Recurrent neural network models for working memory of continuous variables: activity manifolds, connectivity patterns, and dynamic codes

Christopher J. Cueva[*] CCUEVA@GMAIL.COM
Department of Brain and Cognitive Sciences
MIT, Cambridge, MA, USA

Adel Ardalan[*] ADEL.ARDALAN@PRINCETON.EDU
Princeton Neuroscience Institute
Princeton University, Princeton, NJ, USA

Misha Tsodyks MTSODYKS@GMAIL.COM
Simons Center for Systems Biology
School of Natural Sciences
Institute for Advanced Study
Princeton, NJ, USA

Ning Qian NQ6@COLUMBIA.EDU
Department of Neuroscience, Zuckerman Institute
Department of Physiology & Cellular Biophysics
Columbia University, New York, NY, USA

Abstract

Many daily activities and psychophysical experiments involve keeping multiple items in working memory. When the items take continuous values (e.g., orientation, direction, contrast, length, weight, loudness) they must be stored in a continuous structure of appropriate dimensions. We investigate how such a structure might be represented in neural circuits by training recurrent networks to report two previously flashed stimulus orientations. We find that the activity manifold for the two orientations resembles a Clifford torus. Although a Clifford torus and a standard torus (the surface of a donut) are topologically equivalent, they have important functional differences. A Clifford torus treats the two orientations equally and keeps them in orthogonal subspaces, as demanded by the task, whereas a standard torus does not. We further find that the Clifford-torus-like manifold is realized by two different sets of locally-excitatory/globally-inhibitory connectivity patterns. Moreover, in addition to attractors that store information via persistent activity, our networks also use a dynamic coding scheme such that many units change their tuning to prevent the new sensory input from overwriting the previously stored one. We argue that such dynamic codes are generally required whenever multiple inputs enter a memory system via shared connections. Finally, we apply our framework to a human psychophysics experiment in which subjects reported two remembered orientations. We demonstrate that not all RNNs reproduce human behavior. By varying the training conditions of the RNNs, we test and support the hypothesis that human behavior is a product of both neural noise and reliance on the more stable and behaviorally relevant memory of the ordinal relationship between the two orientations. This suggests that suitable inductive biases in RNNs are important for uncovering how the human brain implements working memory. Together, these results

[*]. Equal contribution.

offer an understanding of the neural computations underlying a class of visual decoding tasks, bridging the scales from human behavior to synaptic connectivity.

1. Introduction

Humans can keep a few items of interest in their working memory and recall them a little later (Ma et al., 2014). A typical example would be to read off a few measurements from a ruler and then jot them down. Many psychophysical studies are similarly designed to test working memory (Miller, 1956; Cowan, 2001). When the items are discrete in nature (e.g., words), they could be stored in point attractors of neural circuits (Hopfield, 1984). In contrast, if the items take continuous values (e.g., orientations) (Bae et al., 2015; Bae and Luck, 2017; Ding et al., 2017; Luu et al., 2020), then they would have to be stored in a continuous structure that could represent any of the possible values. Indeed, a line or ring attractor has been proposed to store an item whose value varies along one dimension (Zhang, 1996; Compte et al., 2000; Machens et al., 2005; Itskov et al., 2011). Conceptually, this approach could be readily extended to storing multiple items of continuous values.

There are, however, a few key questions regarding how a neural circuit of working memory may store multiple items of continuous values. Consider the typical psychophysical task of remembering two successively presented stimulus orientations (Bae and Luck, 2017; Ding et al., 2017). Since a 1D structure of neural activities (such as a ring attractor) is needed to represent any value of each orientation, a manifold of (at least) 2 dimensions is required to store any combination of the two orientations. It is, however, unclear which specific 2D manifold should be used and how the choice of such a manifold depends on the task. Another question is that given an appropriate activity structure for storing multiple continuous values, what connectivity patterns among the memory units could generate the desired activity structure. Additionally, biologically relevant variables are often encoded by broadly tuned cells. For example, orientation tuning curves of V1 neurons have a full width at half height of about 40 degrees (Schiller et al., 1976). If two stimulus orientations differing by a few degrees are presented successively at the same location (Ding et al., 2017), then they provide very similar inputs, via the same set of connections, to the same set of the memory units. How, then, does the system prevent the memory of the first orientation from being overwritten by the arrival of the second input?

We investigated these questions by training recurrent neural networks (RNNs) to store two successively flashed orientations, and by analyzing the activity and connectivity patterns of the trained networks. RNNs are a natural choice since the recurrent connections are needed to keep the stimuli in memory after their disappearance. Given the periodic nature of each orientation, one might expect that the activity manifold for storing any two orientations would resemble a standard torus (the surface of a donut), a curved 2D surface embedded in 3D. Instead, we found that the activity manifold is more like a Clifford torus, a flat 2D surface embedded in 4D. Although there are smooth, one-to-one mappings between them, the two types of tori differ in their metric properties. In particular, a Clifford torus treats the two orientations equally and keeps them in orthogonal subspaces, which prevents interference as demanded by the task, whereas a standard torus does not. We further found that the Clifford-torus-like manifold of the memory activities is realized by two different sets of connections that exhibit locally excitatory and globally inhibitory connectivity patterns in the orientation domain. Moreover, the units storing the memory of the first orientation change their tuning over time (Masse et al., 2020; Panichello and Buschman, 2021; Wan et al., 2021), transitioning from one attractor-like structure to another, to prevent the memory from being overwritten by the second orientation. Dynamic codes like this could be a general solution to the overwriting problem whenever there is significant overlap among multiple inputs to a memory system (Rademaker et al., 2019).

