories approached the end of the interval their speeds decreased (Fig. 3c), reaching d_1-dependent terminal states that appear in orderly alignment (Fig. 3b, inset). Correspondingly, the percentage of variance explained by the first principal component approached 100% (Fig. 3c, inset), showing that the trajectory arrived to a one-dimensional quasi-attractor region. However, this result does not specify which estimator of d_1 is maintained by the network when it approaches the quasi-attractor. Addressing this question, as well as what geometrical properties of the trajectories encode the estimator, requires connecting behavior with population activity and will be dealt with later (section "State-space geometry is related to behavior").

If the networks were to implement a Bayesian computation, integrating the noisy measure of the current d_1 with prior knowledge, then, when the same value of d_1 occurs in two different contexts, the population activity should follow distinct trajectories. Since classes 5 and 6 are shared by the horizontal and diagonal blocks (Fig. 1b), we wondered whether the networks learnt that these blocks define different contexts (h and d). To verify this supposition we first defined relative distances following the Kinematic analysis of Neural Trajectory (KiNeT) methodology [19]. Briefly, for each class shared by the blocks h and d (that we indicate by $j = 5$ or 6) we considered its trajectories in the two contexts (trajectories j_h and j_d). Then we selected one of them (e.g., j_d) as a reference and defined corresponding states between the reference and the other trajectory (j_h). If at time t after the input

pulse (SO1 or SF1, see Fig. 1a) the reference trajectory was in the state $s_{ref}(t)$, the corresponding state in trajectory j_h was taken as the state s_j closest to $s_{ref}(t)$. Then, for each time t, we computed the Euclidean distance $D_j(t)$ between the pair of corresponding states. During the presentation of the first time interval, trajectories of classes shared by the two contexts (Fig. 3d, left. Example network 1) evolved disjointly, keeping a significantly non-zero relative distance from each other (Fig. 3d, right). During the delay period, those trajectories (Fig. 3e, left) also maintained significantly non-zero distances (Fig. 3e, right). These results indicate that the networks acquired knowledge about the context during the training stage. It should be noted that the networks were not trained specifically on the horizontal and diagonal contexts, since all the classes of the set in Fig. 1b, chosen randomly, were used for that purpose (Methods).

Neural activity is affected by sensory history. The results in Figs. 3d,e point to the influence of sensory history on the current trial. We now study this question in a systematic and quantitative way. We first checked whether the neural activity in the current trial maintained information about the interval d_1 presented in the preceding ones. To do so, we began with an analysis of the trajectories in state space, and then performed a more quantitative analysis of mutual information. Note that if the neural population coded the value of the current d_1 throughout the delay period, then trajectories corresponding to different durations should stay at significantly non-zero relative distances. Likewise, if the representation of the first interval persisted to the next trial, trajectories should still maintain non-zero distances during the inter-trial-interval (ITI). Again, we employed the KiNeT methodology [19]; for each pair of trajectories (j and $j+1$) we selected trajectory $j+1$ as a reference and, for each time t, we computed the Euclidean distance $D_j(t)$ between the pair of corresponding states (Methods). All pairs of consecutive trajectories in the diagonal set maintained non-zero relative distances during the delay period (Fig. 4a) and ITI (Fig. 4b). This implies that the neural representation of the first interval is maintained beyond the epoch in which it is presented.

To further support this conclusion we analyzed the mutual information (MI) that the activity had on the presented time intervals. Most neurons maintained significant information about the current d_1 during the delay period and even during the following ITI (Supplementary Fig. S3). During the current trial there was a fraction of neurons showing significant MI about the first interval in the current (Fig. 4c, red lines), the preceding (blue lines) and the second previous trial (green lines). Soon after SO1 (after a time equivalent to the shortest first interval), the fraction of neurons with MI about the current d_1 increased rapidly and remained high during both the entire delay period and the ITI. The population of neurons with significant MI on the first interval of the preceding trial (blue), was quite large before SO1, during the delay period it stayed at a similar level as the population with MI on the current d_1, although it became smaller during the ITI. In contrast, the fraction of neurons with MI on the second previous d_1 (green) decreased rapidly throughout the current trial. There was also a fraction of neurons with significant MI about the second interval (d_2) in the current (Fig. 4d, red lines), previous (blue) and second previous (green) trials.

These results differ from those obtained with recordings in PPC which showed that neurons only had information about the value of stimuli in previous trials [6]; information about the current one supposedly is processed in other areas. Our model assumes an area in which information about the current and preceding stimulus has converged.

Ideal observer model. We verified above that the candidate causes for the contraction bias are indeed present in the neuronal activity of the trained networks. Now, taking advantage of the fact that normative models can help identify which computations are involved in producing the observed behavior, we formulate a model containing the possible causes of the bias. In doing so, we expect to understand how those factors operate and interact to generate the bias. To model behavior, we took the viewpoint of an ideal observer who makes noisy measurements of the intervals d_1 and d_2, has knowledge about the stimulus set, has access to the sensory history and can construct a belief about the duration of the presented interval. Measurements of the two time intervals are assumed to be Gaussian with means d_1 and d_2 and standard deviations that obey the scalar law ($\sigma_1 = W_1 d_1$ and $\sigma_2 = W_2 d_2$, respectively, where the parameters W_1 and W_2 are the corresponding Weber fractions. Methods) [44], [45]. We start by describing the sensory history of the first interval. The observation ($o_{1,n}$) made at trial n combines the noisy measurement of the current interval ($d'_{1,n}$) with the measurements of a few (m) recent past intervals, and with the observation made at a longer temporal scale ($o_{1,n-m-1}$). That is,

$$o_{1,n} = (1-\eta)\, d'_{1,n} + (1-\eta)\, \eta d'_{1,n-1} + \ldots + (1-\eta)\, \eta^m d'_{1,n-m} + \eta^{m+1} o_{1,n-m-1}, \tag{1}$$

where η weights the contribution of preceding intervals (see Methods for further details) and $d'_{1,j}$ ($j = n-1, ..., n-m$) is the noisy measurement of the first interval presented in trial j. We assumed that the long-term sensory history, o_{n-m-1}, is in a Gaussian stationary regime, with its mean $\mathcal{M}_{st}$ and variance $\Sigma^2_{1,st}$ obtained self-consistently (Methods). With these assumptions, the observation of the interval d_1 turns out to also be a Gaussian variable ($\mathcal{O}_1$), with mean $\mathcal{M}$ and variance Σ^2_1. The observation of the second interval is simply the noisy measurement $d'_{2,n}$ (for a model that considers the sensory history of the second interval see Supplementary Text S1). The prior probabilities of d_1 and d_2 were assumed to be discrete (Fig. 1b). Decisions, "$d_1 > d_2$" or "$d_1 < d_2$", were made from the posterior probabilities using the *maximum a posteriori* decision rule (Methods). The general model in Eq. (1) depends on only three parameters: W_1, W_2 and η. We can now determine the observation probability of the first interval in the current trial n. Although this can be done for an arbitrarily long transient history (any m. Methods), in this work we will consider only the versions of the model with $m = 0$ and $m = 1$. The reason for not taking larger values of m is that the population with significant information about $d_{1,n-2}$ is of a smaller size than the populations with significant information about the intervals in the current and the immediately preceding trials (Fig. 4c). For $m = 0$ the observation probability has mean and variance

$$\mathcal{M} = d_{1,n} + \eta(\langle d_1 \rangle - d_{1,n}) \tag{2a}$$

$$\Sigma^2_1 = \Sigma^2_1(W_1,\, \eta,\, \langle d_1 \rangle,\, \sigma^2_{d_1},\, d_{1,n}), \tag{2b}$$

where $\langle d_1 \rangle$ is the first moment of d_1 in the stimulus set and σ_{d_1} is its variance (see Methods for the full expression of Σ^2_1). Note that both $\mathcal{M}$ and Σ_1 could contribute to the contraction bias: The mean indicates explicitly that the first interval is contracted toward the interval mean, by an amount proportional to the difference between the true value of the interval and the d_1 mean. The variance provides a Bayesian explanation of the bias, which operates when the uncertainty in the first interval

exceeds the uncertainty in the second interval, i.e., when $\Sigma_1 > \sigma_2$ [15]. Interestingly, the sensory history parameter η modulates the variance contribution to the bias (Methods). For each network, the parameters η, W_1 and W_2 were fitted to optimize the similarity between the accuracy curves of the model and the network (Methods). The fit for the example network 1 is shown in Fig. 5a (see Extended Data Fig. 3a for example network 2) and the network root-mean-square error (RMSE) distribution of the fits appears in Fig. 5b. A measure of the contraction bias, B_{fit}, obtained from the fitted performance, is given in Fig. 5c.

The model version $m = 0$ considers the effect of the long-term sensory history on the observation of the first interval, but neglects the contribution from recent trials. To examine this issue, we now turn to the case $m = 1$; the mean and variance of the observation probability are

$$\mathcal{M} = d_{1,n} + \eta\Big[d_{1,n-1} - d_{1,n} + \eta\left(\langle d_1\rangle - d_{1,n-1}\right)\Big] \tag{3a}$$

$$\Sigma_1^2 = \Sigma_1^2(W_1,\, \eta,\, d_{1,n},\, d_{1,n-1},\, \left\langle d_1^2\right\rangle,\, \sigma_{d_1}^2), \tag{3b}$$

where $\left\langle d_1^2\right\rangle$ is the second moment of d_1 and $\sigma_{d_1}^2$ is its variance (Methods). The observation of the current interval contracts towards $d_{1,n-1}$ which, in turn, appears contracted towards $\langle d_1\rangle$. Again, there is a Bayesian source for the bias when $\Sigma_1 > \sigma_2$. Note that the preceding interval affects the observation variance and, hence, short-term sensory history modulates the Bayesian contribution to the bias. Thus, in the model, all the sources of the contraction bias appear to be intricately related.

To fit the model parameters (η, W_1 and W_2), we considered the 10 accuracy curves that result from conditioning on each of the possible preceding classes (Fig. 2a for the example network 1 and Extended Data Fig. 1c for the network 2), and used them all to define the similarity between model and data (Methods). The fits for the example networks 1 and 2 are shown in Fig. 5d and Extended Data Fig. 3b, respectively. The network distribution of the RMSEs of the fits is in Fig. 5e, and the relationship between the phenomenological bias B and the bias quantified from these fits appears in Fig. 5f. The average accuracy for each current class, obtained by marginalizing over the preceding first interval (Methods and Supplementary Text S2), is in Fig. 5g (Extended Data Fig. 3c for network 2).

The model in Eq. (1) describes the sensory history of the first interval. However, the activity during the delay period also has information about the second interval in the preceding trial, $d_{2,n-1}$ (Fig. 4d). A more general model that considers the combined sensory history of d_1 and d_2 is provided in Supplementary Text S1. Surprisingly, the quality of the fit with this model is identical to that obtained with the model in which the contribution of the second interval is not considered (Extended Data Fig. 4 and Supplementary Table S1). Since the inclusion of the second interval did not improve the fit, for the sake of simplicity, in the following we will not take it into account.

Before we showed that a weak contraction bias is still present in networks trained resetting the network activity at the end of each trial (Fig. 2c). We now wonder about the model that describes this situation, and about the origin of the bias in these networks. Since the current trial is not affected by the sensory history, the observation model in Eq. (1) should be excluded. Instead, we considered the possibility that neural circuits may have knowledge of the stimulus statistics (e.g., knowledge of $\langle d_1\rangle$), which presumably could have been learned during training and kept stored in the network synapses.

We then contrasted the hypothesis that the circuits have knowledge about $\langle d_1 \rangle$ with the hypothesis that they do not. In the first case the observations have mean $\mathcal{M} = d_{1,n} + \mu(\langle d_1 \rangle - d_{1,n})$ and variance $\Sigma_1^2 = \sigma_1^2$, while in the second these are $\mathcal{M} = d_{1,n}$ and $\Sigma_1^2 = \sigma_1^2$, respectively. The parameters of the first model are μ and σ_1 (but note that μ is not a sensory history parameter), while the only parameter of the second model is σ_1. We fitted the behavioral data for each network with both models obtaining comparable results in terms of the fit quality (Extended Data Fig. 5). Comparison of the two models using BIC or AIC tends to favor the model that does not require knowledge of $\langle d_1 \rangle$ (Supplementary Table S2). This means that the Bayesian explanation could be enough to account for the contraction bias in networks trained with activity reset, without the need to store the interval mean in the synapses.

Correlations between the phenomenological bias and the model parameters. We now use the normative model to examine how the sensory history and the uncertainty in the observations operate and interact to generate the bias. Since the variance of the observation of d_1 depends on sensory history, this analysis is not immediate. First, we noted that the condition $\Sigma_1 > \sigma_2$ determines whether the uncertainty in the observations contributes to the bias [15]. We will say that in networks where this condition is satisfied the Bayesian contribution to the bias is operative. Then, we sorted the 35 networks into two groups: a group in which that condition was satisfied (the positive group) and a group in which it was not (the negative group). To obtain further insight into the role of each parameter, for each network group we evaluated the correlation between the phenomenological measure of the bias (B) and the model parameters: the sensory history parameter (η) and the two observation uncertainties (Σ_1 and W_2). Note the use of Σ_1 instead of W_1, this is because the quantity relevant for determining the bias is the total variance of the observation probability of d_1.

We start with the simplest version of the model ($m = 0$, Eqs. (2a) and (2b)). About half of the networks (18/35) belonged to the positive group ($\Sigma_1 > \sigma_2$), signaling a contribution from observation uncertainty to the contraction bias. As expected, in positive networks B was significantly and positively correlated with Σ_1 (Fig. 6a, left and Supplementary Fig. S4a). Instead, B was not significantly correlated with η (Fig. 6a, middle). This is surprising, given that the mean $\mathcal{M}$ explicitly describes the mechanism by which the sensory history contributes to the formation of the bias, Eqs. (3a) and (3b). The explanation for this paradox is that the sensory history also contributes to Σ_1, and does so in such a way that η is negatively correlated with the variance (Fig. 6b and Supplementary Fig. S5). This reduces the overall effect of η on the bias. Finally, B was not significantly correlated with W_2 (Fig. 6a, right), a parameter that only becomes relevant after the end of the delay period. We observed that in most networks the difference between the noise parameters, $W_1 - W_2$, was positive (Fig. 6c). In negative networks the situation was the other way around: B and Σ_1 were not significantly correlated (Fig. 6d, left, and Supplementary Fig. S4b), in contrast, B and η were positively correlated (Fig. 6d, middle). In negative networks B and W_2 were also not significantly correlated (Fig. 6d, right).

A similar pattern of correlations held for the model version with $m = 1$, Eqs. (3a) and (3b). For most networks (28/35), the fit satisfied that $\Sigma_1 > \sigma_2$, meaning that the Bayesian contribution to the bias was present more often than in the simpler version. Again, in positive networks the bias B was favorably correlated with Σ_1 (Fig. 6e, left and Supplementary Fig. S6). Instead, its correlation with η was not significant (Fig. 6e, middle). The reason for this feature was again the competition between the contributions of the sensory history to the mean $\mathcal{M}$ and to the variance Σ_1^2 (Fig. 6f). Again, B

was not significantly correlated with W_2 (Fig. 6e, right). Finally, the difference $W_1 - W_2$ was still positive (Fig. 6g). As for the negative networks, the model version $m = 1$ gives predictions similar to those found with the version $m = 0$ (Fig. 6h and Supplementary Fig. S7), although the positive correlation between B and η was somewhat below significance level (Fig. 6h, middle).

In summary, inspired by data taken from the trained sRNNs, we began by proposing a normative model, Eq. (1), which included the factors supposedly responsible for the contraction bias, namely, the sensory history and the observation uncertainty relative to *a priori* knowledge. Fitting the data with this model confirmed that both factors were necessary: The measurement of d_1 is affected by a fixed noise (W_1), yet, the uncertainty of its observation is also influenced by the sensory history. We then conclude that the contraction bias results from the sensory history, the uncertainty in interval measurements, and an interaction between these two factors.

State-space geometry is related to behavior. So far we have analyzed behavior and neural activity separately. Guided by a normative model of behavior we have 1) distinguished two groups of networks in which the Bayesian contribution to the bias would have different relevance and 2) noted that the sensory history affects the Bayesian contribution to the bias. Nevertheless, those analyses do not inform us about the neural mechanisms that could implement the computations suggested by the model. To examine this issue, we now investigate how state-space geometry relates to behavior. The task requires that during the presentation of d_1 the neural activity represents an estimate of its duration and then maintains it during the delay period, until the end of this interval. Accordingly, we expect that over these two epochs an estimate of d_1 may be represented in the geometric properties of the trajectories [20], [21]. We have seen that during interval presentation the trajectories define a one-dimensional manifold characterized by its curvature (Fig. 3a) while after the completion of each interval (SF1) the trajectories remain distant, reaching a quasi-attractor region at the end of the delay epoch (Figs. 3b,c). So, we will now assume that estimations of d_1 are represented, first, in the curvature of the manifold (the "curved manifold hypothesis" [21]) then, during the delay period, in the relative distances of the trajectories and, finally, in the distances between the terminal states on the quasi-attractor. However, the nature of the estimator is unknown, since in delayed comparison tasks the estimate of the first stimulus is not reported, only the comparison choices are available. This contrasts with magnitude reproduction tasks, in which the reproduced magnitude can be taken as a noisy version of the estimate [11]. Given the limited information about what estimator of d_1 employed the networks, we considered, apart from the true d_1, two other alternatives, the Bayesian and Maximum Likelihood (ML) estimators ($d_{1,Bayes}$ and $d_{1,ML}$, respectively. Methods), and studied how they related with curvature, relative distances between neural trajectories and distances between terminal states on the attractor.

