Changes in System Dynamics Reflect Consolidation of Memory during Formation and Interference

Bachelor Thesis Research

Szabolcs Máté Mészáros *

22 May 2022

*Supervised by Dr. Mario Senden,
Raphael Stolpe,
Tonio Weidler

Abstract

Recent studies have proposed and tested multiple neural network models to capture some characteristics of perceptual learning in the primary visual area, such as specificity, transfer and interference between visual memory traces. The present thesis aims to complement these results by qualitatively describing the behaviour of the nonlinear recurrent model used by Lange, Senden, Radermacher, and De Weerd (2020). We replicated their simulations of an interference condition, and assessed the dynamics of the model throughout the network's training by computing the kinetic energy, and applied principal component analysis to identify the terminal states of the system after each trial. We present evidence that two principal components (PC1 and PC2) can capture the system's end-states, where PC2 was associated with the system's decision, and PC1 was associated with the kinetic energy of the system. Throughout the training, for each reference angle we observed a decrease in system velocity, and a divergence between terminal states along PC1. Similar results were obtained from a linearised version of the model, which however traded accuracy for computational efficiency. Based on our results, we suggest that anti-Hebbian learning in recurrent networks increases performance by further separating equilibrium points in phase-space, thereby allowing the system to reach a terminal state that is closer to the equilibrium point associated with the correct decision.

1 Introduction

Cognitive neuroscience research has been increasingly focused on the computational description of neuron populations involved in perceptual processing and learning. Knowledge from physiological recordings in humans and animals has inspired neural network architectures and computational rules that allow the modeling and analysis of neuronal dynamics, which in turn may guide further physiological experiments (Blohm, Kording, & Schrater, 2020). Recent studies have proposed and tested multiple neural network models to capture some of the characteristics of perceptual learning in the primary visual area (V1), such as specificity, transfer and interference between visual memory traces(B. Dosher & Lu, 2017; B. A. Dosher, Jeter, Liu, & Lu, 2013; Douglas, Koch, Mahowald, Martin, & Suarez, 1995; Lange et al., 2020; Teich & Qian, 2003). In the present thesis, we aim to provide complementary analyses to the neural network simulations conducted by Lange et al. (2020), in order to describe the behaviour of their network from the dynamical systems approach. We also contrast the original nonlinear model with a linearised version based on the simplifications suggested by Weidler et al. (2021), which may provide a computationally efficient alternative.

1.1 Modeling the Formation of Visual Memory

Neural networks are often comprised of multiple layers with each layer transforming and passing information to the subsequent ones through weighted connections that mimic the widely accepted theory of hierarchical processing in the brain (Biederman, 1987; Herzog & Clarke, 2014). Within this conceptual framework, visual processing is described as a series of extracting increasingly complex features, starting with the detection of contrast by neurons in the retina and lateral geniculate nucleus (Rodieck, 1965). Later processing stages build upon these clues; columns of neurons in V1 have been found to extract information regarding the orientation of detected edges (Hubel & Wiesel, 1959), while V2 neurons are assumed to recognize angles by pooling information from multiple V1 orientation columns (Ito & Komatsu, 2004). Feedforward neural networks with a hierarchical organization are powerful tools in computer vision, and can describe some characteristics of perceptual learning in V1, such as specificity and transfer of learning between tasks (B. Dosher & Lu, 2017). The integrated reweighting model having both retinal location-specific, and invariant layers, can account for location-specific learning, as well as for the transfer of learning between retinal locations (B. A. Dosher et al., 2013). However, feedforward networks do not fully capture the biological mechanisms of perception and perceptual learning, since lateral connections within each layer are completely absent from such networks. Physiological, behavioural and computational studies have proposed that lateral excitatory and inhibitory connections may facilitate the extraction of simple visual objects, such as differently oriented edges (Blakemore, Carpenter, & Georgeson, 1970; Das & Gilbert, 1999; Lange et al., 2020; Sato, Katsuyama, Tamura, Hata, & Tsumoto, 1996). In particular, aminobutyric acid (GABA) modulated lateral inhibitions between orientation columns in V1 have been thought to sharpen orientation tuning curves, enhancing the ability of neuronal populations to discriminate between similar orientations (Blakemore et al., 1970; Khan et al., 2018; Lange et al., 2020; Seriès, Latham, & Pouget, 2004).

1.2 Modelling Lateral Connections in V1

In contrast to feedforward models, recurrent neural networks (RNNs) show promise in exploring the role of lateral connections and feedback, as they implement weighted connections between neurons of the same layer. RNNs have already been proposed as physiologically inspired models of perceptual learning that may be able to explain observations of psychophysical experiments, such as orientation discrimination tasks (Douglas et al., 1995; Lange et al., 2020; Teich & Qian, 2003). In the model proposed by Teich and Qian (2003), the sharpening of tuning curves during learning may result from decreasing excitatory lateral connections, while the increase of both excitatory and inhibitory lateral connections could lead to the peak-shift of tuning curves. This model was based on physiological findings by Schoups, Vogels, Qian, and Orban (2001) and Dragoi, Sharma, and Sur (2000), and utilized signal

detection theory to quantify perceptual consequences of learning, and to compare these consequences to psychophysical observations. While Teich and Qian (2003) claimed their model conformed to psychophysical observations, their study was implemented exclusively *in silica* without direct comparison to participant data.

However, these limitations have been addressed later by Lange et al. (2020), who collected psychophysical data from 8 participants. They conducted 3 experiments (Figure 1), each involving a series of orientation discrimination tasks with Gabor stimuli. Experiments included an ABA (Experiment 2) and a BAB (Experiment 3) design with different reference angles expected to interfere, and a control experiment (Experiment 1). They compared the collected data to a set of model simulations reflecting the same experiments. These simulations were carried out in MATLAB (2022) and utilized the ode45 algorithm to simulate the time-course of neural activations until a stable pattern is reached. Their neural network was based on Teich and Qian (2003), adjusted to only increase inhibitory connection strengths when learning. This modification was made to reflect the findings of Khan et al. (2018), who showed changes in the selectivity of GABA-ergic interneurons as a result of perceptual learning. Lange et al. (2020) reported that the learning curve of their network closely resembled the learning curve of human participants. Moreover, they showed that the sharpening of tuning curves through modification of inhibitory connections could account for patterns of interference and transfer of learning when participants are subsequently trained in discrimination tasks with interfering reference orientations Lange et al. (2020).

Physiological and computational studies have revealed that lateral connections in V1 have a 'Mexican hat' shaped profile, where each orientation shows the strongest excitatory connections with its neighboring columns, but expresses inhibitory connections to the vicinity of its neighbors (Müller, Mollenhauer, Rösler, & Kleinschmidt, 2005; Sato et al., 1996). This has been suggested to increase the acuity of orientation discrimination, as inhibition around the target orientation may reduce ambiguity (Müller et al., 2005; Teich & Qian, 2003). In the models used by Lange et al. (2020); Teich and Qian (2003), this characteristic shape was estimated by combining inhibitory and excitatory connection distributions expressed by Gaussian modulated periodic functions. In contrast, Weidler et al. (2021) expressed the initial recurrent weight distribution in their model via the Ricker-wavelet, establishing excitatory and inhibitory connections with a single pair of parameters.

1.3 The Dynamics of RNNs

The qualitative description of the dynamics of RNNs is a notoriously difficult task, since the result of feedback loops present in such systems is hard to predict (Wang, 2008). While previous research was successful in producing and quantitatively validating models of perceptual learning, they only offered speculations as to how they worked (Lange et al., 2020; Teich & Qian, 2003). The dynamical systems perspective offers a way to complement previous studies with a qualitative description of the temporal dynamics, by mapping and describing the stability of critical points, such as fixed-points and slow-points (Archer, Köster, Pillow, & Macke, n.d.; Sussillo & Barak, 2013). This may be achieved with the help of an auxiliary function that represents the speed, or 'kinetic energy' of the system (Equation 10). A fixed point, by these terms, can be defined as a state where the speed of the system equals zero, and similarly a slow point can be defined as a state where the speed is close to zero (Sussillo & Barak, 2013). It is assumed that, whenever the activity of a dynamical system is simulated for a sufficiently long time, the system will reach a stable fixed point, from where it cannot escape.