Finally, we considered the specific study of Ding et al. (2017) in which subjects reported two remembered line orientations and (implicitly) their ordinal relationship (whether the second orientation is clockwise or counterclockwise from the first). In this paradigm, absolute orientations of individual lines and their ordinal relationship are the lower- and higher-level features, respectively. Ding et al. showed that their data contradict the widely assumed low-to-high-level decoding hierarchy which decodes the absolute orientations first, and then compares them to determine the ordinal relationship. Instead, they explained their data with a high-to-low-level decoding hierarchy that uses the ordinal relationship to constrain the decoding of the absolute orientations in working memory. Ding et al. argued that the high-to-low-level decoding is advantageous because working memory is noisy and so the brain should prioritize the more stable discrete/ordinal memory to constrain the less stable continuous memories of orientations. In addition, higher-level more categorical memories should be prioritized because they are often more behaviorally relevant (Peelen and Kastner, 2014; Ekman and Friesen, 2003). We reasoned that if this is true then not all RNNs trained to implement the psychophysics task of Ding et al., by reporting two absolute orientations and their ordinal relationship, should display the behavioral biases of humans. We hypothesized that injecting noise into RNNs during memory periods helps the networks to learn the high-to-low-level decoding scheme, and our simulations confirmed this hypothesis. We also found that the activity manifolds and connectivity patterns of the trained networks are similar to those discussed above. But because the outputs now include the ordinal relationship between the two absolute orientations, the subspaces for the two orientations deviate somewhat from orthogonality, producing a slightly deformed Clifford torus as the activity structure of the memory units. Overall, our results help understand how recurrent neural networks store multiple continuous variables, and shed light on possible neural mechanisms underlying many psychophysical and daily tasks that require working memory.

2. Methods

We first describe our standard setup and then the variations. The training task and network architecture are shown schematically in Figure 1. Two arbitrary input orientations (θ and ϕ) were presented successively to the RNN with a variable delay (Delay 1) between them. After another variable delay (Delay 2), input go-cues were presented to indicate when to decode the absolute orientations, and in some simulations, their ordinal relationship (see Figure 5), at the output. We used 32 orientation-tuned input units whose preferred orientations cover the full 180° range, and modeled their activities after V1 orientation selective cells (Teich and Qian, 2003). The go-cue inputs were binary 0 and then 1 to initiate the RNN's response. The RNN consisted of 100 fully-connected units. We used the cos and sin of 2θ and 2ϕ as absolute-orientation outputs so that a line oriented at 0° had the same output as a line oriented at 180°. When the ordinal output was included, it was a binary unit for the

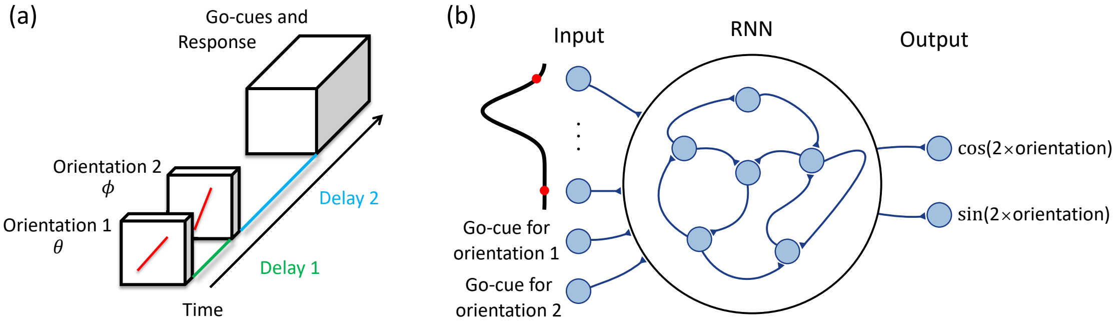

Figure 1: RNN task and architecture. **(a)** The RNN was trained to remember the orientations of two lines and then report these orientations sequentially when prompted with go-cues. Each input line was followed by a delay. **(b)** The RNN consisted of 100 recurrently connected units which received 32 orientation-tuned inputs and two go-cues that switched from 0 to 1 to initiate the RNN's responses. The target outputs were cos and sin of two times the line orientations.

clockwise/counterclockwise relationship between the two input lines (see Section 3 for more details).

The dynamics $u_i(t)$ of each recurrent unit in the network was governed by the standard continuous-time RNN equations:

$$\tau \frac{dx_i(t)}{dt} = -x_i(t) + \sum_{j=1}^{N^{\text{rec}}} W^{\text{rec}}_{ij} u_j(t) + \sum_{k=1}^{N^{\text{in}}} W^{\text{in}}_{ik} I_k(t) + b_i \tag{1}$$

$$u_i(t) = f(x_i(t)) + \xi_i(t) \tag{2}$$

for $i = 1, \ldots, N^{\text{rec}}$. The activity $u_i(t)$ of unit i at time t was computed based on $x_i(t)$ through a rectified tanh nonlinearity $f(x) = \max(0, \tanh(x))$. Each unit received input from other units through recurrent connections with weights determined by the matrix W^{rec}, initialized orthogonally (Saxe et al., 2014). The units also received input $I(t)$ that entered the RNN through input weights determined by the matrix W^{in}. Each unit had two sources of bias: (1) b_i which was learned and (2) $\xi_i(t)$ which represented noise intrinsic to the network and was taken to be white Gaussian (independently sampled at each time step) with zero mean. The network was simulated using the Euler method for $T = 500$ timesteps, each of duration $\tau/10$ (Mante et al., 2013).

To perform the psychophysics task with the RNN, we linearly combined the activity of recurrent units to decode the output $y_j(t)$ according to:

$$y_j(t) = \sum_{i=1}^{N^{\text{rec}}} W^{\text{out}}_{ji} u_i(t). \tag{3}$$

We optimized the network parameters W^{in}, W^{rec}, b and W^{out} to minimize the mean squared error between the target outputs and the network outputs:

$$E = \frac{1}{MTN^{\text{out}}} \sum_{m,t,j=1}^{M,T,N^{\text{out}}} (y_j(t,m) - y_j^{\text{target}}(t,m))^2 \tag{4}$$

Parameters were updated with the Hessian-free algorithm (Martens and Sutskever, 2011) using mini-batches of size $M = 500$ trials. We also varied many aspects of the above standard setup to confirm robustness of our conclusions. Most importantly, we used different output representations, different activation functions, different learning algorithm (Adam, Kingma and Ba (2015)), and simultaneous instead of sequential reporting for the two orientations. The results reported below remained consistent in the above variants.