To find out which estimator best explained the curvature in the neural activity, we projected the neural states in the curved manifold onto an encoding vector $\mathbf{u}$, defined as the unit vector pointing from the state associated with the shortest to the longest d_1 in the set [21] (Methods). For the example network 1 (Fig. 7a) the distances P_j between the projections (See inset; $j = 1, \ldots, 5$ labels the values of d_1 in the diagonal block) gave a warped representation of d_1. We then asked whether the P_j more closely reflected the true d_1 values or one of the two estimates, $d_{1,Bayes}$ or $d_{1,ML}$. To confront these possibilities, we first separately fitted a linear expression ($aP_j + b$) to the true d_1 and to the two

estimators (the three regressions appear in Fig. 7b). Then, we compared the corresponding goodness of fit using their RMSEs (Fig. 7c); for this network the test favored the Bayesian estimator hypothesis, and the same happened for most positive networks (Extended Data Figure 6a). To determine which estimator of d_1 best accounted for the relative distances between the neural trajectories during the delay epoch, we followed a similar approach. The relative distance D_j(t) between each trajectory $j = 1, \ldots, 5$ and the reference trajectory d_1=370 ms, was defined following the KiNeT method [19] (Methods). The time courses of these distances for the example network 1 (Fig. 7d) and their temporal means $D_j \equiv \langle D_j(t) \rangle$ (inset), also appeared warped as a function of d_1. We then asked whether D_j coded better the true d_1 values or their Bayesian or ML estimators. To confront those alternatives, we separately fitted a linear expression ($aD_j + b$) to these three estimates (Fig. 7e) and we compared their RMSEs [20] (Fig. 7f). Again, for this network the test favored the Bayesian estimator hypothesis (see also Extended Data Fig. 6b). Finally, to test the estimators on the quasi-attractor we computed the distances δ_j between terminal points of the trajectories on the attractor region and made an analysis similar to the two previous ones. Again, the test favored the Bayesian estimator (Figs. 7g-i and Extended Data Fig. 6c). The results of the three tests provide us with links between behavior and state-space geometry, for those networks in which the bias can be explained by Bayesian arguments. Since the example network 2 does not fulfill $\Sigma_1 > \sigma_2$ we do not expect it to pass the tests well. To verify this expectation, we performed the same analyses on network 2. Although the 1-dimensional manifold did present curvature, in this network the Bayesian estimator did not explain better than the other estimators the geometry of neural trajectories (Extended Data Fig. 7).

Earlier, we quantified the contraction bias in terms of the discrimination curves (B). Now, we can quantify it using the estimator $d_{1,Bayes}$, which resulted from the behavioral fit (Methods). This gave us a measure for the bias (B_{Bayes}) that correlated well with the empirical measure B (C=0.96, p<10-6). In addition, B_{Bayes} had the same pattern of correlations with the model parameters as B (Supplementary Fig. S8).

Effect of sensory history on state-space geometry. We saw that the sensory history modulates the contraction bias (Fig. 2a). So, according to the results in Fig. 7, it can be expected to modulate geometry as well, and to do so in a way consistent with the change it produces on the bias. For example, in network 1, when varying the previous class along the set's upper diagonal (i.e., from class 1 to 5. Fig. 1b), the bias decreases approximately by half (B=0.12 vs. 0.06. Fig. 2a). Therefore, if our expectation were correct, the geometric features should exhibit a more warped representation of the current d_1 in trials preceded by class 1 than in trials preceded by class 5. To verify this, we first checked whether the geometry of the trajectories still codes the Bayesian estimator when conditioning on the class of the preceding trial. The tests for the curvature, relative distances and positions on the quasi-attractor, obtained conditioning separately on the previous classes 1 or 5, showed that $d_{1,Bayes}$ was still the preferred estimator (Extended Data Fig. 8). Next, we related each of the three geometric features with the contraction bias, defining the compressions $C[Z]$ ($Z = P_j, D_j$ or δ_j) of the associated d_1 representations (Methods, Eq. 52). Finally, we computed each $C[Z]$ in the two conditions (previous class 1 or 5). In all three cases the compression was higher when the preceding class was class 1. The compressions from the curved manifold were $C[P] = 0.49$ (previous class 1) and $C[P] = 0.42$ (previous class 5) with p-value=0.004 (Mann-Wilcoxon test over two bootstrap distributions, each

with 100 resamples). For the mean relative distances, $C[D] = 0.33$ (class 1) and $C[D] = 0.30$ (class 5), p-value=0.09. Finally for the positions on the attractor, $C[\delta] = 0.43$ (class 1) and $C[\delta] = 0.21$ (class 5), p-value=6e-21. The decrease in significance during the delay period and its important recovery at the end of this epoch could be related to the initial decrease and subsequent rise in the information about the previous intervals that occur during the memory period (Fig. 4c,d). Network 2 has a similar bias B as network 1 and undergoes a similar change in bias when conditioning on prior classes 1 or 5 (B=0.14 vs. 0.06. Extended Data Fig. 1c). However, since this network does not satisfy that $\Sigma_1 > \sigma_2$, $d_{1,Bayes}$ was not unambiguously the preferred estimator (data not shown).

Discussion

The contraction bias is influenced by the stimulation history [6], [25]. But the action of preceding stimuli is not the only possible explanation for this bias, which could also be generated from the integration of prior knowledge with noisy observations of the stimuli [9], [15], [46]. Here we addressed the issue of how the contraction bias results from the joint action of these factors by training sRNNs on a delayed comparison task. The analyses of the behavioral and firing activity data allowed us to identify the sources of the bias and to propose a normative model that assisted us in determining the network computations responsible for this bias. A Bayesian explanation of the contraction bias was operative in most trained networks (80%), where it played a preponderant role, with sensory history modulating the uncertainty in the observation of the first stimulus. Although the sensory history was always present, in practice it was correlated with the bias in a low fraction of the networks (the remaining 20%).

During the execution of the task, successive geometric properties of the state-space trajectories showed compressed representations of the duration of the first interval. During the presentation of that stimulus this happened on (a linear readout of) the curvature of a 1-dimensional manifold; then, during the delay period, on the relative distances of the trajectories; finally, at the end of that epoch, on a one-dimensional quasi-attractor. These compressions were matched by the Bayesian estimator resulting from the normative model, indicating that the contraction bias emerged from the geometry of the trajectories in state space.

Several aspects of our results can be compared with recent experimental studies of neural activity in both delayed comparison tasks and reproduction tasks. Akrami et al. found that during a discrimination task neurons in PPC carried information about recent sensory history [6]. However, this area did not present information about the stimulus of the current trial, so it does not correspond to the model considered in our work, which assumes an area that has information about current and preceding stimuli. On the other hand, in a temporal reproduction task Sohn et al. found that a curved one-dimensional manifold exhibited a compressed representation of the time interval, that was matched by the compression predicted by the Bayesian estimator [21] (see also [47]). Our results showed that a similar relationship between population activity and behavior may occur even when an estimate of the interval is not reported in the task.

The analysis of the time interval discrimination task presented here can be adapted for use in monkeys or rodents, and we anticipate that the results will also apply to other delayed response tasks, including discrimination of auditory or tactile stimuli. Our work may inspire the adaptation of such

experiments to verify our findings. In addition, our use of spiking RNNs offers practical advantages, such as the possibility of a straightforward application of the analysis techniques developed here to real recordings of neural activity.

Our trained spiking RNNs exhibited the contraction bias. However, other training procedures yielded network models that failed to reproduce the biases exhibited by behaving animals [31], [48]. The question of how neural networks should be trained to obtain models consistent with observed behavior has received some attention [49], but is not yet well understood. One possible approach is the use of noise; indeed, it helped to obtain animal-like behavior [30] and to account for biases due to temporal expectation [4] with trained RNNs of rate neurons. However, networks of noisy GRUs optimized in a frequency discrimination task achieved perfect performance and thus did not exhibit the contraction bias [31]. A possible cause of such good performance of the networks trained in [31], [48] is the use of neuron and network models with computational capabilities that are beyond those of real brains, and thus can outperform animal behavior, even in the presence of noise [31]. Although our training technique is not biologically realistic, the resulting spiking RNNs are plausible; the networks were trained without explicitly introducing noise, but the spiking noise could be responsible for the agreement with behavioral data in two-interval forced-choice tasks. Given the usefulness that RNNs are showing in understanding how cognitive tasks can be solved in biological networks, fully understanding how factors such as the neural model, various types of noise, or pre-training procedures lead to models in agreement with animal behavior is a crucial problem that deserves further research.

This work advances the understanding of the origin of the contraction bias by proposing a unified description of the potential sources of this bias and showing that the Bayesian explanation is prevalent in recurrent networks trained to discriminate temporal intervals. The results raise the question of whether they occur in the brain; in this regard, one of the most pressing questions is the elucidation of in which areas and in what way the contraction bias is generated.

References

[1] F. Carnevale, V. de Lafuente, R. Romo, and N. Parga, "Internal signal correlates neural populations and biases perceptual decision reports," Proceedings of the National Academy of Sciences, vol. 109, no. 46, pp. 18 938–18 943, 2012.

[2] K. M. Bausenhart, O. Dyjas, and R. Ulrich, "Temporal reproductions are influenced by an internal reference: Explaining the vierordt effect," Acta psychologica, vol. 147, pp. 60–67, 2014.

[3] C. Summerfield and F. P. De Lange, "Expectation in perceptual decision making: Neural and computational mechanisms," Nature Reviews Neuroscience, vol. 15, no. 11, pp. 745–756, 2014.

[4] F. Carnevale, V. de Lafuente, R. Romo, O. Barak, and N. Parga, "Dynamic control of response criterion in premotor cortex during perceptual detection under temporal uncertainty," Neuron, vol. 86, no. 4, pp. 1067–1077, 2015.

[5] Z. Shi and D. Burr, "Predictive coding of multisensory timing," Current Opinion in Behavioral Sciences, vol. 8, pp. 200–206, 2016.

[6] A. Akrami, C. D. Kopec, M. E. Diamond, and C. D. Brody, "Posterior parietal cortex represents sensory history and mediates its effects on behaviour," Nature, vol. 554, no. 7692, pp. 368–372, 2018.

[7] F. P. De Lange, M. Heilbron, and P. Kok, "How do expectations shape perception?" Trends in cognitive sciences, vol. 22, no. 9, pp. 764–779, 2018.

[8] A. Hermoso-Mendizabal, A. Hyafil, P. E. Rueda-Orozco, S. Jaramillo, D. Robbe, and J. De la Rocha, "Response outcomes gate the impact of expectations on perceptual decisions," Nature communications, vol. 11, no. 1, pp. 1–13, 2020.

[9] S. Sarno, M. Beirán, J. Falcó-Roget, et al., "Dopamine firing plays a dual role in coding reward prediction errors and signaling motivation in a working memory task," Proceedings of the National Academy of Sciences, vol. 119, no. 2, e2113311119, 2022. DOI: https://doi.org/10.1073/pnas.2113311119.

[10] Z. Shi, R. M. Church, and W. H. Meck, "Bayesian optimization of time perception," Trends in cognitive sciences, vol. 17, no. 11, pp. 556–564, 2013.

[11] M. Jazayeri and M. N. Shadlen, "Temporal context calibrates interval timing," Nature neuroscience, vol. 13, no. 8, pp. 1020–1026, 2010.

[12] F. H. Petzschner, S. Glasauer, and K. E. Stephan, "A bayesian perspective on magnitude estimation," Trends in cognitive sciences, vol. 19, no. 5, pp. 285–293, 2015.

[13] K. Thurley, "Magnitude estimation with noisy integrators linked by an adaptive reference," Frontiers in Integrative Neuroscience, vol. 10, p. 6, 2016.

[14] H. L. Hollingworth, "The central tendency of judgment," The Journal of Philosophy, Psychology and Scientific Methods, vol. 7, no. 17, pp. 461–469, 1910.

[15] P. Ashourian and Y. Loewenstein, "Bayesian inference underlies the contraction bias in delayed comparison tasks," PloS one, vol. 6, no. 5, e19551, 2011.

[16] O. Dyjas, K. M. Bausenhart, and R. Ulrich, "Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence," Attention, Perception, & Psychophysics, vol. 74, no. 8, pp. 1819–1841, 2012.

[17] O. Dyjas and R. Ulrich, "Effects of stimulus order on discrimination processes in comparative and equality judgements: Data and models," Quarterly Journal of Experimental Psychology, vol. 67, no. 6, pp. 1121–1150, 2014.

[18] M. Jazayeri and M. N. Shadlen, "A neural mechanism for sensing and reproducing a time interval," Current Biology, vol. 25, no. 20, pp. 2599–2609, 2015.

[19] E. D. Remington, D. Narain, E. A. Hosseini, and M. Jazayeri, "Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics," Neuron, vol. 98, no. 5, pp. 1005–1019, 2018.

[20] S. W. Egger, E. D. Remington, C.-J. Chang, and M. Jazayeri, "Internal models of sensorimotor integration regulate cortical dynamics," Nature neuroscience, vol. 22, no. 11, pp. 1871–1882, 2019.

[21] H. Sohn, D. Narain, N. Meirhaeghe, and M. Jazayeri, "Bayesian computation through cortical latent dynamics," Neuron, vol. 103, no. 5, pp. 934–947, 2019.

[22] N. Meirhaeghe, H. Sohn, and M. Jazayeri, "A precise and adaptive neural mechanism for predictive temporal processing in the frontal cortex," Neuron, vol. 109, no. 18, pp. 2995–3011, 2021.

[23] J. Henke, D. Bunk, D. von Werder, S. Häusler, V. L. Flanagin, and K. Thurley, "Distributed coding of stimulus magnitude in rodent prefrontal cortex," bioRxiv, pp. 2020–04, 2021.

[24] M. Beiran, N. Meirhaeghe, H. Sohn, M. Jazayeri, and S. Ostojic, "Parametric control of flexible timing through low-dimensional neural manifolds," Neuron, 2023.

[25] O. Raviv, M. Ahissar, and Y. Loewenstein, "How recent history affects perception: The normative approach and its heuristic approximation," PLoS computational biology, vol. 8, no. 10, e1002731, 2012.

[26] R. Akaishi, K. Umeda, A. Nagase, and K. Sakai, "Autonomous mechanism of internal choice estimate underlies decision inertia," Neuron, vol. 81, no. 1, pp. 195–206, 2014.

[27] J. Fischer and D. Whitney, "Serial dependence in visual perception," Nature neuroscience, vol. 17, no. 5, pp. 738–743, 2014.

[28] C. Preuschhof, T. Schubert, A. Villringer, and H. R. Heekeren, "Prior information biases stimulus representations during vibrotactile decision making," Journal of Cognitive Neuroscience, vol. 22, no. 5, pp. 875–887, 2010.

[29] C. Papadimitriou, A. Ferdoash, and L. H. Snyder, "Ghosts in the machine: Memory interference from the previous trial," Journal of neurophysiology, vol. 113, no. 2, pp. 567–577, 2015.

[30] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome, "Context-dependent computation by recurrent dynamics in prefrontal cortex," nature, vol. 503, no. 7474, pp. 78–84, 2013. DOI: https://doi.org/10.1038/nature12742.

[31] H. F. Song, G. R. Yang, and X.-J. Wang, "Reward-based training of recurrent neural networks for cognitive and value-based tasks," Elife, vol. 6, e21492, 2017. DOI: https://doi.org/10.7554/eLife.21492.

[32] W. Chaisangmongkon, S. K. Swaminathan, D. J. Freedman, and X.-J. Wang, "Computing by robust transience: How the fronto-parietal network performs sequential, category-based decisions," Neuron, vol. 93, no. 6, pp. 1504–1517, 2017.

[33] J. Wang, D. Narain, E. A. Hosseini, and M. Jazayeri, "Flexible timing by temporal scaling of cortical responses," Nature neuroscience, vol. 21, no. 1, pp. 102–110, 2018.

[34] G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, and X.-J. Wang, "Task representations in neural networks trained to perform many cognitive tasks," Nature neuroscience, vol. 22, no. 2, pp. 297–306, 2019.

[35] A. Dubreuil, A. Valente, M. Beiran, F. Mastrogiuseppe, and S. Ostojic, "The role of population structure in computations through neural dynamics," Nature Neuroscience, pp. 1–12, 2022.

[36] B. DePasquale, M. M. Churchland, and L. Abbott, "Using firing-rate dynamics to train recurrent networks of spiking model neurons," arXiv preprint arXiv:1601.07620, 2016. DOI: https://doi.org/10.48550/arXiv.1601.07620.

[37] L. F. Abbott, B. DePasquale, and R.-M. Memmesheimer, "Building functional networks of spiking model neurons," Nature neuroscience, vol. 19, no. 3, pp. 350–355, 2016.

