Title

Sensory-memory interactions via modular structure explain errors in visual working memory

Authors

Jun Yang[1†], Hanqi Zhang[2,3,4], Sukbin Lim*[2,3,4]

[1] Weiyang College, Tsinghua University, Beijing, 100084, People's Republic of China

[2] Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning, NYU Shanghai, 567 West Yangsi Road, Shanghai, 200126, People's Republic of China

[3] Neural Science, NYU Shanghai, 567 West Yangsi Road, Shanghai, 200126, People's Republic of China

[4] NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, 3663 Zhongshan Road North, Shanghai, 200062, People's Republic of China

† Current affiliation:

Interdisciplinary Graduate Program in Quantitative Biosciences, Georgia Institute of Technology, Atlanta, GA 30332; School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332

* Correspondence: sukbin.lim@nyu.edu

Author Contributions

J. Y., H. Z., and S.L. designed and performed research. All authors prepared the figures and wrote the manuscript.

Acknowledgments

We appreciate X. Wei for sharing the code for Bayesian inference models. J. Y. was supported by the NYU Shanghai Summer Undergraduate Research Program (SURP). S. L. received STI2030-Major Projects, No.2021ZD0203700/2021ZD0203705. H. Z. and S. L. also acknowledge the support of the Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning and the NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai.

Competing Interests

The authors declare no competing financial interests.

Abstract

Errors in stimulus estimation reveal how stimulus representation changes during cognitive processes. Repulsive bias and minimum variance observed near cardinal axes are well-known error patterns typically associated with visual orientation perception. Recent experiments suggest that these errors continuously evolve during working memory, posing a challenge that neither static sensory models nor traditional memory models can address. Here, we demonstrate that these evolving errors, maintaining characteristic shapes, require network interaction between two distinct modules. Each module fulfills efficient sensory encoding and memory maintenance, which cannot be achieved simultaneously in a single-module network. The sensory module exhibits heterogeneous tuning with strong inhibitory modulation reflecting natural orientation statistics. While the memory module, operating alone, supports homogeneous representation via continuous attractor dynamics, the fully connected network forms discrete attractors with moderate drift speed and nonuniform diffusion processes. Together, our work underscores the significance of sensory-memory interaction in continuously shaping stimulus representation during working memory.

Introduction

The brain does not faithfully represent external stimuli. Even for low-level features like orientation, spatial frequency, or color of visual stimuli, their internal representations are thought to be modified by a range of cognitive processes, including perception, memory, and decision (Geisler 2008, Webster 2015, Bays, Schneegans et al. 2022). Experimental studies quantified such modification by analyzing behavior data or decoding neural activities. For instance, biases of errors, the systematic deviation from the original stimuli, observed in estimation tasks have been used as indirect evidence to infer changes in the internal representations of stimuli (Wei and Stocker 2017).

One important source of biases is adaptation to environmental statistics, such as the nonuniform stimulus distribution found in nature or the limited range in specific settings. Cardinal repulsion, which refers to the systematic shift away from the horizontal and vertical orientations observed in many perceptual tasks, is one of the examples (de Gardelle, Kouider et al. 2010). Theoretical works suggest that such a bias pattern reflects the prevalence of the cardinal orientations in natural scenes (Girshick, Landy et al. 2011). Similarly, the variance of errors for orientation stimuli was found to be inversely proportional to the stimulus statistics, minimum at cardinal and maximum at oblique orientations (van Bergen, Ma et al. 2015). It was postulated that the dependence of biases and variance of errors on natural statistics results from sensory encoding optimized to enhance precision around the most common stimuli (Ganguli and Simoncelli 2014, Wei and Stocker 2015, Wei and Stocker 2017).

On the other hand, there is a growing body of evidence indicating that error patterns are not solely influenced by sensory encoding but are also shaped by memory processes. In delayed estimation tasks, where participants are presented with stimuli followed by a delay period during which they rely on their working memory for estimation, it has been observed that representations of orientation or color stimuli undergo gradual and continuous modifications throughout the delay

period (Panichello, DePasquale et al. 2019, Bae 2021, Gu, Lee et al. 2023). Such dynamic error patterns are inconsistent with sensory encoding models, most of which only establish a static relationship between stimuli and internal representations. Traditional working memory models are not suitable either. Most of them are constructed to faithfully maintain information about stimuli during the delay period, and thus, the memory representation has a similar geometry as that of the stimuli (Wang 2001, Khona and Fiete 2022). For continuous stimuli such as orientation, location, direction, or color, all stimuli are equally maintained in ring-like memory activities, predicting no biases (Zhang 1996, Compte, Brunel et al. 2000, Burak and Fiete 2009).

How can we explain error patterns in working memory tasks that are similar to those observed in perception tasks? Here, we claim that not a single module but a two-module network with recursive interaction is required. Each module has a distinct role — sensory encoding and memory maintenance. To illustrate this, we use orientation stimuli and examine how their representations change during the delayed estimation tasks. We employ two approaches to find solutions for generating correct error patterns. The first extends previously suggested sensory encoding models, while the second modifies low-dimensional memory models based on attractor dynamics. These approaches are integrated into the network models, which link network connectivity to neuronal tuning properties and behavioral error patterns and reveal the attractor dynamics through low-dimensional projection. Our results show that the sensory-memory interacting networks outperform single-module networks with better control over the shapes and evolution of dynamic error patterns. Furthermore, our network models emphasize the importance of inhibitory tuning in sensory circuits for generating correct error patterns under typical associative learning of natural statistics. Finally, we provide testable predictions regarding the effect of perturbations in sensory-memory interactions on error patterns in delayed estimation tasks.

Results

Low-dimensional attractor models

In natural images, cardinal orientations are the most prevalent (**Figure 1A**). Error patterns in estimation tasks show dependence on such natural statistics, such as biases away from cardinal orientations where the variance of errors is nonetheless minimal (**Figure 1B,C**). In delayed estimation tasks, such a bias pattern is consolidated in time (**Figure 1B**). Also, experimental data suggested that estimation errors increase with a longer delay (Wimmer, Nykamp et al. 2014, Schneegans and Bays 2018), while the precision is still highest at cardinal orientations (van den Berg, Shin et al. 2012, Bays 2014, van Bergen, Ma et al. 2015). Thus, we assumed that the variance of errors increases as keeping its characteristic shape (**Figure 1C**). To explain these errors across orientations and over time, we first explored the underlying working memory mechanism. We considered low-dimensional attractor models with input noise that describe the drift and diffusion of the memory states. Here, we show that two prominent classes of previously suggested models are inconsistent with experimental observations and examine what modification to the models is required.

The most widely accepted model for working memory of orientation stimuli has continuous attractor dynamics, which assumes that all orientations are equally encoded and maintained

(**Figure 1D-F**). Each attractor corresponds to the memory state for different stimuli and forms a continuous ring following the geometry of orientation stimuli. The dynamics along continuous attractors are conceptually represented as movement along a flat energy landscape (**Figure 1D**). Without external input, there is no systematic shift of mean activity, that is, no drift during the delay period (**Figure 1E**). Also, under the assumption of equal influence of noise for all orientations, the variance of errors is spatially flat with constant diffusion along the ring, while the overall magnitude increases over time due to the accumulation of noise (**Figure 1F**).

While such continuous attractor models have been considered suitable for memory storage of continuous stimuli, they cannot capture drift dynamics observed during the delay period. Instead, discrete attractor models with uneven energy landscapes have been suggested with the energy wells corresponding to discrete attractors (**Figure 1G-I**). As evolution toward a few discrete attractors creates drift dynamics, the bias increases during the delay (**Figure 1H**). Also, discrete attractor models naturally produce nonuniform variance patterns. Even with constant noise along the ring, variance becomes minimum/maximum at the attractors/repellers due to the drift dynamics (**Figure 1I**). However, discrete attractor models with constant noise yield inconsistent results when inferring the locus of attractors from the bias and variance patterns observed in the data. Cardinal orientations should be the repeller to account for cardinal repulsion. In contrast, the minimum variance observed at the cardinal orientations suggests they should be the attractors.

How can such inconsistency be resolved? One possible solution is discrete attractor models with nonuniform noise amplitude (**Figure 1J**). Let's consider that attractors are formed at oblique orientations to generate correct bias patterns (**Figure 1K**). Additionally, we assumed that noise has the highest amplitude at the obliques. When the difference in the noise amplitude is large enough to overcome the attraction toward the obliques, the models can produce correct variance patterns, maximum at the obliques and minimum at cardinal orientations (**Figure 1L**). In sum, unlike two prominent memory models, continuous attractors or discrete attractors with constant noise, discrete attractors with maximum noise at the obliques could reproduce experimentally observed error patterns of orientation stimuli. Note that these attractor models often simplify the full network dynamics. Namely, the drift and diffusion terms are derived by projecting network dynamics onto low-dimensional memory states (Burak and Fiete 2012, Darshan and Rivkind 2022). Thus, it is still in question whether there exist memory networks that can implement attractor dynamics with correct drift and diffusion terms.

Bayesian sensory model and extension

Before exploring full memory network models, we note that previous theoretical works for sensory processing suggested that Bayesian inference with efficient coding could generate the repulsive bias and the lowest variance at cardinal orientations (Wei and Stocker 2015, Wei and Stocker 2017). Efficient coding theory suggests the sensory system should enhance the sensitivity around more common stimuli. For orientation stimuli, precision should be highest around cardinal directions, which could be achieved by sharpening the likelihood functions. Equipped with Bayesian optimal readout, such a sensory system could reproduce correct error patterns observed in perceptual tasks for various visual stimuli, including orientations (**Figure 2**).

However, such models only account for the relationship between external and perceived stimuli during sensory processing, resulting in static error patterns. Here, we extended the framework so that the system can maintain information about the stimulus after its offset while bias and variance of errors grow in time (**Figure 2A**). We added a memory stage to Bayesian sensory models such that the memory stage receives the output of the sensory stage and returns it as the input after the maintenance. For instance, let's denote the external orientation stimulus given during the stimulus period as θ_1. The sensory stage receives θ_1 as input and generates the perceived orientation, $\hat{\theta}_1$, which varies from trial to trial with sensory noise (**Figure 2B**). Through the memory stage, $\hat{\theta}_1$ is returned as the input to the sensory stage for the next iteration with the addition of memory noise ξ_1.

Such a recursive process mimics interactions between sensory and memory systems where the sensory system implements efficient coding and Bayesian inference, and the memory system faithfully maintains information. As the recursive process iterates, the distribution of the internal representation of orientation broadens due to the accumulation of noise from the sensory and memory systems. This leads to an increase in bias and variance at each step while keeping their characteristic shapes (**Figure 2C**). Thus, recurrent interaction between sensory and memory systems during the delay period, each of which meets different demands, successfully reproduces correct error patterns observed in memory tasks.

Network models with sensory and memory modules

Next, we construct network models capturing the sensory-memory interactions formalized under the Bayesian framework. We consider two-module networks where each module corresponds to the sensory and memory systems. To generate orientation selectivity, both modules have a columnar architecture where neurons in each column have a similar preference for orientation (**Figure 3A**). However, their connectivity structures are different (**Figure 3B**). The memory module in isolation resembles the traditional ring attractor network with a strong and homogeneous recurrent connection. This enables the memory module in isolation to maintain information about all orientations equally during the delay period (**Figure 3B-F**, right). Conversely, the recurrent connectivity strengths in the sensory module are relatively weak, such that without connection to the memory module, the activities during the delay period decay back to the baseline levels (**Figure 3B**, left). Furthermore, the connectivity strengths across columns are heterogeneous, particularly stronger at the obliques. As a result, the tuning curves near cardinal orientations can be sharper and denser, consistent with experimental observations showing a larger number of cardinally tuned neurons (Li, Peterson et al. 2003, Shen, Tao et al. 2014) and their narrower tuning (Li, Peterson et al. 2003, Kreile, Bonhoeffer et al. 2011) (**Figure 3C-F**, left). Different response activities of the two modules in isolation are demonstrated in their response manifolds as more dispersed representations around cardinal orientations in the sensory module, compared to the ring-like geometry of the memory module (**Figure 3F**).

For sensory-memory interacting networks, we connected the two modules with inter-module connections set to be stronger between neurons with similar orientation selectivity (**Figure 4A**). Activity profiles in both modules follow that of the sensory module — heterogeneous with

narrower and denser tuning curves around cardinal orientations, leading to higher sensitivity (**Figure 4B**). Such activity pattern is maintained even during the delay period when recurrent connections in the memory module support activities of both sensory and memory modules (**Figure 4B**, right). Note that while sensory activities convey stimulus information during the delay period, their overall firing rates are much lower than those during the stimulus period with weak interconnection strengths. Such low firing rates may lead to both positive and negative evidence of sustained activity in early sensory areas (Leavitt, Mendoza-Halliday et al. 2017).

When the internal representation of the orientation stimulus is read from the memory module using a population vector decoder mimicking Bayesian optimal readout (Fischer 2010), the sensory-memory interacting network exhibits repulsive bias and minimum variance at cardinal orientations, inheriting from efficient sensory coding (**Figure 4C,D**). Similar error patterns were observed when decoded from activities of the sensory module (**Figure 4** – **Figure supplement 1**). Such bias increases during the delay period with increasing asymmetry of tuning widths despite lower firing rates than the stimulus period (**Figure 4 – Figure supplement 2**). At the same time, errors gradually increase due to noise accumulation in time, as in typical memory networks (Compte, Brunel et al. 2000, Burak and Fiete 2012). Note that the variance of errors is negligible during stimulus presentation when the external input overwhelms internal noise, which may not fully account for the variability observed during perception tasks (see **Discussion**). We obtained Fisher information measuring sensitivity at each orientation from the neural responses (see **Methods**). Opposite to the variance of errors, Fisher information is highest at cardinal orientations, while it decreases during the delay period (**Figure 4E**). Thus, the sensory-memory interacting network model that mechanistically embodies the extension of the Bayesian sensory model correctly reproduces the error patterns observed in delayed estimation tasks.