3. Results

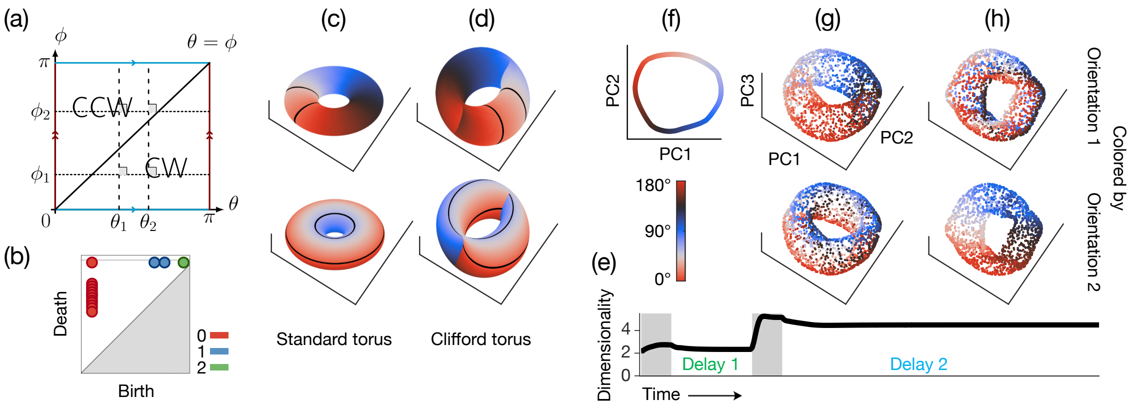

Figure 2: Standard vs. Clifford tori and memory geometry in the RNN. **(a)** Fundamental polygon of both types of tori. Identifying the blue edges together and the red edges together results in a standard torus if done in 3D ambient space, while it results in a Clifford torus if done in 4D ambient space. **(b)** Persistence diagram for both types of tori is the same. **(c)** Example of a standard torus colored by the two angular parameters. The black curves show rings obtained by fixing one of the two angular parameters, e.g. rings corresponding to the dashed vertical line at θ_1 or dotted horizontal line at ϕ_2 in panel (a). **(d)** Example 3D projections of a Clifford torus colored by the two angular parameters. **(e)** Dimensionality of the RNN activity over time and across all trials. Shaded regions denote the intervals when orientation 1 and orientation 2 were presented. The dimensionality, quantified by the participation ratio (Rajan et al., 2010; Abbott et al., 2011), is near 2 after the presentation of the first orientation and near 4 after both orientations have been presented. **(f)** The RNN activity at the end of the first delay period is shown projected onto the first two principal components (2 PCs explained 92% of variance). The first angular variable is stored around a ring. **(g-h)** The RNN activity is shown at two different timepoints after the presentation of the second orientation, after projecting onto the first three principal components (3/4 PCs explained 74%/93% of variance at the end of Delay 2). "Switching" of the subspaces encoding orientation 1 and orientation 2 between (g) and (h) is consistent only with the Clifford torus geometry.

Low-dimensional manifolds of recurrent-unit activities To understand how neural circuits may represent multiple items of continuous values, we trained RNNs to remember two successively presented input orientations and decoded them later (Figure 1). After training we tested the RNNs using various combinations of the two input orientations and recorded the activities of the recurrent units over time. We then applied PCA to the activities to examine their low-dimensional structure. When interrogating the RNN with various line 1 stimuli, the activity manifold settles into a ring attractor during Delay 1 as expected (Figure 2f). We next investigated the activity manifold after the presentation of both lines.

To see what results may be expected, consider the square region of Figure 2a which represents all possible combinations of the two input orientations in a trial, with each axis covering the full 180° range of each orientation. Since each orientation is periodic, each pair of opposite edges of the square should be identified. If we do the two identifications in $\mathbb{R}^3$, we need to distort the square to form a standard torus (the surface of a donut) shown in Figure 2c. This distortion, however, has undesirable consequences for the task as the two orientations are treated very differently. The first orientation could be stored on the toroidal rings (as shown in the top panel of Figure 2c) and the second orientation could be stored on the poloidal rings (Figure 2c bottom panel), or vice versa. Since toroidal rings have different circumferences, they represent the same orientation differently when the other orientation changes. In contrast, the task requires equally accurate recall of both orientations. When we consider ordinal output later, we will show yet another undesirable consequence of the distortion.

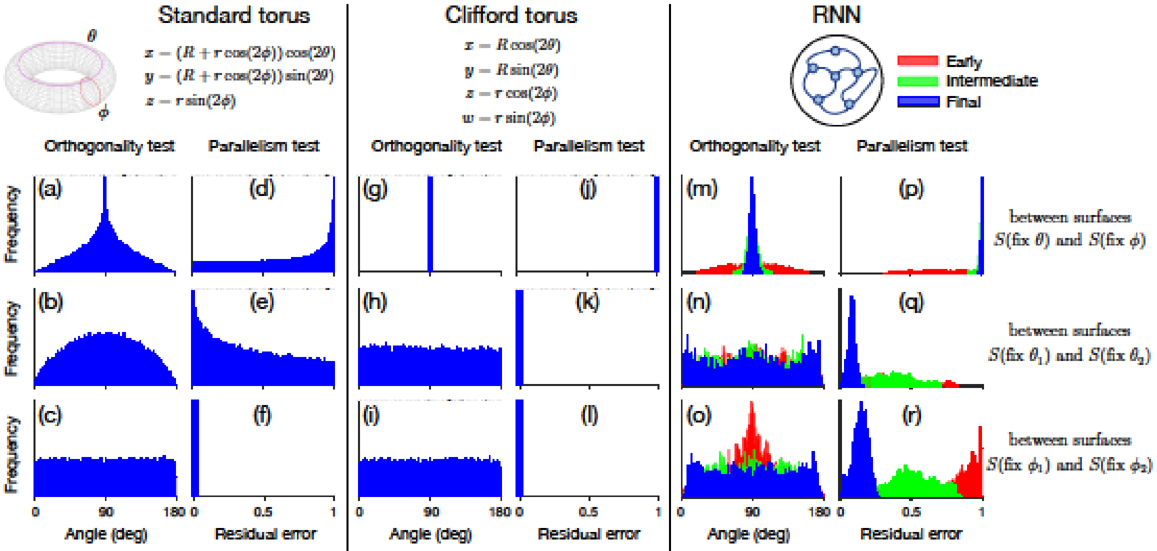

Figure 3: Orthogonality and parallelism tests for **(a-f)** a standard torus, **(g-l)** Clifford torus, and **(m-r)** trained RNN. The scales of the y-axes are chosen to make the shapes of the distributions easier to see and are not all the same. The Clifford torus and RNN store the two orientations in orthogonal subspaces whereas the standard torus does not (see first row). See text for explanations of the orthogonality and parallelism tests. Comparing the second and third rows, we see the standard torus treats each angle asymmetrically, e.g. the orthogonality tests for the standard torus are not the same for both angles. Similarly, the parallelism tests are also not the same for both angles. In contrast, the Clifford torus and RNN store the two angles in a similar manner.