Based on this assumption, Weidler et al. (2021) showed that it is possible to reverse-engineer the dynamics of RNNs by estimating recurrent feedback via a linear feedforward step that utilizes the Jacobian of the weight matrix, and directly solves for the equilibrium point (Equation 9). This requires the linearisation of the recurrent network, since the Jacobian only captures the linear term of the Taylor-expansion. This approach is expected to have higher computationally efficiency than frequently used simulation-based approaches, such as the one used by Lange et al. (2020). However to our knowledge, it has not yet been evaluated in biologically plausible models. Qualitative description of V1 neural dynamics may help guide future research into cortical stimulation procedures by highlighting struc-

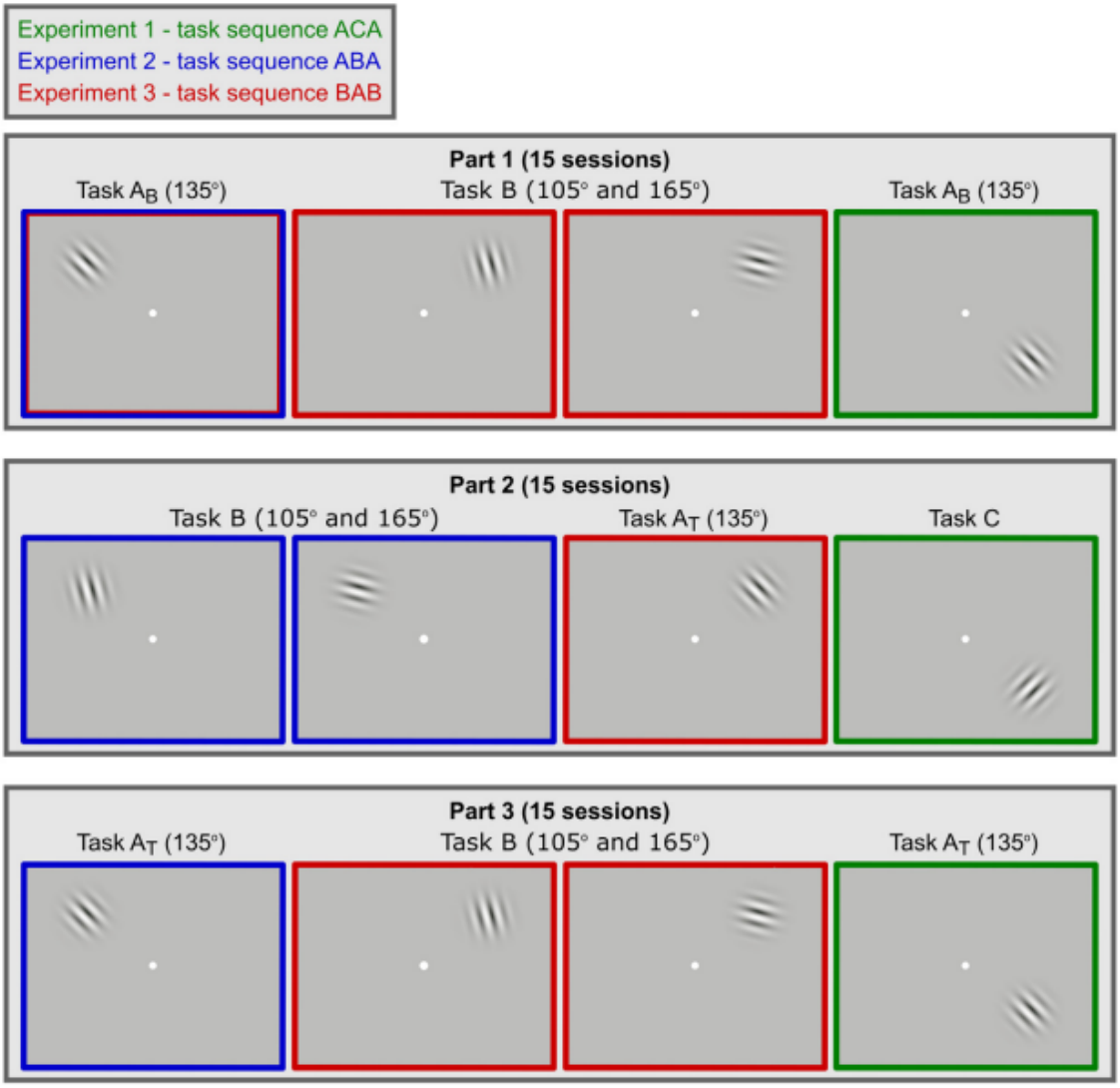

Figure 1: Study design of Lange et al. (2020). Reprinted from "Interfering with a memory without erasing its trace" by Lange, G., Senden, M., Radermacher, A., & De Weerd, P., 2020, *Neural Networks*, *24*(13), 3313. Copyright 2020 by The Authors. Reprinted with permission.

tural influences on visual perception and perceptual learning. Cortical stimulation shows promise in the treatment of numerous neurological disorders, such as Parkinson's disease or Migraine (Lefaucheur, 2009). Specifically, microstimulation of V1 areas specifically has been proposed as a potential treatment for cortical blindness (Tehovnik, Slocum, Smirnakis, & Tolias, 2009). With the present research, we aimed to explore the behaviour of the recurrent neural network modelling perceptual learning in V1 used by Lange et al. (2020) from the dynamical systems perspective. We hypothesized that a slow point or fixed point would be reached by the model after a sufficient amount of simulation time. Furthermore, we expected the geometric relationship of stable points to change over the course of learning, allowing better representations of the input stimuli. Our second objective was to evaluate the formulation proposed by Weidler et al. (2021) in a biologically plausible model of orientation selectivity. We expected to obtain comparable, if slightly worse performance from this system.

2 Methods

As the present study aimed to complement the experiments conducted by Lange et al. (2020), we have adopted their recurrent model, simulation parameters and analyses. We have chosen to implement Experiment 2 in our simulations (see Figure 1) in order to explore the dynamics both during formation and interference of memory traces Lange et al. (2020). Here, baseline training at the 135° reference angle (A_b) is followed by training at interfering reference angles counterclockwise (105°) and clockwise (165°) to the previous angle (B_left and B_right respectively). Afterwards, the interference effect is tested again at 135° (A_t). In line with Lange et al. (2020), our network was trained over 8 sessions, with each session consisting of 480 trials. Just-noticable-differences (JNDs) averaged over the trials were calculated for each session at each reference angle. The direction of the probe compared to the reference (left or right) was determined pseudo-randomly, therefore we averaged the mean JNDs over 5 repetitions to counterbalance potential random effects. We also obtained tuning curves (TCs) after training at each reference in order to replicate the findings of Lange et al. (2020) by presenting stimuli to the model with orientations between 45°and 225°with 1°difference between stimulus orientations. Both neural network models and all analyses were coded in MATLAB (2022) (scripts and recurrent model classes available at https://github.com/skandar98/BTR).

2.1 Nonlinear Recurrent Model

Our nonlinear recurrent model (RM-N) of V1 orientation tuning was identical to the one used by Lange et al. (2020). It consists of a single layer of $N = 512$ neurons, each tuned to an orientation covering 180 degrees, so that their firing rate $R(\theta, \phi, t)$ at time t is maximal, when the presented orientation ϕ is equal to the preferred orientation θ. Calculation of the firing rate introduces the nonlinear term into the system by transforming membrane potential (V) via the Rectified Linear Unit (ReLU) function and gain factor (α):

$$R = \alpha \max(V, 0) \tag{1}$$

Membrane potential (V) in the recurrent model evolves according to the differential equation

$$\tau \Delta V = -V + I_f + I_r \tag{2}$$

where τ is the membrane time constant, I_f is the feedforward input received by the neuron and I_r is the recurrent input from neighboring neurons. Since no additional layers are involved, the feedforward input received by a neuron with preferred orientation θ as a result of presenting orientation ϕ is given by

$$I_f(\theta, \phi) = J_f exp\left[-\frac{\left[\angle e^{2i(\theta-\phi)}\right]^2}{8\sigma_f^2}\right] \tag{3}$$

where J_f determines the strength and σ determines the width of the input.

$$I_{r,j}(\theta_j, t) = \sum_{i=1}^{N} w_{j,i} R(\theta_j, t) \tag{4}$$

We chose to initialize recurrent weights with the difference-of-Gaussians method, the same way as Lange et al. (2020), since this is a close approximation to to the Ricker-wavelet formulation used by Weidler et al. (2021), but allows direct control over the inhibitory and excitatory connections separately. The net connection strength from neuron i and to neuron j is given by ($W_{ji} = W_{exc,ji} - W_{inh,ji}$). Excitatory and inhibitory connections from between neuron j and i, having preferred orientations θ_j and θ_i respectively, have been initialized as

$$\begin{aligned} \mathrm{W_{exc,ji}} &= J_r E(\theta_i - \theta_j) \\ \mathrm{W_{inh,ji}} &= J_r I(\theta_i - \theta_j) \end{aligned} \tag{5}$$

where J_r is scaling the strength of recurrent connections and E_θ and I_θ represent the probability distribution of excitatory and inhibitory connections respectively, expressed by periodic functions

$$\begin{aligned} E_\theta &= c_e \left[cos(2\theta) + 1\right]^{a_e} \\ I_\theta &= c_i \left[cos(2\theta) + 1\right]^{a_i} \end{aligned} \tag{6}$$

where c_e and c_i serve to normalize the distributions and exponents a_e and a_i control their sharpness. Learning in the model is implemented only at incorrect decisions through increasing inhibitory weights, which is mathematically equivalent to the anti-Hebbian learning rule. The change in inhibitory weights is described by

$$\Delta W_{inh,ji} = \mu R(\theta_j, \phi_p robe) R(\theta_i, \phi_p robe) \tag{7}$$

where μ denotes the learning rate.

2.2 Linearised Recurrent Model

Our linear recurrent model (RM-L) was based on the previously described nonlinear version with small modifications that made it possible to avoid using the ode45 algorithm in estimating the stable state of the system. In particular, Equation 4 was replaced by the following equation, where the recurrent input to neuron j is given by the linear formula:

$$I_{r,j}(\theta_j) = \sum_{i=1}^{N} w_{j,i} V(\theta_j) \tag{8}$$

The ReLU function was still utilized in this version of the model to calculate firing rates for the purposes of decision making, however this does not influence the behaviour of the model. Based on Weidler et al. (2021), the following equation was used to directly solve for the stable state of a neuron with preferred orientation θ for presented orientation ϕ

$$V(\theta, \phi) = -(\mathbf{W} - \mathbf{I})^{-1} I_f(\theta, \phi) \tag{9}$$

where $\mathbf{I}$ is the identity matrix, and $\mathbf{J} = (\mathbf{W} - \mathbf{I})$ is the Jacobian of the dynamical system. $I_f(\theta, \phi)$ was calculated according to equation 3. Furthermore, we increased the the learning rate for this network to 10^4 times higher than what we used in the nonlinear model in order to obtain weight gradients of similar magnitude.