Analysis of low-dimensional memory states

To further understand the mechanisms of generating the correct error patterns in sensory-memory interacting networks, we analyzed the network dynamics during the delay period. For this, we identified the low-dimensional manifold that has slow dynamics during the delay period, which corresponds to the memory states (**Figure 5A**). We projected the dynamics along this manifold to obtain the drift and diffusion terms (**Figure 5A-C**; **Figure 5 – Figure supplement 1**). The drift term shows similar patterns to cardinal repulsion (**Figure 5B,E**). Integrating this drift for orientation yields the energy function, which is minimum at the obliques (**Figure 5D**). This suggests that the network implements discrete attractor dynamics with attractors formed at the obliques. The diffusion term is also uneven — the noise amplitude is maximum at the obliques so that despite attraction toward them, the variance of errors can be maximum (**Figure 5C,F**). Note that while we use Poisson noise in all units to replicate neuronal spike variability, the pattern of noise coefficients remains unchanged even with constant Gaussian noise (**Figure 5– Figure supplement 2**). This lower variance near cardinal orientations arises from more dispersed representations of stimuli, as the noise coefficient is inversely proportional to the distance between stimulus representations (**Eq. 21**). Thus, the nonuniform characteristics of both drift and diffusion processes stem from the heterogeneous connections within the sensory module and align with the solution identified in low-dimensional memory models (**Figure 1J-L**).

Next, we examined how heterogeneity of connectivity in the sensory module affects the dynamics along the memory states. The magnitude of heterogeneity is denoted as α, and larger α represents a larger asymmetry of connectivity strengths at cardinal and oblique orientations (**Figure 3B**, left). When α increases, the asymmetry of drift and energy levels becomes more prominent, leading to a more rapid increase in bias (**Figure 5B,D,E**). The diffusion term is also more asymmetric, compensating for stronger attraction to the obliques (**Figure 5C**). Thus, for larger α, the variability of errors is still higher at the obliques (**Figure 5F**). Another important parameter influencing error patterns is the intermodal connectivity strengths (**Figure 6**). Similar to the effect of increasing α, increases in feedforward or feedback strengths cause the energy levels to become more asymmetrical (**Figure 6A,E**), leading to a larger bias (**Figure 6B,F**). Conversely, the noise coefficient is less affected (**Figure 6C,G**), and the variance of errors decreases as the drift force becomes stronger (**Figure 6D,H**). Note that bias and variance patterns depend on the product of feedforward and feedback connections, denoted as γ, such that for a fixed γ, the error patterns remain similar (**Figure 6I,J**). In sum, the bias and variability of errors are determined by the degree of heterogeneity in the sensory module (α) and intermodal connectivity strengths (γ) as both α and γ affect the asymmetry of drift term similarly, while the asymmetry of diffusion term is more strongly influenced by α (**Figure 6K,L**).

Importance of heterogeneously tuned inhibition

We showed that network models realizing sensory-memory interactions reproduce correct error patterns, where each module has a different connectivity structure. Previous work suggested that such a heterogeneous connection of the sensory system may arise from experience-dependent synaptic modification (Olshausen and Field 1996, Zylberberg, Murphy et al. 2011). For example, typical Hebbian learning is thought to potentiate connectivity strengths between neurons whose preferred stimuli are more frequently encountered. For orientations, cardinal directions are predominant in natural scenes. Thus, if experience-dependent learning occurs mainly at the excitatory synapses, the excitatory connections near cardinal orientations become stronger in the sensory module. This is opposite to the previously discussed case where the sensory module has the strongest connection at the obliques. With the strongest excitatory connections at cardinal orientations, the error patterns are reversed, resulting in cardinal attraction instead of repulsion, and the lowest variance occurs at the obliques.

Inhibitory synaptic connections can also be modified through learning (Vogels, Froemke et al. 2013, Khan, Poort et al. 2018, Larisch, Gonner et al. 2021). Here, we considered that experience-dependent learning exists in both excitatory and inhibitory pathways and similarly shapes their connectivity (**Figure 7A**). We assumed that excitatory and inhibitory connections are segregated and stronger near cardinal orientations (**Figure 7B**). We modulated the heterogeneity degree of both excitatory and inhibitory connections, denoted as α and β, respectively (**Figure 7B-D**). The ratio between α and β determines the direction and magnitude of bias and variance patterns (**Figure 7C,D**). For relatively larger α, the network shows cardinal attraction and minimum variance of errors at the obliques (**Figure 7E**). Reversely, for relatively larger β with stronger modulation in inhibitory connections, the network reproduced cardinal repulsion and minimum variance of errors at cardinal orientations, consistent with experiments (**Figure 7F**). With a larger

difference between α and β, such patterns of bias and variance are potentiated and minimum Fisher information across orientations decreases, corresponding to memory loss (**Figure 7C,D**; **Figure 7 – Figure supplement 1**). Thus, this emphasizes the important role of heterogeneously tuned inhibition in shaping the sensory response for higher precision at cardinal orientations and enabling the sensory-memory interacting network to generate correct error patterns.

Comparison to alternative circuit structures

So far, we have shown the sufficiency of sensory-memory interacting networks with different connectivity structures featuring heterogeneous-homogeneous recurrent connections within each module. Here, we explore whether such architecture is necessary by comparing its performance with alternative circuit structures for sensory-memory interactions. One candidate mechanism involves having the heterogeneous sensory network maintain memory with a long intrinsic time constant, similar to having autapses (Seung, Lee et al. 2000). However, this model fails to replicate the evolution of error patterns during the delay period as a long intrinsic time constant slows down the overall dynamics, thus hindering the evolution of error patterns (**Figure 7** – **Figure Supplement 2**). Alternatively, we focused on a two-module network with variations in connectivity structure. We assumed that sensory and memory modules still serve their distinctive functions, namely, sensory encoding and memory maintenance, with weak/strong recurrent connections in sensory/memory modules. On the other hand, the heterogeneity of connections in other circuits might differ as homogeneous-homogeneous, homogeneous-heterogeneous, and heterogeneous-heterogeneous connections for sensory-memory modules.

Circuits with homogeneous connections in both sensory and memory modules are similar to previous continuous attractor models for working memory, such that the energy landscape and noise amplitude are uniform for all orientations (**Figure 1D-F**). Such architecture is not suitable as it generates no bias in errors and flat variance patterns. This leaves the latter two types of configurations, which require heterogeneous connections within the memory module. With a strong recurrent connection within the memory module, its heterogeneous activity pattern dominates overall activities in sensory-memory interacting networks, which makes it analogous to an isolated memory module. Thus, we examined the property of the memory module alone, which can maintain memory while generating heterogeneous responses without connection to the sensory module (**Figure 8**).

To generate the correct bias pattern, we assumed that excitatory and inhibitory pathways in the memory module are stronger near cardinal orientations, as we previously considered for the sensory module in the sensory-memory interacting network (**Figure 8A,B**). However, memory circuits with heterogeneous connections have problems in maintaining the information and reproducing correct error patterns (**Figure 8C-E**). First, memory circuits alone require fine-tuning of heterogeneity whose range generating a moderate drift speed is at least one order of magnitude smaller than that of the two-module network (**Figure 8C,D**). Deviation from this range results in a fast drift toward oblique orientations, leading to rapid loss of information during the delay period (**Figure 7 – Figure supplement 1**). Second, despite the correct bias direction, the variance pattern is reversed such that the variance of errors is minimal at the oblique orientations (**Figure 8E**).

Varying the heterogeneity in excitatory and inhibitory connections shows that such rapid drift and reversed error patterns are prevalent across different parameters (**Figure 8C,D**).

To understand why a heterogeneous memory circuit alone fails to reproduce correct error patterns, we compared its low-dimensional dynamics along the memory states to that of the sensory-memory interacting networks. For the network with a similar range of bias and variance on average, we compared their energy landscape and noise amplitude, which vary similarly in both networks with minimum energy level and maximum noise at the oblique orientations (**Figure 9A-F**). However, the energy difference between cardinal and oblique orientations in a single memory-circuit model is bigger than that in a sensory-memory interacting network (**Figure 9C,** left in **G,H**). In contrast, the difference in noise amplitude is smaller (**Figure 9D-F,** right in **G,H**). The attraction at the obliques is much stronger, leading to the correct bias patterns, but too rapid an increase. Also, smaller differences in noise amplitude cannot overcome strong drift dynamics, leading to the minimum variance of errors at the obliques and reversed variance patterns. Even for different types or levels of noise, such as Gaussian noise with varying amplitude, distinctive error patterns in one-module and two-module networks are maintained (**Figure 9** – **Figure supplement 1**).

For an intuitive understanding of how connectivity heterogeneity affects the degrees of asymmetry in drift and diffusion differently in one-module and two-module networks, consider a simple case where only the excitatory connection exhibits heterogeneity, the degree of which is denoted by α. For memory maintenance, the overall recurrent connections need to be strong enough to overcome intrinsic decay, simplified to $w = 1$. In the one-module network, α in the memory module causes deviations from perfect tuning, creating potential differences at cardinal and oblique orientations as $1 \pm \alpha$. In the two-module network, with $w = 1$ fulfilled by the memory module, α in the sensory module acts as a perturbation. The effect of α is modulated by the intermodal connectivity strengths, denoted by γ, and potential differences at cardinal and oblique orientations can be represented as $1 \pm \gamma\alpha$. Thus, while a relatively large α leads to too fast drift in the one-module network, the drift speed in the two-module network could remain modest with small $\gamma<1$. Conversely, even with small γ, the asymmetry of noise coefficients can be large enough to produce correct variance patterns because the noise coefficient is more strongly influenced by α in the two-module network (**Figure 6**). In sum, compared to a heterogeneous memory circuit alone, interactions between heterogeneous sensory and homogeneous memory modules are advantageous due to an additional degree of freedom, intermodal connectivity strengths, which allows better control of energy and noise difference at cardinal and oblique orientations.

Discussion

While higher association areas have long been considered as a locus of working memory (Roussy, Mendoza-Halliday et al. 2021, Mejias and Wang 2022), recent human studies found memory signals in early sensory areas, prompting a re-evaluation of their role in working memory (Xu 2020, Adam, Rademaker et al. 2022). Our work extends the traditional memory models (Wang 2001, Khona and Fiete 2022) with novel insights into the significance of stimulus-specific sensory areas. We showed how sensory-memory interactions can elucidate changes in the internal representation of orientation stimuli and their behavioral readout during memory tasks. The observed error patterns suggest that the network meets two demands simultaneously: efficient

encoding that reflects natural statistics and memory maintenance for successful retrieval of stimuli after a delay. Achieving both demands for orientation stimuli conflicts in a one-module network. Efficient encoding necessitates asymmetrical connections, resulting in inconsistent bias and variance patterns and overly rapid drift in the one-module network unless fine-tuned. In contrast, connecting sensory and memory modules can generate error patterns correctly and with less need for fine-tuning heterogeneity for slow drift. Efficient coding of natural statistics in the sensory module underscores the role of inhibitory plasticity. Low-dimensional projection onto memory states reveals that drift and diffusion processes governing working memory dynamics closely resemble the bias and variance patterns derived under Bayesian sensory models. It also elucidates how the magnitudes of bias and variance change depending on the heterogeneity of sensory connections and intermodal connectivity strengths.

Our model makes testable predictions to differentiate two-module and one-module networks using perturbation, such as transcranial magnetic stimulation (TMS). Many studies have found that during the delay period, TMS can intervene with the feedforward signal from sensory areas through which working memory is consolidated (van de Ven, Jacobs et al. 2012) (but see (Adam, Rademaker et al. 2022) for mixed effects of TMS and related debate). Under such perturbations, the ability to maintain information in the memory module will not be affected due to strong recurrent connections in both two-module and one-module networks. However, we expect different effects on bias patterns — in the two-module network, the bias will stop systematically drifting towards the obliques, reducing systematic repulsion (**Figure 10**). This accompanies the nonincreasing heterogeneity of tuning curves after the disruption, marked by their tuning width indices (see **Methods**). In contrast, in the one-module network, perturbation does not incur changes in error patterns as memory activities are less dependent on the sensory module during the delay period. Thus, perturbation studies can be used to reveal the role of the sensory module in shaping the error patterns during working memory. Note that our model cannot predict the effects of distractors during working memory, as such effects do not experimentally lead to changes in error patterns (Rademaker, Chunharas et al. 2019). The effect of distractors and direct intervention in the inter-module connections may differ due to potential differences in the encoding of distractors compared to task-relevant stimuli. More advanced models are required to comprehensively understand the influence of distractors and the processing of ongoing visual stimuli or the storage of multiple stimuli.

Our work suggests biologically plausible network mechanisms for the previously postulated efficient coding and Bayesian inference principles, relating network connectivity to tuning properties and error patterns. Previous normative explanations for systematic bias observed in perception tasks also suggested possible neural substrates for efficient coding, such as asymmetrical gain, width, or density of tuning curves across stimulus features (Ganguli and Simoncelli 2014, Wei and Stocker 2015). Our work narrowed the mechanism to denser and narrower tuning curves at cardinal orientations, consistent with neurophysiological recordings in the visual cortex (Li, Peterson et al. 2003, Kreile, Bonhoeffer et al. 2011, Shen, Tao et al. 2014). We implemented a population vector decoder reflecting neuronal preferred orientations, which approximates Bayesian optimal readout (Fischer 2010). Compared to a previous work adapting efficient coding theories with static tuning curves to account for error patterns in working memory tasks (Taylor and Bays 2018), our extension to memory processes demonstrated how neural

activities and behavior readout change dynamically during the delay period. Notably, recent work combined dynamic change of signal amplitude with static tuning curves to capture different time courses of estimation precision during sensory encoding and memory maintenance (Tomić and Bays 2023). Our network models embody such phenomenological models as the networks exhibit changes in overall firing rates after the stimulus offset.

Like our study, a few recent studies have employed attractor dynamics to explain dynamic error patterns observed for visual color memory (Panichello, DePasquale et al. 2019, Pollock and Jazayeri 2020, Eissa and Kilpatrick 2023). Behavior studies showed attractive bias and minimum variance around the prevalent colors, which one-module discrete attractor models could reproduce. However, these models cannot be generalized to other visual stimuli, such as orientations, spatial locations, or directions, of which the responses show repulsive bias away from the common stimuli (Wei and Stocker 2017). Also, a one-module network storing color memory requires fine-tuned heterogeneity for moderate drift speed. While the desired low-dimensional manifold and drift dynamics can be engineered in the one-module network (Pollock and Jazayeri 2020), its biological mechanism needs further investigation. The two-module network considered in our study also requires fine-tuning of homogeneity in the memory module and heterogeneity in the sensory module. However, the condition of asymmetrical connections in the sensory module is less stringent as they have a weaker influence on the entire dynamics than those in the memory module. Fine-tuning of homogeneous connections in the memory module can be mediated through activity-dependent plasticity, such as short-term facilitation (Itskov, Hansel et al. 2011, Hansel and Mato 2013, Seeholzer, Deger et al. 2019) or long-term plasticity (Renart, Song et al. 2003, Gu and Lim 2022). Also, recent work showed that continuous attractors formed under unstructured, heterogeneous connections are robust against synaptic perturbations (Darshan and Rivkind 2022). Thus, the two-module networks can control the drift speed better with possible additional mechanisms that promote homogeneous memory states. It needs further exploration whether they can be generalized to other stimuli like color, possibly involving additional categorical structures (Hardman, Vergauwe et al. 2017, Pratte, Park et al. 2017).