Alternatively, we can identify the two sets of opposite edges of the square in $\mathbb{R}^4$ without distorting the square at all, to produce what is known as the Clifford torus. It is a two-dimensional manifold embedded in $\mathbb{R}^4$ defined as the Cartesian product of two circles each embedded in $\mathbb{R}^2$ (Wikipedia contributors, 2020). Although topologically equivalent to a standard torus (Figure 2a) a Clifford torus is intrinsically flat and treats the two orientations, and different values of the same orientation, equally, as demanded by the task (Weeks, 2019). Intuitively, one cannot fit the product of two circles in $\mathbb{R}^3$ without distortion (because each 1D circle requires a 2D space to be embedded in) but can do so in $\mathbb{R}^4$.

We now show our analyses for differentiating these two geometric objects. When we first plotted the activity manifold in the space of the first 3 PCs, it looked like a standard torus with the two orientations represented along the toroidal and poloidal rings, respectively (Figure 2g). However, when we followed the manifold over time, we noticed that the manifold appeared to run though itself such that at a later time, the two orientations represented by the toroidal and poloidal rings swapped (Figure 2h). Since the set of differential equations governing the system is autonomous when these activities are recorded (no time dependent inputs), the solution should be unique, suggesting that the actual manifold is embedded in a higher-dimensional space without running through itself. We confirmed that the embedding space was indeed roughly four dimensional (Rajan et al., 2010; Abbott et al., 2011; Mazzucato et al., 2016; Gao et al., 2017): the top 3 and 4 PCs explained 74% and 93% of the variance at the end of the second delay period (Figure 2e).

Next, standard and Clifford tori make different predictions about the geometry (orthogonality and parallelism) of the subspaces formed by fixing one orientation while varying the other. We tested these predictions. For convenience, we denote the subspace $\mathbf{S}_\theta$ when we fixed the first orientation at θ and varied the second orientation, and $\mathbf{S}_\phi$ when we fixed the second orientation at ϕ and varied the first orientation. For example, the two black curves on the torus in Figure 2d (e.g., top panel) could correspond to $\mathbf{S}_{\theta=100^\circ}$ and $\mathbf{S}_{\theta=125^\circ}$. A Clifford torus predicts that different $\mathbf{S}_\theta$s (or different $\mathbf{S}_\phi$s) should be parallel to each other (Figure 3h-3i) whereas $\mathbf{S}_\theta$s and $\mathbf{S}_\phi$s should be orthogonal to each other[1] (Figure 3g). These predictions clearly do not hold for a regular torus (Figures 3a-3c).

To test the orthogonality prediction, we calculated the angles between the first two PCs of $\mathbf{S}_\theta$'s and the first two PCs of $\mathbf{S}_\phi$'s. These angles distributed around 90 degrees (Figure 3m) for the RNN, supporting the Clifford torus hypothesis. As a control, we also calculated the angles between the PCs spanning different $\mathbf{S}_\theta$s (or similarly, different $\mathbf{S}_\phi$s); they distributed near uniformly in the interval $[0, 180)$ degrees (Figure 3n-3o).

To test the parallelism prediction, we calculated the residuals of reconstructing one $\mathbf{S}_\theta$'s PCs using another $\mathbf{S}_\theta$'s PCs, or one $\mathbf{S}_\phi$'s PCs using another $\mathbf{S}_\phi$'s PCs. For near-parallel subspaces, we expect these residuals to be close to 0 whereas for orthogonal subspaces they should be close to 1 (for unit-length PCs). The results in Figures 3q-3r show that the residuals distributed near 0, again supporting the Clifford torus hypothesis. As a comparison, we also calculated the residuals of reconstructing an $\mathbf{S}_\theta$'s PCs using an $\mathbf{S}_\phi$'s PCs and vice versa. As expected, the residuals distributed near 1 for the Clifford torus (Figure 3j) and RNN (Figure 3p).

The above analyses indicate that the low-dimensional manifold of the recurrent unit activities, after the presentation of both orientations, resembled a Clifford torus more than a standard torus. We then examined how the Clifford-torus-like structure emerged during

[1] Two (vector) subspaces A and B of an inner product space V are called orthogonal subspaces if each vector in A is orthogonal to each vector in B.

the training process. The colored panels in Figure 3 show the orthogonality and parallelism results for the RNN at three stages during training. Early in training, the results did not look like those of a Clifford torus (red) but they evolved with further training (green) until the final geometry of the memory representation resembled a Clifford torus (blue). We posit that storing the two orientations in nearly orthogonal subspaces makes the memories more robust to noise.

To check whether formation of the Clifford tori is specific to our choice of unit activation function, noise level and learning algorithm, we experimented with a different output representation for orientations (same as the input), other activation functions (tanh and ReLU), various noise levels, and a different learning algorithm (Adam, Kingma and Ba (2015)). In all of the cases, we obtained similar results which suggests the storage of the two orientations on near-orthogonal subspaces was a rather general strategy for solving the task.

Topological data analysis (TDA) has been applied to study activity manifolds in both real and artificial neural networks. As we noted previously, Clifford and standard tori are topologically equivalent; consequently, TDA cannot tell them apart. In contrast, the orthogonality and parallelism tests we developed can. Although TDA is a powerful and useful tool, our work demonstrates the importance of going beyond TDA, allowing us to answer new questions about the geometry of working memory.

Analysis of Network Connectivity. We next looked at the tuning properties and connectivity motifs in the trained network to understand how they could support the activity structure described above. The neural activities of all 100 units in the RNN are shown in Figures 4a and 4b. Each small heatmap represents one recurrent unit's activity as a function of the first orientation (x-axis) and the second orientation (y-axis), with yellow indicating high activity and blue indicating no activity. For each unit, the tuning is shown as orientation 1 and orientation 2 vary between 0 and 180 degrees.