2.3 Analysis of system dynamics

After each trial, we sampled the orientation of the probe and the firing rate of each unit resulting from both the presentation of the reference and probes, for a total of 3840 samples (480 trials * 8 sessions) for each reference angle condition. Sampling was done only in the first repetition of the simulation, since the pseudo-random nature of the presented stimuli does not have a qualitative influence on system dynamics. Firing rates were sampled at the end of the differential equitation simulation for RM-N. We simulated the activity of

the nonlinear system between 0 and 0.5 seconds after confirming that the firing rates do not change past 0.5 seconds, indicating that the system reached a stable state. The linear system was assumed to be at a stable state at the solution for Equation 8). The kinetic energy of both systems at these stable states was calculated by

$$q = \frac{1}{N}(\Delta \mathbf{V}^T \Delta \mathbf{V}) \tag{10}$$

where $\Delta \mathbf{V}$ is given by Equation (2).

Finally, we applied principal component analysis (PCA) to the resulting 512-dimensional states in order to explore their relationships to each other in 2 or 3 dimensions. The first principal components able to explain 99.99% variance were selected for analysis.

3 Results

3.1 Evolution of Stable Points during Learning and Interference

First, to analyze the stability of nonlinear and linear systems' terminal states we assessed the kinetic energy of both systems in each session averaged over the trials. For RM-N, we found a steady decrease in kinetic energy (q-values) in all 4 reference conditions over the learning sessions (Figure 2). Notably, the trajectory and magnitude of kinetic energy associated with the presentation of references were almost identical to the trajectory and magnitude associated with the presentation of probes. The lowest q-values both in the initial and terminal phases of learning were achieved by the system in the AT condition, and the highest q-values can be observed in the BL condition. The magnitude of q-values obtained from the AB and BR conditions were similar. In contrast, the evolution of kinetic energy in the RM-L model was the opposite as observed in the RM-N model, as it steadily increased throughout the training sessions (Figure 3). The relationship of the different conditions with regards to kinetic energy was also reversed, the highest q-values were reached in AT, followed by BR, then BL, and finally AB. Moreover, the magnitude of q-values in all sessions was 10^5 times larger than what we observed in the RM-N.

The application of PCA to the 512-dimensional terminal states resulting from the presentation of reference and probe angles resulted in the cumulative explained variances by principal components shown in Figure 4. In all reference conditions in both the nonlinear and linear models, 99.99% explained variance was achieved by the first 2 principal components. Although in most conditions the first principal component was already able to explain 99.99% variance, we selected the first 2 principal components for analysis in all conditions.

Projecting the stable states of the RM-N model to principal components 1 (PC1) and 2 (PC2) resulted in a characteristic pattern, where the states representing the reference orientations and probes align to vertical lines along PC1 (Figure 5 and Figure 6). The position of states along PC1 is associated with the kinetic energy obtained at the respective states, as q-values tend to decrease as PC1 decreases.The variation of training sessions also seems to follow this trend, with later sessions positioned towards the bottom of the plots. Furthermore, a separation of the vertical lines is observable along the PC2 axis. Along this principal component, a trend in orientation angles is observable, with the reference angle positioned at 0, more negative values depicting orientations clockwise from the reference, and more positive values corresponding to orientations counterclockwise from the reference. While a similar pattern can be observed in the results of the linearised model (Figures 7 and 8), the relationship between PC1 and the kinetic energy of the system is, again, the opposite as was observed in RM-N. Furthermore, in RM-N a slight divergence of points along PC2 can be noticed as PC1 decreases, increasing the distance between the reference and probe orientations, while this relationship seems to be absent in the linearised version. In both models, training in the AB and AT conditions (Figure 5 and 7) resulted in a symmetrical plot, whereas the interfering conditions (Figure 6 and 8) resulted in asymmetrical plots. Particularly, during training at left interference (105°), the terminal states resulting from the presentation of probes with angles smaller than 105°are positioned higher along the PC1 axis, and probes with angles larger than 105°are positioned lower along the PC1 axis, with states reflecting the reference angle positioned between the two. The opposite pattern

can be observed in the right interference condition (165°), although the asymmetry is less pronounced in this condition.

3.2 Learning and Tuning Curves in Nonlinear and Linearised Recurrent Models

Our nonlinear model showed an increase in performance over 8 sessions in all conditions, indicated by the decrease in JNDs (Figure 9). The best performance was achieved during the AB condition (JND = 3.1°), followed by the AT condition (JND = 3.48°). In contrast, the best performance reached by the linear model (Figure 10) was 11.57°JND, which is considerably higher than the performance of RM-N. Furthermore, the performance of RM-L in the AB and AT learning curves showed large differences, whereas in RM-N, the initial and final performance in AB and AT were highly similar. The trajectory of performance in both models was similar between the BL and BR conditions.

The tuning curves obtained from RM-N (Figure 11) were considerably sharper than the tuning curves obtained from RM-L (Figure 12). In both models, a depression of the tuning curves of neurons with a preferred orientation near the reference angle is observable after training. In the naive states, all tuning curves have equal height and width. After baseline training at 135°, the peaks of tuning curves around this angle are lower compared to further orientations. Training at 105°(BL) afterwards produces a shift in this depression to the left side, while successive training at 165°(BR) shifts the depression to the right, seemingly negating the changes from the BL condition. Repeating the training at 135°(AT) further reduces the height of peaks around this angle. Interestingly, these tendencies are observable in both RM-N and RM-L, but changes are more pronounced in the linear version.

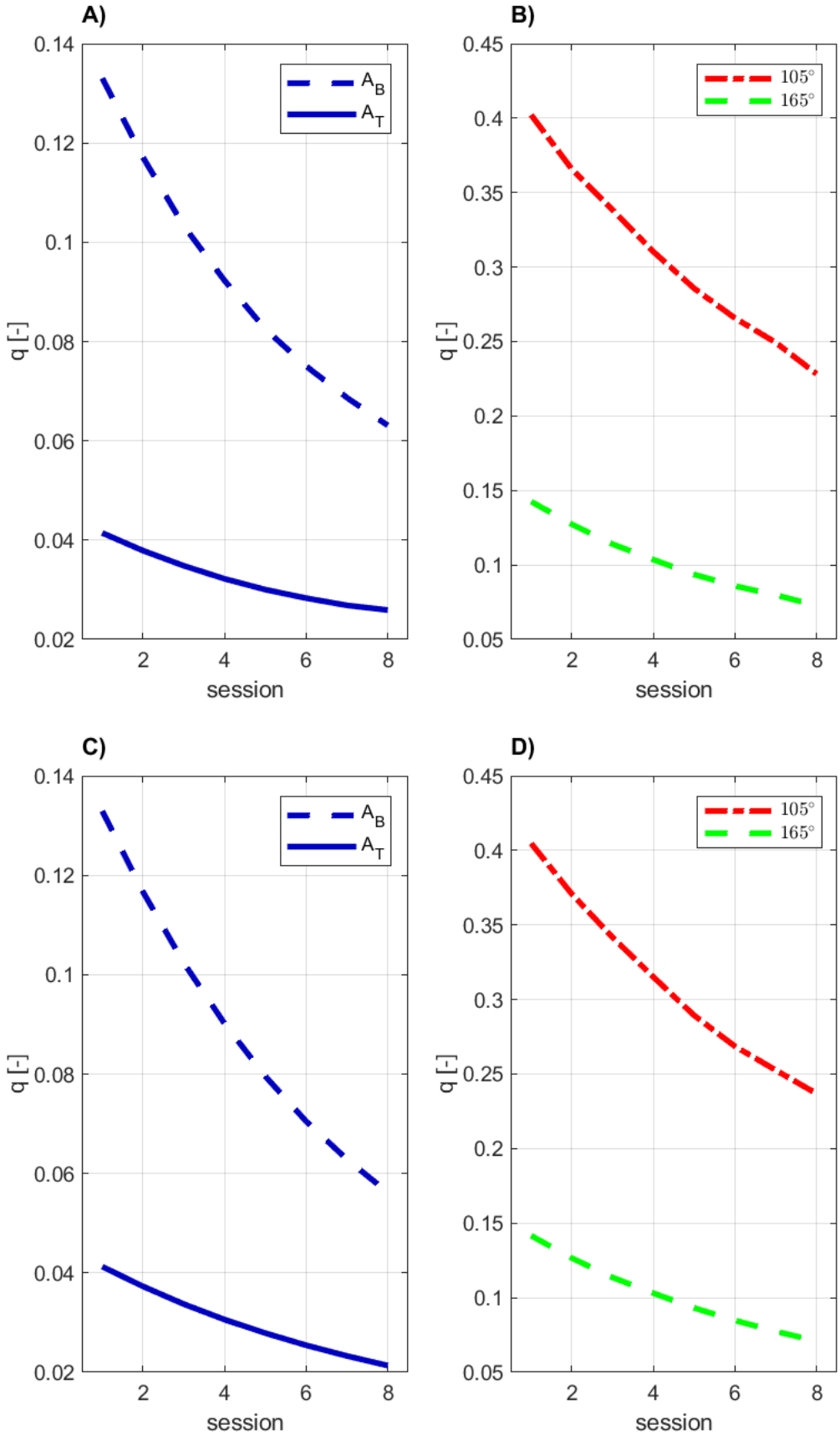

Figure 2: Evolution of kinetic energy over the training session in the **RM-N** model. Top row (**A)** and **B)**) depicts average q-values obtained from presentation of the probe angles, bottom row (**C)**, **D)**) depicts q-values obtained from presentation of the reference angles. AB and AT conditions (135°reference) are represented by dashed and solid blue lines respectively in plots **A)** and **C)**. BL (105°reference) and BR (165°reference) conditions are shown by the red and green lines respectively in plots **B)** and **D)**.