Our current study is limited to the dynamic evolution of memory representation for a single orientation stimulus and its associated error patterns, which does not capture nuanced error patterns in broader experimental settings (Hahn and Wei 2024). For instance, while shorter stimulus presentations with no explicit delay led to larger biases experimentally, our current model, which starts activities from a flat baseline, shows an increase in bias throughout the stimulus presentation (de Gardelle, Kouider et al. 2010). Additionally, the error variance during stimulus presentation is almost negligible compared to that during the delay period, as the external input overwhelms the internal noise. These mismatches during stimulus presentation have minimal impact on activities during the delay period when the internal dynamics dominate. Nonetheless, the model needs further refinement to accurately reproduce activities during stimulus presentation, possibly by incorporating more biologically plausible baseline activities. Also, a recent Bayesian perception model suggested different types of noise like external noise or variations in loss functions that adjust tolerance to small errors may help explain various error patterns observed across different modalities (Hahn and Wei 2024). Even for memories involving multiple items, noise can be critical in determining error patterns, as encoding more items might cause higher noise for each individual item (Chunharas, Rademaker et al. 2022).

The modularity structure in the brain is thought to be advantageous for fast adaptation to changing environments (Simon 1995, Cole, Reynolds et al. 2013, Frankland and Greene 2020). Recent works showed that recurrent neural networks trained for multiple cognitive tasks form clustered neural activities and modular dynamic motifs to repurpose shared functions for flexible computation (Yang, Joglekar et al. 2019, Driscoll, Shenoy et al. 2022). Resonant with these computational findings, an fMRI study showed that shared representation across distinct visual stimuli emerges during the delay period (Kwak and Curtis 2022). Although our work focuses on a single task, it highlights the necessity of having dedicated sensory and memory modules, and a memory module with ring geometry can be repurposed for various visual stimuli such as motion, spatial location, and color. It is reminiscent of the flexible working memory model, which proposes connections between multiple sensory modules and a control module (Bouchacourt and Buschman 2019). However, a key distinction lies in the role of the control module. Unlike the flexible working memory model that loses memory without sensory-control interactions, our work suggests that the memory module can independently maintain memory, while interaction with the sensory module continuously shapes the internal representation, potentially consolidating prior beliefs regarding natural statistics. The sensory-memory interaction and network architecture derived from dynamic changes of single stimulus representation can be a cornerstone for future studies in more complex conditions, such as under the stream of visual inputs (Xu 2020, Adam, Rademaker et al. 2022) or with high or noisy memory loads (Bays, Schneegans et al. 2022).

Methods

Low-dimensional attractor models. To illustrate error patterns in different low-dimensional attractor models shown in **Figure 1**, we considered a one-dimensional stochastic differential equation given as

$$\mathrm{d}\theta_t = \mu(\theta_t)\mathrm{d}t + \sigma(\theta_t)\mathrm{d}W_t, \tag{1}$$

where θ_t and W_t are orientation and standard Brownian motion at time t. We assumed that the drift and noise coefficients μ and σ only depend on θ_t where $\sigma = \sqrt{2\mathcal{D}}$ with diffusion coefficient $\mathcal{D}$.

For continuous attractor models in **Figure 1D-F**, μ and σ were set to be constant as $\mu = 0$ and $\sigma = 2°$. For discrete attractor models in **Figure 1G-L**, we assumed that the energy function $U(\theta_t)$ is proportional to $\cos(4\theta_t)$ (**Figure 1G,J**) so that the drift term $\mu(\theta_t) = \sin(4\theta_t)$ with $\mu(\theta_t) = -\frac{dU}{d\theta_t}$. In these attractor models, the constant noise in **Figure 1G-I** is $\sigma = 2°$ and the nonuniform noise in **Figure 1J-L** is $\sigma = 2°(1 - \cos(4\theta_t))$. The biases and standard deviation (SD) of errors were plotted at T = 1, 2, and 3 with 50,000 iterations. For the numerical simulation, dt = 0.01.

Bayesian sensory models and extension. In **Figure 2,** we first constructed the sensory inference process, which receives orientation input θ, forms a corresponding noisy sensory representation m given θ, and then infers $\hat{\theta}$ as an estimate of the input orientation from the encoded representation m. This inference is made in a Bayesian manner based on likelihood function $p(m|\theta)$ and orientation prior $q(\theta)$.

To construct $p(m|\theta)$, we followed the procedure given in (Wei and Stocker 2015), and the summary is as follows. We started from the sensory space of $\tilde{\theta}$ where both discriminability and Fisher Information $J(\tilde{\theta})$ are uniform, and all likelihood functions $p(m|\tilde{\theta})$ are homogeneous von Mises functions. And since $J(\theta) \propto (q(\theta))^2$ under the efficient coding condition, the sensory space of $\tilde{\theta}$ and the stimulus space of θ can be mapped by forward and backward mappings $F(\theta)$ and $F^{-1}(\tilde{\theta})$ where $F(\theta)$ is the cumulative distribution function of prior $q(\theta)$. Thus, likelihood functions $p(m|\theta)$ can be obtained by taking homogeneous von Mises likelihoods in the sensory space and transforming them back to the stimulus space using F^{-1}. To sum up the upper half of the procedural diagram in **Figure 2A**, the sensory module receives θ, encodes it in m following $p(m|\theta)$, and decodes $\hat{\theta}$ using likelihood functions and prior $q(\theta)$.

As an extension to include a memory process, the decoded $\hat{\theta}$ is passed on to the memory module, where $\hat{\theta}$ is maintained with the addition of memory noise ξ. The output of the memory module, $\hat{\theta} + \xi$, is fed back to the sensory module as the new input. This completes one iteration of sensory-memory interaction. The whole process is then repeated recursively, resulting in increased biases and standard deviations in the θ statistics at subsequent iterations (call them θ_i for the input of iteration i).

For **Figure 2B-C**, we set the von Mises sensory-space likelihoods to be $p(m|\tilde{\theta}) \propto \exp\,(\kappa_m \cos\,(m - \tilde{\theta}))$, with $\kappa_m = 250$. These likelihood functions are transformed by $F^{-1}(\tilde{\theta}) = \{\int q(\theta)\}^{-1}$ where $q(\theta) = 3 + \cos\,(4\theta)$. Each internal representation m is sampled from $p(m|\theta)$, after which $\hat{\theta}$ is estimated as the mean of the posterior $p(\theta|m)q(\theta)$. With the parameters chosen above, the inferred samples of $\hat{\theta}$ after the first sensory iteration have a circular standard deviation of $\sigma_\theta \approx 1.3°$ at cardinal orientations. To have comparable memory and sensory noise levels, we set the memory noise as $\xi \sim \mathcal{N}(0, (1.3°)^2)$ which is added on top of the sensory outputs. Thus, the memory outputs of the first iteration $\theta_1 = \hat{\theta}_1 + \xi$ have a standard deviation of 1.84° at the cardinals. The first three iterations' memory output statistics are plotted in **Figure 2C**, i.e., bias(θ_1), bias(θ_2), bias(θ_3), and SD(θ_1), SD(θ_2), SD(θ_3). The statistics were computed from 10,000 iterations of the simulation. The magnitude of biases and standard deviations vary for different sensory or memory noise levels, while the overall patterns and the increasing temporal trend are unchanged (not shown).

Firing rate models. For network models, we considered sensory circuits with heterogeneous connections (**Figure 3**), memory circuits with homogeneous connections (**Figure 3**) and heterogeneous connections (**Figure 8,9**), and sensory-memory interacting circuits (**Figure 4-7, 9, 10**). In all cases, the activities of neurons are described by their firing rates and synaptic states, denoted by r and s. For columnar structure encoding orientation stimuli, we indexed the neurons by uniformly assigning them indices $\psi_i = \frac{(i-1)}{N} \times 180°$ for i from 1 to N where N is the number of neurons in each population. For sensory or memory networks alone, the dynamics of neuron i are described by the following equations,

$$\begin{aligned} r_k^i &= f_k\left(\sum_j W_k^{ij} s_k^j + I_{\text{ext},k}^i\right) \\ \tau \dot{s}_k^i &= -s_k^i + r_k^i + \xi_k^i \end{aligned}, \tag{2}$$

where the superscripts i and j are the neuronal indices, and the subscript k is either s or m, representing sensory or memory circuits. For the sensory-memory interacting network, the dynamics are given as

$$\begin{aligned} \boldsymbol{r}_{\mathrm{s}} &= f_{\mathrm{s}}\left(\boldsymbol{W}_{\mathrm{s}}\boldsymbol{s}_{\mathrm{s}} + \boldsymbol{W}_{\mathrm{b}}\boldsymbol{s}_b + \boldsymbol{I}_{\text{ext,s}}\right) \\ \boldsymbol{r}_{\mathrm{m}} &= f_{\mathrm{m}}\left(\boldsymbol{W}_m\boldsymbol{s}_{\mathrm{m}} + \boldsymbol{W}_{\mathrm{f}}\boldsymbol{s}_{\mathrm{f}} + \boldsymbol{I}_{\text{ext,m}}\right) \\ \tau\dot{\boldsymbol{s}}_k &= -\boldsymbol{s}_k + \boldsymbol{r}_{\mathrm{s}} + \boldsymbol{\xi}_k, \quad \text{for } k = \text{s or f} \\ \tau\dot{\boldsymbol{s}}_l &= -\boldsymbol{s}_l + \boldsymbol{r}_{\mathrm{m}} + \boldsymbol{\xi}_l, \quad \text{for } l = \text{m or b} \end{aligned}, \tag{3}$$

where activities and synaptic inputs are represented in the vector and matrix multiplication form, shown in bold cases. The additional subscripts f and b represent feedforward and backward connections between sensory and memory modules.

In both Eqs. (2) and (3), $s(t)$ is the low pass filtered $r(t)$ with synaptic time constant τ and with the addition of ξ approximating Poisson noise. We modeled ξ as the Gaussian process with covariance $\langle \xi^i(t)\xi^j(t')\rangle = r^i(t)\delta_{ij}\delta(t-t')$, following (Burak and Fiete 2012). We assumed that the rate dynamics are relatively fast such that $r(t)$ equals the input current-output rate transfer function *f*. The input current is the sum of external input I_{ext} and the synaptic currents from other neurons in the network, which are the postsynaptic states s^j weighted by synaptic strengths W^{ij}. The transfer function *f* has the Naka–Rushton form (Wilson 1999) given as

$$f(x) = f_{max}\frac{(x-T)^q}{w^q + (x-T)^q} \cdot [x-T]_+, \tag{4}$$

where $[\cdot]_+$ denotes the linear rectification function. The transfer functions differ in the sensory and memory modules, denoted as f_{s} and f_{m}, respectively.

Synaptic inputs in network models. Note that for all network models, we only considered excitatory neurons under the assumption that the inhibitory synaptic pathways have relatively fast dynamics. Thus, recurrent connectivity strengths, W_{s} and W_{m} within sensory and memory modules, reflect summed excitation and inhibition, and thus, can have either positive or negative signs. On the other hand, we assumed that intermodal interactions, W_{f} and W_{b}, are dominantly excitatory and, thus, can be only positive.

All W's can be defined using neuronal indices of post- and presynaptic neurons as

$$W^{ij} = \frac{1}{N}J(\psi_i, \psi_j). \tag{5}$$

For W_{s} without segregating excitation and inhibition in **Figure 3-6**, N is the population size of sensory module, N_{s}, and J_{s} is the sum of a constant global inhibition and a short-range excitatory connection as

$$J_{\mathrm{s}}(\psi_i, \psi_j) = -J_{\mathrm{I,s}} + J_{\mathrm{E,s}}(1 - \alpha\cos 4\psi_i)\mathrm{e}^{-\frac{(\psi_i - \psi_j)^2}{\lambda_{\mathrm{E,s}}^2}}, \tag{6}$$

where $\alpha > 0$ represents the heterogeneity degree of excitatory connectivity, and λ_{E} is the width of local excitatory connections.

When we segregated excitation and inhibition and considered the heterogeneity of inhibitory connection in **Figure 7-10**, Eq. (6) is replaced with

$$J_{\mathrm{s}}(\psi_i, \psi_j) = -J_{\mathrm{I,s}}(1 + \beta\cos 4\psi_i)\mathrm{e}^{-\frac{(\psi_i-\psi_j)^2}{\lambda_{\mathrm{I,s}}^2}} + J_{\mathrm{E,s}}(1 + \alpha\cos 4\psi_i)\mathrm{e}^{-\frac{(\psi_i-\psi_j)^2}{\lambda_{\mathrm{E,s}}^2}}, \tag{7}$$

where $\beta > 0$ is the degree of heterogeneity of inhibitory connections. Note the signs of modulation change in Eqs. (6) and (7) such that when only excitation is modulated in Eq. (6), the connectivity strengths near the obliques are strong. In contrast, when excitation and inhibition are both modulated in Eq. (7), the connectivity strengths near cardinal orientations are strong.

For the memory module, N is the population size of the memory module, N_{m} in Eq. (5). Without heterogeneity in **Figure 3-7** and **10**, J_{m} is defined as

$$J_{\mathrm{m}}(\psi_i, \psi_j) = -J_{\mathrm{I,m}}\mathrm{e}^{-\frac{(\psi_i-\psi_j)^2}{\lambda_{\mathrm{I,m}}^2}} + J_{\mathrm{E,m}}\mathrm{e}^{-\frac{(\psi_i-\psi_j)^2}{\lambda_{\mathrm{E,m}}^2}}. \tag{8}$$

In contrast, for the one-module network model in **Figure 8**, the connectivity of the memory module is heterogeneous, as in the sensory module in Eq. (7), and is defined as

$$J_{\mathrm{m}}(\psi_i, \psi_j) = -J_{\mathrm{I,m}}(1 + \beta\cos 4\psi_i)\mathrm{e}^{-\frac{(\psi_i-\psi_j)^2}{\lambda_{\mathrm{I,m}}^2}} + J_{\mathrm{E,m}}(1 + \alpha\cos 4\psi_i)\mathrm{e}^{-\frac{(\psi_i-\psi_j)^2}{\lambda_{\mathrm{E,m}}^2}}. \tag{9}$$

The feedforward and feedback connectivity are similarly defined as

$$\begin{aligned} W_{\mathrm{f}}^{ij} &= \frac{1}{N_{\mathrm{s}}} J_{\mathrm{f}}\, \mathrm{e}^{-(\psi_{\mathrm{m}i}-\psi_{\mathrm{s}j})^2/\lambda_{\mathrm{f}}^2} \\ W_{\mathrm{b}}^{ij} &= \frac{1}{N_{\mathrm{m}}} J_{\mathrm{b}}\, \mathrm{e}^{-(\psi_{\mathrm{s}i}-\psi_{\mathrm{m}j})^2/\lambda_{\mathrm{b}}^2}. \end{aligned} \tag{10}$$

Note the connectivity strength is normalized by the size of the presynaptic population so that the total synaptic current remains the same for different population sizes.