[38] B. DePasquale, C. J. Cueva, K. Rajan, G. S. Escola, and L. F. Abbott, "Full-force: A target-based method for training recurrent networks," PLOS ONE, vol. 13, no. 2, pp. 1–18, Feb. 2018. DOI: 10.1371/journal.pone.0191527. [Online]. Available: https://doi.org/10.1371/journal.pone.0191527.

[39] S. Vyas, M. D. Golub, D. Sussillo, and K. V. Shenoy, "Computation through neural population dynamics," Annual Review of Neuroscience, vol. 43, p. 249, 2020.

[40] S. Saxena and J. P. Cunningham, "Towards the neural population doctrine," Current opinion in neurobiology, vol. 55, pp. 103–111, 2019.

[41] R. B. Ebitz and B. Y. Hayden, "The population doctrine in cognitive neuroscience," Neuron, vol. 109, no. 19, pp. 3055–3068, 2021.

[42] F. Mastrogiuseppe and S. Ostojic, "Linking connectivity, dynamics, and computations in low-rank recurrent neural networks," Neuron, vol. 99, no. 3, pp. 609–623, 2018.

[43] M. Jazayeri and S. Ostojic, "Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity," Current opinion in neurobiology, vol. 70, pp. 113–120, 2021.

[44] C. R. Gallistel and J. Gibbon, "Time, rate, and conditioning.," Psychological review, vol. 107, no. 2, p. 289, 2000.

[45] B. C. Rakitin, J. Gibbon, T. B. Penney, C. Malapani, S. C. Hinton, and W. H. Meck, "Scalar expectancy theory and peak-interval timing in humans.," Journal of Experimental Psychology: Animal Behavior Processes, vol. 24, no. 1, p. 15, 1998.

[46] I. Lieder, V. Adam, O. Frenkel, S. Jaffe-Dax, M. Sahani, and M. Ahissar, "Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia," Nature neuroscience, vol. 22, no. 2, pp. 256–264, 2019.

[47] H. Sohn and D. Narain, "Neural implementations of bayesian inference," Current Opinion in Neurobiology, vol. 70, pp. 121–129, 2021.

[48] M. Molano-Mazón, Y. Shao, D. Duque, G. R. Yang, S. Ostojic, and J. de la Rocha, "Recurrent networks endowed with structural priors explain suboptimal animal behavior," Current Biology, 2023.

[49] G. R. Yang and M. Molano-Mazón, "Towards the next generation of recurrent network models for cognitive neuroscience," Current Opinion in Neurobiology, vol. 70, pp. 182–192, 2021.

Methods

Temporal Interval Discrimination (TID) task. Spiking Recurrent Neural Networks (sRNNs) were trained to decide which of two time intervals (d_1, d_2), selected from the stimuli set (Fig. 1b) and presented sequentially separated by a delay epoch, was the shortest. The task scheme (Fig. 1a) began with the Start Cue event. After a 1s-delay, the first time interval (d_1) was presented (SO1 and SF1 are the stimulus onset and offset of this interval, respectively). A second time interval (d_2) of variable duration (Fig. 1b) was displayed after a 1.5 s delay period (SO2). Both d_1 and d_2 were introduced as step signals. In training trials, one second after the offset of d_2 (SF2) a target signal, $z_{tgt}(t)$, was presented for a time T_{tgt} (0.5 s). This target was a beta distribution whose sign indicated the correct decision, i.e., positive for decision d_1 Longer, $L \equiv$ "$d_1 > d_2$", and negative for decision d_1 Shorter, $S \equiv$ "$d_1 < d_2$". Finally, the duration of the inter-trial-interval (ITI) was variable and it was defined so that all trials have the same duration (5.5 s).

Network model. We considered sRNNs of N=500 leaky integrate-and-fire (LIF) neurons. We will use bold characters to denote arrays and subscripts to denote array elements. The dynamics of the membrane potential of a spiking neuron i ($i = 1, \ldots, N$), $V_i(t)$, is

$$\tau_m \frac{dV_i(t)}{dt} = V_{rest} - V_i(t) + \sum_{j=1}^{N} \left(W^s_{i,j} s_j(t) + W^f_{i,j} f_j(t) \right) + \sum_{k=1}^{N_{in}} W^{in}_{i,k} F^{in}_k(t) + I, \tag{4}$$

where $\tau_m = 10$ ms is the membrane time constant and $V_{rest} = -65$ mV is the rest potential. Once the membrane potential reaches a threshold value ($V_{th} = -55$ mV) an action potential (spike) is produced. Immediately after an action potential, the membrane potential is reset to $V_{reset} = V_{rest}$. Whenever a neuron j fires a spike, its corresponding synaptic currents, $s_j(t)$ and $f_j(t)$, are incremented by 1, otherwise

$$\tau_s \frac{ds_j}{dt} = -s_j(t) \quad \text{and} \quad \tau_f \frac{df_j}{dt} = -f_j(t), \tag{5}$$

where $\tau_s = 100$ ms for the slow current, $s_j(t)$, and $\tau_s = 5$ ms for the fast current, $f_j(t)$. Slow (fast) current flows from a pre-synaptic neuron j to a post-synaptic neuron i through the hidden recurrent connections $W^s_{i,j}$ ($W^f_{i,j}$). An input layer of $N_{in} = 2$ neurons introduces both stimuli (d_1, d_2) into the network as external input signals, $F^{in}_k(t)$ ($k = 1, ..., N_{in}$), through the input weights $W^{in}_{i,k}$, defined from a uniform distribution. I is an external current.

Full-FORCE training algorithm. The full-FORCE algorithm was used as a target-based supervised learning method to train sRNNs [36]. The full-FORCE scheme trains a sRNN (called task-performing network) to generate the desired target output, $z_{tgt}(t)$, using target functions for the recurrent currents, $F^{\bar{W}}_i(t)$ (where $i = 1, ..., N$), obtained from an auxiliary RNN of rate neurons (known as target-generating network) (Supplementary Fig. S1). These target functions provide the necessary information about the currents to train the recurrent connections of the sRNN.

Target-generating network. The objective of this network of $\bar{N}$=500 recurrently connected firing-rate units is to provide targets to the sRNN. Therefore, the target-generating network must be driven by

the desired target signals, $z_{tgt}(t)$, in addition to the input events, $\boldsymbol{F^{in}}(t)$. The dynamics of the activity $\bar{x}_i(t)$ of neuron i ($i = 1, \ldots, \bar{N}$) from the target-generating network is governed by

$$\tau_x \frac{d\bar{x}_i(t)}{dt} = -\bar{x}_i(t) + \bar{g} \sum_{j=1}^{\bar{N}} \bar{W}_{i,j} H(\bar{x}_j(t)) + \sum_{k=1}^{\bar{N}_{in}} \bar{W}^{in}_{i,k} F^{in}_k(t) + \bar{U}_i z_{tgt}(t). \tag{6}$$

The non-plastic weights $\bar{W}_{i,j}$ are built from a Gaussian distribution of zero mean and variance $1/\bar{N}$. $H(\bar{x}_i) = tanh(\bar{x}_i)$ is the rate activity. $\bar{W}^{in}_{i,k}$ ($k = 1, ..., \bar{N}_{in}$ where $\bar{N}_{in} = 2$) and $\bar{U}_i$ (only one output neuron is considered) are non-plastic synapses that introduce the input signal, $\boldsymbol{F^{in}}(t)$, and the target signal, $z_{tgt}(t)$, respectively, into the network. Both synapses were built from uniform distributions. Finally, $\bar{g} = 1.4$ is a gain factor.

The auxiliary target functions, $\boldsymbol{F^{\bar{W}}}(t)$, are transferred as targets to the task-performing network

$$\boldsymbol{F^{\bar{W}}}(t) = \boldsymbol{v}\Big(\bar{g}\bar{\boldsymbol{W}} H(\bar{\boldsymbol{x}}(t)) + \bar{\boldsymbol{U}} z_{tgt}(t)\Big). \tag{7}$$

The $N \times \bar{N}$ matrix $\boldsymbol{v}$ is used to solve a possible mismatch between the dimensions of $\boldsymbol{F^{\bar{W}}}$, which is N, and of the target-generating network, which is $\bar{N}$.

Training of task-performing networks and learning rule. All the output weights $\boldsymbol{W^{out}}$ and the post-synaptic weights (slow and fast) of $n = 200$ randomly chosen neurons were trained (Supplementary Fig. S1, red lines). The slow and fast currents flowing through the plastic recurrent synapses were concatenated in a $2n$-dimensional vector $\boldsymbol{c}(t)$ while the corresponding matrix $\boldsymbol{W}$ was an Nx$2n$ array. In each training trial both $\boldsymbol{W}$ and $\boldsymbol{W^{out}}$ were adjusted to generate the optimal network response. Considering the auxiliary functions, $\boldsymbol{F^{\bar{W}}}(t)$, and the internal plastic currents, $\boldsymbol{c}(t)$, the recurrent hidden weights were modified so that the following condition was met

$$\boldsymbol{W}\boldsymbol{c}(t) \approx \boldsymbol{F^{\bar{W}}}(t), \tag{8}$$

while for the output weights it was imposed that

$$z(t) = \boldsymbol{W^{out}}\boldsymbol{c}(t) \approx z_{tgt}(t), \tag{9}$$

where $z(t)$ is the output signal of the task-performing network given by an output layer of $N_{out} = 1$ neuron. The recurrent hidden weights were trained minimizing the cost function

$$C_W = \langle (\boldsymbol{W}\boldsymbol{c}(t) - \boldsymbol{F^{\bar{W}}}(t))^2 \rangle, \tag{10}$$

while the output synapses, $\boldsymbol{W^{out}}$, were modified minimizing

$$C_{W^{out}} = \langle (z(t) - z_{tgt}(t))^2 \rangle. \tag{11}$$

The optimization and update were done using the Recursive Least Squares Algorithm (RLS) [50].

Test trials. In test trials, a network output was considered correct if $z_{tgt}(t)$ and $z(t)$ fulfilled the following condition during the presentation of the target [51]

$$\frac{\sum_{t \in T_{tgt}} z(t) z_{tgt}(t)}{\sqrt{\sum_{t \in T_{tgt}} z(t)^2} \sqrt{\sum_{t \in T_{tgt}} z_{tgt}(t)^2}} > 0.5. \tag{12}$$

Conversely, a trial was taken as wrong if $z(t)$ satisfied Eq. (12) with $z_{tgt}(t)$ replaced by $-z_{tgt}(t)$. In case $z(t)$ did not satisfy any of these two conditions, then it was considered as a noisy response. These trials were a small fraction of the test trials and were not used in the analyses.

Numerical experiments and simulation details. *Training and simulations in the main numerical experiment.* Thirty five different networks (obtained from different random initializations of the network parameters) were trained in a block of 5,000 training trials, presented in a continuous manner, without perturbing the network activity at the end of the trial. Each training trial followed the scheme in Fig. 1a. During the training stage, classes were selected randomly from Fig. 1b, excluding the central class (370, 370) ms. Once a network was trained, test trials were collected in the same way as the training trials, except that the target was not presented. During the testing stage, we considered three different trial blocks: the full block, the horizontal block and the diagonal block. Test trials in the full block used all the classes shown in Fig. 1b (including the central class). Trials in the horizontal block only considered classes with $d_2 = 370$ ms. Finally, trials in the diagonal block only used the diagonal classes (classes labeled from 1 to 10, Fig. 1b). Horizontal classes from the full block were used to compute psychometric curves (explained later). The horizontal and the diagonal blocks were employed to verify that the networks learned that these blocks define different contexts (Fig. 3d-e). All the other analyses were based on the diagonal block. Behavior was analyzed with 50,000 test trials per network. Analyses of neural activity were done with 4,000 test trials (when the sensory history was not considered) or with 20,000 test trials (for studies conditioning on the preceding class).
Training and simulations with several delay-period durations. In a second numerical experiment (Fig. 2b), fifteen networks were trained to perform the task with four different values of the delay period. In each training trial, one of the four possible durations was randomly selected from the set: 1,000 ms, 1,500 ms, 2,000 ms and 2,500 ms. Once a network was trained, a block of 4,000 consecutive test trials was generated for each duration of the delay-period, selecting classes from the diagonal block.
Training and simulations with neural activity reset. In a third numerical experiment (Fig. 2c), 2,000 ms before the Start Cue of the following trial, the neural activity was reset to a random initial state, both on training and test trials. Seven networks were trained and tested with this experiment. A dataset of 4,000 test trials from the diagonal block and another dataset of 4,000 test trials from the full block were obtained for each network.

Data analyses: Behavior. *Psychometric function*: This curve plots the probability $P(S)$ to decide "$d_1 < d_2$" as a function of the values of first interval, d_1 (Fig. 1d; Fig. 2c, left; Extended Data Fig. 1a; Extended Data Fig. 2a; Supplementary Fig. S2). It is computed using test trials with horizontal classes from the full block. It was fitted to a logistic sigmoid

$$P(S) = \frac{1}{1 + e^{A_1 + A_2 d_1}}, \tag{13}$$

where A_1 and A_2 are adjustable parameters.
Discrimination function (accuracy): This curve is the percentage of correct trials as a function of the first interval (Fig. 1e; Fig. 2a and c, middle; Extended Data Fig. 1b and c; Extended Data Fig. 2b). It uses test trials with classes from the diagonal block. Its purpose is to analyze the network

discrimination properties.

Phenomenological measure of the contraction bias. Denoting the probability of deciding S given a presented class C as $P(S|C)$, a phenomenological measure B of the contraction bias (Fig. 1g; Fig. 2c, right; Fig. 5c and f; Fig. 6) was defined as the root-mean-square deviation of $P(S|C)$ on diagonal classes, relative to the classes with $d_1 = 370$ ms (classes 3 and 8 in Fig. 1b)

$$B^2 = \frac{1}{10}\left[\sum_{uC=1}^{5}\Big(P(S|uC) - P(S|3)\Big)^2 + \sum_{lC=6}^{10}\Big(P(S|lC) - P(S|8)\Big)^2\right], \tag{14}$$

where uC and lC denote classes on the upper and the lower diagonals of the stimulus set (Fig. 1b), respectively. Note that the value of B^2 corresponding to the maximum possible bias equals 2/5, while for the minimum possible bias it is 0.

Data analyses: Neural Activity. *Firing Rates.* Trials were aligned at the relevant task events: SO1 (presentation of the first stimulus), SF1 (delay period) or SF2 (inter-trial-interval, ITI). Firing rates were calculated by sliding a 150 ms window by 25 ms intervals, except in the analysis of the relationship between behavior and activity (see below), where a time window of size 100 ms sliding every 10 ms was used. Trial-averaged firing rates were estimated conditioning over the value of d_1.

Neural Tuning. During the delay period, the firing rates of a large population of neurons coded the duration of the first time interval parametrically, either in a monotonically increasing or decreasing fashion. For this classification, we fitted the firing rate of each neuron with a linear function $\alpha d_1 + \beta$. If the slope of the fit was significant and positive, then the neuron was said to be positively tuned and it was classified as *plus*. Conversely, if the slope was significant and negative, the neuron was negatively tuned and it was classified as *minus*. From the 500 neurons of the example network 1, 137 neurons were classified as plus and 97 as minus, using test trials from the diagonal block. In the full block there were 135 plus neurons and 90 minus neurons. Example network 2 had 116 plus neurons and 114 minus neurons using the diagonal block while with the full block there were 115 plus neurons and 97 minus neurons.

Mutual Information. The information that neurons carried about the duration of the intervals was measured by the mutual information (MI) [52]. The single neuron response was quantified as a discrete array $r = \{r_1, ..., r_L\}$, where each r_t indicates the number of spikes fired in the t-th time bin and L is the total number of bins. The bins, of size $\Delta t = 250$ ms, were shifted by steps of 25 ms. D_1 collected all possible values of d_1 and R collected all values of r. The MI about a time interval d_1 carried by a neuron is

$$I(D_1|R) = \sum_{d_1} P(d_1) \sum_{r} P(r|d_1) \log_2 \frac{P(r|d_1)}{P(r)}, \tag{15}$$

where $P(d_1)$ is the probability of a time interval d_1 and $P(r|d_1)$ is the conditional probability of observing a response r given that the presented interval was d_1. To compute Eq. (15) we used a MATLAB-based toolbox [53] that includes corrections for a bias in the estimate of the MI due to limited-sampling of the neural response [54]. This bias was estimated using an extrapolation quadratic in the number of trials, N_{trial},

$$bias = \frac{a}{N_{trial}} + \frac{b}{(N_{trial})^2}, \tag{16}$$

where a and b are parameters that depend on the stimulus-response probabilities.

Once the MI from each single neuron was obtained (Supplementary Fig. S3), we analysed its significance. For that, we used a bootstrap procedure given by the toolbox, which consists in pairing randomly stimuli and responses in order to destroy its correlation [53]. The MI of a single neuron at a given time was significant if its value was higher than the sum of the corresponding bootstrapped MI and the standard deviation over permutations used in the bootstrap procedure. Then, the percentage of neurons with significant MI through time was calculated (Fig. 4c-d). We implemented a second shuffling procedure to destroy causality between the responses of consecutive trials. This shuffling consisted of putting together all firing rates of all trials of all neurons and performing 25 random permutations without replacement. From this shuffled dataset, we calculated, bin-by-bin, the percentage of neurons with significant mutual information and its mean ± CI's (grey horizontal shadings in Fig. 4c-d).