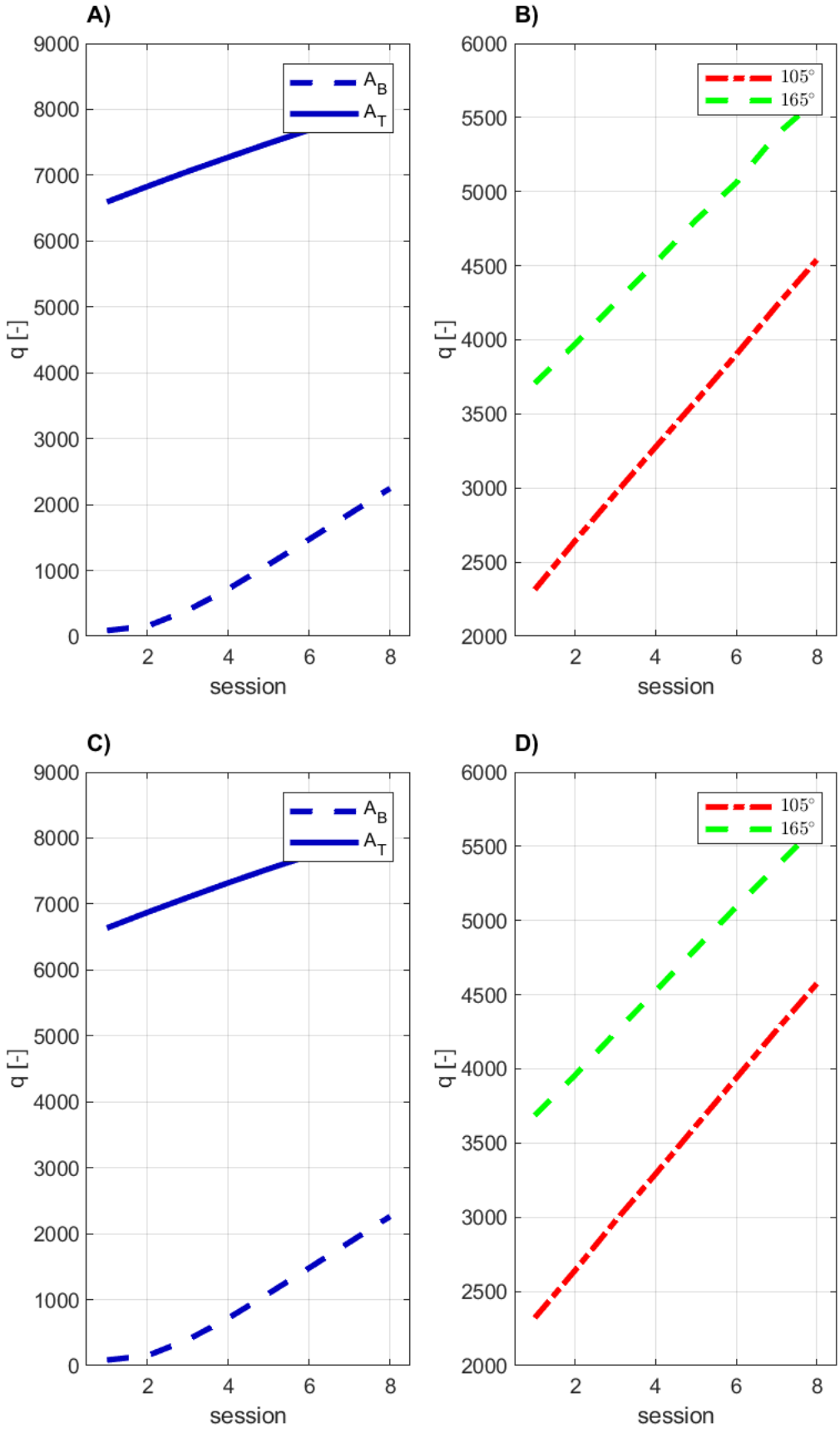

Figure 3: Evolution of kinetic energy over the training session in the **RM-L** model. Top row (**A)** and **B)**) depicts average q-values obtained from presentation of the probe angles, bottom row (**C)**, **D)**) depicts q-values obtained from presentation of the reference angles. AB and AT conditions (135°reference) are represented by dashed and solid blue lines respectively in plots **A)** and **C)**. BL (105°reference) and BR (165°reference) conditions are shown by the red and green lines respectively in plots **B)** and **D)**.

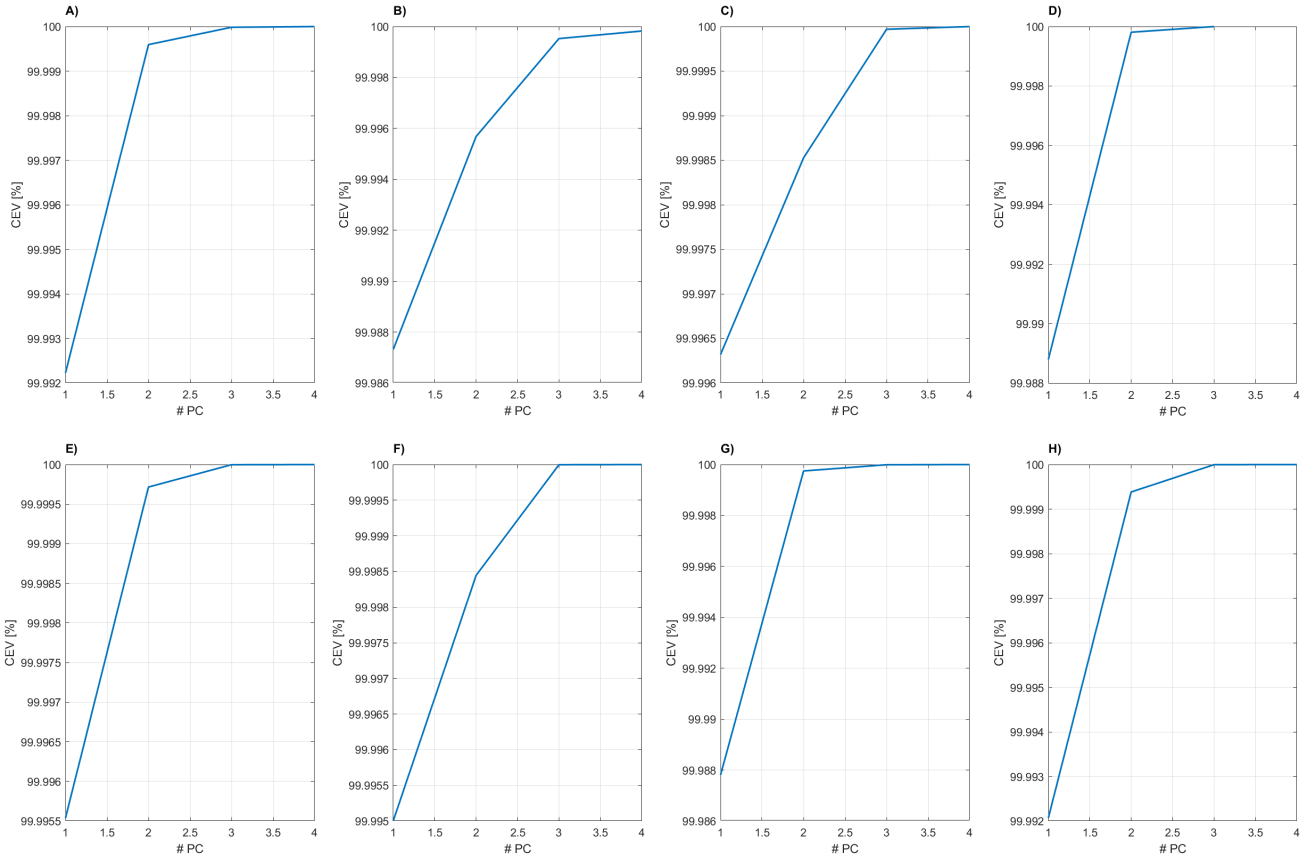

Figure 4: Cumulative explained variance (CEV) of terminal firing rates by principal components 1-4. In the top row, CEV from the RM-N model is shown, while the bottom row depicts CEV from the RM-L. **A)** and **E)** represent the AB condition (135°), **B)** and **F)** represent the BL condition (105°), **C)** and **G)** represent the BR condition (165°), and **D)** and **H)** represent the AT condition (135°).