For the external inputs with orientation θ, $I_{\mathrm{ext,s}}$ in the sensory module is modeled as

$$I_{\mathrm{ext,s}}^{i}(\theta) = C\left(1 - 2\varepsilon + 2\varepsilon\mathrm{e}^{-(\psi_i-\theta)^2/\lambda_{\mathrm{ext,s}}^2}\right), \tag{11}$$

where $\varepsilon \in (0, 0.5]$ determines the stimulus tuning of the input, $\lambda_{\mathrm{ext,s}}$ determines the width, and C describes the contrast (Hansel and Sompolinsky 1998).

For the memory network not connected to the sensory module in **Figure 3** and **8**, we assumed stimulus-specific input as

$$I_{\mathrm{ext,\,m}}^{i}(\theta) = \frac{1}{2}\left(\cos\left(2(\psi_i - \theta)\right) + 1\right) + I_{\mathrm{c,m}}, \tag{12}$$

where $I_{\mathrm{c,m}}$ is a constant background input. When the memory module receives the inputs from the sensory population in **Figure 4-7** and **10**, we assumed $I_{\mathrm{ext,\,m}}^{i}(\theta)$ is constant as $I_{\mathrm{c,m}}$.

Analysis of network activities. We used population vector decoding to extract the internal representation of orientation and quantified how such representation deviated from the original stimulus. We also examined how tuning properties and Fisher information change during the delay period.

Note that while we indexed neurons uniformly with ψ_i between $0°$ and $180°$, the maximum of the tuning curve of neuron ψ_i can change dynamically and differ from ψ_i. We defined the preferred feature (PF) of neuron i as the maximum of its tuning curve when the tuning curve reaches a steady state in the presence of external input. For numerical estimation, we set the stimulus-present encoding epoch to 5 seconds to obtain the steady states of tuning curves. The tuning width is given as the full width at half maximum (FWHM) of the tuning curve. To estimate PF and FWHM, we did a cubic spline interpolation to increase the number of sample orientations to 1000. The tuning width index (WI) is given as

$$\mathrm{WI} = \frac{\mathrm{FWHM}(\psi = 45°) - \mathrm{FWHM}(\psi = 0°)}{\mathrm{FWHM}(\psi = 45°) + \mathrm{FWHM}(\psi = 0°)}. \tag{13}$$

To estimate the internal representation of orientation in the network models, denoted as $\hat{\theta}$, we utilized the population vector decoder (PVD) (Georgopoulos, Schwartz et al. 1986)

$$\hat{\theta}(t) = \frac{1}{2}\mathrm{Arg}\left(\sum_{j=1}^{N} \exp\left\{2\mathrm{i}r^j(t)\tilde{\theta}_j\right\} / \sum_{j=1}^{N} r^j\right), \tag{14}$$

where N denotes the number of neurons and $\tilde{\theta}_j$ denotes the PF of neuron j. The orientation is always decoded from the memory network tuning curves $r_{\mathrm{m}}(t)$ except for **Figure 10A**. The estimation bias $b(\theta, t) = \mathbb{E}\left[\hat{\theta}(t)\right] - \theta$. Since the bias is typically small enough, we computed the estimation standard deviation (SD) as the SD of bias using linear statistics. The SD index is defined as

$$\mathrm{SD\ \ index} = \frac{\mathrm{SD}(\theta = 45°) - \mathrm{SD}(\theta = 0°)}{\mathrm{SD}(\theta = 45°) + \mathrm{SD}(\theta = 0°)}. \tag{15}$$

The Fisher information (FI) is estimated by assuming that the probability density function $p(r \mid \theta)$ is Gaussian as

$$p\left(r_{\mathrm{m}}^i \middle| \theta\right) = \frac{1}{\sqrt{2\pi}\sigma_i(\theta)} \mathrm{e}^{\frac{\left(r_{\mathrm{m}}^i(\theta) - \mathbb{E}\left[r_{\mathrm{m}}^i(\theta)\right]\right)^2}{2\sigma_i^2(\theta)}}, \tag{16}$$

where $\sigma_i^2(\theta) = \mathrm{Var}\left(r_{\mathrm{m}}^i(\theta)\right)$ denotes the variance of the firing rate of memory neuron i. Thus, we can estimate the FI of memory neuron i based on the empirical mean and variance of the firing rate at time t as

$$\mathrm{FI}(\psi_i, t) = \frac{\left(\partial \mathbb{E}[r_{\mathrm{m}}^i(\theta, t)] / \partial\theta\right)^2}{\sigma_i^2(\theta, t)}, \tag{17}$$

and the total FI is the summation of the FI of all memory neurons, given as $\mathrm{FI}(t) = \sum_i \mathrm{FI}\,(\psi_i, t)$.

Drift and diffusivity in network models. Although the modulation breaks the continuity of the ring attractor and forms two discrete attractors at the obliques, there is still a one-dimensional trajectory $\bar{\boldsymbol{s}}(\theta)$ to which the noise-free dynamics quickly converge. We can linearize the system in the vicinity of this trajectory if the noise is small (Burak and Fiete 2012). Note that the dynamics of the synaptic variables in Eq. (3) can be put into the following form

$$\tau\dot{\boldsymbol{s}} = -\boldsymbol{s} + \boldsymbol{\phi}(\boldsymbol{W}\boldsymbol{s} + \boldsymbol{h}) + \boldsymbol{\xi}, \tag{18}$$

and by linearizing around the stable trajectory $\boldsymbol{s} = \bar{\boldsymbol{s}}$, we get

$$\tau\dot{\delta\boldsymbol{s}} = K\delta\boldsymbol{s} + \boldsymbol{\xi}, \tag{19}$$

where we have ignored the zeroth- and higher-order terms. The drift velocity $\mu(\theta)$ is estimated by projecting the noise-free dynamics along the normalized right eigenvector u of K with the largest real part of the eigenvalue

$$\mu(\theta) = \frac{1}{\tau \parallel \bar{\boldsymbol{s}}'(\theta) \parallel} u^{\mathrm{T}}(\theta)\big[-\bar{\boldsymbol{s}}(\theta) + \boldsymbol{\phi}\big(\boldsymbol{W}\bar{\boldsymbol{s}}(\theta) + \boldsymbol{h}(\theta)\big)\big]. \tag{20}$$

The coefficient of diffusion can be obtained in the same way

$$2\mathcal{D}(\theta) = \frac{1}{(\tau \parallel \bar{\boldsymbol{s}}'(\theta) \parallel)^2} \sum_i u_i^2\,(\theta)\phi_i\left(\sum_j W_{ij}\,\bar{s}_j(\theta) + h_i\right). \tag{21}$$

The noise coefficient is given as $\sigma = \sqrt{2\mathcal{D}}$. Hence, we have reduced the high-dimensional dynamics to a simple one-dimensional stochastic differential equation as in Eq. (1) as

$$\mathrm{d}\theta = \mu(\theta)\mathrm{d}t + \sigma(\theta)\mathrm{d}W_t,$$

and the potential $U(\theta)$ is obtained by the relation $\frac{\mathrm{d}U}{\mathrm{d}\theta} = -\mu(\theta)$. To quantitatively measure the heterogeneity of noise coefficient across different orientations, we define the noise coefficient index as follows:

$$\text{Noise Coef. index} = \frac{\sigma(\theta = 45^\circ) - \sigma(\theta = 0^\circ)}{\sigma(\theta = 45^\circ) + \sigma(\theta = 0^\circ)}. \tag{22}$$

Network parameters and simulations. Unless otherwise specified, $N_\mathrm{s} = N_\mathrm{m} = 300$, $\tau = 10$ ms. The connectivity parameters are $J_{\mathrm{E,s}} = 0.6, J_{\mathrm{I,s}} = 0.35, J_{\mathrm{E,m}} = 1, J_{\mathrm{I,m}} = 0.17, J_\mathrm{f} = 0.1, J_\mathrm{b} = 0.25, \lambda_{\mathrm{E,s}} = 0.36\pi, \lambda_{\mathrm{I,s}} = 1.1\pi, \lambda_{\mathrm{E,m}} = 0.2\pi, \lambda_{\mathrm{I,m}} = 0.6\pi, \lambda_\mathrm{f} = \lambda_\mathrm{b} = 0.17\pi$. For the external input, we set $C = 4, \varepsilon = 0.2$, and $\lambda_{\mathrm{ext,s}} = 0.3\pi$. For the modulation of the sensory network, unless otherwise specified, we set $\alpha = 0.04$ when only the excitatory plasticity is considered, and $\alpha = 0.03, \beta = 0.08$ when the inhibitory plasticity is added. As for the modulation of the single-layer memory network, we set $\alpha =$

$5 \times 10^{-4}, \beta = 2.4 \times 10^{-3}$. For the transfer function, $f_{\text{max}} = 100, T = 0.1, q = 2, w = 6$ for sensory f_{s}, and $f_{\text{max}} = 100, T = 0.1, q = 1.5, w = 6.6$ for memory f_{m}.

We uniformly sampled 50 cue orientations in $[0°, 180°]$. The visual cue lasts for 0.5 seconds except for the estimation of the PFs. In the grid parameter search figures, the delay epochs last for 1 s. In **Figure 3,** we set $\alpha = 0.07$. In **Figure 5A**, the manifold corresponds to the synaptic variables at 4s into the delay with $\alpha = 0.05$. We uniformly sampled 100 cue orientations for the manifold.

To compute the drift velocity and noise coefficient in Figure 5, 6, and 9, we use the stable trajectory $\bar{\boldsymbol{s}}(\theta)$ at 1s into the delay to ensure the fast transient dynamics induced by stimulus offset fully decays. The stable trajectory is parameterized by the 50 cue orientations to numerically compute $\bar{\boldsymbol{s}}'(\theta)$.

All simulations of ordinary or stochastic differential equations of the network models were done using the Euler method with $\mathrm{d}t = 1\ \mathrm{ms}$. We checked that similar results hold for smaller $\mathrm{d}t$. Example bias and standard deviation patterns were estimated from 1000 independent realizations. The Fisher information patterns were estimated from 3000 independent realizations. The grid search of maximum bias at $\theta = 22.5°$ and standard deviation index were computed from 3000 realizations.

All simulations were run in Matlab. The code is available at https://github.com/KYang-N/Cardinal-Repulsion.git.

Acknowledgments

We appreciate X. Wei for sharing the code for Bayesian inference models. J. Y. was supported by the NYU Shanghai Summer Undergraduate Research Program (SURP). S. L. received STI2030-Major Projects, No.2021ZD0203700/2021ZD0203705. H. Z. and S. L. also acknowledge the support of the Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning and the NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai.

Competing Interests

The authors declare no competing financial interests.

Figure legends

Figure 1. Error patterns of orientation stimuli in delayed-estimation tasks and low-dimensional attractor models. (**A-C**) Characteristic patterns of natural statistics of orientation stimuli θ (**A**), bias (**B**), and standard deviation (SD; **C**) during the delay period observed experimentally. Cardinal orientations are predominant in natural images (**A**). Bias and SD increase during the delay period, keeping patterns of repulsive bias (**B**) and minimum variance (**C**) around cardinal orientations. These characteristic patterns are visualized using trigonometric functions, and the range is normalized by their maximum values. Red vertical lines correspond to representative cardinal and oblique orientations, and with a periodicity of the error patterns, we only show the grey-shaded range in the remaining panels. (**D-L**) Comparison of different attractor models. (**D-F**) Continuous attractors with constant noise. Energy potential is flat (**D**), resulting in no bias (**E**) and uniform SD with uniform noise (**F**). (**G-L**) Discrete attractors with constant (**G-I**) and nonuniform noise (**J-L**). The discrete attractor models have potential hills and wells at cardinal and oblique orientations, respectively (**G,J**). While the bias patterns depend only on the energy landscape (**H,K**), SD representing variability also depends on noise (**I,L**). For the correct SD pattern (**L**), uneven noise with its maxima at the obliques (**J**) is required. Bias and SD patterns in the attractor models were obtained by running one-dimensional drift-diffusion models (see **Methods**).

Figure 2. Extension of Bayesian sensory models. (**A**) Schematics of extension to memory processing. We adapted the previous Bayesian models (Wei and Stocker 2015) for sensory encoding where θ and $\hat{\theta}$ are the input and output of sensory modules. We added a memory module where it maintains $\hat{\theta}$ with the addition of memory noise ξ. The output of the memory module, $\hat{\theta} + \xi$, is fed back to the sensory module as the input for the next iteration. (**B**) Illustration of the first iteration of sensory-memory interaction. Prior distribution follows the natural statistics (top), resulting in a sharper likelihood function near cardinal orientations (middle). Combining prior and likelihood functions leads to the posterior distribution of decoded $\hat{\theta}$ (light colors at the bottom), which is broadened with the addition of memory noise (dark colors at the bottom). Different curves correspond to different initial θ. (**C**) Bias (top) and SD (bottom) patterns obtained from decoded $\hat{\theta}$ for the 1st, 2nd, and 3rd iterations.