The joint tuning is shown at the steady state after line 1 presentation (Delay 1, Figure 4a) and after line 2 presentation (Delay 2, Figure 4b). We found that at a given time, most units were tuned to either the first or second orientation (vertical or horizontal stripes, respectively), consistent with the near-orthogonal subspaces for the two orientations.

We wondered what connectivity patterns among the units could support the attractors we found above. We first divided the units into two classes according to which line they were tuned to, and determined each unit's preferred orientation that yields maximum activity. Then for each class of units, we plotted the mean connection strength between the units as a function of the difference in their preferred orientations, and found a highly structured connectivity pattern of local excitation and global inhibition as shown in Figures 4c and 4d. Specifically, within a class, units that had similar preferred orientations were connected through positive weights, and units with more different preferred orientations were connected through negative weights. This local-excitation/global-inhibition connectivity pattern is well known to generate ring attractors for representing periodic variables such as orientation and heading direction (Somers et al., 1995; Zhang, 1996; Teich and Qian, 2003; Cueva et al., 2019). As a control, we verified that this connectivity did not exist between the two classes of units (Figure 4d bottom row), which may support the near orthogonal subspaces for the two orientations.

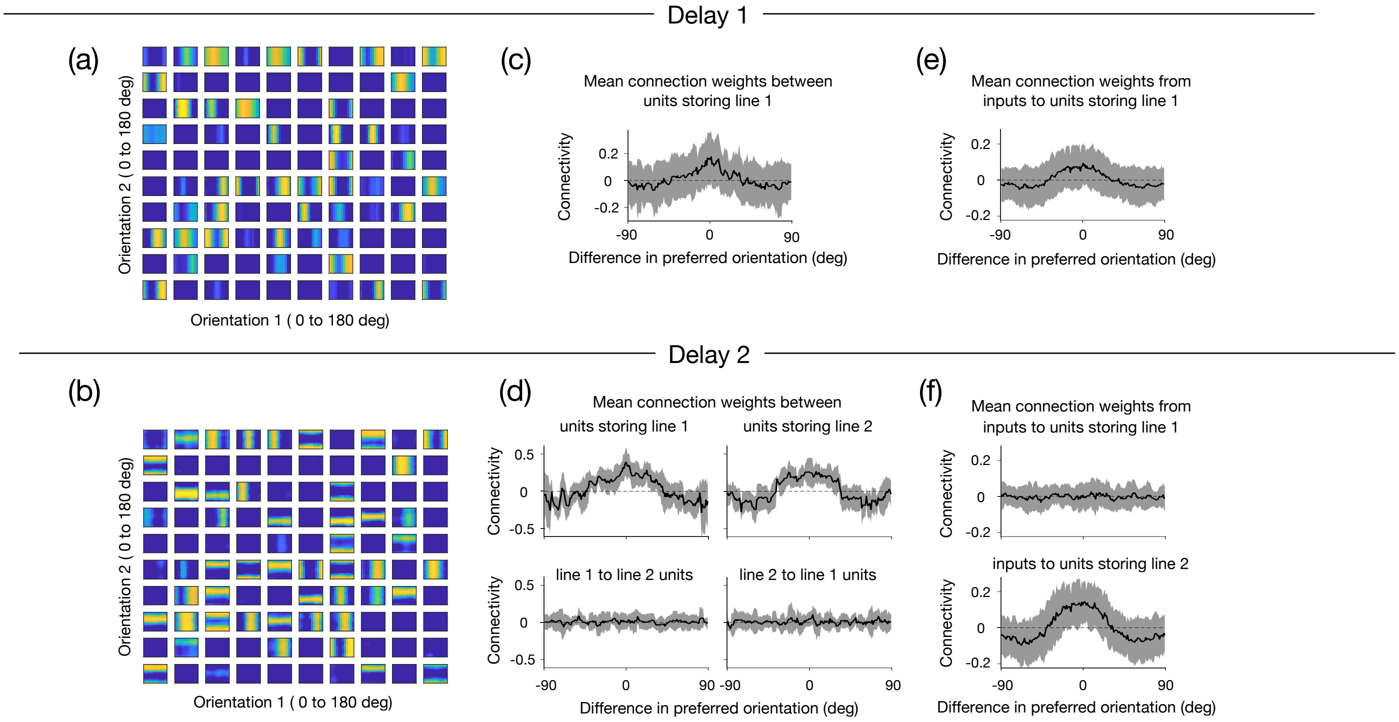

Figure 4: Tuning (leftmost column) and connectivity (other columns) in a trained RNN during the first delay period after the presentation of the first line (top row, delay 1) and during the second delay period after the presentation of the second line (bottom row, delay 2). **(a-b)** Joint tuning of all 100 recurrent units to the two orientations are shown at their steady state values, for the two delay periods. The activity of each unit is shown as a function of line 1 orientation (x-axis) and line 2 orientation (y-axis), with yellow indicating high activity and blue indicating no activity. **(c-d)** Mean connection weights between the recurrent units as a function of their preferred-orientation difference during the two delay periods. The classic local-excitation/global-inhibition connectivity motif developed between units tuned to the same line (panel c and the top row of panel d), but not between units tuned to different lines (bottom row of panel d). **(e-f)** Mean connection weights from the inputs to the recurrent units as a function of their preferred-orientation difference during the two delay periods. The input connections to the recurrent units tuned to line 1 in the first delay period had a local-excitation/global-inhibition pattern (panel e). This pattern disappeared during the second delay period (top row of panel f) because recurrent units changed their tuning to line 1. However, the same connectivity pattern emerged from the inputs to the recurrent units tuned to line 2 (bottom row of panel f). In panels c-f, error bars show one standard deviation.

Dynamic coding. We noticed that the tuning of some units in the RNN changed over time (Figures 4a and 4b). For example, a unit that stored information about line 1 during the first delay period did not always continue to store information about line 1 during the second delay period. Do these changes in tuning confer some computational benefit? In particular, both input orientations entered the RNN through the exact same set of fixed connection weights, namely W^{in} in Equation (1); could the tuning change help prevent line 2 orientation from overwriting the memory of line 1 orientation?