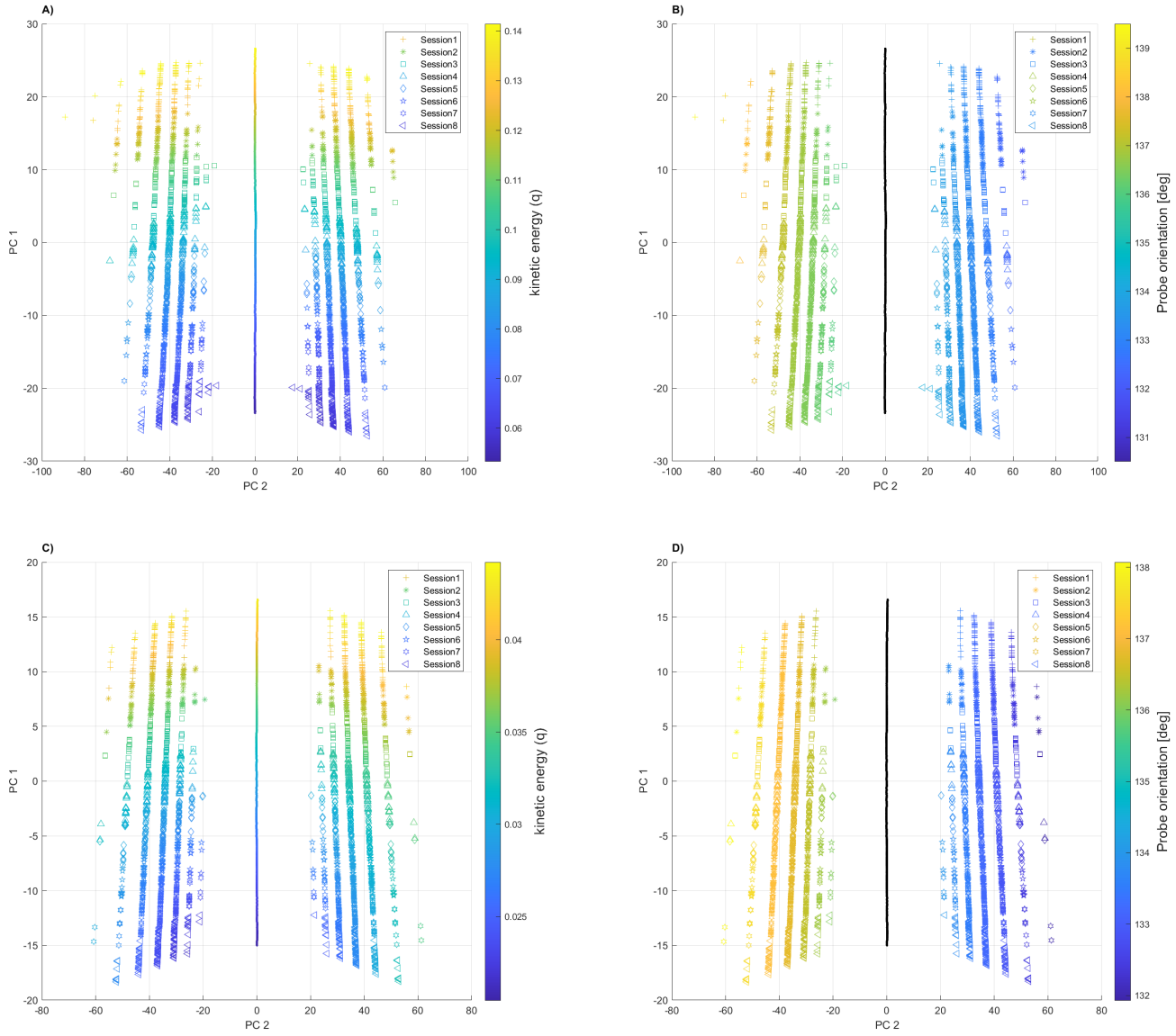

Figure 5: Terminal firing rates of the **RM-N** model in the interference conditions projected into two principal components. Plots in the top row represent the BL condition (105°), while the bottom row represents the BR condition (165°). Points in the left column are colored by the magnitude of kinetic energy (q-values), and points in the right column are colored by the orientation of the presented probes in respective trials, with black points representing the reference angle. Sessions 1-8 of the training cycle are depicted with different markers, and full circles represent the reference orientation in all sessions.

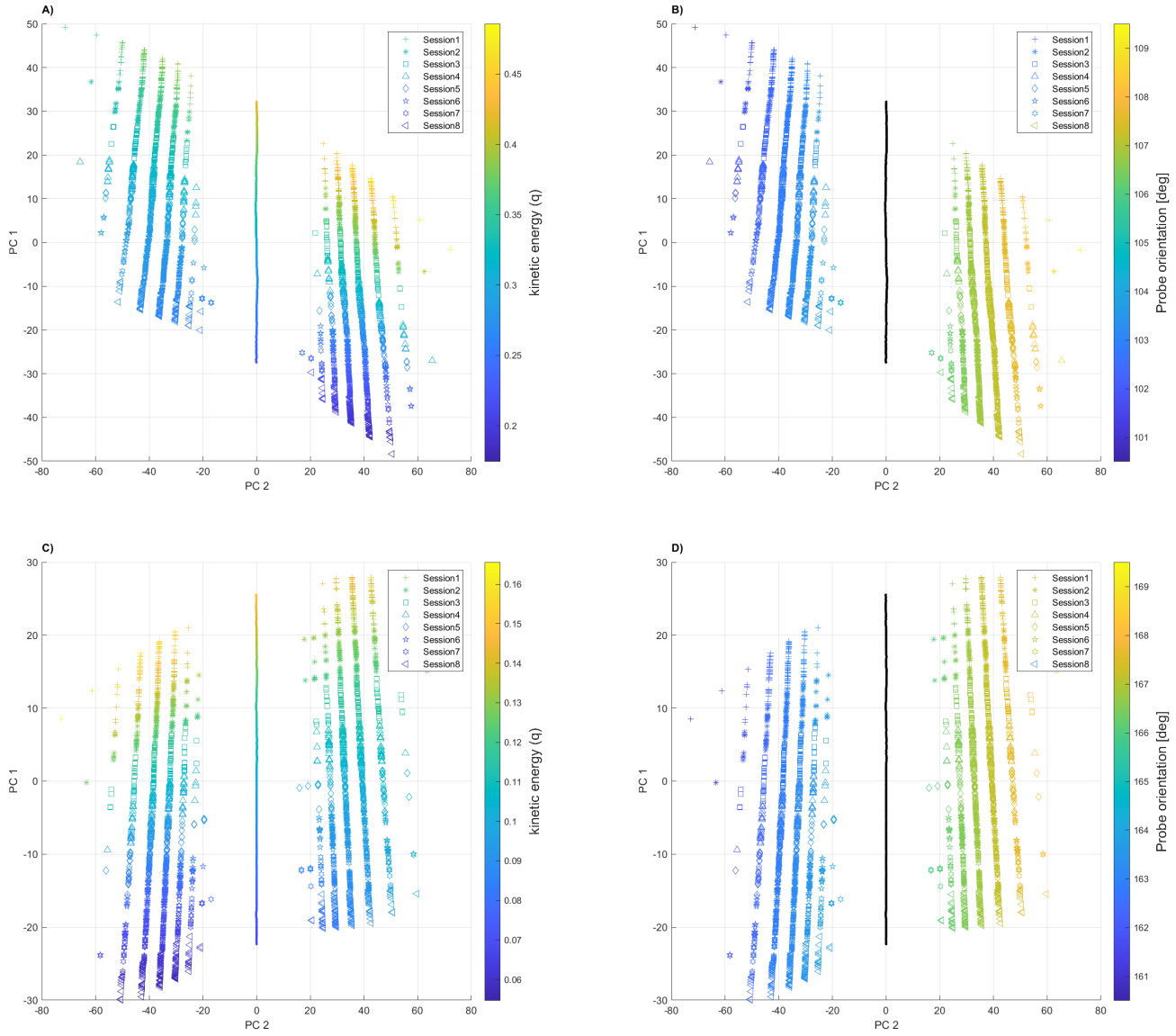

Figure 6: Terminal firing rates of the **RM-L** model in the reference conditions (135°) projected into two principal components. Plots in the top row represent the AB condition, while the bottom row represents the AT condition. Points in the left column are colored by the magnitude of kinetic energy (q-values), and points in the right column are colored by the orientation of the presented probes in respective trials, with black points representing the reference angle. Sessions 1-8 of the training cycle are depicted with different markers, and full circles represent the reference orientation in all sessions.