Figure 3. Network models of sensory and memory circuits in isolation, implementing efficient coding and ring-attractor dynamics, respectively. (**A**) Schematics of columnar architecture for orientation selectivity. Neurons in the same column have similar preferred orientations, and recurrent connections are a combination of local excitation and global inhibition, represented as triangles and circles, respectively. (**B**-**F**) Connectivity and tuning properties of the sensory network (left column) and memory network (right column). (**B**) Example connectivity strengths. We indexed neurons by ψ ranging uniformly from 0° to 180°. The connectivity strengths depend only on ψ's of the presynaptic and postsynaptic neurons. Each curve shows the connectivity strengths from presynaptic neuron ψ to an example postsynaptic neuron. Unlike the homogeneous connectivity in the memory network (right), the sensory connectivity is heterogeneous, and its degree is denoted by α. (**C**) Heterogeneous tuning curves for different stimulus θ in the sensory network in the

stimulus period (left) and homogeneous ones in the memory network in the delay period (right). The memory network can sustain persistent activity in isolation, while the sensory network cannot. (**D**) Histograms of the preferred orientations. We measured the maximum of the tuning curve of each neuron, denoted as $\tilde{\theta}$ (**Methods**). The heterogeneous sensory network has more cardinally tuned neurons. (**E**) Widths of tuning curves measured at the half maximum of the tuning curves (**Methods**). The sensory tuning curves sharpen around cardinal orientations. Each neuron is labeled with its index ψ as in (**B**). (**F**) Neural manifolds projected onto the first two principal components of activities during the stimulus period (left) and during the delay period (right). The neural manifold of the sensory network resembles a curved ellipsoid, while the manifold corresponding to the homogeneous memory network is a perfect ring.

Figure 4. Network model with interacting sensory and memory modules generates correct error patterns in delayed estimation tasks. (**A**) Schematic of two-module architecture. The sensory and memory modules are connected via feedforward and feedback connectivity to form a closed loop. The sensory module receives external input with orientation θ while internal representation is decoded from the memory module, denoted as $\hat{\theta}$. (**B**) Tuning curves of sensory (upper panels) and memory (lower panels) modules at the end of the stimulus epoch (i.e., the beginning of the delay epoch; left panels) and during the delay period (right panels). Note that while both modules can sustain persistent activity in the delay period, the firing rates of the sensory module are significantly lower than those in the stimulus period (upper right). (**C-E**) Bias (**C**), standard deviation (SD; **D**), and Fisher information (FI; **E**) patterns. Error patterns evaluated at 1, 2.5, and 4 seconds into the delay are consistent with the characteristic patterns observed experimentally in delayed estimation tasks (**Figure 1A-C**). However, the low SD right after the stimulus offset in (**D**) deviates from error patterns seen in perception tasks (see **Discussion**). While FI decays due to noise accumulation, it is largest around cardinal orientations, corresponding to a smaller discrimination threshold (**E**). In (**C**) and (**D**), shaded areas mark the $\pm$s.e.m. of 1000 realizations.

Figure 5. Low-dimensional dynamics along memory manifold and their dependence on heterogeneity degrees in the sensory module. (**A**) Low-dimensional projection along the memory states. Left panel: The memory manifold projected to the first two PCs associated with the vector fields. Right panel: Example drift-diffusion trajectories along the memory manifold starting at $\theta = 112.5^\circ$. (**B**,**C**) Velocity (**B**) and noise coefficients (**C**) corresponding to drift and diffusion processes. Different grey scales represent different heterogeneity degrees in the sensory module, α in **Figure 3B**. The velocity with which the remembered orientation drifts to the obliques in a noise-free network (**B**). A larger noise coefficient around the obliques overcomes the underlying drift dynamics and causes the standard deviation pattern to reach its maxima at the obliques (**C**). (**D**) Equivalent one-dimensional energy potential derived from the velocity in (**B**). (**E,F**) Example bias (**E**) and standard deviation (**F**) patterns at 4s into the delay. The shaded areas mark the $\pm$s.e.m. of 1000 realizations.

Figure 6. Error patterns and low-dimensional dynamics for different intermodal connectivity strengths. (**A-J**) Low-dimensional dynamics and error patterns with varying feedforward and feedback connection strengths, denoted by J_f and J_b. (**K,L**) Potential differences and noise coefficient indices comparing low-dimensional dynamics at cardinal and oblique orientations for changing J_b and heterogeneity degree, α. Increasing both feedforward (**A-D**) and feedback (**E-H**) connection strengths deepens the potential difference (**A,E,K**) and increases the bias (**B,F**), similar to the effects of α increases in **Figure 5D,E**. In contrast, the profile of noise coefficients is less affected (**C,G,L**) and the SD pattern gets flattened with stronger drift (**D,H**). Bias and SD patterns depend on the product of feedforward and feedback connection strengths (**I,J**). Bias and SD are estimated at 4s (**B,D,F,H**) or 1s (**I,J**) into the delay and shaded areas mark the $\pm$s.e.m. of 1000 realizations.

Figure 7. Stronger inhibitory synaptic modulation is required for correct error patterns. (**A**) Segregation of excitatory (blue) and inhibitory (red) synaptic pathways. (**B**) Example excitatory (left) and inhibitory (right) connectivity strengths of the sensory module. The heterogeneity degrees of excitatory and inhibitory connections are denoted by α and β, respectively. Unlike combined excitation and inhibition in **Figure 3B**, the connectivity strengths are maximal around cardinal orientations. (**C,D**) Bias with stimulus at $22.5°$ (**C**) and standard deviation (SD) index (**D**) estimated at 1s into the delay for different values of α and β. SD index compares the SD at the cardinal and oblique orientations (**Methods**). (**E,F**) Example bias (left) and SD (right) patterns when excitatory modulation overwhelms inhibitory modulation ($\alpha = 0.068, \beta = 0.04$; **E**) and when inhibitory modulation is stronger ($\alpha = 0.03, \beta = 0.08$; **F**). In (**C**) and (**D**), green (yellow) pentagrams mark the parameters used in (**E**) and (**F**). Stronger inhibitory modulation is required for correct bias and variance patterns (**F** and green regions in **C** and **D**). In (**E**) and (**F**), shaded areas mark the $\pm$s.e.m. of 1000 realizations.

Figure 8. Network model with memory module only cannot reproduce correct error patterns. (**A**) Schematics of one-module network with heterogeneous and strong recurrent connections that enable both efficient coding and memory maintenance. (**B**) Example tuning curves at the end of the stimulus epoch (left) and at 4s into the delay epoch (right). (**C,D**) Bias with stimulus at $22.5°$ (**C**) and standard deviation (SD) index (**D**) estimated at 1s into the delay for different heterogeneity degrees of excitatory and inhibitory connections, denoted by α and β. For the parameters that generate reasonable bias patterns, the SD index is always negative, which indicates that the SD pattern is inconsistent with experimental findings. (**E**) Bias (left), and SD (right) patterns in the delay. While the bias pattern is correct, the SD reaches maxima around cardinal orientations, unlike the experiments. In (**C**) and (**D**), the yellow pentagram marks the parameters used in (**E**).

Figure 9. Comparison of low-dimensional dynamics between two-module and one-module network models. (**A,B**) Bias and standard deviation (SD) patterns of two-module (**A**) and one-module (**B**) networks, adapted from **Figure 7F** and **Figure 8E**, respectively. The averages of bias and SD over different θ at 4s into the delay are similar in the two networks. (**C-F**) Low-dimensional

dynamics of two-module (black) and one-module (red) networks. In both networks, the energy potential (**C**), the distance between stimulus representation, $\| \bar{\boldsymbol{s}}'(\theta) \|$ and its inverse determining noise coefficients (**D,E**; Eq. (21)), and the noise coefficients (**F**) exhibit similar profiles. However, the two-module network has a shallower potential (**C**) but larger heterogeneity in $\| \bar{\boldsymbol{s}}'(\theta) \|$ and the noise coefficient profile (**D-F**). These differences make it possible for the SD to become smaller around cardinal orientations in the two-module network (right in **A**), while drift dynamics overwhelm and the SD pattern is opposite to that of the noise coefficient in the one-module network (right in **B**). (**G,H**) Potential difference (left) and index of noise coefficients (right) comparing low-dimensional dynamics at the cardinal and oblique orientations in two-module (**G**) and one-module (**H**) networks. The two-module network shows a smaller potential difference and more heterogeneous noise coefficients over a broad range of heterogeneity (see the color bars in **G** and **H**).

Figure 10. Effect of perturbations in sensory-memory interaction on error patterns. (**A,B**) Example bias (**A**) and standard deviation (**B**) patterns when we assumed that TMS is applied to interrupt the feedforward signal from 2.5s into the delay. Shaded areas mark the $\pm$s.e.m. of 1000 realizations. (**C,D**) Evolution of bias with example cue orientation at $\theta = 18°$ (**C**) and the tuning width indices in the memory network (WI; **C**) representing the asymmetry of tuning widths at cardinal and oblique orientations (**Methods**). Two vertical dashed lines mark the end of the stimulus epoch and the beginning of TMS disruption, respectively. Solid and dashed curves correspond to with and without perturbations, respectively. Both bias (**C**) and WI (**D**) stop increasing when TMS is on (**C,D**).

Figure 4 – Figure supplement 1. Bias (**A**) and SD (**B**) patterns decoded from activities of sensory module. All parameters are the same as in **Figure 4**.

Figure 4 – Figure supplement 2. Dynamics of bias and tuning properties of sensory-memory interacting network models. (**A**) Maximum firing rate of $\psi = 22.2°$ for all stimulus orientations. The vertical grey line represents the end of the stimulus presentation. Both sensory and memory modules show lower but sustained activities during the delay period. (**B**) Bias evolution to input orientation $\theta = 18°$. The bias increases both in the stimulus and delay periods, while its increasing speed is reduced during the delay period. (**C**) Tuning width indices (WI) measuring the asymmetry of tuning widths at cardinal and oblique orientations (**Methods**). WI also increases in the whole process, indicating the tuning curves of the neural population become more heterogeneous. All parameters are the same as in **Figure 4**.

Figure 5 – Figure supplement 1. Comparison between bias and standard deviation (SD) patterns of the full network model (orangish) and low-dimensional projection (bluish curves). From top to bottom, each row corresponds to sensory-memory interacting networks in **Figure 5** with $\alpha = 0.03$ (**A,D**), 0.04 (**B,E**), and 0.05 (**C,F**), respectively. We projected the dynamics onto the left (**A-C**) and right (**D-F**) eigenvectors of the Jacobian matrix obtained from local dynamics along the memory

states (**Methods**). The initial orientations of the low-dimensional model were set to be the orientations decoded from the full model at 1s into the delay. We compared the increase of bias from then on, i.e., at 1.2s, 1.5s, and 2s into the delay for the full model, but 0.2s, 0.5s, and 1s for the low-dimensional model. Low-dimensional projection captures characteristic patterns well despite relatively larger deviation in the SD compared to bias, and we found that projecting to the right eigenvector (**D-F**) generally yields better predictions than projecting to the left eigenvector (**A-C**). All parameters are the same as in **Figure 5** except for α. Shaded areas (too narrow to be seen) mark the $\pm$s.e.m. of 3000 realizations.

Figure 5 – Figure supplement 2. SD pattern remains consistent under different noise types. When Poisson noise is replaced by uniform Gaussian noise with $\boldsymbol{\xi} \sim N(0, {\sigma_\xi}^2)$ in Eq. (3), the diffusion term and variability of errors are affected (**A-C**), but the drift and bias patterns remain unchanged (not shown). The distance between stimulus representation $\| \bar{\boldsymbol{s}}'(\theta) \|$ determining noise coefficients (**A**; Eq. (21)), noise coefficients (**B**), and SD (**C**) exhibit similar profiles and dependence on α as those under Poisson noise (see **Fig. 5C,F**). σ_ξ = 0.2 and all remaining parameters are the same as in **Figure 5**.

Figure 7 – Figure supplement 1. Relationship between drift speed and memory loss in two-module (**A-C**) and one-module (**D-F**) networks. (**A,D**) Drift speed for different heterogeneity degrees, α and β. (**B,E**) Minimum Fisher information (FI). The upper panels were obtained with a coarse parameter grid, and lower panels were obtained with a fine grid but only along parameters for the smallest bias increase (red circle) and the orthogonal direction (red triangle). When bias speed is large with unbalanced excitation and inhibition strengths (triangle), the minimum FI decreases quickly in both two-module and one-module networks, suggesting memory loss. On the other hand, along the direction with the smallest bias increase (circle), the minimum FI is relatively high. The FI was estimated at 4s into the delay epoch using 1000 realizations. (**C,F**) Negative correlation between minimum FI and drift speed.

Figure 7 – Figure supplement 2. Error patterns in sensory networks with long intrinsic time constants. In the sensory module alone ($J_\mathrm{f} = J_\mathrm{b} = 0$), with an intrinsic time constant τ=1s in Eq. (3), the bias (**A**) and SD (**B**) remain nearly unchanged during a 5-second delay period. Preferred orientations are estimated 10s after stimulus onset, and all other parameters are the same as those in **Fig. 7F**.

Figure 9 – Figure supplement 1. Error patterns remain unchanged under different levels of noise. Bias (left) and SD (right) in two-module (**A,B**) and one-module (**C,D**) networks with uniform Gaussian noise $\boldsymbol{\xi} \sim N(0, {\sigma_\xi}^2)$ at varying σ_ξ levels. For each heterogeneity level α, the error patterns are unchanged across different σ_ξ values. Bias and SD are estimated 1 second into the delay period from 6000 realizations. All other parameters are the same as in **Figure 7** and **8**.