To help answer this question, we plotted the mean connection weights from the 32 inputs to the 100 recurrent units as a function of the difference between their preferred orientations, in Figures 4e and 4f. We found that during the first delay period, there is a local-excitation/global-inhibition connectivity pattern from the inputs to the recurrent units tuned to line 1. However, during the second delay period, this pattern disappeared as the recurrent units changed their tuning (Libby and Buschman, 2021). Meanwhile, a similar local-excitation/global-inhibition connectivity pattern emerged from the inputs to the RNN units tuned to line 2. Thus, the tuning change appeared to both "move" the working memory of line 1 orientation and "hide" it from the inputs, while "making room" for the network to encode new sensory information of line 2.

Reproducing Psychophysics Results of Ding et al. With the above understanding of how an RNN stores two orientations, we now present our RNN implementation of Ding et al.'s (Ding et al., 2017) decoding scheme that explains their psychophysical data. In their experiment, when subjects rotated markers to report two remembered, absolute orientations, they also implicitly indicated the ordinal relationship between the orientations. The key findings were that the two reported absolute orientations in a trial were correlated and that the second line repelled the first line (backward aftereffect) as much as the first line repelled the second line (forward repulsion). To explain the results, Ding et al. used the ordinal relationship as a Bayesian prior to constrain the decoding of the absolute orientations. They argued that such a high-to-low-level decoding scheme is advantageous when the working memory is noisy. We thus hypothesized that if human behavior in this task relies on making higher-level, more categorical judgments and these judgements are prioritized due to noisy working memory, then RNNs trained to make categorical judgements in the presence of noise should more accurately capture these behavioral patterns. To test this, we trained different RNNs to solve this task and then compared the behavior of humans and RNNs. Specifically, we trained RNNs either with or without an ordinal output as an additional optimization goal, and with or without injecting noise into the RNN, particularly during Delay 2 when both orientations are stored.

To test the hypotheses, we trained three main versions of the RNN. To remove the confound of an asymmetric motor response, and probe the sufficiency of memory noise plus the ordinal constraint to explain the behavioral results, we trained RNNs to report all orientations simultaneously. However, our findings hold for RNNs with both sequential and simultaneous reports. Version 1 was the same as the RNN above which was only trained to output the two absolute orientations, but not their ordinal relationship. We probed the trained network with line 1 and line 2 oriented at 50° and 53°, as in Ding et al.'s experiment, with noise added to the firing rates to create trial-to-trial variability. The network generated a distribution of predictions, spherically centered around the true values as shown in Figure 5c, quite different from the actual data (see panel A of Figure 3 in Ding et al. (2017)). A spherical distribution was obtained for all RNNs with no ordinal output, regardless of whether they were trained with noise or without.

To examine the effect of the ordinal relationship between the lines, we incorporated an additional ordinal output in version 2 as shown in Figure 5a. This output was $+1$ if the orientation of the second line was clockwise from the first line, and -1 if counterclockwise. For this version, we did not add any firing rate noise $\xi_i(t)$ during training. When it was tested on inputs of 50° and 53° (again, with a small amount of noise added to the firing rates to create trial-to-trial variability) the output distribution was, once more, spherical and centered around the true values as in Figure 5c. Even though the ordinal relationship was stored by the RNN, it was not used to constrain the decoding of the two absolute orientations.

To reproduce the psychophysics results, we needed to force the RNN to use the ordinal memory by introducing noise into the network during training, making the continuous orientation memories less stable than the binary ordinal memory. Version 3 of the network included the ordinal output as in version 2, but additionally, noise was injected into the firing rates of the network during training via $\xi_i(t)$ in Equation (2). As shown in Figure 5b, in this case the RNN outputs of the two absolute orientations are correlated with, and repelled from, each other, and display similar forward and backward aftereffects (Figure 5d-5e) in agreement with the psychophysics results of Ding et al. (2017). Intuitively, the noise injection made it difficult to get the ordinal output correct when the two absolute orientations were nearly identical (i.e., near the diagonal line in the joint space of Figure 2a). The ordinal training, then, must force a coordinated shift of the absolute orientations away from the diagonal line to produce the correlation and repulsion.

Version 3 is also consistent with the original interpretation of Ding et al. (2017) that the retrospective high-to-low-level decoding is advantageous when the decoding occurs in noisy working memory. Interestingly, we did not have to train the network on the higher-level, ordinal relationship before the lower-level, absolute orientations. Instead, training all these outputs together produced the result that the ordinal relationship constrained the absolute orientations, but not vice versa. This is likely because the ordinal relationship is categorical and thus easier to learn and maintain in the noisy recurrent units, compared with the continuous, absolute orientations.

We also analyzed the memory geometry of the RNN that explains the psychophysics data (i.e., the RNN with the ordinal output and noise added). We found that the activity manifold after the presentation of the second orientation appeared to be a distorted Clifford torus: the subspaces for the different orientations were still nearly orthogonal but not as close to orthogonal as in Figure 3, and the subspaces for two values of the same orientation were no longer parallel. This distortion must be for accommodating the correlation and repulsion between the two lines. Interestingly, the manifold still resembled a Clifford torus more than a standard torus. The reason is likely that the metric structure of the standard torus is also undesirable for the ordinal output. Consider the diagonal line in Figure 2a, which is the decision boundary for the ordinal output. When the opposite sides of the square are identified in 3D to form the standard torus, the decision boundary becomes a curve not contained in a plane (this can be shown by demonstrating that the torsion of the curve is not 0 everywhere). Consequently, the two opposite ordinal outcomes are not linearly separable. In contrast, when the opposite sides of the square are identified in 4D to form a Clifford torus, the decision boundary does not change and the ordinal outcomes remain linearly separable.

To summarize, there are two requirements during training for reproducing Ding et al's psychophysical data: memory noise and ordinal output. We believe this is an interesting finding because previous literature may have created the (improper) expectation that training neural networks to perform a task well automatically produces brain-like computation (Yamins et al., 2014; Cadieu et al., 2014; Yamins and DiCarlo, 2016). In contrast, we found that simply training a network to accurately report absolute orientations does not generate a brain-like system. We also have to consider factors that constrain neural computation (Sussillo et al., 2015), such as memory noise and ordinal output.

In addition, our network model allowed us to explicitly probe the hypothesis that memory noise coupled with the ordinal constraint was sufficient to reproduce the psychophysics behavior. We trained models where noise was only injected during the second delay period when both memories must be stored, demonstrating that motor and sensory noise were not necessary to explain human behavior on this task.