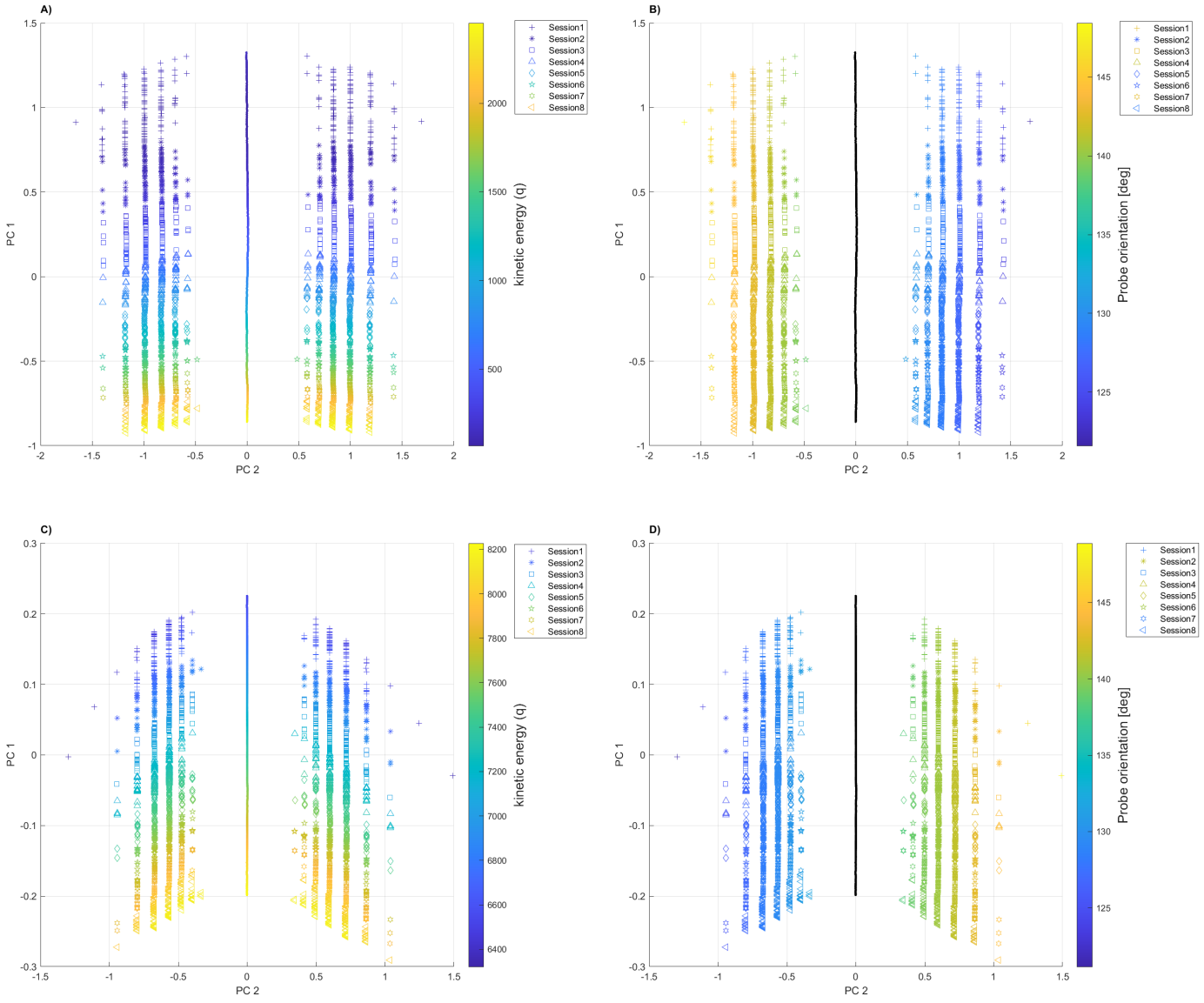

Figure 7: Terminal firing rates of the **RM-L** model in the interference conditions projected into two principal components. Plots in the top row represent the BL condition (105°), while the bottom row represents the BR condition (165°). Points in the left column are colored by the magnitude of kinetic energy (q-values), and points in the right column are colored by the orientation of the presented probes in respective trials, with black points representing the reference angle. Sessions 1-8 of the training cycle are depicted with different markers, and full circles represent the reference orientation in all sessions.

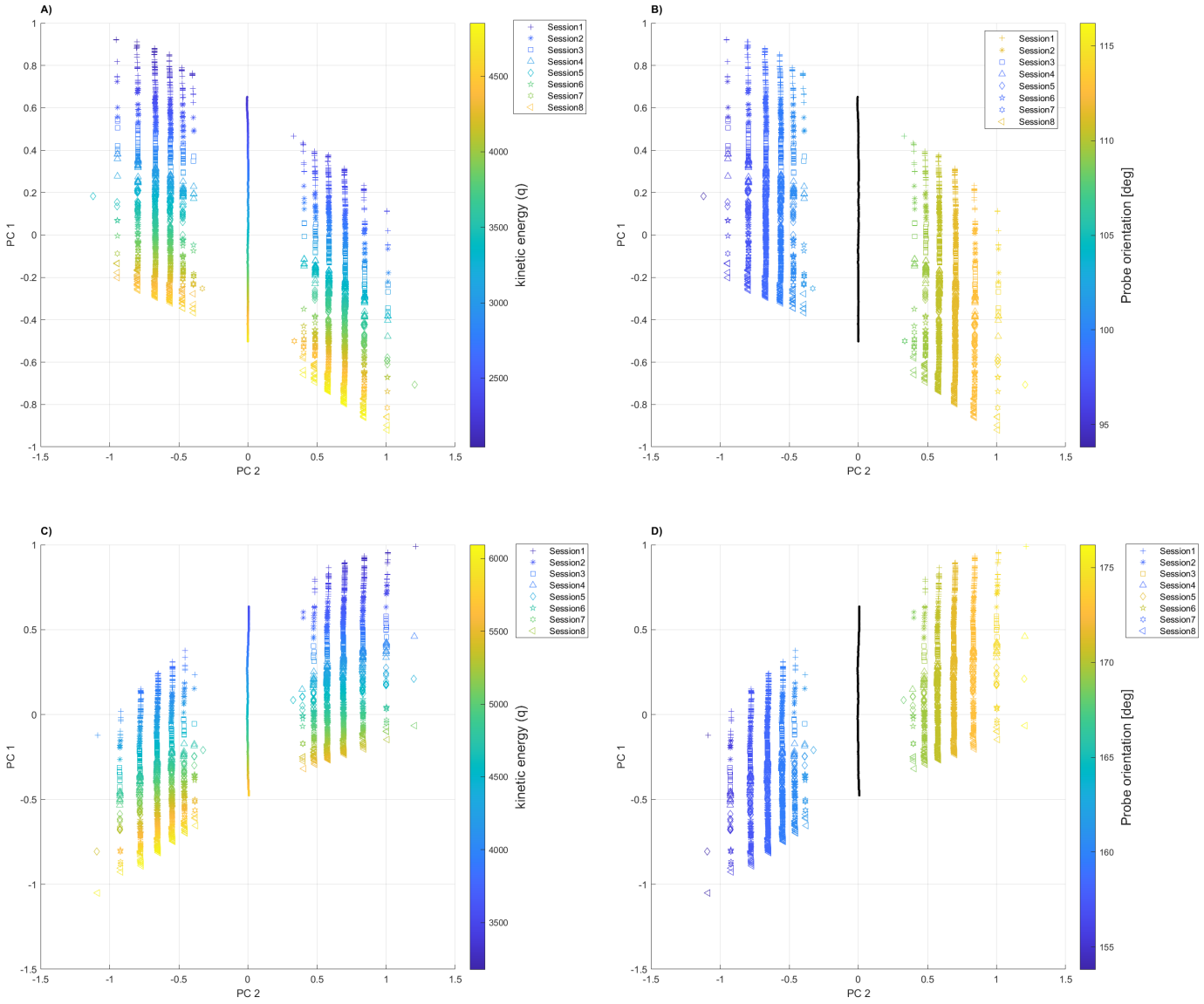

Figure 8: Terminal firing rates of the **RM-L** model in the reference conditions (135°) projected into two principal components. Plots in the top row represent the AB condition, while the bottom row represents the AT condition. Points in the left column are colored by the magnitude of kinetic energy (q-values), and points in the right column are colored by the orientation of the presented probes in respective trials, with black points representing the reference angle. Sessions 1-8 of the training cycle are depicted with different markers, and full circles represent the reference orientation in all sessions.

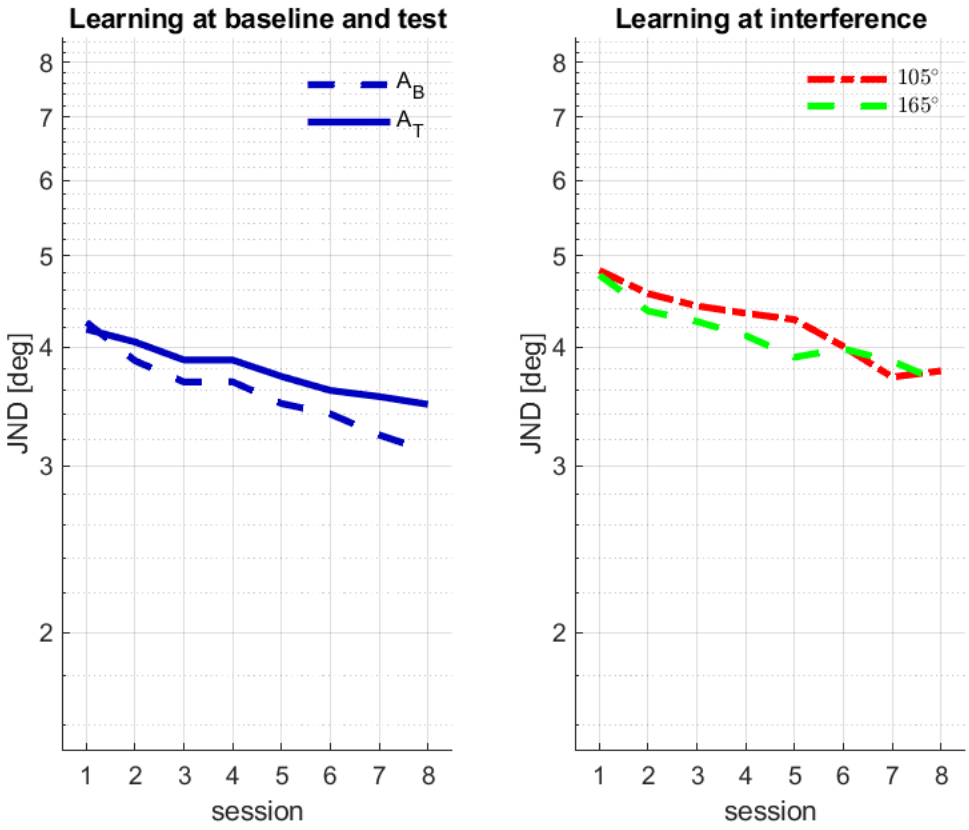

Figure 9: Learning curves of the **RM-N** model over 8 session. In the left plot, dashed and solid blue lines represent the baseline and test reference conditions (135°). In the right plot, red and green lines represent left interference and right interference conditions, respectively. The performance of the system was evaluated using the JNDs averaged over 5 repetitions.