Adam, K. C. S., R. L. Rademaker and J. T. Serences (2022). Evidence for, and challenges to, sensory recruitment models of visual working memory. Visual Memory. T. F. Brady and W. A. Bainbridge, Routledge**:** 5-25.
Bae, G. Y. (2021). "Neural evidence for categorical biases in location and orientation representations in a working memory task." Neuroimage **240**: 118366.
Bays, P., S. Schneegans, W. Ma and T. F. Brady (2022). Representation and computation in working memory, PsyArXiv.
Bays, P. M. (2014). "Noise in neural populations accounts for errors in working memory." J Neurosci **34**(10): 3632-3645.
Bouchacourt, F. and T. J. Buschman (2019). "A Flexible Model of Working Memory." Neuron **103**(1): 147-160 e148.
Burak, Y. and I. R. Fiete (2009). "Accurate path integration in continuous attractor network models of grid cells." PLoS Comput Biol **5**(2): e1000291.
Burak, Y. and I. R. Fiete (2012). "Fundamental limits on persistent activity in networks of noisy neurons." Proc Natl Acad Sci U S A **109**(43): 17645-17650.
Chunharas, C., R. L. Rademaker, T. F. Brady and J. T. Serences (2022). "An adaptive perspective on visual working memory distortions." J Exp Psychol Gen **151**(10): 2300-2323.
Cole, M. W., J. R. Reynolds, J. D. Power, G. Repovs, A. Anticevic and T. S. Braver (2013). "Multi-task connectivity reveals flexible hubs for adaptive task control." Nat Neurosci **16**(9): 1348-1355.
Compte, A., N. Brunel, P. S. Goldman-Rakic and X. J. Wang (2000). "Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model." Cereb Cortex **10**(9): 910-923.
Darshan, R. and A. Rivkind (2022). "Learning to represent continuous variables in heterogeneous neural networks." Cell Rep **39**(1): 110612.
de Gardelle, V., S. Kouider and J. Sackur (2010). "An oblique illusion modulated by visibility: non-monotonic sensory integration in orientation processing." J Vis **10**(10): 6.
Driscoll, L., K. Shenoy and D. Sussillo (2022). "Flexible multitask computation in recurrent networks utilizes shared dynamical motifs." bioRxiv: 2022.2008.2015.503870.
Eissa, T. L. and Z. P. Kilpatrick (2023). "Learning efficient representations of environmental priors in working memory." PLoS Comput Biol **19**(11): e1011622.
Fischer, B. J. (2010). Bayesian estimates from heterogeneous population codes. The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain.
Frankland, S. M. and J. D. Greene (2020). "Concepts and Compositionality: In Search of the Brain's Language of Thought." Annu Rev Psychol **71**: 273-303.
Ganguli, D. and E. P. Simoncelli (2014). "Efficient sensory encoding and Bayesian inference with heterogeneous neural populations." Neural Comput **26**(10): 2103-2134.
Geisler, W. S. (2008). "Visual perception and the statistical properties of natural scenes." Annu Rev Psychol **59**: 167-192.
Georgopoulos, A. P., A. B. Schwartz and R. E. Kettner (1986). "Neuronal population coding of movement direction." Science **233**(4771): 1416-1419.
Girshick, A. R., M. S. Landy and E. P. Simoncelli (2011). "Cardinal rules: visual orientation perception reflects knowledge of environmental statistics." Nat Neurosci **14**(7): 926-932.
Gu, H., J. Lee, S. Kim, J. Lim, H.-J. Lee, H. Lee, M. Choe, D.-G. Yoo, J. H. Ryu, S. Lim and S.-H. Lee (2023). "Decision-consistent bias mediated by drift dynamics of human visual working memory." bioRxiv: 2023.2006.2028.546818.
Gu, J. and S. Lim (2022). "Unsupervised learning for robust working memory." PLoS Comput Biol **18**(5): e1009083.
Hahn, M. and X. X. Wei (2024). "A unifying theory explains seemingly contradictory biases in perceptual estimation." Nat Neurosci **27**(4): 793-804.

Hansel, D. and G. Mato (2013). "Short-term plasticity explains irregular persistent activity in working memory tasks." J Neurosci **33**(1): 133-149.
Hansel, D. and H. Sompolinsky (1998). Modeling Feature Selectivity in Local Cortical Circuits. Methods in Neuronal Modeling: From Ions to Networks. C. Koch and I. Segev, MIT Press**:** 499-567.
Hardman, K. O., E. Vergauwe and T. J. Ricker (2017). "Categorical working memory representations are used in delayed estimation of continuous colors." J Exp Psychol Hum Percept Perform **43**(1): 30-54.
Itskov, V., D. Hansel and M. Tsodyks (2011). "Short-Term Facilitation may Stabilize Parametric Working Memory Trace." Front Comput Neurosci **5**: 40.
Khan, A. G., J. Poort, A. Chadwick, A. Blot, M. Sahani, T. D. Mrsic-Flogel and S. B. Hofer (2018). "Distinct learning-induced changes in stimulus selectivity and interactions of GABAergic interneuron classes in visual cortex." Nat Neurosci **21**(6): 851-859.
Khona, M. and I. R. Fiete (2022). "Attractor and integrator networks in the brain." Nat Rev Neurosci **23**(12): 744-766.
Kreile, A. K., T. Bonhoeffer and M. Hubener (2011). "Altered visual experience induces instructive changes of orientation preference in mouse visual cortex." J Neurosci **31**(39): 13911-13920.
Kwak, Y. and C. E. Curtis (2022). "Unveiling the abstract format of mnemonic representations." Neuron **110**(11): 1822-1828 e1825.
Larisch, R., L. Gonner, M. Teichmann and F. H. Hamker (2021). "Sensory coding and contrast invariance emerge from the control of plastic inhibition over emergent selectivity." PLoS Comput Biol **17**(11): e1009566.
Leavitt, M. L., D. Mendoza-Halliday and J. C. Martinez-Trujillo (2017). "Sustained Activity Encoding Working Memories: Not Fully Distributed." Trends Neurosci **40**(6): 328-346.
Li, B., M. R. Peterson and R. D. Freeman (2003). "Oblique effect: a neural basis in the visual cortex." J Neurophysiol **90**(1): 204-217.
Mejias, J. F. and X. J. Wang (2022). "Mechanisms of distributed working memory in a large-scale network of macaque neocortex." Elife **11**.
Olshausen, B. A. and D. J. Field (1996). "Emergence of simple-cell receptive field properties by learning a sparse code for natural images." Nature **381**(6583): 607-609.
Panichello, M. F., B. DePasquale, J. W. Pillow and T. J. Buschman (2019). "Error-correcting dynamics in visual working memory." Nat Commun **10**(1): 3366.
Pollock, E. and M. Jazayeri (2020). "Engineering recurrent neural networks from task-relevant manifolds and dynamics." PLoS Comput Biol **16**(8): e1008128.
Pratte, M. S., Y. E. Park, R. L. Rademaker and F. Tong (2017). "Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory." J Exp Psychol Hum Percept Perform **43**(1): 6-17.
Rademaker, R. L., C. Chunharas and J. T. Serences (2019). "Coexisting representations of sensory and mnemonic information in human visual cortex." Nat Neurosci **22**(8): 1336-1344.
Renart, A., P. Song and X. J. Wang (2003). "Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks." Neuron **38**(3): 473-485.
Roussy, M., D. Mendoza-Halliday and J. C. Martinez-Trujillo (2021). "Neural Substrates of Visual Perception and Working Memory: Two Sides of the Same Coin or Two Different Coins?" Front Neural Circuits **15**: 764177.
Schneegans, S. and P. M. Bays (2018). "Drift in Neural Population Activity Causes Working Memory to Deteriorate Over Time." J Neurosci **38**(21): 4859-4869.
Seeholzer, A., M. Deger and W. Gerstner (2019). "Stability of working memory in continuous attractor networks under the control of short-term plasticity." PLoS Comput Biol **15**(4): e1006928.
Seung, H. S., D. D. Lee, B. Y. Reis and D. W. Tank (2000). "The autapse: a simple illustration of short-term analog memory storage by tuned synaptic feedback." J Comput Neurosci **9**(2): 171-185.

Shen, G., X. Tao, B. Zhang, E. L. Smith, 3rd and Y. M. Chino (2014). "Oblique effect in visual area 2 of macaque monkeys." J Vis **14**(2).
Simon, H. A. (1995). Near-decomposability and complexity: How a mind resides in a brain. The Mind, the Brain, and Complex Adaptive Systems. H. Morowitz and J. Singer, Addison-Wesley**:** 25-43.
Taylor, R. and P. M. Bays (2018). "Efficient Coding in Visual Working Memory Accounts for Stimulus-Specific Variations in Recall." J Neurosci **38**(32): 7132-7142.
Tomić, I. and P. M. Bays (2023). "A dynamic neural resource model bridges sensory and working memory." bioRxiv: 2023.2003.2027.534406.
van Bergen, R. S., W. J. Ma, M. S. Pratte and J. F. Jehee (2015). "Sensory uncertainty decoded from visual cortex predicts behavior." Nat Neurosci **18**(12): 1728-1730.
van de Ven, V., C. Jacobs and A. T. Sack (2012). "Topographic contribution of early visual cortex to short-term memory consolidation: a transcranial magnetic stimulation study." J Neurosci **32**(1): 4-11.
van den Berg, R., H. Shin, W. C. Chou, R. George and W. J. Ma (2012). "Variability in encoding precision accounts for visual short-term memory limitations." Proc Natl Acad Sci U S A **109**(22): 8780-8785.
Vogels, T. P., R. C. Froemke, N. Doyon, M. Gilson, J. S. Haas, R. Liu, A. Maffei, P. Miller, C. J. Wierenga, M. A. Woodin, F. Zenke and H. Sprekeler (2013). "Inhibitory synaptic plasticity: spike timing-dependence and putative network function." Front Neural Circuits **7**: 119.
Wang, X. J. (2001). "Synaptic reverberation underlying mnemonic persistent activity." Trends Neurosci **24**(8): 455-463.
Webster, M. A. (2015). "Visual Adaptation." Annu Rev Vis Sci **1**: 547-567.
Wei, X. X. and A. A. Stocker (2015). "A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts." Nat Neurosci **18**(10): 1509-1517.
Wei, X. X. and A. A. Stocker (2017). "Lawful relation between perceptual bias and discriminability." Proc Natl Acad Sci U S A **114**(38): 10244-10249.
Wilson, H. R. (1999). Spikes, decisions, and actions : the dynamical foundations of neuroscience. Oxford, Oxford University Press.
Wimmer, K., D. Q. Nykamp, C. Constantinidis and A. Compte (2014). "Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory." Nat Neurosci **17**(3): 431-439.
Xu, Y. (2020). "Revisit once more the sensory storage account of visual working memory." Vis cogn **28**(5-8): 433-446.
Yang, G. R., M. R. Joglekar, H. F. Song, W. T. Newsome and X. J. Wang (2019). "Task representations in neural networks trained to perform many cognitive tasks." Nat Neurosci **22**(2): 297-306.
Zhang, K. (1996). "Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory." J Neurosci **16**(6): 2112-2126.
Zylberberg, J., J. T. Murphy and M. R. DeWeese (2011). "A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields." PLoS Comput Biol **7**(10): e1002250.

Figure 1

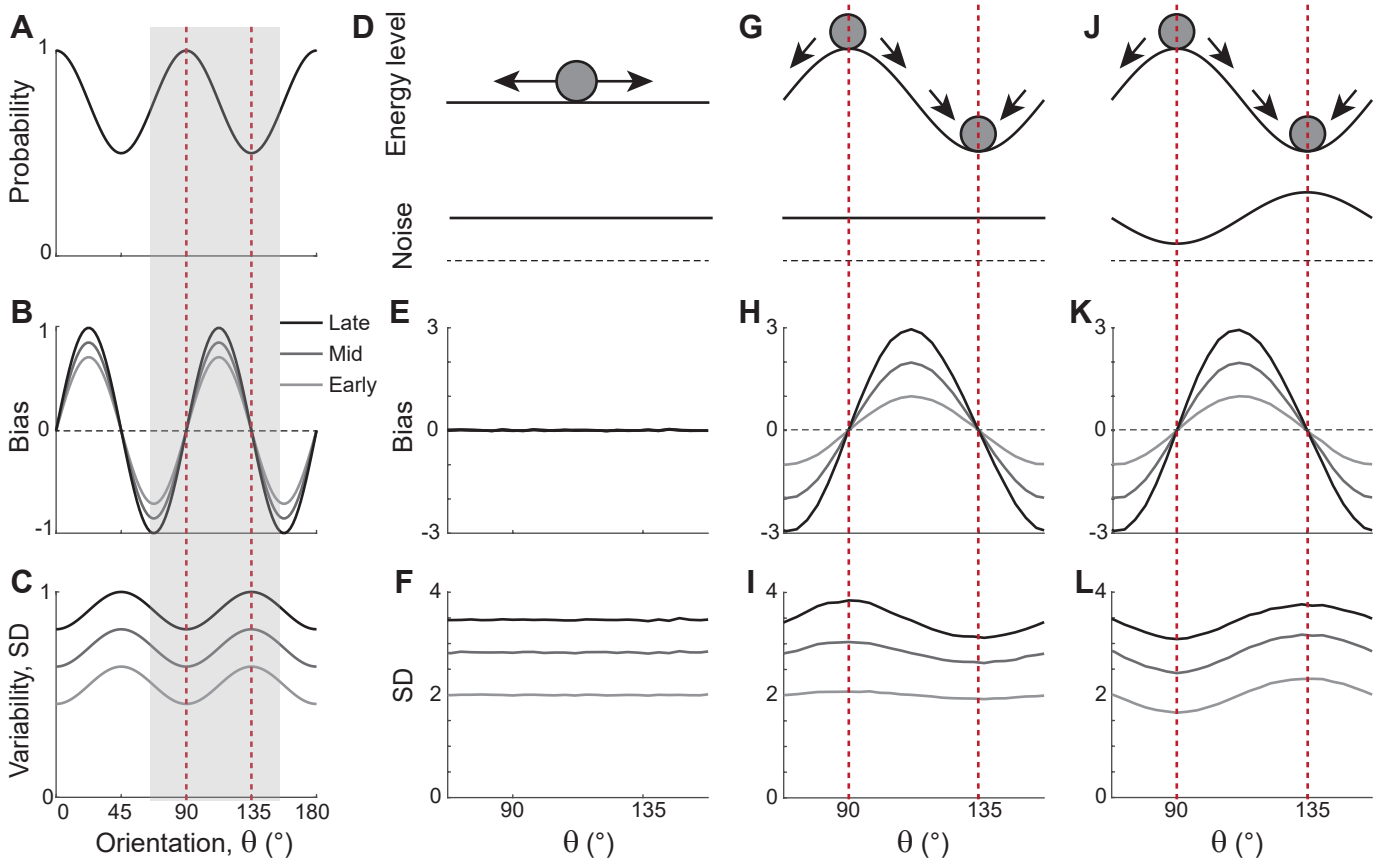
A
B
C
D
E
F
G
H
I
J
K
L
Probability
Bias
Variability, SD
Orientation, θ (°)
Late
Mid
Early
Energy level
Noise
SD
θ (°)
0
45
90
135
180
1
-1
3
-3
4
2