Data analyses: State-space Analysis and Geometry. *Principal Component Analysis* (PCA). We used PCA to examine the trial-averaged firing rates and to identify the low-dimensional space that captured at least 90% of data variance. Neural activity described curved trajectories embedded in this low-dimensional space; the associated speed was computed as the norm of the covered distance by a time window of 150 ms, which slides 25 ms (Fig. 3c).

KiNeT Method for relative distances. Relative distance between a pair of trajectories (i and j) was defined following the Kinematic analysis of Neural Trajectory (KiNeT) methodology [19]. We first selected one of the two trajectories as a reference (e.g., $ref{=}i$) and then defined corresponding states between the reference and the other trajectory (j) as follows: If at time t after the aligning event (SO1 for the presentation epoch of d_1, SF1 for the delay period and SF2 for the ITI) the reference trajectory was in the state $s_{ref}(t)$, then the corresponding state $s_j(t)$ on trajectory j was the one at the minimum Euclidean distance to $s_{ref}(t)$. This is the relative distance $D_j(t)$ between the two trajectories. Relative distances were computed in a PC space which accounted for at least 90% of the total variance during the investigated epoch. This analysis has been used in the following situations:

1) State-space trajectories depend on context. To verify that classes common to horizontal and diagonal blocks (classes 5 and 6) defined different trajectories (Fig. 3d-e), their relative distances were computed, taking the trajectory in the diagonal context as a reference. Standard errors were computed using the mean ± 95% confidence intervals (CI), calculated from the PCA and the corresponding KiNeT analysis performed on each of the 100 bootstrap resamples with replacement.

2) State-space trajectories depend on d_1. As a first check whether the duration of the first interval was transmitted to the delay interval and the ITI, we verified that the relative distances between trajectories with different values of d_1 were significantly nonzero (Fig. 4a-b). Standard errors were computed as in the previous situation.

3) Relative distances have been also used to examine whether they encoded an estimator of d_1 during the delay period (Fig. 7d; Extended Data Fig. 7d; Extended Data Fig. 8g and j). See details below.

Bayesian Model. *Formulation and Likelihood.* Behavioral data were fitted using a Bayesian model. We assumed that at the current trial n the Bayesian observer makes an observation of the first interval ($o_{1,n}$) given by a combination of the current-trial noisy measurement of d_1 ($d'_{1,n}$) and the observation in the preceding trial

$$o_{1,n} = (1-\eta)\, d'_{1,n} + \eta o_{1,n-1}. \tag{17}$$

Here η is a parameter that measures the contribution of the observation in the preceding trial (if $\eta = 0$ there is no contribution from sensory history while if $\eta = 1$ the current stimulus does not contribute). To work with analytical expressions, instead of iterating this equation we expressed $o_{1,n}$ by extracting the interval measurements made in the m past trials, named as the short-term sensory history (Eq. (1) in the main text)

$$o_{1,n} = \sum_{j=0}^{m} \eta^j (1-\eta)\, d'_{1,n-j} + \eta^{m+1} o_{1,n-(m+1)}. \tag{18}$$

We assumed that the noisy measurements of the intervals can be described by normal distributions ($j = 0, 1, ..., m$)

$$d'_{1,n-j} \sim \mathcal{N}\left(d_{1,n-j}, W_1^2 d_{1,n-j}^2\right), \tag{19}$$

where W_1 is the Weber fraction, a parameter of the Bayesian model to be determined. In addition, we assumed that trials prior to the last m constitute a Gaussian stationary regime (the long-term sensory history). The mean and variance of the stationary variable are obtained by self-consistency arguments

$$M_{1,st} = \langle d_1 \rangle, \tag{20}$$

$$\Sigma_{1,st}^2 = \frac{1-\eta}{1+\eta}\left(W_1^2 \langle d_1^2 \rangle + \sigma_{d_1}^2\right), \tag{21}$$

where $\langle d_1 \rangle$ and $\sigma_{d_1}^2$ are the mean and variance of the prior probability of d_1. Within this Gaussian approximation the observation in the current trial is also a Gaussian variable, with mean $\mathcal{M}$ and variance Σ_1^2, that we denote by $\mathcal{O}_1$

$$\mathcal{O}_1 \sim \mathcal{N}\left(\mathcal{M}, \Sigma_1^2\right). \tag{22}$$

The distribution of this variable defines the likelihood of an observation $\mathcal{O}_1$, given the short-term sensory history, that we denote by $P\left(\mathcal{O}_1 | d_{1,n}, \ldots, d_{1,n-m}\right)$. The mean and variance of the observation probability for the general case, in which the transient consists of the previous m trials, is given by

$$\mathcal{M} = d_{1,n} + \sum_{j=1}^{m} \eta^j \left[d_{1,n-j} - d_{1,n-(j-1)}\right] + \eta^{m+1}\left[\langle d_1 \rangle - d_{1,n-m}\right], \tag{23}$$

$$\Sigma_1^2 = \sum_{j=0}^{m} \eta^{2j} (1-\eta)^2 W_1^2 d_{1,n-j}^2 + \eta^{2(m+1)} \Sigma_{1,st}^2. \tag{24}$$

We considered two cases of the model, $m = 0$ and $m = 1$. In the first case all past trials contribute to the stationary regime. The likelihood $P\left(\mathcal{O}_1 | d_{1,n}\right)$ has mean and variance

$$\mathcal{M} = d_1 + \eta(\langle d_1 \rangle - d_1), \tag{25}$$

$$\Sigma_1^2 = (1-\eta)^2 W_1^2 d_1^2 + \eta^2 \Sigma_{1,st}^2. \tag{26}$$

In the second case ($m = 1$) the transient regime contains only the preceding trial. The likelihood $P(\mathcal{O}_1 | d_{1,n}, d_{1,n-1})$ has mean and variance

$$\mathcal{M} = d_{1,n} + \eta \Big[d_{1,n-1} - d_{1,n} + \eta \left(\langle d_1 \rangle - d_{1,n-1} \right) \Big], \tag{27}$$

$$\Sigma_1^2 = (1-\eta)^2 W_1^2 d_{1,n}^2 + \eta^2 (1-\eta)^2 W_1^2 d_{1,n-1}^2 + \eta^4 \Sigma_{1,st}^2. \tag{28}$$

A more general model, which considers the contribution of the noisy measurements of the second interval in the previous trial ($d'_{2,n-1}$), is given in Supplementary Text S1.

Beliefs. The computation of posterior probabilities closely follows a Bayesian model for tactile frequency discrimination [9]. For the sake of completeness, here we adapt it to the time interval discrimination task. The main difference is that the likelihoods are given by the expressions derived above. We present the equations only for the case $m = 1$. The discrete prior probabilities of the first interval are denoted as $P(d_1^i)$ ($i = 1, \ldots, 5$ labels the possible values of d_1 in the stimulus set, Fig. 1b). Bayes theorem establishes that the posterior distribution $P(d_1^i | \mathcal{O}_1, d_{1,-1}^j)$, conditioned to the preceding interval $d_{1,-1}^j$ ($j = 1, \ldots, 5$), results from combining the noisy observation with the prior distribution,

$$b_1(d_{1,0}^i | d_{1,-1}^j) \equiv P(d_{1,0}^i | \mathcal{O}_1, d_{1,-1}^j) \propto P(\mathcal{O}_1 | d_{1,0}^i, d_{1,-1}^j) P(d_{1,0}^i). \tag{29}$$

Notice that the notation was simplified denoting $d_{1,n}$ and $d_{1,n-1}$ as $d_{1,0}$ and $d_{1,-1}$, respectively. The left hand side indicates that this probability is also known as the belief $b_1(d_{1,0}^j | d_{1,-1}^j)$ about the duration of the first time interval, conditioned to the duration in the previous trial. One may need to marginalize over $d_{1,-1}^j$, summing over all values of j.

Denoting the discrete prior probabilities of the second time interval as $P(d_2^l)$ ($l = 1, \ldots, 9$ labels the possible values of d_2 in the stimulus set, Fig. 1b), the posterior probability $P(d_2^l | \mathcal{O}_2)$, or belief $b_2(d_2^l)$ about the duration of the second time interval, is

$$b_2(d_2^l) \equiv P(d_2^l | \mathcal{O}_2) \propto P(\mathcal{O}_2 | d_2^l) P(d_2^l). \tag{30}$$

Let us denote class k as $C^k = (d_1^k, d_2^k)$, where $k = 1, \ldots, 10$ is a class label (Fig. 1b). The joint posterior probabilities (or joint beliefs) that class C^k had been presented in the current trial (the observations are $\mathcal{O}_1$ and $\mathcal{O}_2$), conditioned to the interval in the previous trial are [9]

$$B(C^k | \mathcal{O}_1, \mathcal{O}_2, d_{1,-1}^j) = \frac{P(\mathcal{O}_2)}{P(\mathcal{O}_2 | \mathcal{O}_1)} \frac{P(d_2 = d_2^k | d_1 = d_1^k)}{P(d_2^k)} b_1(d_1^k | d_{1,-1}^j) b_2(d_2^k). \tag{31}$$

The first quotient is a normalization factor. $P(d_2 = d_2^k | d_1 = d_1^k)$ is the transition probability matrix of d_2 being presented after d_1. If the network has good knowledge of the structure of the set, the only non-zero transition matrix elements correspond to the 10 classes in the set. Note that this matrix depends on the block under consideration (full, diagonal or horizontal blocks, Fig. 1b). For example,

in the full block, for a given d_1 the second interval can take either one or two or three different values. Thus, the non-zero transition probabilities would be 1, 1/2 or 1/3, respectively (Fig. 1c).

The Bayesian observer makes a decision based on which of the choices, "$d_1 < d_2$" or "$d_1 > d_2$", has a higher belief. These beliefs are computed by adding the class beliefs $B(k|\mathcal{O}_1^*, \mathcal{O}_2)$ over classes k with a specific choice

$$\begin{cases} b(L|d_{1,-1}^j) = \sum_{d_1>d_2} B(k|\mathcal{O}_1, \mathcal{O}_2, d_{1,-1}^j) & \text{belief about } d_1 \text{ Longer (L="}d_1 > d_2\text{")}, \\ b(S|d_{1,-1}^j) = 1 - b(L|d_{1,-1}^j) & \text{belief about } d_1 \text{ Shorter (S="}d_1 < d_2\text{")}. \end{cases} \tag{32}$$

Probability of deciding "$d_1 < d_2$". For the model version $m = 0$, the probability that the Bayesian observer, with knowledge of $\mathcal{O}_1$ and $\mathcal{O}_2$, chooses "$d_1 < d_2$" when the current class is $C_0^k = (d_{1,0}^k, d_{2,0}^k)$ is

$$P(S|C_0^k) = \sum_{\mathcal{O}_1,\mathcal{O}_2} P(\mathcal{O}_1|d_{1,0}^k)P(\mathcal{O}_2|d_{2,0}^k)G(b(S)), \tag{33}$$

where $G(b(S)) = 1$ if $b(S) > 0.5$ and zero otherwise [9], [15].

Likewise, for the model version $m = 1$, the conditional probability to choose "$d_1 < d_2$" when, in addition, the previous class is $C_{-1}^l = (d_{1,-1}^l, d_{2,-1}^l)$ is

$$P(S|C_0^k, d_{1,-1}^l) = \sum_{\mathcal{O}_1,\mathcal{O}_2} P(\mathcal{O}_1|d_{1,0}^k, d_{1,-1}^l)P(\mathcal{O}_2|d_{2,0}^k)G(b(S|d_{1,-1}^l)). \tag{34}$$

Choice bias. The trained networks showed a tendency to repeat the choice of the previous trial when it had been correct or make the opposite choice when it had not, that is, they followed a win-stay/loss-shift (WSLS) strategy. We modelled this choice bias by modifying the argument of the decision factor G in Eq. (34) which now depends on the outcome of the previous trial, U_{-1}, that takes the values 1 or 0 for correct and incorrect previous choices, respectively. Moreover, the choice bias depended on which diagonal of the stimulus set (Fig. 1b) the previous class was on. So, instead of Eq. (34) we now have

$$P(S|C_0^k, C_{-1}^l, U_{-1}) = \sum_{\mathcal{O}_1,\mathcal{O}_2} P(\mathcal{O}_1|d_{1,0}^k, d_{1,-1}^l)P(\mathcal{O}_2|d_{2,0}^k)G(b(S|d_{1,-1}^l), U_{-1}, q), \tag{35}$$

where $q = 1$ ($q = 2$) denotes the upper (lower) diagonal. Now the function G is

$$G(b(S|d_{1,-1}^l), U_{-1}, q) = \begin{cases} 1 & \text{if } (1-\gamma_q)b(S|d_{1,-1}^l) + \gamma_q > 0.5 \text{ and } U_{-1} = 1 \\ 0 & \text{if } (1-\gamma_q)b(S|d_{1,-1}^l) + \gamma_q < 0.5 \text{ and } U_{-1} = 1 \\ 1 & \text{if } (1-\gamma_q)b(S|d_{1,-1}^l) > 0.5 \text{ and } U_{-1} = 0 \\ 0 & \text{if } (1-\gamma_q)b(S|d_{1,-1}^l) < 0.5 \text{ and } U_{-1} = 0 \end{cases} \tag{36}$$

where the parameters γ_1, γ_2 bias choices S and L according to the WSLS strategy. Notice that Eq. (35) yields 20 performance curves defined on current classes (one per previous class and outcome), which can be obtained from the simulated behavioral data (Fig. 2a).

Correlations between the bias B and the model parameters. Correlations (between the model parameters with each other and with the bias B) in Fig. 6 were computed with the MATLAB function

corrcoef. Correlations were considered significant if the p-value was smaller than 0.05.

Fitting the Bayesian Model to Data. Both versions of the model ($m = 0$ and $m = 1$) were fitted by minimizing mean squared differences between the model and network performances. Minimizations were done using MATLAB's *simulannealbnd* function which implements a simulated annealing algorithm. For the case $m = 0$, the model fit (Fig. 5a) was done by minimizing the mean squared difference between the performance curve in Eq. (33) and the corresponding behavioral curve obtained from simulations. For the case $m = 1$ (conditioning over the previous class), the model fit was done using the 20 performance curves in Eq. (35). The minimized error resulted from summing over the 20 squared differences between the model and the network performances (one difference per previous class and outcome). Then, from the obtained fit we computed the 10 accuracy curves conditioned to the previous class (Fig. 5d) and the marginal accuracy (Fig. 5g), by marginalization, as explained next.

Marginalization over previous class and outcome. From the simulated behavioral data we know the performance curves $P(S|C_0^k,\, C_{-1}^l,\, U_{-1} = 1)$, conditioned on the current class C_0^k, the previous class C_{-1}^l and the previous outcome U_{-1}. From these curves, the parameters of the version $m = 1$ of the model were obtained as described above. Now, we want to obtain the fitted marginal discrimination curve $P(U_0 = 1|C_0^k)$ ($k = 1, \ldots, 10$) by marginalizing $P(S|C_0^k,\, C_{-1}^l,\, U_{-1} = 1)$ on the previous class and outcome. For this, it is necessary to solve a system of 10 equations with 10 unknowns (for further details see Supplementary Text S2). The first 5 equations of this system correspond to the 5 current classes on the upper diagonal ($k = 1, \ldots, 5$)

$$\sum_{l=1}^{10} \left[P(S|C_0^k,\, C_{-1}^l,\, U_{-1} = 1) - P(S|C_0^k,\, C_{-1}^l,\, U_{-1} = 0)\right] \cdot\, P(U_0 = 1|C_0^l) \\ - 10P(U_0 = 1|C_0^k) = -\sum_{l=1}^{10} P(S|C_0^k,\, C_{-1}^l,\, U_{-1} = 0), \quad (37)$$

with 10 unknowns $P(U_0 = 1|C_0^l)$ ($l = 1, \ldots, 10$). The other 5 equations correspond to the 5 current classes on the bottom diagonal ($k = 6, \ldots, 10$)

$$\sum_{l=1}^{10} \left[P(S|C_0^k,\, C_{-1}^l,\, U_{-1} = 1) - P(S|C_0^k,\, C_{-1}^l,\, U_{-1} = 0)\right] \cdot\, P(U_0 = 1|C_0^l) \\ + 10P(U_0 = 1|C_0^k) = 10 - \sum_{l=1}^{10} P(S|C_0^k,\, C_{-1}^l,\, U_{-1} = 0). \quad (38)$$

Solving the system in Eqs. (37-38) we obtained the marginal accuracy curve, $P(U_0 = 1|C_0^l)$ ($l = 1, \ldots, 10$) (Fig. 5g). Once this is known, the probability of deciding $S \equiv$ "$d_1 < d_2$" conditioned on the current and previous classes, $P(S|C_0,\, C_{-1})$ (Fig. 5d), can be computed from

$$P(S|C_0^k,\, C_{-1}^l) = P(S|C_0^k,\, C_{-1}^l,\, U_{-1} = 1) \cdot P(U_{-1} = 1|C_{-1}^l) + \\ + P(S|C_0^k,\, C_{-1}^l,\, U_{-1} = 0) \cdot P(U_{-1} = 0|C_{-1}^l). \quad (39)$$

Model comparison. *Akaike Information Criteria* (AIC).

$$AIC = -2log(\mathcal{L}) + 2N_p, \tag{40}$$

where N_p is the number of model parameters and $\mathcal{L}$ is the likelihood of obtaining the observed choice ("$d_1 < d_2$" or "$d_1 > d_2$") across all trials. For given class and choice, the likelihood function takes the values $l(x = \text{"}d_1 < d_2\text{"}) = P(\text{"}d_1 < d_2\text{"})$ or $l(x = \text{"}d_1 > d_2\text{"}) = P(\text{"}d_1 > d_2\text{"})$. Then, $log(\mathcal{L})$ is the sum over all trials

$$log(\mathcal{L}) = \sum_{n=1}^{N_{trials}} log(l(x_i)). \tag{41}$$

Bayesian Information Criteria (BIC):

$$BIC = -2log(\mathcal{L}) + N_p\, log(N_{Obs}), \tag{42}$$

where N_{Obs} is the sample size (number of observations).