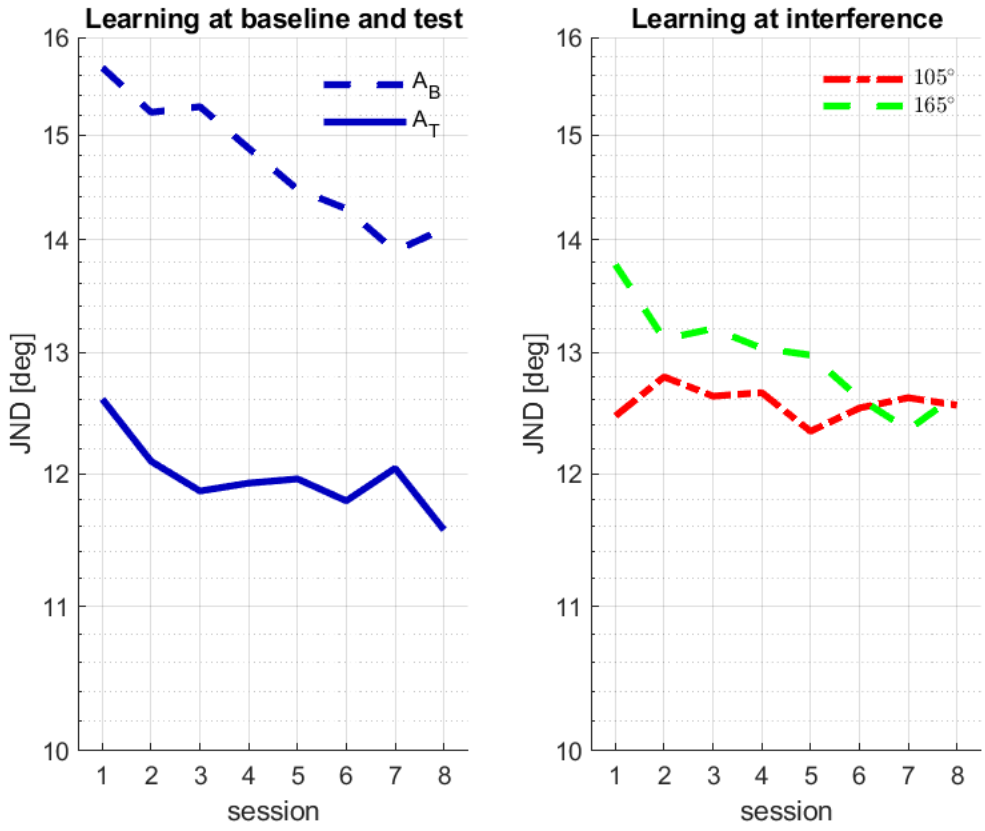

Figure 10: Learning curves of the **RM-L** model over 8 session. In the left plot, dashed and solid blue lines represent the baseline and test reference conditions (135°). In the right plot, red and green lines represent left interference and right interference conditions, respectively. The performance of the system was evaluated using the JNDs averaged over 5 repetitions.

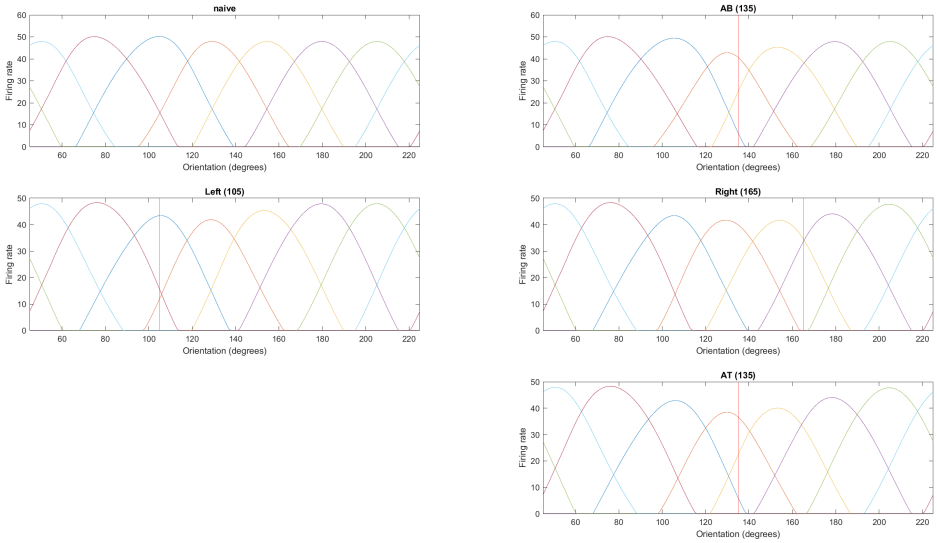

Figure 11: Tuning curves of 7, linearly spaced neurons in the RM-N model obtained from the naive state, after baseline training (AB), after left and right interference training and after test training (AT). Firing rates of each selected neuron were obtained by presenting 180 angles spanning from 45°to 225°. Red vertical lines indicate the reference angle at the respective conditions.

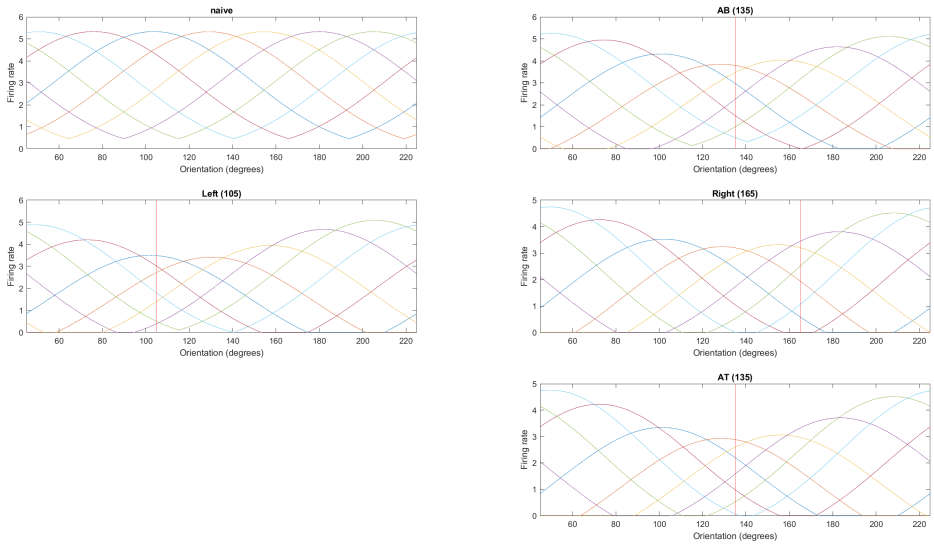

Figure 12: Tuning curves of 7, linearly spaced neurons in the RM-N model obtained from the naive state, after baseline training (Ab), after left and right interference training and after test training (AT). Firing rates of each selected neuron were obtained by presenting 180 angles spanning from 45°to 225°.Red vertical lines indicate the reference angle at the respective conditions.

4 Discussion

The present study aimed to replicate, and provide complementary analyses for the V1 recurrent model simulations used by Lange et al. (2020). Furthermore, we compared the original nonlinear model to a linearised version, which was expected to have higher computational efficiency, as it circumvented the necessity of using the ode45 algorithm. By taking a dynamical system approach, we aimed to provide a different, but complementary explanation for the behaviour of the recurrent networks modelling the learning and interference patterns observed in orientation discrimination tasks by human participants. In their paper, Lange et al. (2020) argued that learning and interference patterns result from the characteristic peak-shifts of tuning curves, which in turn emerge as a consequence of increased inhibitory connections around the neuron with preferred orientation equalling the presented reference orientation. In line with this explanation, in both models we found a depression in tuning curves at baseline around the reference angle compared to the naive state, and a shift in this depression resulting from training in left and right interference conditions towards the respective sides (Figures 11 and 12). Replicating the learning curves from Lange et al. (2020) by assessing the changes in mean JNDs over 8 sessions in RM-N produced similar learning trajectories as in Lange et al. (2020). This may be regarded as further confirmation of their results, although we have not compared this learning curve to that of human subjects.

We also described the learning process in terms of the speed, or kinetic energy (q) of the systems at their final stage in each trial. We assumed that when the state of the system ceases to change during simulation (at 0.5 seconds in our case), the system would arrive at a stable equilibrium point, where its kinetic energy equals zero, or is close to zero. However, we found that the kinetic energy of the nonlinear system does not reach zero at these points, although it approaches zero during the training cycle (Figure 2). It is unclear why the firing rates remained unchanged upon simulating further when the system had not reached zero kinetic energy, however, we speculate that the system may be circling such an equilibrium point, but is unable to reach it because of a high momentum that may result from the attraction by neighboring equilibrium points. Future research could further explore the dynamics of the system with other methods, such as the identification of its actual fixed points, where $q \approx 0$ via an optimization algorithm, and the assessment of the stability of identified fixed points through linearisation around the fixed points.

However, most importantly Figure 2 suggests that a lower kinetic energy at terminal states

may be associated with better performance. This suggestion is further supported by Figures 5 and 6, where a clear association between PC1, session number and kinetic energy can be observed. Notably, these plots also show a divergence of probe-associated terminal states from the reference-associated states along PC2, although this effect has not been quantified. The divergence of terminal states may reflect the divergence of the actual equilibrium points. We suggest based on these observations, that as the equilibrium points representing probe orientations diverge from the equilibria representing the reference orientation, attraction by neighboring stable points may weaken, which could result in the system's ability to approach the correct stable point further during simulation, resulting in lower kinetic energy at these states. This effect could explain the increased performance of the system, as it reduces ambiguity between different orientations.