Figure 2

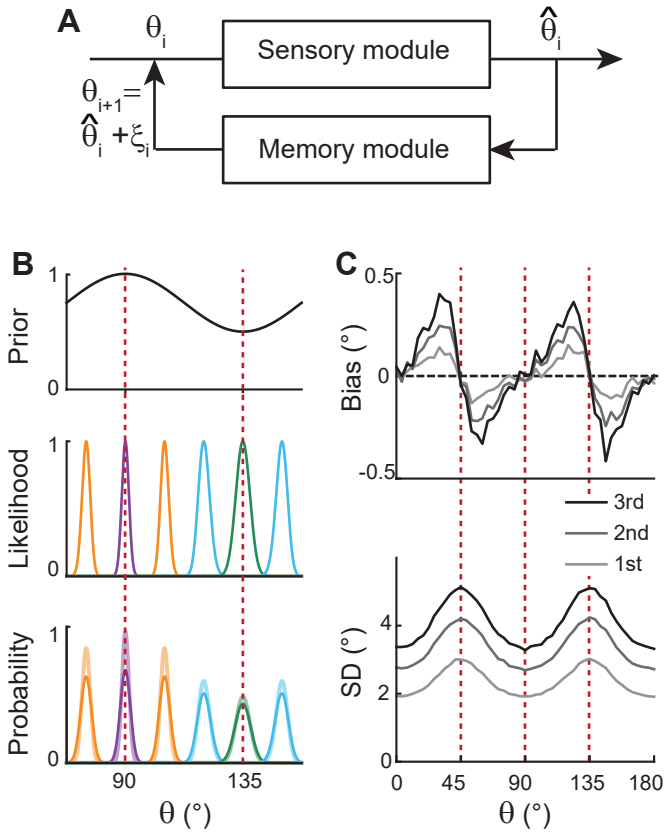
A
θi
Sensory module
θ̂i
θi+1 =
θ̂i +ξi
Memory module
B
Prior
Likelihood
Probability
1
0
90
135
θ (°)
C
Bias (°)
0.5
0
-0.5
3rd
2nd
1st
SD (°)
4
2
0
45
180
θ (°)

Figure 3

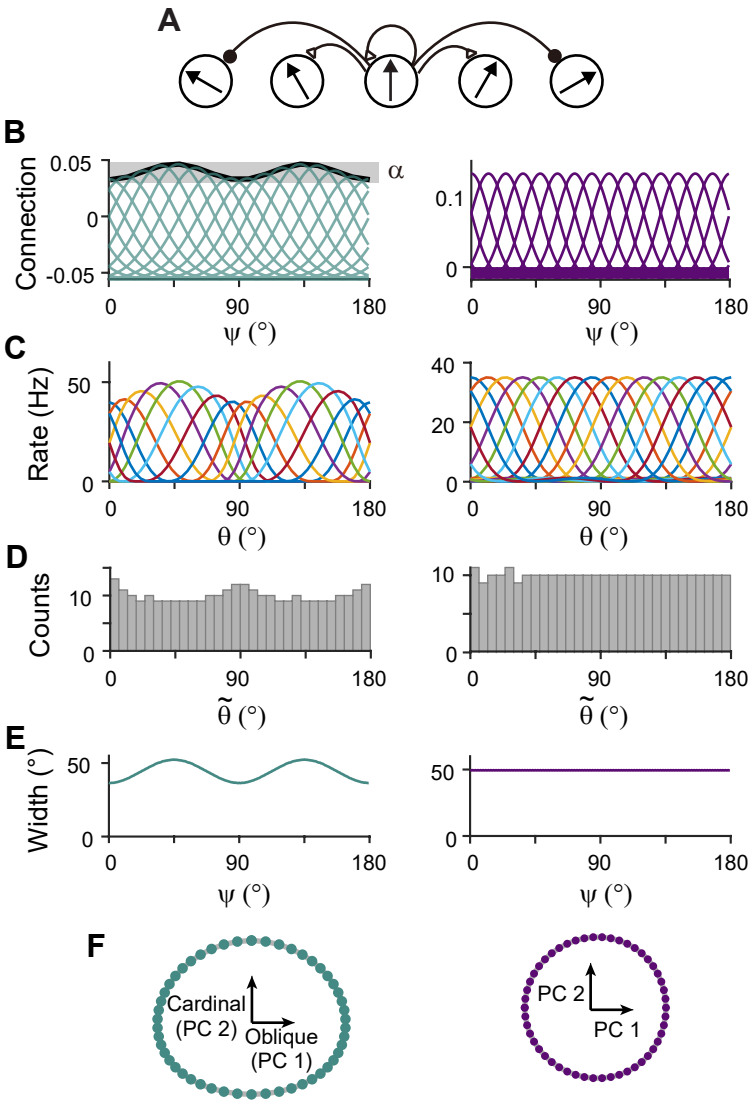
A
B
Connection
0.05
0
-0.05
α
0.1
0
0
90
180
ψ (°)
C
Rate (Hz)
50
0
40
20
θ (°)
D
Counts
10
θ̃ (°)
E
Width (°)
F
Cardinal
(PC 2)
Oblique
(PC 1)
PC 2
PC 1

Figure 4

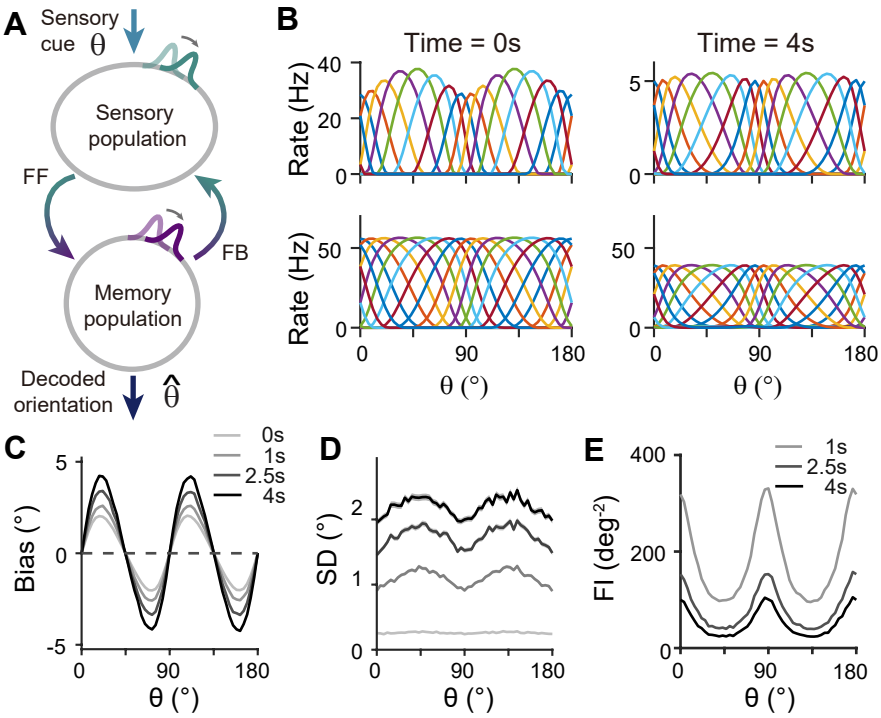
A
Sensory cue θ
Sensory population
FF
FB
Memory population
Decoded orientation θ̂
B
Time = 0s
Time = 4s
Rate (Hz)
θ (°)
C
Bias (°)
0s
1s
2.5s
4s
D
SD (°)
E
FI (deg⁻²)

Figure 5

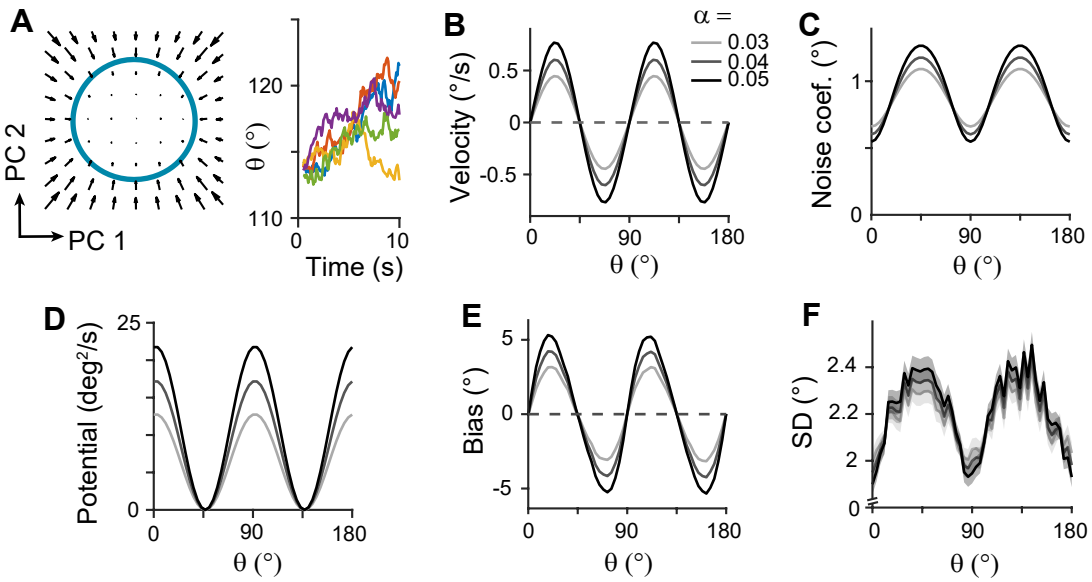
A
PC 2
PC 1
120
110
θ (°)
0
10
Time (s)
B
α =
0.03
0.04
0.05
Velocity (°/s)
0.5
0
-0.5
0
90
180
θ (°)
C
Noise coef. (°)
1
0
0
90
180
θ (°)
D
Potential (deg²/s)
25
0
0
90
180
θ (°)
E
Bias (°)
5
0
-5
0
90
180
θ (°)
F
SD (°)
2.4
2.2
2
0
0
90
180
θ (°)

Figure 6

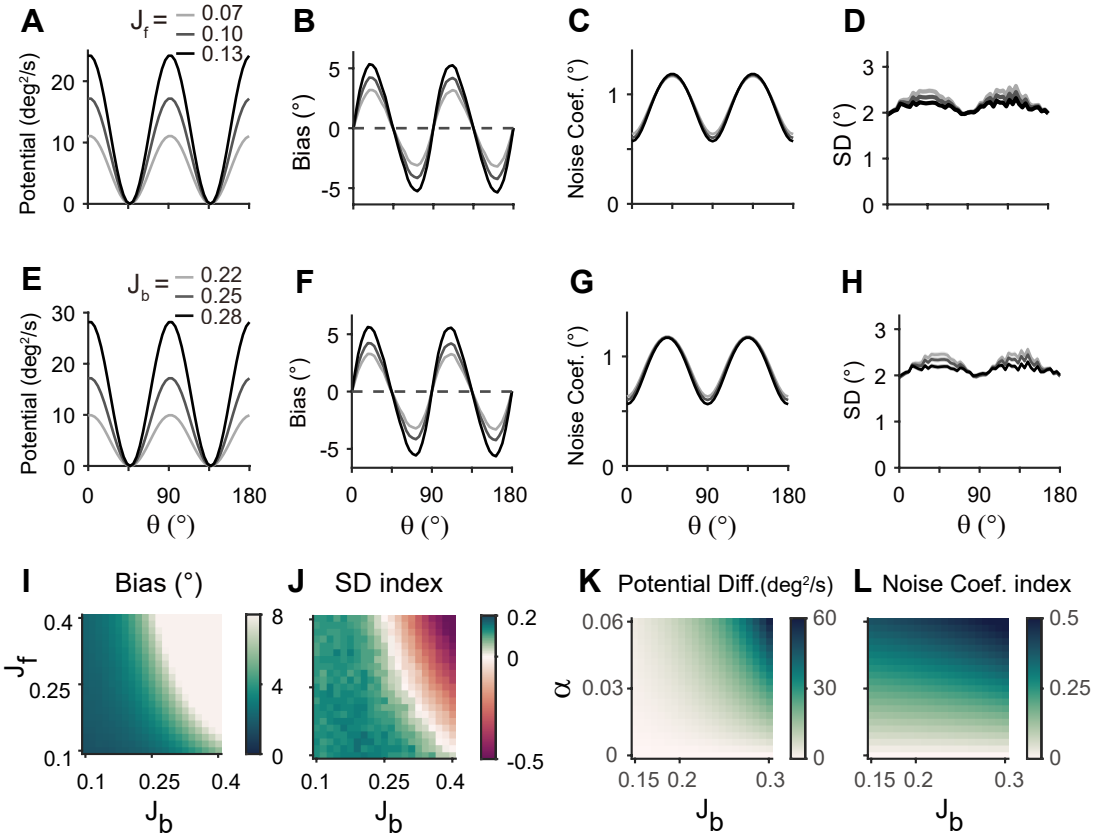
A
J_f = 0.07
0.10
0.13
Potential (deg²/s)
B
Bias (°)
C
Noise Coef. (°)
D
SD (°)
E
J_b = 0.22
0.25
0.28
F
G
H
θ (°)
I
Bias (°)
J
SD index
K
Potential Diff.(deg²/s)
L
Noise Coef. index
J_f
J_b
α

Figure 7

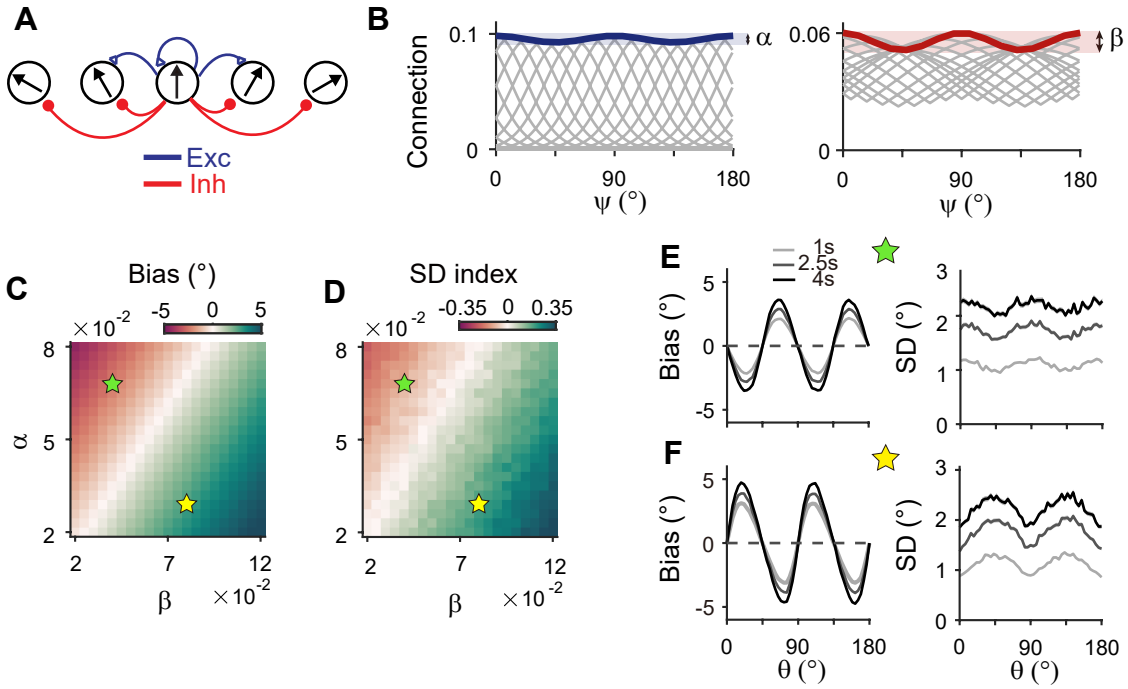
A
Exc
Inh
B
Connection
0.1
0
α
0.06
β
0
90
180
ψ (°)
C
Bias (°)
×10^-2
-5
0
5
8
5
2
α
2
7
12
β
×10^-2
D
SD index
×10^-2
-0.35
0
0.35
E
1s
2.5s
4s
Bias (°)
SD (°)
F
θ (°)