Estimation of d_1 and its relationship with state-space geometry. To check whether geometric features of the trajectories encoded an estimate of the first interval we compared separately three estimators (the true value of d_1, the Bayesian estimator $d_{1,Bayes}$ and the ML estimator $d_{1,ML}$) with measures of the curvature, relative distances and spatial structure of the quasi-attractor. To do this, we first computed the estimators in single trials simulating the model (version $m = 1$, Eqs.(27,28)). Then, we obtained the means $d_{1,Bayes}$ and $d_{1,ML}$ as an average over trials.

Bayesian and Maximum Likelihood estimators. For a given trial n, the Bayesian estimator of $d_{1,n}$ ($d_{1,n}^{Bayes}$) results from a loss function; taking it as the squared error

$$(d_{1,n}^{Bayes} - d_{1,n})^2,$$

the Bayesian estimator is the mean of d_1 weighted with the posterior probability $b_1(d_{1,n}^i)$ or $b_1(d_{1,n}^i|d_{1,n-1}^j)$,

$$d_{1,n}^{Bayes} = \sum_{i=1}^{5} d_1^i \cdot b_1(d_{1,n}^i), \tag{43}$$

$$d_{1,n}^{Bayes}(d_{1,n-1}^j) = \sum_{i=1}^{5} d_1^i \cdot b_1(d_{1,n}^i|d_{1,n-1}^j), \tag{44}$$

for the marginal and conditional cases, respectively, from model version $m = 1$. The ML estimator ($d_{1,n}^{ML}$) is the value of d_1 that maximizes the likelihood function,

$$d_{1,n}^{ML} = \arg\max_{d_1} P(\mathcal{O}_1|d_{1,n}^i), \tag{45}$$

$$d_{1,n}^{ML}(d_{1,n-1}^j) = \arg\max_{d_1} P(\mathcal{O}_1|d_{1,n}^i, d_{1,n-1}^j). \tag{46}$$

Once the model parameters have been determined, we simulated 10,000 trials (marginal case) and 10,000 trials for each $d_{1,n-1}^j$ (conditional case) and computed the corresponding Bayesian (Eqs. (43-44)) and ML (Eqs. (45-46)) estimators. Their trial averages yielded $d_{1,Bayes}$ and $d_{1,ML}$, respectively. The bias of the Bayesian estimator (Supplementary Fig. S8) was obtained from

$$B_{Bayes} = \sqrt{\frac{1}{5}\sum_{i=1}^{5}\left(d_1^i - d_{1,Bayes}^i\right)^2}. \tag{47}$$

Curvature of the one-dimensional curved manifold. To investigate the state-space representation of d_1 during its presentation epoch we analyzed the curvature of the one-dimensional manifold developed during that period [21]. We first projected the neural trajectories onto an encoding axis, $\mathbf{u}$, defined as a unit vector between the state $s(\text{t=SF1}_1)$ associated to the shortest SF1, and the state $s(\text{t=SF1}_5)$ associated to the longest SF1. Then, for each trajectory j ($j = 1, \ldots, 5$) we computed the projection, $Proj_j$, of the state $s(\text{t=SF1}_j)$ onto the encoding axis

$$Proj_j = \frac{\overrightarrow{s(SF1_1)s(SF1_j)} \cdot \mathbf{u}}{||\mathbf{u}||^2}\,\mathbf{u} + s(SF1_1). \tag{48}$$

To test which of the three estimators best explained the encoding axis projection data during the presentation of d_1 we fitted them, separately, to

$$d_1;\ d_{1,Bayes};\ d_{1,ML} = aP_j + b, \tag{49}$$

where P_j ($j = 1, \ldots, 5$) are the distances between projections on an encoding axis [21], taking as reference the projection of the shortest d_1. Again, $d_{1,Bayes}$ and $d_{1,ML}$ are the average of the estimators over the simulated trials obtained either with Eqs. (43) and (45) (marginalizing over the previous stimulus) or with Eqs. (44) and (46) (conditioning over the previous stimulus). Firing rates were computed using a 100 ms analysis window, with a slide of 10 ms throughout the presentation of the first interval. We kept the number of PCs needed to explain, at least, 95% of the variance (between 3 and 5 PCs). This analysis was done for the marginal case (Fig. 7a-c and Extended Data Fig. 7a-c) and the conditional case (Extended Data Fig. 8a-f).

Mean relative distances during the delay period. To test which of the three estimators best explained the relative distances between neural trajectories (taking as reference the trajectory corresponding to d_1=370 ms; KiNeT method) during the delay period, we compared the three hypotheses

$$d_1;\ d_{1,Bayes};\ d_{1,ML} = aD_j + b, \tag{50}$$

where D_j ($j = 1, \ldots, 5$) is the temporal mean of the relative distances [20]. We used an analysis window of 100 ms with a slide of 10 ms throughout the delay period. We kept the number of PCs needed to explain, at least, 90% of the variance (between 5 and 11 PCs). This analysis was done for the marginal case (Fig. 7d-f and Extended Data Fig. 7d-f) and the conditional case (Extended Data Fig. 8g-l).

Distances between terminal states on the quasi-attractor: Likewise, to check which estimator is encoded by the attractor we considered the Euclidean distances δ_j ($j = 1, \ldots, 5$) between the terminal states on the attractor region of each d_1 trajectory and the shortest d_1 trajectory.

$$d_1;\ d_{1,Bayes};\ d_{1,ML} = a\delta_j + b. \tag{51}$$

The test was done in the last 100 ms of the delay epoch. Although for network 1 the first PC explained about 96% of the variance, the quality of the regressions improved by including the second PC (which

contributed 3.5% of the variance). This analysis was done for the marginal case (Fig. 7g-i and Extended Data Fig. 7g-i) and the conditional case (Extended Data Fig. 8m-r).

In these three tests the goodness of the fits was compared using their Root Mean Square Errors (RMSEs). Significance was evaluated using a bootstrap procedure. From the original neural population we created 100 resamples by repeatedly sampling with replacement. Then we repeated the tests in each resample and computed the corresponding dispersions.

Compression of geometric features. Denoting by Z_j ($j = 1, \ldots, 5$) any of the three geometric features considered in this work (projections on the encoding axis during presentation of d_1, P_j; mean relative distances during the delay period, $D_j = \langle D_j(t) \rangle$; distances to terminal states on the attractor region, δ_j), we defined the compression $C[Z]$ as

$$C[Z] = 1 - min\left(\frac{Z_3 - Z_1}{Z_5 - Z_3}, \frac{Z_5 - Z_3}{Z_3 - Z_1}\right). \tag{52}$$

Code and Data Availability

Training with full-FORCE was done adapting to the time interval discrimination task the MATLAB codes from [36], available in https://github.com/briandepasquale/supervised_learning_recurrent_spiking_networks.

Custom MATLAB scripts used to implement the model and produce the figures will be available upon publication. The simulated data that support the findings in this study are available from the corresponding author upon reasonable request.

References

[50] D. Sussillo and L. F. Abbott, "Generating coherent patterns of activity from chaotic neural networks," Neuron, vol. 63, no. 4, pp. 544–557, 2009.

[51] A. Ingrosso and L. Abbott, "Training dynamically balanced excitatory-inhibitory networks," PloS one, vol. 14, no. 8, e0220547, 2019.

[52] C. E. Shannon, "A mathematical theory of communication," The Bell system technical journal, vol. 27, no. 3, pp. 379–423, 1948.

[53] C. Magri, K. Whittingstall, V. Singh, N. K. Logothetis, and S. Panzeri, "A toolbox for the fast information analysis of multiple-site lfp, eeg and spike train recordings," BMC neuroscience, vol. 10, pp. 1–24, 2009.

[54] A. Treves and S. Panzeri, "The upward bias in measures of information derived from limited data samples," Neural Computation, vol. 7, no. 2, pp. 399–407, 1995.

Acknowledgements

This work was supported by the grant PGC2018-101992-B-I00 from the Spanish Ministry of Science, Innovation and Universities. We thank Joan Falcó-Roget for discussions.

Author contributions

L.S.-F. and N.P. designed the study. L.S.-F. implemented the training methods and performed the numerical analysis. L.S.-F. and N.P. developed the analysis techniques. M.B. and N.P. designed the normative model. L.S.-F., M.B. and N.P. performed research and contributed to writing the manuscript.

Figures

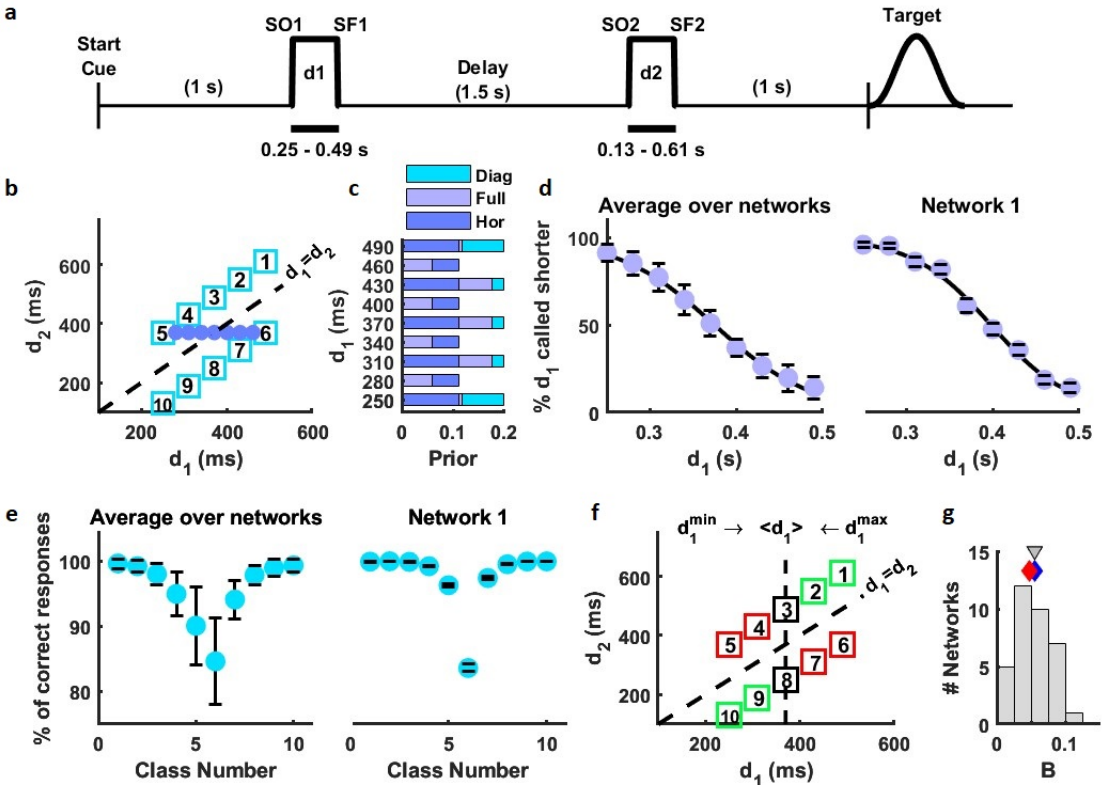


Fig. 1. Task and network behavior. **a**. Timeline of the time interval discrimination task. The Start Cue indicates the beginning of the trial. One second later, two time intervals (d_1 and d_2) are presented sequentially, separated by a delay period of 1,500 ms (SO1 and SO2 are their respective onset times, while SF1 and SF2 denote their offset times). During the training epoch, a target (a beta distribution) is presented one second after SF2. **b**. Stimulus set of possible classes or pairs (d_1, d_2), indicated by squares or blue circles. There are three block types: horizontal block (all classes with d_2=370 ms), diagonal block (classes labeled from 1 to 10) and full block, defined by all the classes. Classes 5 and 6 are common to the three blocks. **c**. Prior distributions of d_1 in each block. **d.** Psychometric curves, computed using test trials with horizontal classes from the full block. **Left**: Averaged psychometric curve over 35 networks trained and tested on consecutive trials without activity reset. Error bars are standard deviations over networks. **Right**: Psychometric curve for example network 1. Error bars were computed with 1,000 bootstrap resamples. Black lines in both psychometric curves are sigmoidal fits (Methods). **e.** Percentage of correct test trials, computed on the diagonal block. **Left**: Average accuracy over 35 networks. Error bars indicate standard deviations computed over networks. **Right**: Accuracy for example network 1. Error bars were computed with 1,000 bootstrap resamples. **f**. Contraction bias: the first interval (d_1) is coded contracted towards the mean of its distribution. This bias induces errors in some classes (red squares) while favors others (green squares). **g**. Histogram over the networks of the phenomenological bias (B) (Methods, Eq. (14)). Gray triangle indicates the histogram mean. Diamonds indicate B for example network 1 (blue) and example network 2 (red).

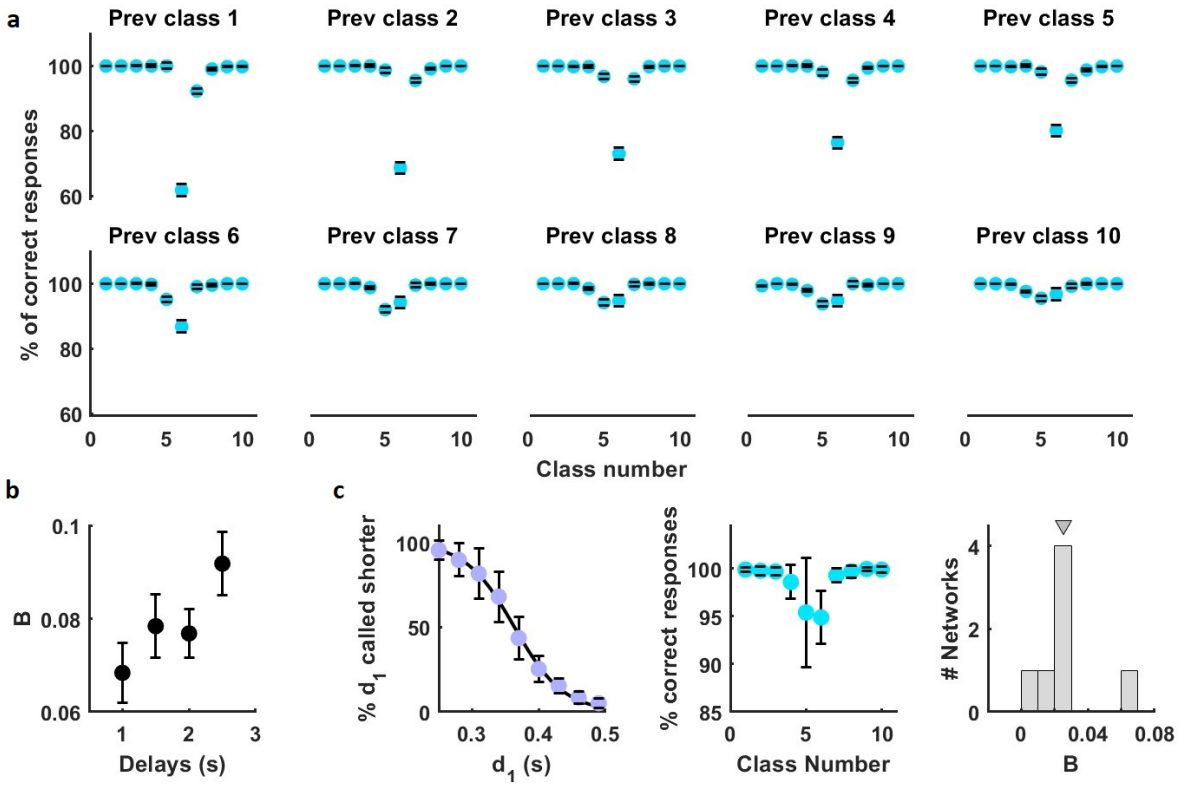


Fig. 2. Potential causes of the contraction bias. **a**. Accuracy depends on the preceding class (main numerical experiment). Accuracy conditioned on each of the ten possible previous classes (example network 1; test trials in the diagonal block). Error bars computed with 1,000 bootstrap resamples. **b**. Phenomenological bias (B) increases with delay-period duration (second numerical experiment). B vs. delay-period duration for a network trained by randomly interleaving four different delay durations (1,000, 1,500, 2,000 and 2,500 ms) and tested in four blocks of test trials, each one with a fixed delay duration (diagonal block). Error bars computed with 100 bootstrap resamples. **c**. Contraction bias in the absence of sensory history, such that the neural activity was reset to a different random initial state in all training and test trials (Methods). **Left**: Average over 7 networks of the psychometric curves. Horizontal classes collected from the full block. Black line is a sigmoidal fit (see Extended Data Fig. 2a for a single network). **Middle**: Average over 7 networks of the accuracy. Classes from the diagonal block. Error bars are standard deviations over networks (see Extended Data Fig. 2b for an example network). **Right**: Histogram of the bias B (Methods, Eq. (14)), from all 7 reset networks. Gray triangle indicates the histogram's mean.