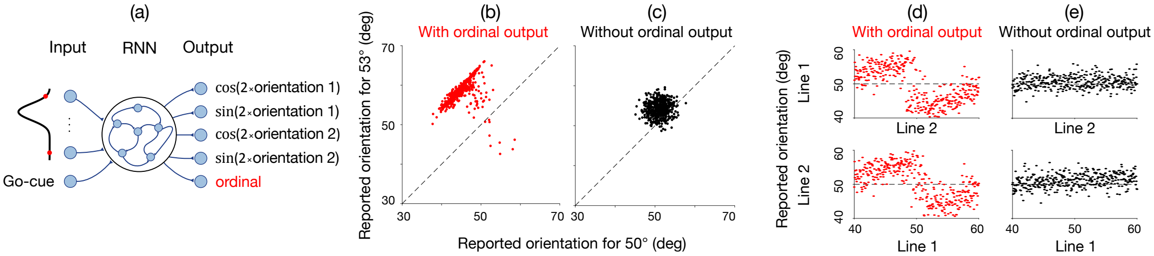

Figure 5: Human behavior was not captured by all RNNs. **(a)** RNN optimized in the presence of memory noise to report both absolute orientations and their ordinal relationship, i.e. whether the second orientation is a clockwise or counterclockwise rotation from the first. The RNN consisted of 100 recurrently connected units which received 32 orientation-tuned inputs and a go-cue that switched from 0 to 1 to initiate the RNN's responses. The target outputs were cos and sin of two times the line orientations, and a binary indicator for their ordinal relationship. To remove potential biases due to an asymmetric motor response, and probe the sufficiency of memory noise plus the ordinal constraint to explain the behavioral results, all outputs were reported simultaneously. However, our results held in both sequential and simultaneous settings. **(b)** After training, the network in (a) was probed with line orientations of 50° and 53° as in Ding et al. (2017). The RNN's outputs of the two lines' orientations were correlated and repelled from the diagonal line ($\theta = \phi$) as observed in the psychophysics results of Ding et al. **(c)** In contrast, RNNs trained without this inductive bias (i.e. without the ordinal output) did not recapitulate human behavior. These results are consistent with the hypothesis that humans use higher-level strategies to solve this task. **(d-e)** Forward and backward aftereffect. Each dot indicates the result from a separate trial where the orientations of line 1 and line 2 were varied. In this example the target output for the RNN is always 50° (dotted black line). For the RNN with the ordinal output (d), its reported orientations were biased by the orientation of the other line in way that is consistent with the ordinal memory, and with the human psychophysics of Ding et al. Notice the RNN's report for the orientation of line 1 was affected by line 2 which appeared *afterwards*, a result not predicted by a standard feedforward network (top panel of d). In contrast, for the RNN without the ordinal output the reported orientations were not affected by the orientation of the other line (e).

4. Discussion and Conclusions

We investigated how neural circuits might store multiple items of continuous values in working memory by training RNNs to report orientations of two previously presented lines. We analyzed the dimensionality of the embedding space, and the relationships between subspaces of the recurrent-unit activities, and found that after the presentation of the second line, the low-dimensional activity manifold resembled a Clifford torus more than a standard torus. In order to disambiguate these two memory geometries we could not rely on tools from topological data analysis (Figure 2b), as the Clifford and standard tori are topologically equivalent. Therefore we introduced the orthogonality and parallelism tests to disambiguate these two possibilities and reveal the geometry of working memory. We argued that the Clifford torus better matches the task demands because it treats the two line orientations equally, and stores them in orthogonal subspaces to avoid interference. We then examined the connections among the units, and found that for units tuned to the same line (first or second), those preferring similar orientations excited each other and those preferring different orientations inhibited each other. There were thus two sets of locally-excitatory/globally-inhibitory connectivity patterns, one for each line. Such connectivity patterns did not exist between units tuned to the different lines. Therefore, the overall connectivity supported the Clifford-torus-like activity manifold that stored the two line orientations in nearly orthogonal subspaces.

We further found that the recurrent units changed their orientation preferences over time. By analyzing the connectivity patterns from the input units to the recurrent units at different times, we provided evidence that this dynamic code appeared to safeguard the memory of the first orientation by “moving” it to a different subspace while the network encoded the second orientation, resolving the conflict between preserving the memory of old information and encoding new sensory input. This “dynamic memory” mechanism must be generally required when a working memory system receives a sequence of inputs through the same sensory channel.

Finally, we tested the hypothesis that visual perception in the task of Ding et al. (2017) is actually high-to-low-level decoding in noisy working memory, with higher-level features constraining the decoding of lower-level features. Specifically, we simulated Ding et al.’s task of reporting the absolute orientations of, and the ordinal relationship between, two successively flashed lines. We demonstrated that noise injected into the recurrent units naturally led to the higher-level ordinal relationship constraining the decoding of the lower-level absolute orientations to reproduce the key psychophysical findings of Ding et al. (2017). Noise was necessary; without it the RNN did not use the higher-level ordinal memory to constrain the remembered values of the two orientations.

Although our learning tasks were very simple, these tasks and their many variants have been widely used in visual psychophysics literature (Jazayeri and Movshon, 2007; Stevenson and Koerding, 2009; Zhang and Luck, 2009; Brady and Alvarez, 2011; Ma et al., 2014; Bae et al., 2015; Green and Swets, 1966). Our study thus provides a thorough understanding of a neural network implementation for a major class of psychophysical paradigms, from the behavioral level to the activity and connectivity levels. Additionally, although high-to-low-level decoding is a property of many auto-encoders, standard feedforward networks do not model the effect of noisy working memory on decoding hierarchy, which was necessary for reproducing human behavior in our RNN model. Our work may thus inspire future efforts on understanding visual decoding hierarchy from a machine learning perspective.

Acknowledgments

This work was supported by NSF grant 1754211, NSF NeuroNex Award DBI-1707398, and the Gatsby Charitable Foundation. Simulations were partially performed using computing facilities on CloudLab (Duplyakin et al., 2019).

References

L.F. Abbott, Kanaka Rajan, and Haim Sompolinsky. Interactions between intrinsic and stimulus-evoked activity in recurrent neural networks. *Oxford university press*, pages 65–82, 2011.