Figures 6 and 8 also support the findings of Lange et al. (2020) with regards to the interference of reference angles, and provides an alternative description of the models' behaviour in these conditions. The asymmetry of the plots suggest that, in the left interference condition (105°), the system can better identify angles to the right side of the reference, which are closer to the previously trained reference (135°), while in the right interference condition, the opposite is true. The asymmetry in the BL condition may be more pronounced because this condition preceded the BR condition, therefore training in the AB condition has not been negated yet by an interference. This is in line with our tuning curve results, where the shift in tuning curve peaks in BR is less marked than in BL (Figures 11 and 12).

Finally, when comparing our results from RM-N and RM-L, it is clear that the linearised version traded considerable amount of accuracy for computational efficiency. With our code running on our system, this version was simulated about 10 times faster. However, comparing Figure 9 and Figure 10 clearly shows that RM-L was not able to achieve the performance of RM-N, and therefore it is likely not comparable to human performance, thereby limiting the usefulness of the simplification suggested by Weidler et al. (2021) in biologically plausible models, at least in the form implemented in this thesis. Besides being less accurate, RM-L also does not show the expected interference effect, as the learning curve of AT starts at the performance achieved in the final sessions of AB training (Figure 10). We speculate that the discrepancy between RM-N and RM-L in terms of performance, and the anomalous finding that the kinetic energy of the RM-L increases with learning may be explained 1) by the inability of linear models to estimate nonlinear terms of the Taylor-expansion, 2) by numerical instabilities resulting from the inversion of the Jacobian, or 3) by an incompatibility between the way we solve for equilibrium points (Equation 9) and the learning rule we employ (Equation 7) Although, we also have to keep in mind that the tuning curve shifts in RM-L conformed to those we observed in RM-N, and to the hypothesis of Lange et al. (2020), which suggests that the behaviour of the linear model may also be able to capture biological learning in its current form. Therefore, further research is necessary to explore the exact reason for the anomalous behaviour of the linearised recurrent model.

5 Conclusion

In the present thesis, we provided complementary analyses on the recurrent V1 model simulations done by Lange et al. (2020), and compared the original nonlinear network to a linearised version. While the linear model showed higher computational efficiency, it is unable to model biological learning in its current form. Our tuning and learning curve results from the nonlinear model conformed to the findings of Lange et al. (2020). Furthermore, we showed through principal component analysis and the computation of the model's 'kinetic energy' that the dynamics of the system in its final state slow down as the training progresses. We were able to represent final states with two principal components explaining 99.99% of the variance (PC1 and PC2), which were associated with the kinetic energy or learning progress, and with the model's decision in each trial, respectively. We observed a divergence of probe-associated states along the PC2 dimension over the training period, which indicate that the consolidation of visual memory traces may result from increasing the distance between equilibrium points representing each possible decision, thereby decreasing the ambiguity between decisions.

6 Critical reflection

I believe the most important thing I have learned during this thesis is the necessity of identifying which analyses/methods are possible to carry out during our limited time for research. In the first month I have tried to force a method that we ended up discarding in favor of a much simpler analysis. This could have been easily avoided, and I would have had more time implementing other methods. In the future I will take more care to organize my work and resources (such as scripts, simulation results, etc.), if I am taking a computational approach to studying a topic, as this too could have saved a considerable time.

References

Archer, E., Köster, U., Pillow, J. W., & Macke, J. H. (n.d.). Low-dimensional models of neural population activity in sensory cortical circuits. In *Nips.*

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115-147. https://doi.org/10.1037/0033-295x.94.2.115

Blakemore, C., Carpenter, R., & Georgeson, M. (1970). Lateral inhibition between orientation detectors in the human visual system. *Nature*, *228*, 37-9. Retrieved from `https://www.nature.com/articles/228037a0` https://doi.org/10.1038/228037a0

Blohm, G., Kording, K. P., & Schrater, P. R. (2020). A how-to-model guide for neuroscience. *eneuro*, *7*(1), ENEURO.0352-19. Retrieved from `https://dx.doi.org/10.1523/eneuro.0352-19.2019` https://doi.org/10.1523/eneuro.0352-19.2019

Das, A., & Gilbert, C. D. (1999). Topography of contextual modulations mediated by short-range interactions in primary visual cortex. *Nature*, *399*(6737), 655-661. https://doi.org/10.1038/21371

Dosher, B., & Lu, Z.-L. (2017). Visual perceptual learning and models. *Annual Review of Vision Science*, *3*(1), 343-363. https://doi.org/10.1146/annurev-vision-102016-061249

Dosher, B. A., Jeter, P., Liu, J., & Lu, Z.-L. (2013). An integrated reweighting theory of perceptual learning. *Proceedings of the National Academy of Sciences*, *110*(33), 13678. Retrieved from `http://www.pnas.org/content/110/33/13678.abstract` https://doi.org/10.1073/pnas.1312552110

Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A., & Suarez, H. H. (1995). Recurrent excitation in neocortical circuits. *Science*, *269*(5226), 981-5. https://doi.org/10.1126/science.7638624

Dragoi, V., Sharma, J., & Sur, M. (2000). Adaptation-induced plasticity of orientation tuning in adult visual cortex. *Neuron*, *28*(1), 287-98. https://doi.org/10.1016/s0896-6273(00)00103-3

Herzog, M. H., & Clarke, A. M. (2014). Why vision is not both hierarchical and feedforward. *Frontiers in Computational Neuroscience*, *8*. Retrieved from `https://www.frontiersin.org/article/10.3389/fncom.2014.00135` https://doi.org/10.3389/fncom.2014.00135

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex [Journal Article]. *The Journal of Physiology*, *148*(3), 574-591. https://doi.org/10.1113/jphysiol.1959.sp006308

Ito, M., & Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area v2 of macaque monkeys. *Journal of Neuroscience*, *24*(13), 3313–3324. Retrieved from `https://www.jneurosci.org/content/24/13/3313` https://doi.org/10.1523/JNEUROSCI.4364-03.2004

Khan, A. G., Poort, J., Chadwick, A., Blot, A., Sahani, M., Mrsic-Flogel, T. D., & Hofer, S. B. (2018). Distinct learning-induced changes in stimulus selectivity and interactions of gabaergic interneuron classes in visual cortex [Journal Article]. *Nature Neuroscience*, *21*(6), 851-859. Retrieved from `https://dx.doi.org/10.1038/s41593-018-0143-z` https://doi.org/10.1038/s41593-018-0143-z

Lange, G., Senden, M., Radermacher, A., & De Weerd, P. (2020). Interfering with a memory without erasing its trace. *Neural Networks*, *121*, 339-355.

https://doi.org/10.1016/j.neunet.2019.09.027
Lefaucheur, J. P. (2009). Methods of therapeutic cortical stimulation. *Neurophysiol Clin*, *39*(1), 1-14. https://doi.org/10.1016/j.neucli.2008.11.001
MATLAB. (2022). *version 9.12.0.1927505 (r2022a).* Natick, Massachusetts: The MathWorks Inc.
Müller, N. G., Mollenhauer, M., Rösler, A., & Kleinschmidt, A. (2005). The attentional field has a mexican hat distribution. *Vision Res*, *45*(9), 1129-37. https://doi.org/10.1016/j.visres.2004.11.003
Rodieck, R. W. (1965). Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision research*, *5*(12), 583-601.
Sato, H., Katsuyama, N., Tamura, H., Hata, Y., & Tsumoto, T. (1996). Mechanisms underlying orientation selectivity of neurons in the primary visual cortex of the macaque. *The Journal of Physiology*, *494*(3), 757-771. https://doi.org/10.1113/jphysiol.1996.sp021530
Schoups, A. A., Vogels, R., Qian, N., & Orban, G. A. (2001). Practising orientation identification improves orientation coding in v1 neurons. *Nature*, *412*, 549-553.
Seriès, P., Latham, P. E., & Pouget, A. (2004). Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nature Neuroscience*, *7*(10), 1129-1135. https://doi.org/10.1038/nn1321
Sussillo, D., & Barak, O. (2013). Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, *25*(3), 626-649. https://doi.org/10.1162/NECO$_{a0}$0409
Tehovnik, E. J., Slocum, W. M., Smirnakis, S. M., & Tolias, A. S. (2009). Microstimulation of visual cortex to restore vision. *Prog Brain Res*, *175*, 347-75. https://doi.org/10.1016/s0079-6123(09)17524-6
Teich, A. F., & Qian, N. (2003). Learning and adaptation in a recurrent model of v1 orientation selectivity. *J Neurophysiol*, *89*(4), 2086-100. https://doi.org/10.1152/jn.00970.2002
Wang, X.-J. (2008). Decision making in recurrent neuronal circuits. *Neuron*, *60*(2), 215-234. Retrieved from `https://pubmed.ncbi.nlm.nih.gov/18957215` `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2710297/` https://doi.org/10.1016/j.neuron.2008.09.034
Weidler, T., Lehnen, J., Denman, Q., Seb, D. a., Weiss, G., Driessens, K., & Senden, M. (2021). Biologically inspired semantic lateral connectivity for convolutional neural networks. *arXiv pre-print server*. Retrieved from `https://arxiv.org/abs/2105.09830v1`