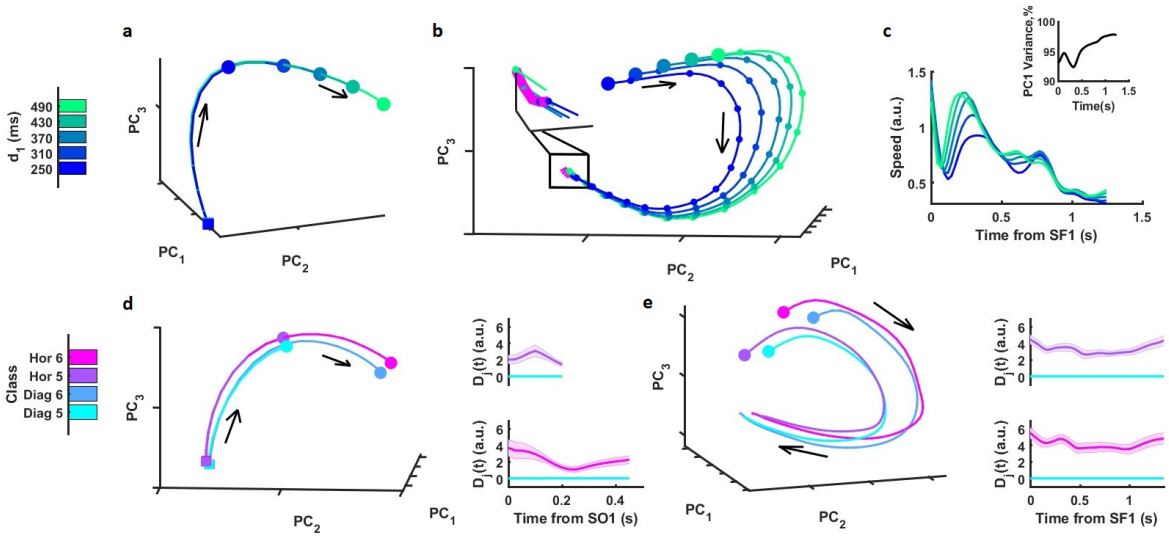


Fig. 3. State-space dynamics and contexts. **a-e**. Example network 1. **a-c**. Projection of neural trajectories onto the first three PCs. **a**. Curved manifold during presentation of d_1. Squares indicate SO1 and circles indicate SF1. **b**. Neural trajectories during the delay period. Big circles indicate the SF1s. Small circles indicate neural states at 100 ms increments. Note that the distance between consecutive circles reflects speed. **Inset** shows the quasi-attractor (magenta line), represented by connecting the terminal states at the end of the delay period. **c**. Speed of neural trajectories versus time, during the delay period (Methods). **Inset** shows the percentage of variance explained by the first PC. By the end of the delay period its variance approached 100%. **d-e**. Comparison of diagonal and horizontal contexts using the classes 5 (d_1=250 ms) and 6 (d_1=490 ms), which are common to both blocks. **d**. Geometry during the presentation of the first interval. **Left**: State-space trajectories of classes 5 and 6. **Right**: Relative distances of the classes 5 (upper) and 6 (bottom). The number of PCs needed to explain at least 90% of variance was two for class 5 and three for class 6. Shadings indicate the median ± 95% CIs, computed from 100 bootstrap resamples. **e**. Geometry during the delay period. **Left**: State-space trajectories of classes 5 and 6. **Right**: Relative distances of classes 5 (upper) and 6 (bottom). Four PCs were needed to explain at least 90% of variance. Shadings are as in (d), right. All the analyses were done with a time window of 150 ms sliding every 25 ms.

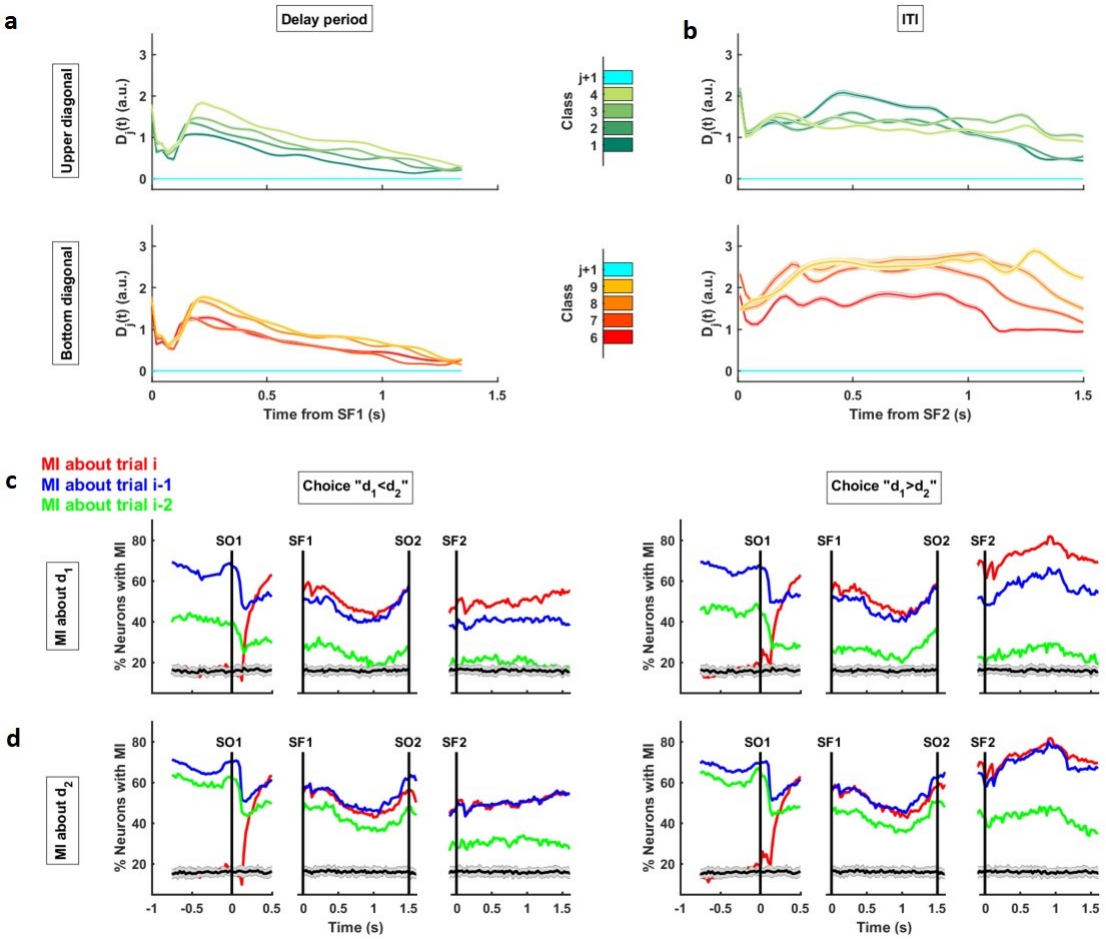


Fig. 4. Mutual Information analyses. a-d Example network 1. **a-b**. Relative distances $D_j(t)$ between pairs of consecutive classes from diagonal block ($j+1$ denotes the reference class) during the delay period (a) and the ITI (b). The (rather thin) shadings indicate the mean $\pm$ 95% CIs computed from 100 bootstrap resamples. **c-d**. Percentage of neurons with significant mutual information about the first (c) and second (d) time interval (red: from current trial i; blue: from trial $i-1$; green: from trial $i-2$), conditioning on correct trials and on choice (**left**: choice "$d_1 < d_2$"; **right**: choice "$d_1 > d_2$"). Horizontal gray shadings refer to the mean and standard deviation computed from shuffled data (Methods).

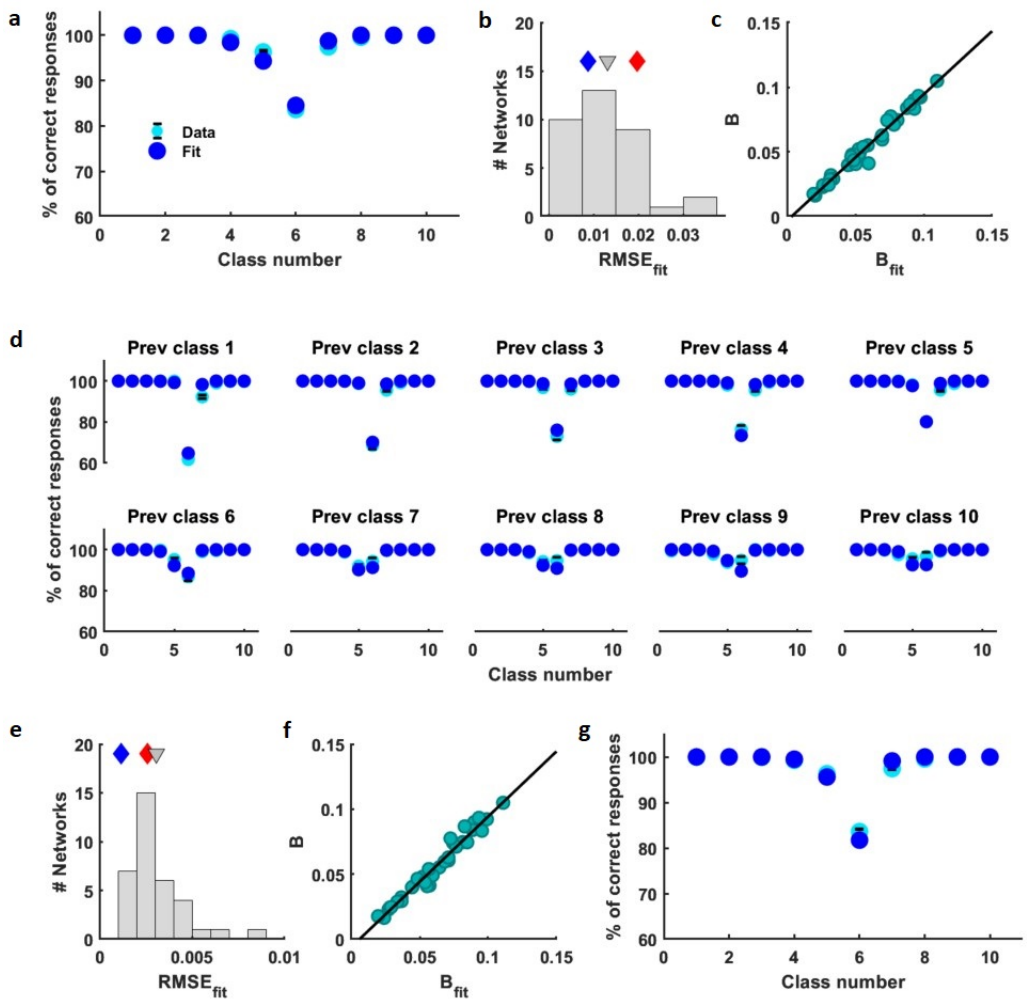


Fig. 5. Bayesian model fits. **a-c**. Fit with model version $m = 0$ (Eqs. (2a,b)). **a**. Fit of percentage of correct test trials (computed on the diagonal block) of example network 1: behavioral data (cyan dots with error bars are the data in Fig. 1e, right) and fit (dark blue circles). **b**. Histogram of RMSEs (Root Mean Square Error from the difference between the behavioral data performance and the model-fitted performance) over 35 networks. Histogram's mean is represented by gray triangle. Diamonds indicate the RMSEs values of network 1 (blue) and network 2 (red). **c**. Correlation between the phenomenological bias (B) and the bias from the model fit (B_{fit}), computed over 35 networks (correlation=0.98; p-value=5.6e-27). Both B and B_{fit} were computed with Eq. (14) (Methods). **d-g**. Fit with model version $m = 1$ (Eqs. (3a,b)). **d**. Fit of the ten accuracy curves conditioned over each possible preceding diagonal class of example network 1 (cyan circles are the behavioral data in Fig. 2a). **e**. Histogram of RMSEs of performance fits over 35 networks. **f**. Correlation between B and B_{fit} (correlation=0.97; p-value=4.2e-26). **g**. Marginal accuracy curve for network 1, obtained by marginalizing over the preceding first interval both the data and fit in panel (d) (see Methods).

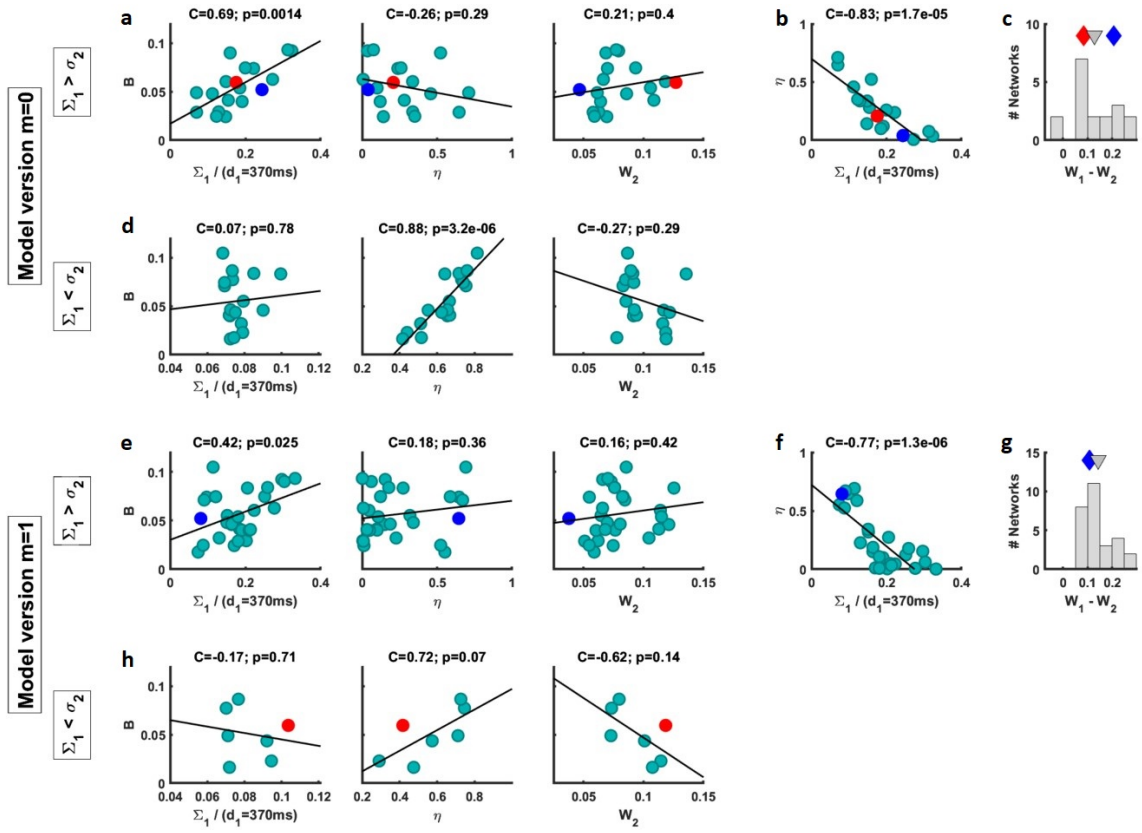


Fig. 6. Correlations of parameters from both model versions $m = 0$ and $m = 1$, sorted by network type ($\Sigma_1 > \sigma_2$ or $\Sigma_1 < \sigma_2$). a-c. Correlations over the 18 networks classified as positive ($\Sigma_1 > \sigma_2$) by model version $m = 0$. **a**. Correlations between the bias B and the model parameters. **Left**: Correlation between B and the total variance of the observation probability (Σ_1) for $d_1 = 370$ ms (for other values of d_1 see Supplementary Fig. S4a). **Middle**: Correlation between B and the sensory history parameter (η). **Right**: Correlation between B and the noise parameter of d_2 measurements (W_2). **b**. Correlation between η and Σ_1 for $d_1 = 370$ ms (see Supplementary Fig. S5). **c**. Histogram of the difference ($W_1 - W_2$). Gray triangle indicates the histogram's mean. **d**. Correlations between B and the model parameters from 17 networks classified as negative ($\Sigma_1 < \sigma_2$) by model version $m = 0$. **Left**: Correlation between B and Σ_1 for $d_1 = 370$ ms (see Supplementary Fig. S4b). **Middle**: Correlation between B and η. **Right**: Correlation between B and W_2. **e-g**. Correlations over the 28 networks classified as positive by model version $m = 1$. **e**. Correlations between B and the model parameters. **Left**: Correlations between B and Σ_1 for the current and previous first stimulus equal to 370 ms (see Supplementary Fig. S6). **Middle**: Correlation between B and η. **Right**: Correlation between B and W_2. **f**. Correlation between η and Σ_1 for the current and previous first stimulus equal to 370 ms (the rest of the cases hardly change). **g**. Histogram of $W_1 - W_2$. **h**. Correlations between B and the model parameters from 7 networks classified as negative by model version $m = 1$. **Left**: Correlation between B and Σ_1 for the current and the previous first interval equal to 370 ms (see Supplementary Fig. S7). **Middle**: Correlation between B and η. **Right**: Correlation between B and W_2. Blue markers refer to network 1 and red markers refer to network 2. C indicates correlation values and p is the p-value.