Gi-Yeul Bae and Steven J. Luck. Interactions between visual working memory representations. *Attention, Perception, & Psychophysics*, 79(8):2376–2395, 2017.

Gi-Yeul Bae, Maria Olkkonen, Sarah Allred, and Jonathan I. Flombaum. Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, 144(4):744–763, 2015.

Timothy F. Brady and George A. Alvarez. Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, 22(3):384–392, 2011.

Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLOS Computational Biology*, 10(12):1–18, 2014.

Albert Compte, Nicolas Brunel, Patricia S. Goldman-Rakic, and Xiao-Jing Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral cortex*, 10(9):910–923, 2000.

Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114, 2001.

Christopher J. Cueva, Peter Y. Wang, Matthew Chin, and Xue-Xin Wei. Emergence of functional and structural properties of the head direction system by optimization of recurrent neural networks. *arXiv preprint arXiv:1912.10189*, 2019.

Stephanie Ding, Christopher J. Cueva, Misha Tsodyks, and Ning Qian. Visual perception as retrospective bayesian decoding from high- to low-level features. *PNAS*, 114(43):E9115–E9124, 2017.

Dmitry Duplyakin, Robert Ricci, Aleksander Maricq, Gary Wong, Jonathon Duerig, Eric Eide, Leigh Stoller, Mike Hibler, David Johnson, Kirk Webb, Aditya Akella, Kuangching Wang, Glenn Ricart, Larry Landweber, Chip Elliott, Michael Zink, Emmanuel Cecchet, Snigdhaswin Kar, and Prabodh Mishra. The design and operation of CloudLab. In *Proceedings of the USENIX Annual Technical Conference (ATC)*, pages 1–14, July 2019.

Paul Ekman and Wallace V. Friesen. *Unmasking the Face: A Guide to Recognizing the Emotions from Facial Cues.* Oxford: Prentice Hall, 2003.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, 2017. doi: 10.1101/214262.

David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics.* Wiley, New York, 1966.

John J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10): 3088–3092, 1984.

Vladimir Itskov, David Hansel, and Misha Tsodyks. Short-term facilitation may stabilize parametric working memory trace. *Frontiers in computational neuroscience*, 5:40, 2011.

Mehrdad Jazayeri and J. Anthony Movshon. A new perceptual illusion reveals mechanisms of sensory decoding. *Nature Cell Biology*, 446(7138):912–915, April 2007.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Alexandra Libby and Timothy J. Buschman. Rotational dynamics reduce interference between sensory and memory representations. *Nature Neuroscience*, 24(5):715–726, 2021.

Long Luu, Mingsha Zhang, Misha Tsodyks, and Ning Qian. Cross-fixation interactions of orientations suggest that orientation decoding occurs in a high-level area of visual working memory. *Journal of Vision*, 20(11):216–216, 2020.

Wei Ji Ma, Masud Husain, and Paul M. Bays. Changing concepts of working memory. *Nature Neuroscience*, 17:347–356, 2014.

Christian K. Machens, Ranulfo Romo, and Carlos D. Brody. Flexible control of mutual inhibition: a neural model of two-interval discrimination. *Science*, 307(5712):1121–1124, 2005.

Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474): 78–84, 2013.

James Martens and Ilya Sutskever. Learning recurrent neural networks with hessian-free optimization. *International Conference on Machine Learning (ICML)*, 46:68, 2011.

Nicolas Y. Masse, Matthew C. Rosen, and David J. Freedman. Reevaluating the role of persistent neural activity in short-term memory. *Trends in Cognitive Sciences*, 24(3): 242–258, 2020.

Luca Mazzucato, Alfredo Fontanini, and Giancarlo La Camera. Stimuli reduce the dimensionality of cortical activity. *Frontiers in Systems Neuroscience*, 10:11, 2016.

George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Matthew F. Panichello and Timothy J. Buschman. Shared mechanisms underlie the control of working memory and attention. *Nature*, 592:601–605, 2021.

Marius V. Peelen and Sabine Kastner. Attention in the real world: toward understanding its neural basis. *Trends in Cognitive Sciences*, 18(5):242–250, 2014.

Rosanne L. Rademaker, Chaipat Chunharas, and John T. Serences. Coexisting representations of sensory and mnemonic information in human visual cortex. *Nature Neuroscience*, 22:1336–1344, 2019.

Kanaka Rajan, L. F. Abbott, and Haim Sompolinsky. Inferring stimulus selectivity from the spatial structure of neural network dynamics. *Advances in Neural Information Processing Systems*, 23, 2010.

Andrew Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *International Conference on Learning Representations (ICLR)*, 2014.

Peter H. Schiller, Barbara L. Finlay, and Susan F. Volman. Quantitative studies of single-cell properties in monkey striate cortex. ii. orientation specificity and ocular dominance. *Journal of neurophysiology*, 39(6):1320–1333, 1976.

David C. Somers, Sacha B. Nelson, and Mriganka Sur. An emergent model of orientation selectivity in cat visual cortical simple cells. *Journal of Neuroscience*, 15(8):5448–5465, 1995.

Ian Stevenson and Konrad Koerding. Structural inference affects depth perception in the context of potential occlusion. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1777–1784. Curran Associates, Inc., 2009.

David Sussillo, Mark M. Churchland, Matthew T. Kaufman, and Krishna V. Shenoy. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature Neuroscience*, 18:1025–1033, 2015.

Andrew F. Teich and Ning Qian. Learning and adaptation in a recurrent model of v1 orientation selectivity. *Journal of Neurophysiology*, 89(4):2086–2100, 2003.

Quan Wan, Jorge A. Menendez, and Bradley R. Postle. Rotational remapping between differently prioritized representations in visual working memory. *bioRxiv 443973*, 2021.

J. R. Weeks. *The Shape of Space, Third Edition.* Textbooks in Mathematics. Taylor & Francis, 2019. ISBN 9781138062931.

Wikipedia contributors. Clifford torus, 2020. URL `https://en.wikipedia.org/wiki/Clifford_torus`. [Online; accessed 17-May-2020].

Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19:356–365, 2016.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS*, 111(23):8619–8624, 2014.

Kechen Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *Journal of Neuroscience*, 16(6):2112–2126, 1996.

Weiwei Zhang and Steven J. Luck. Sudden death and gradual decay in visual working memory. *Psychological science*, 20(4):423–428, 2009.