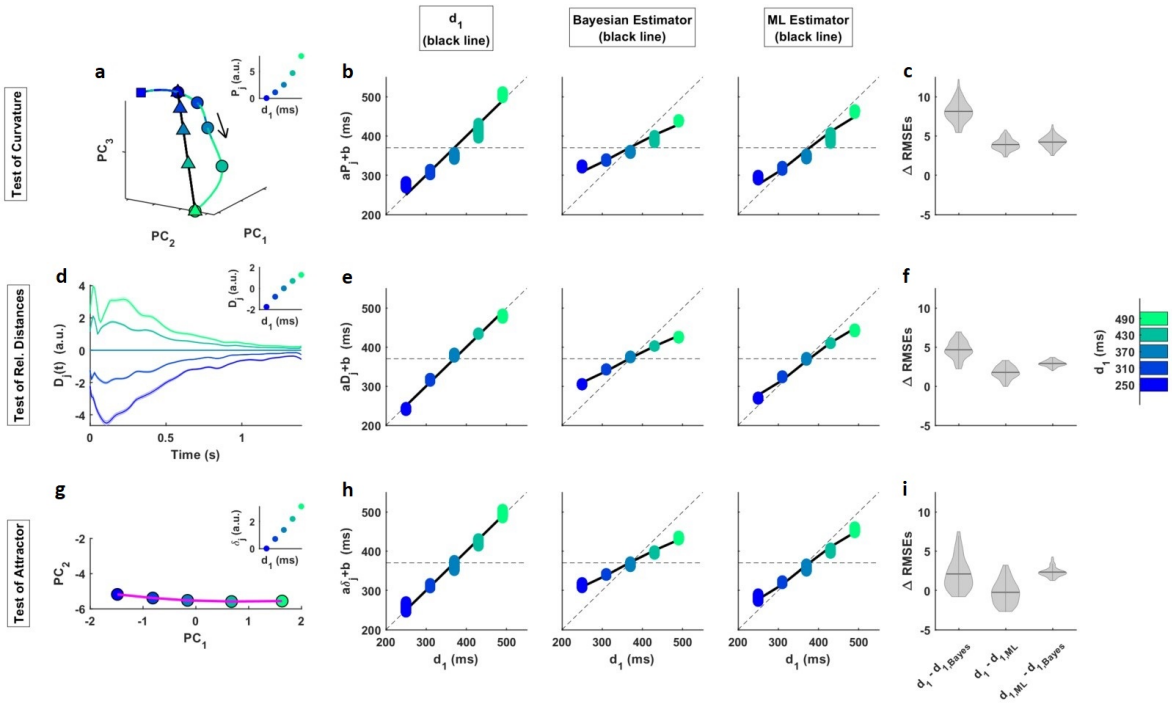
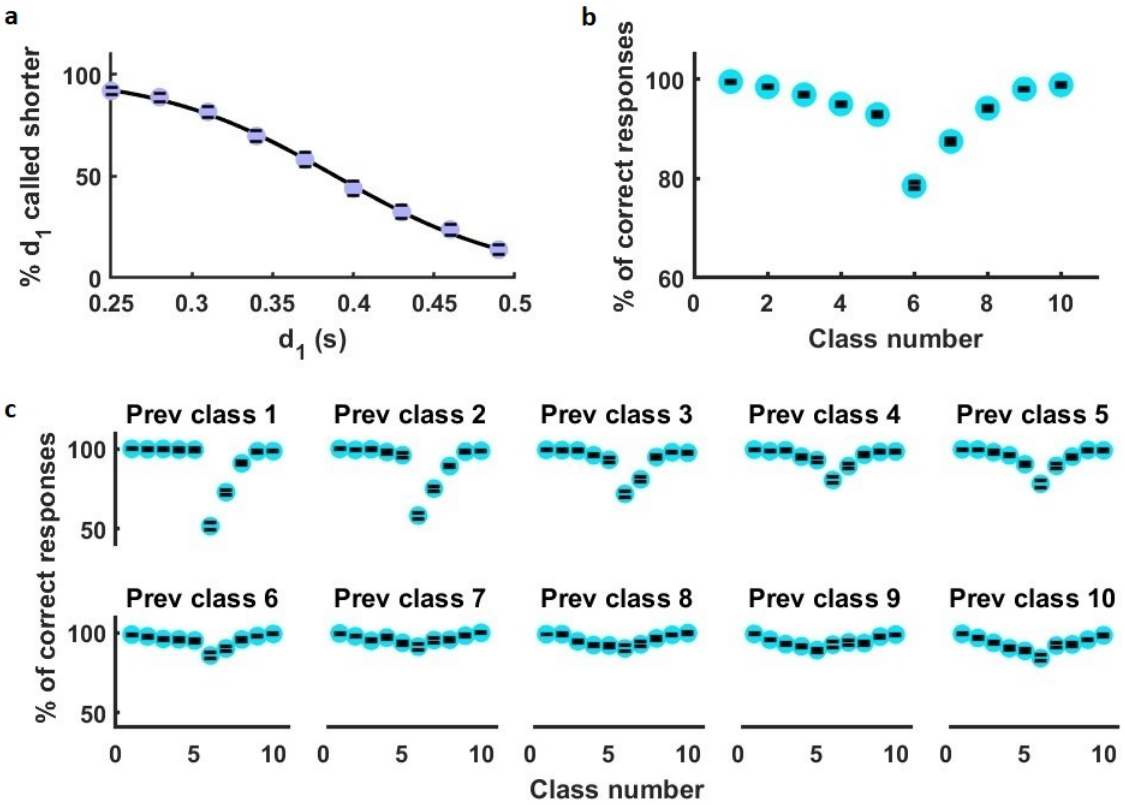
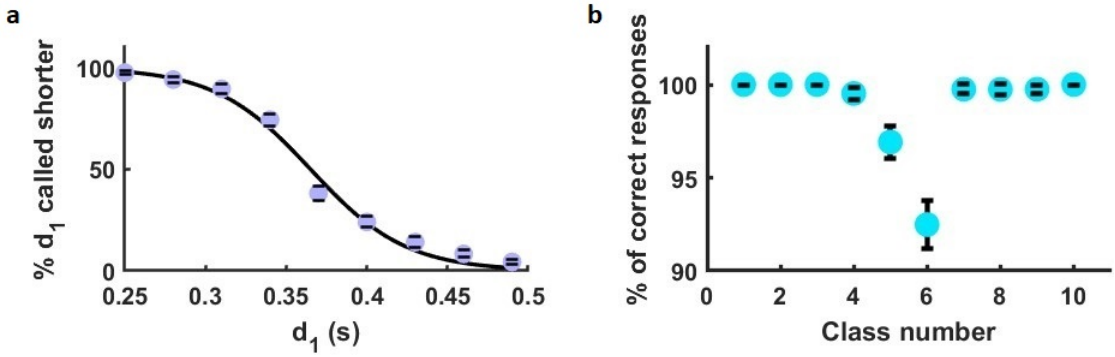


Fig. 7. Several geometric features code the Bayesian estimator of d_1. a-i Example network 1. **a-c**. Curvature of the one-dimensional curved manifold during the presentation of the first stimulus. **a**. Neural trajectories and projections (triangles) of each SF1 terminal state (circles) onto the encoding axis, **u** (black line). Squares represent the beginning of the figure, located within the first stimulus. Neural trajectories are embedded within a state space built with the first 4 PCs, which explain 97% of variance. **Inset** shows, as a function of d_1, the Euclidean distances P_j ($j = 1, \ldots, 5$ labels the values of d_1 in the diagonal block) between projections, taken as reference the projection of the shortest d_1. **b**. Linear regression of the distances between projections ($aP_j + b$) to the true d_1 (left), the Bayesian estimator (middle) and the ML estimator (right), as a function of d_1. The colored dots show multiple estimates of the regression derived from bootstrapping (n=100). Solid lines indicate the true d_1 (left), the Bayesian estimator (middle) and the ML estimator (right) versus d_1. **c**. Violin plots (kernel probability density) on shufflings of the RMSE differences between the three hypotheses (d_1, Bayesian estimator and ML estimator). The width of the gray area indicates the density of the shufflings located there. Horizontal line inside of each violin plot indicates the distribution's mean. The chosen hypothesis is always the Bayesian estimator **d-f**. Relative distances during the delay period. **d**. Relative distance, $D_j(t)$, between each trajectory $j = 1, \ldots, 5$ and the reference trajectory d_1=370 ms. Neural trajectories are embedded in a state space built with 6 PCs, which explain 92% of variance. Shadings indicate the mean ± 95% CIs computed from 100 bootstrap resamples. **Inset** shows the temporal means of relative distances as a function of d_1. **e**. Linear regression of the mean relative distances ($aD_j + b$) to the true d_1 (left), the Bayesian estimator (middle) and the ML estimator (right) as a function of d_1 (same format as in (b)). **f**. Analogous to (c), but for the relative distance analysis. The preferred hypothesis is always the Bayesian estimator. **g-i**. Distances between terminal states on the quasi-attractor. **g**. One-dimensional quasi-attractor region (magenta line), at the end of the delay period, embedded within the first 2 PCs space (99.7% of variance). **Inset** shows the Euclidean distances, δ_j, between two d_1 terminal states (taken the shortest d_1 as reference) on the attractor region, as function of d_1. **h**. Linear regression of the terminal states on attractor ($a\delta_j + b$) to the true d_1 (left), the Bayesian estimator (middle) and the ML estimator (right) as a function of d_1 (same format as in (b)). **i**. Analogous to (c), but for the terminal states analysis. The attractor was analyzed with a time window of size 100 ms at the end of the delay period. In the vast majority of cases, the Bayesian estimator is chosen.

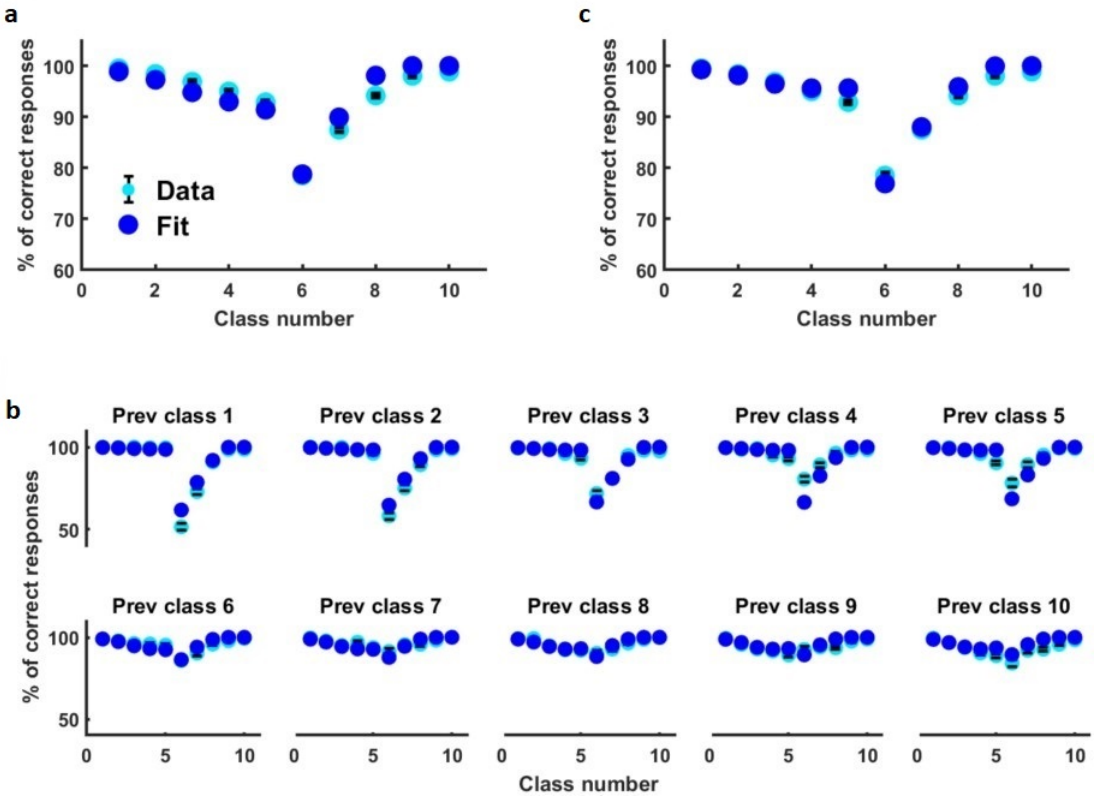
Extended Data Figures



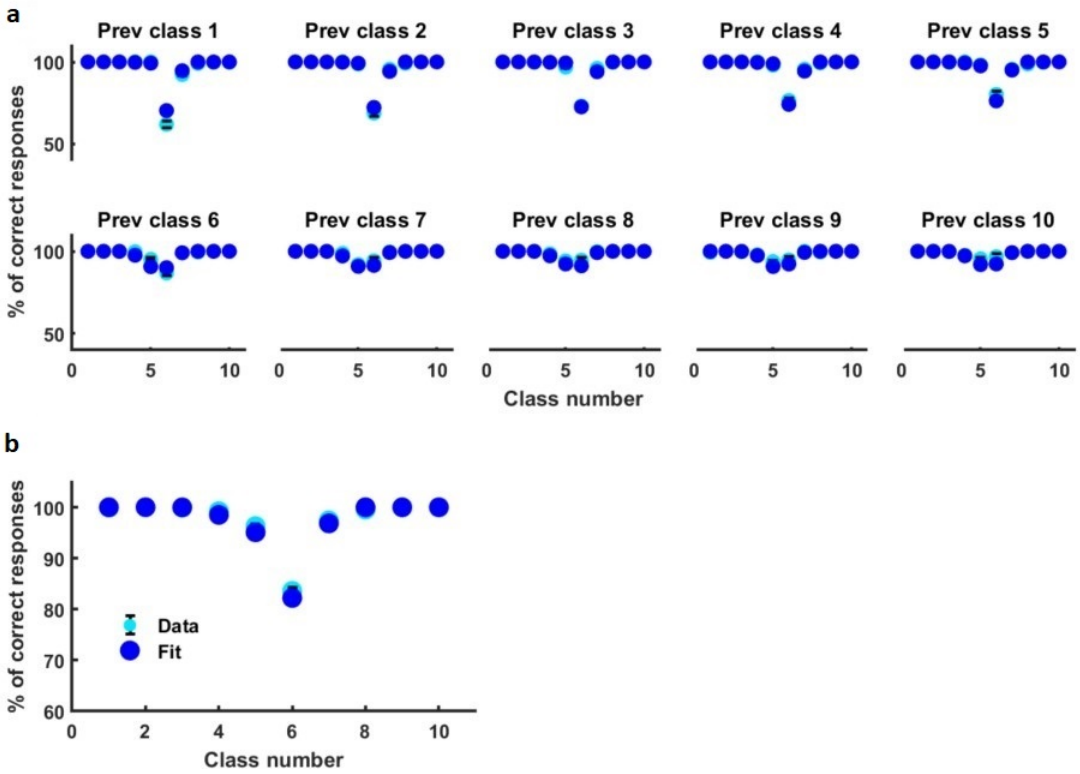
Extended Data Fig. 1. Simulated behavioral data of example network 2. **a**. Psychometric curve (using horizontal classes from the full block). Black line is a sigmoidal fit of the simulated data. **b**. The accuracy curve shows the percentage of correct test trials (using the classes from the diagonal block). **c**. Ten accuracy curves, each conditioned over one of the ten possible previous classes of the diagonal block. Error bars were computed with 1,000 bootstrap resamples.



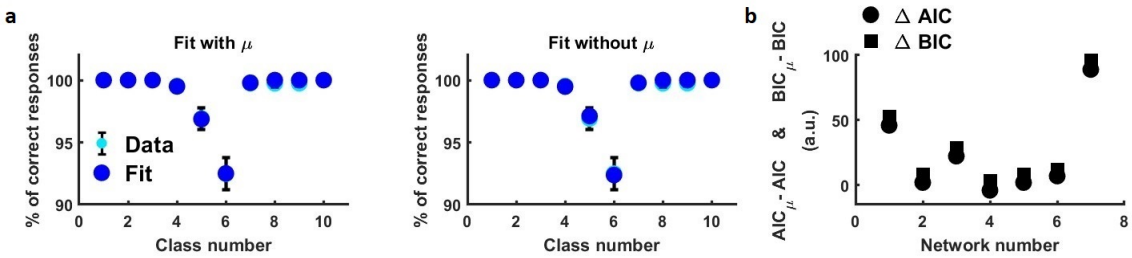
Extended Data Fig. 2. Example network trained and tested resetting its neural activity. **a-b**. Simulated behavioral data. The reset took place 2,000 ms before the end of each training and test trial. **a**. Psychometric curve. Black line is a sigmoidal fit. **b**. Accuracy curve. Error bars were computed with 1,000 bootstrap resamples.



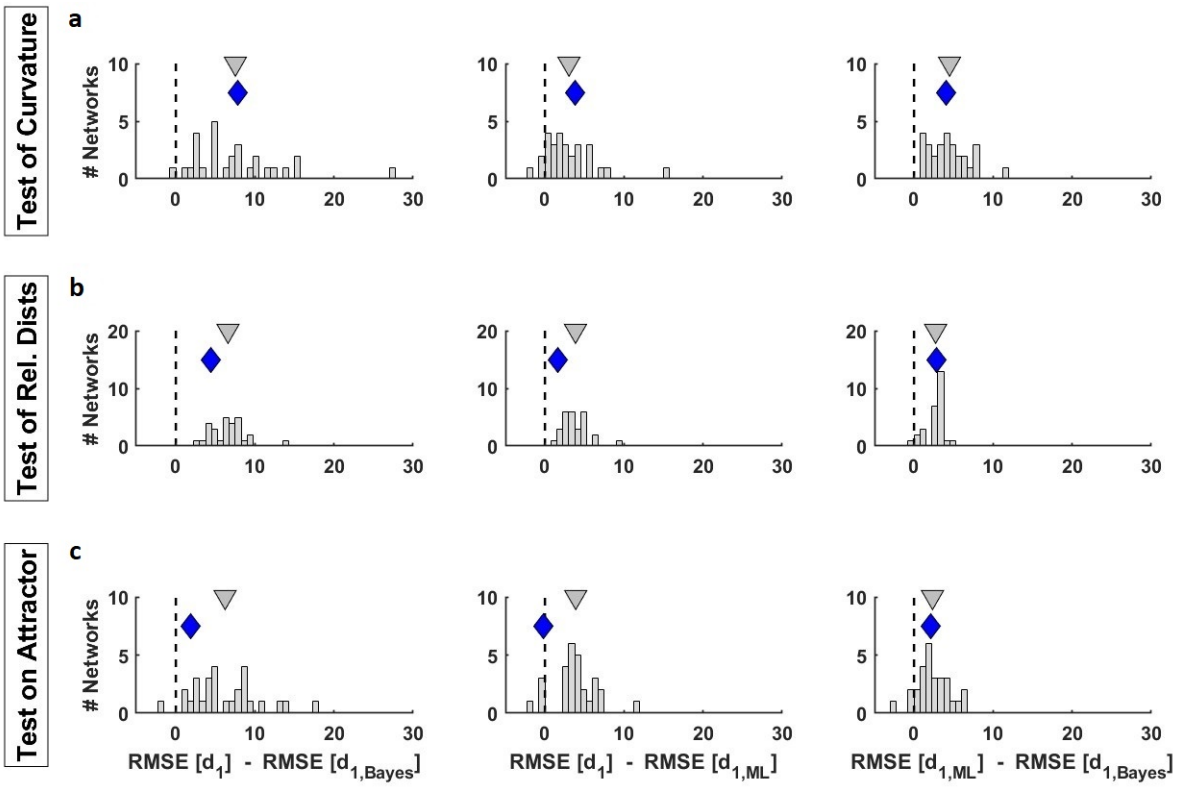
Extended Data Fig. 3. Fits of the behavioral data of example network 2. **a**. Fit of the accuracy curve with the model version $m = 0$ (Eqs. (2a,b)). Cyan circles with error bars are the simulated behavioral data of example network 2 in Extended Data Fig. 1b. Dark blue circles are the fit. **b**. Fit of the ten accuracy curves conditioned over each possible preceding diagonal class, obtained with model version $m = 1$ (Eqs. (3a,b)). Cyan circles are the behavioral data of example network 2 in Extended Data Fig. 1c. **c**. Marginal accuracy curve obtained from the fit in (b) by marginalizing over both the class and the outcome in the previous trial (see Methods).



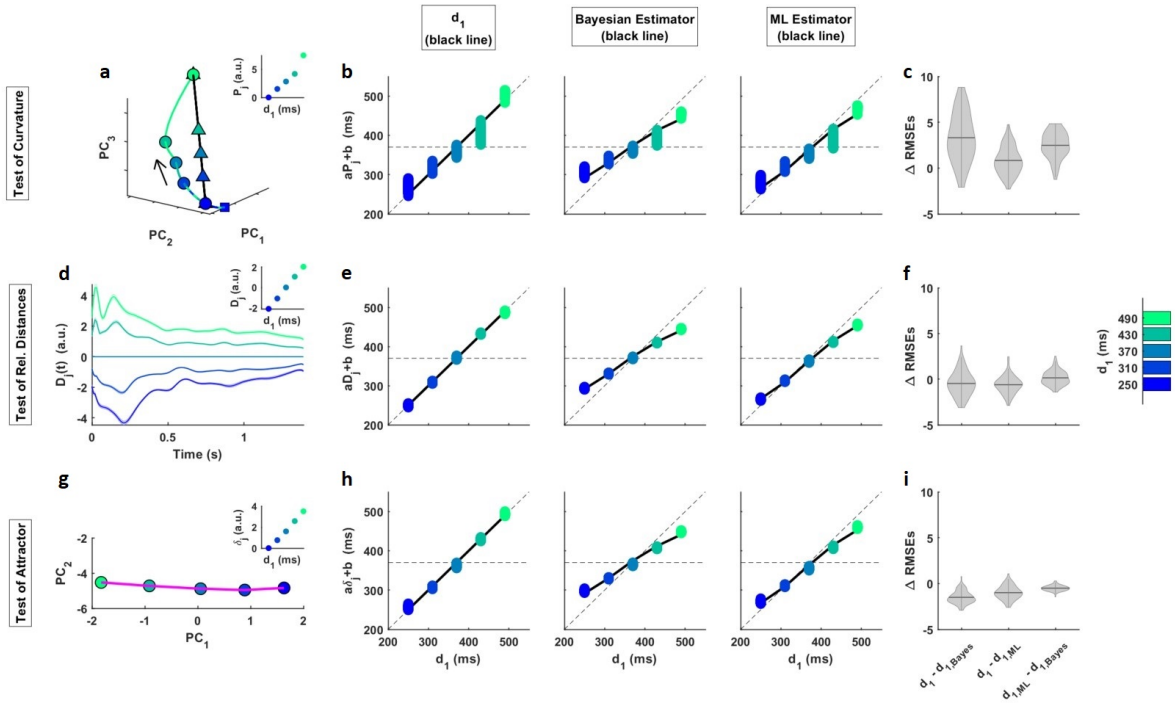
Extended Data Fig. 4. Model that includes both the previous first and second intervals. **a-b**. Fit of the behavioral data of network 1. **a**. Fit of the ten accuracy curves conditioned over each possible preceding diagonal class (Eqs. (S10, S11)). Cyan dots with error bars denote the behavioral data of example network 1 in Fig. 2a. Dark blue circles are the fit. The model is described in Supplementary Text S1. **b**. Marginal accuracy curve, obtained by marginalizing the fitted curves over both the class and the outcome in the previous trial (see Methods). Cyan circles are the behavioral data of example network 1 in Fig. 1e, right.



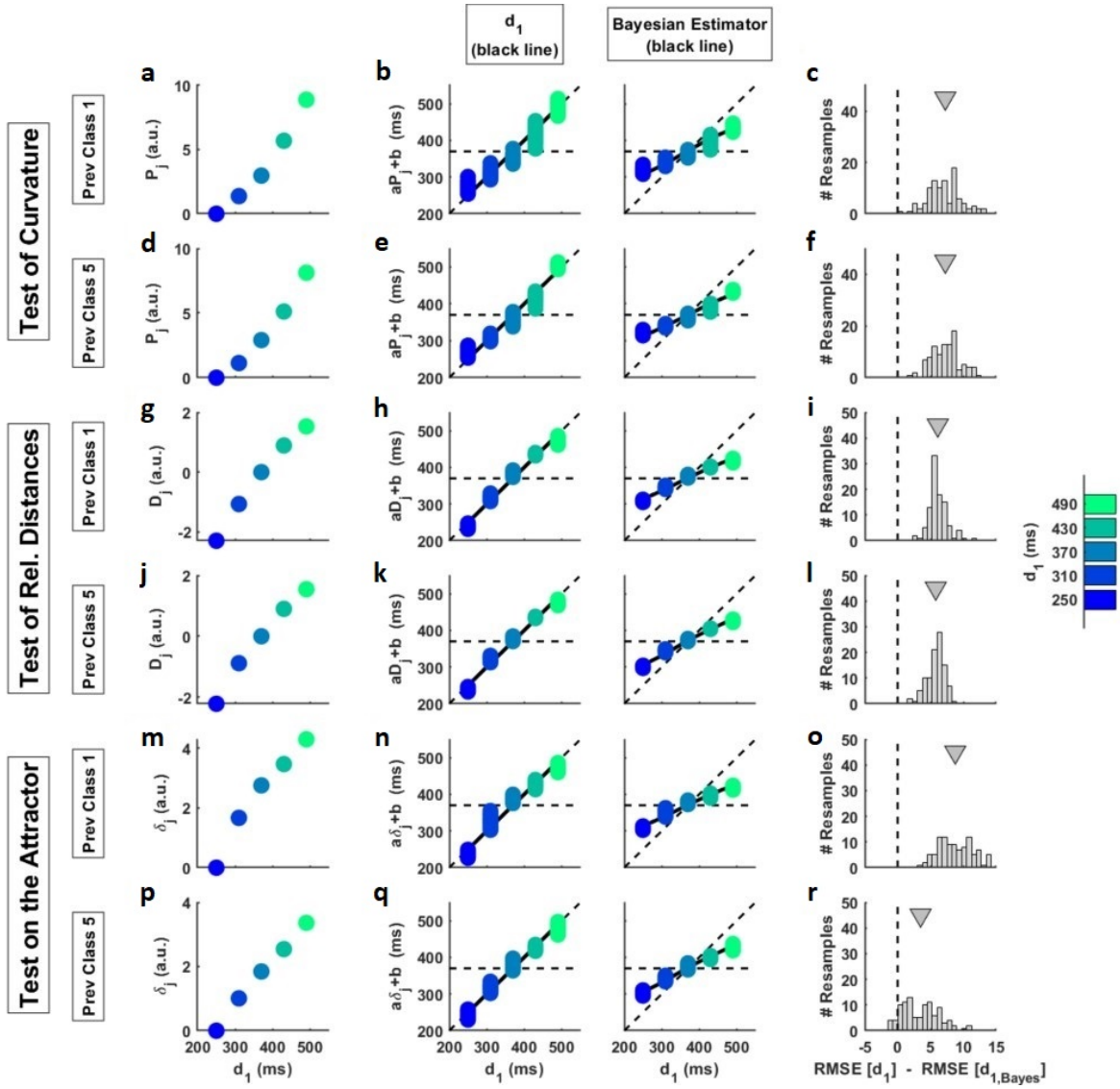
Extended Data Fig. 5. Evaluation of the hypothesis that a reset network has knowledge about the statistics of d_1. a. Two fits of simulated behavioral data of an example network trained and tested resetting its neural activity (Extended Data Fig. 2b). **Left**: Fit with a model that depends on the mean duration $\langle d_1 \rangle$ ($\mu \neq 0$). **Right**: Fit with a model that does not depend on it ($\mu = 0$). **b**. Difference between the AIC measure for the model with parameter μ and the AIC for the model without this parameter, versus network number (circles). Id. for the BIC measures (squares). Comparison of the two models tends to favor the one that does not require knowledge of $\langle d_1 \rangle$.



Extended Data Fig. 6. Geometric features best encode the Bayesian estimator in all positive networks (model version $m = 1$). Histograms over networks of RMSE differences resulting from (**a**) regressions of distances between projections, $aP_j + b$, (**b**) regressions of temporal means of relative distances, $aD_j + b$ and (**c**) regressions of distances between terminal states on the attractor, $a\delta_j + b$. The mean RMSE differences between the Bayesian estimator and the true value of d_1 (left) was around two times larger than that of the ML estimator (middle). In the curvature test only one network favored the true d_1 instead of the Bayesian estimator. Only one network preferred the ML estimator in the test of relative distances and only four positive networks did not choose the Bayesian estimator in the test on the attractor. This indicates that the hypothesis that geometric features (curvature, relative distances or distances on the attractor) encode the Bayesian estimator is better than the hypothesis that it encodes the ML estimator. Gray triangles indicate histogram means. Blue diamonds denote example network 1. Dashed vertical lines identify zero RMSE differences.



Extended Data Fig. 7. Several geometric features code the Bayesian estimator of d_1. a-i Example network 2. **a-c**. Curvature of the one-dimensional curved manifold during the presentation of the first stimulus. **a**. Neural trajectories and projections (triangles) of each SF1 terminal state (circles) onto the encoding axis, **u** (black line). Squares represent the beginning of the figure, located within the first stimulus. Neural trajectories are embedded within a state space built with the first 4 PCs, which explain 95% of variance. **Inset** shows, as a function of d_1, the Euclidean distances between projections, P_j, (where $j = 1, ..., 5$ labels the possible values of d_1 in the diagonal block) taken as reference the projection of the shortest d_1. **b**. Linear regression of the projections ($aP_j + b$) to the true d_1 (left), the Bayesian estimator (middle) and the ML estimator (right), as a function of d_1. The colored dots show multiple estimates of the regression derived from bootstrapping (n=100). Solid lines indicate the true d_1 (left), the Bayesian estimator (middle) and the ML estimator (right) versus d_1. **c**. Violin plots (kernel probability density) on shufflings of the RMSE differences between the three hypotheses (d_1, Bayesian estimator and ML estimator). The width of the gray area indicates the density of the shufflings located there. Horizontal line inside of each violin plot indicates the distribution mean. The mean of the difference ($RMSE_{d_1} - RMSE_{d_1,Bayes}$) was approximately 2.5 times larger for network 1 than for network 2. **d-f**. Mean relative distances during the delay period. **d**. Relative distance, $D_j(t)$, between each trajectory $j = 1, \ldots, 5$ and the reference trajectory d_1=370 ms. Neural trajectories are embedded in a state space built with 7 PCs, which explain 91% of variance. Shadings indicate the mean ± 95% CIs computed from 100 bootstrap resamples. **Inset** shows the temporal means of relative distances as a function of d_1. **e**. Linear regression of the mean relative distances (aD_j+b) to the true d_1 (left), the Bayesian estimator (middle) and the ML estimator (right) as a function of d_1 (same format as in (b)). **f**. Analogous to (c), but for the relative distance analysis. Two thirds of the resamples prefer the real d_1 hypothesis. **g-i**. Distances between terminal states on the quasi-attractor. **g**. One-dimensional quasi-attractor region (magenta line), at the end of the delay period, embedded within the first 2 PCs space (99.6% of variance). **Inset** shows the Euclidean distances, δ_j, between two d_1 terminal states (taken the shortest d_1 as reference) on the attractor region, as function of d_1. **h**. Linear regression of the terminal states on attractor ($a\delta_j + b$) to the true d_1 (left), the Bayesian estimator (middle) and the ML estimator (right) as a function of d_1 (same format as in (b)). **i**. Analogous to (c), but for the terminal states analysis. The attractor was analyzed with a time window of size 100 ms at the end of the delay period. In the vast majority of shufflings, the real d_1 hypothesis is chosen.



Extended Data Fig. 8. Effect of stimulation history on state-space geometry and comparison with behavior. **a-r**. Example network 1. **a-c**. State-space geometry and test during the presentation of the first interval and conditioning over the previous class 1. **a**. Euclidean distances between projections, P_j, (where $j = 1, ..., 5$ labels the possible values of d_1 in the diagonal block) of each SF1 state onto the encoding axis, **u**, as a function of d_1. **b**. Fit of a linear regression of the distances between projections ($aP_j + b$) to the true d_1 (left) and the Bayesian estimator (right) as a function of d_1. Colored circles show multiple estimates of the regression derived from bootstrapping (n=100). Solid lines indicate the true d_1 (left) and the Bayesian estimator (right) as a function of d_1. **c**. Histogram on shufflings of the difference of the RMSEs between the true d_1 hypothesis and the Bayesian estimator hypothesis. Gray triangles indicate the mean of each histogram. Dashed line identifies 0. **d-f**. State-space geometry and test during the presentation of the first interval and conditioning over the previous class 5 (same format as in (a-c)). **g-l**. State-space geometry and test using the temporal mean of the relative distances, D_j, during the delay-period and conditioning over the previous class 1 (**g-i**) and over the previous class 5 (**j-l**). Same format as in (a-c). **m-r**. State-space geometry and test using the distances, δ_j, between consecutive terminal points of the trajectories on the attractor and conditioning over the previous class 1 (**m-o**) and over the previous class 5 (**p-r**). Same format as in (a-c). The tests over curvature and relative distances were done with 5-9 PCs to explain at least 90% of the variance. The attractor tests only required the first 2 PCs to explain 99% of the variance.