Figure 8

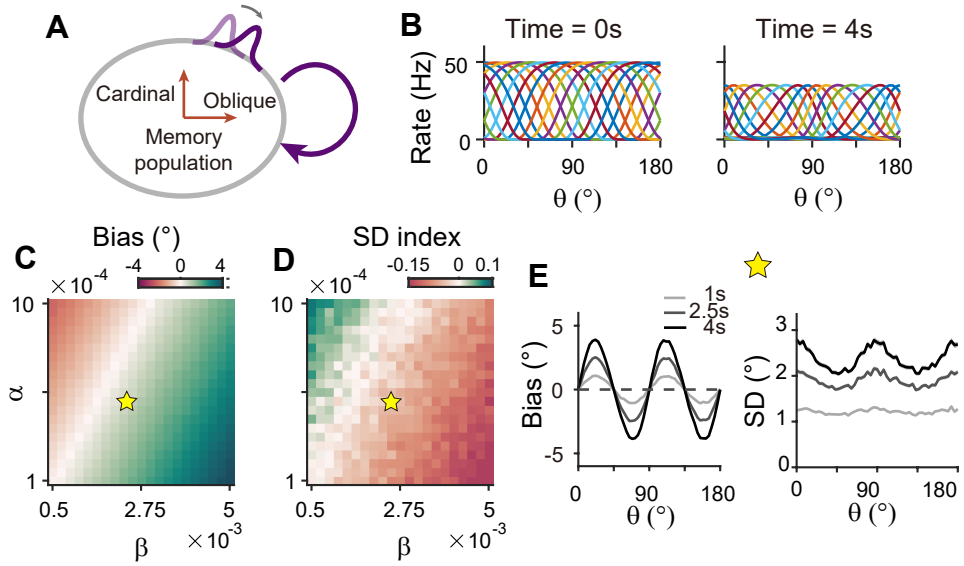
A
Cardinal
Oblique
Memory population
B
Time = 0s
Time = 4s
Rate (Hz)
50
0
0
90
180
θ (°)
C
Bias (°)
-4
0
4
×10^-4
10
α
1
0.5
2.75
5
β
×10^-3
D
SD index
-0.15
0
0.1
E
1s
2.5s
4s
Bias (°)
5
0
-5
SD (°)
3
2
1
0

Figure 9

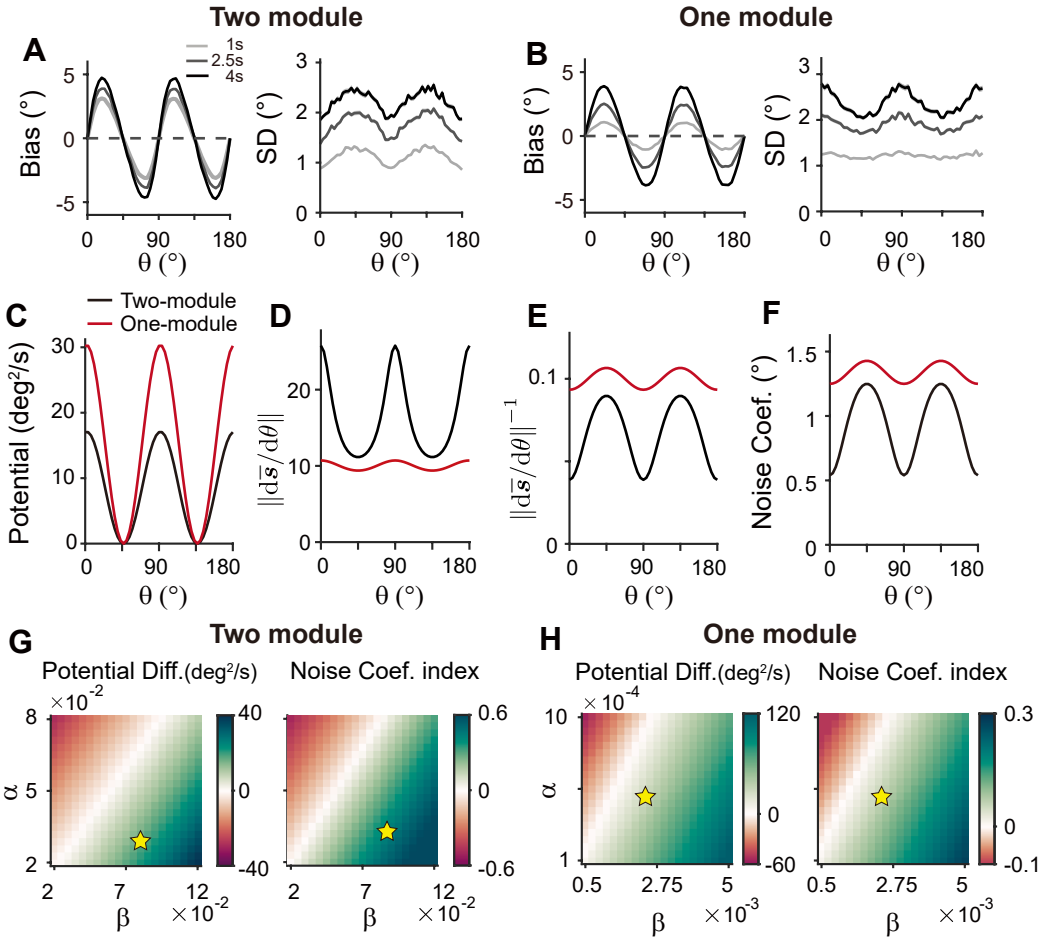
Two module
One module
A
B
1s
2.5s
4s
Bias (°)
SD (°)
θ (°)
C
D
E
F
Two-module
One-module
Potential (deg²/s)
Noise Coef. (°)
G
H
Potential Diff.(deg²/s)
Noise Coef. index
α
β

Figure 10

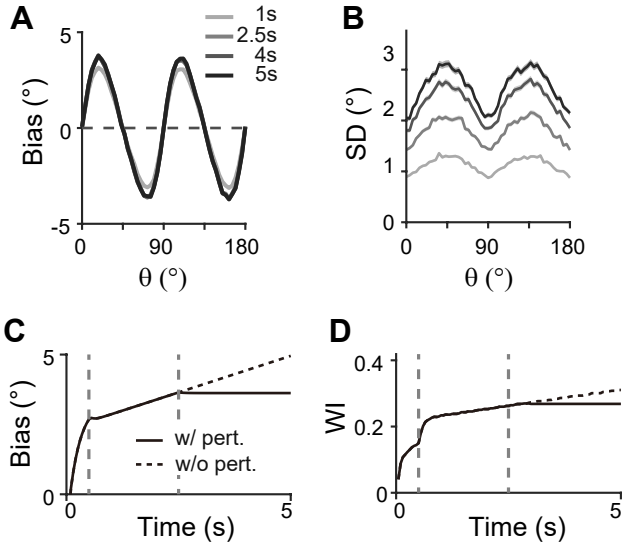
A
1s
2.5s
4s
5s
Bias (°)
5
0
-5
0
90
180
θ (°)
B
SD (°)
3
2
1
0
0
90
180
θ (°)
C
Bias (°)
5
0
w/ pert.
w/o pert.
0
5
Time (s)
D
WI
0.4
0.2
0
0
5
Time (s)

Figure 4 - Figure Supplement 1

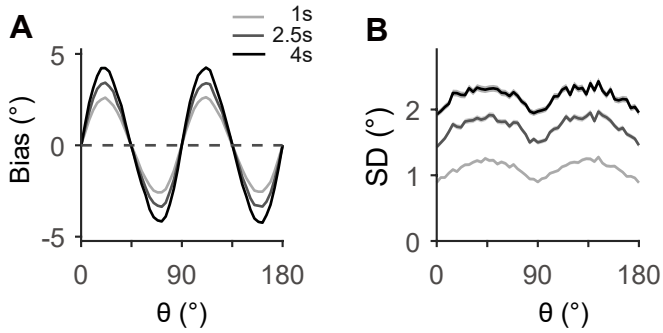

Figure 4 – Figure supplement 1. Bias (**A**) and SD (**B**) patterns decoded from activities of sensory module. All parameters are the same as in **Figure 4**.

Figure 4 - Figure Supplement 2

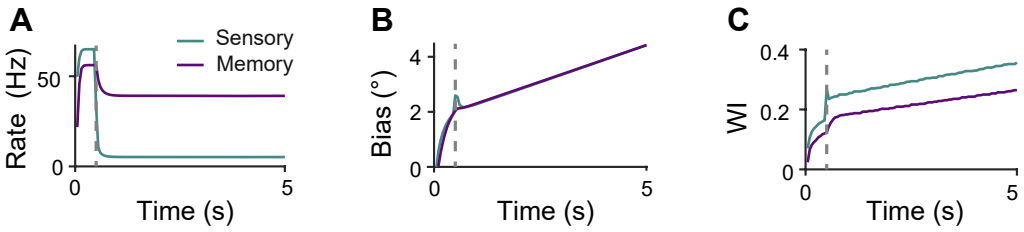

Figure 4 – Figure supplement 2. Dynamics of bias and tuning properties of sensory-memory interacting network models. (**A**) Maximum firing rate of $\psi = 22.2°$ for all stimulus orientations. The vertical grey line represents the end of the stimulus presentation. Both sensory and memory modules show lower but sustained activities during the delay period. (**B**) Bias evolution to input orientation $\theta = 18°$. The bias increases both in the stimulus and delay periods, while its increasing speed is reduced during the delay period. (**C**) Tuning width indices (WI) measuring the asymmetry of tuning widths at cardinal and oblique orientations (**Methods**). WI also increases in the whole process, indicating the tuning curves of the neural population become more heterogeneous. All parameters are the same as in **Figure 4**.

Figure 5 - Figure Supplement 1

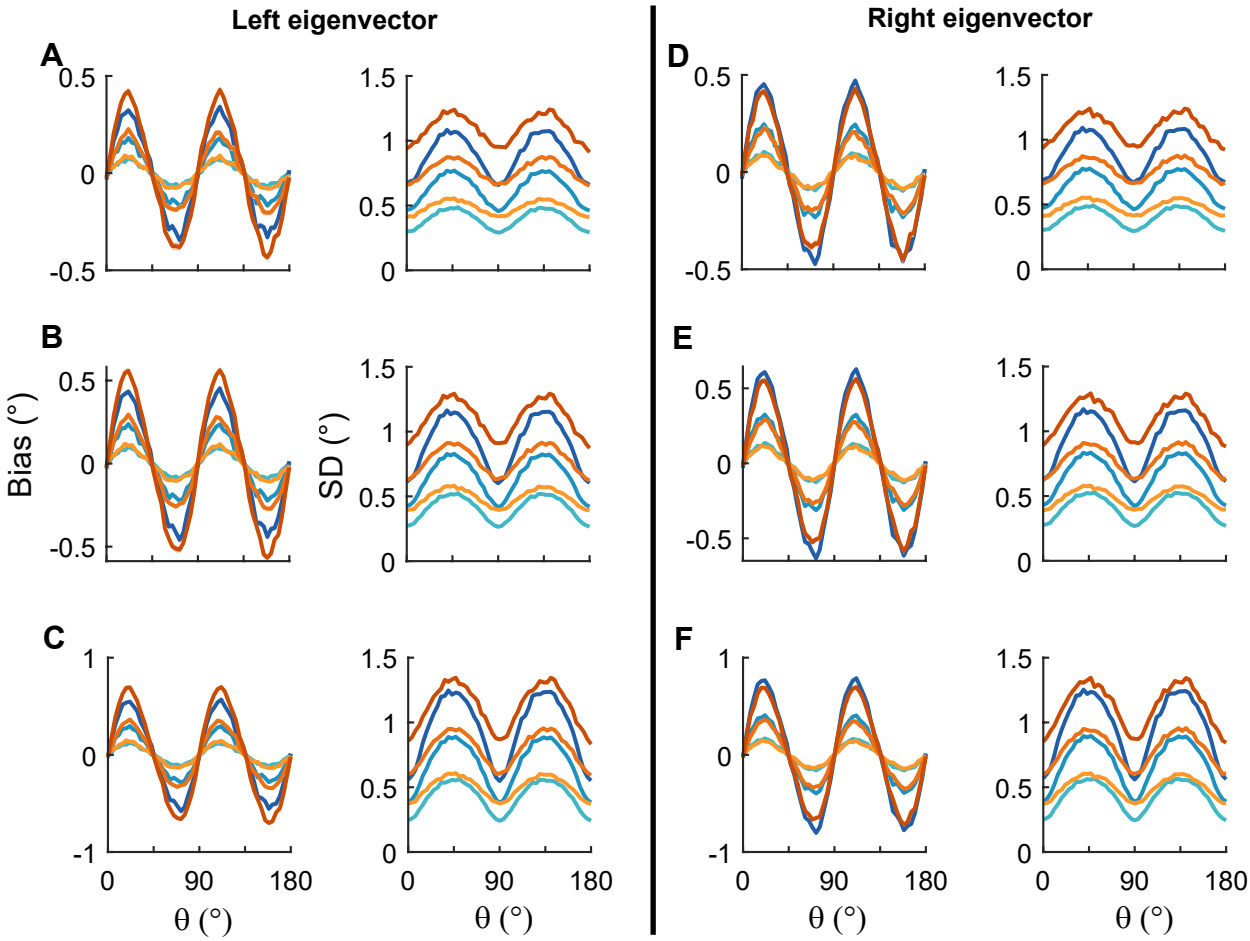

Figure 5 – Figure supplement 1. Comparison between bias and standard deviation (SD) patterns of the full network model (orangish) and low-dimensional projection (bluish curves). From top to bottom, each row corresponds to sensory-memory interacting networks in **Figure 5** with $\alpha = 0.03$ (**A,D**), 0.04 (**B,E**), and 0.05 (**C,F**), respectively. We projected the dynamics onto the left (**A-C**) and right (**D-F**) eigenvectors of the Jacobian matrix obtained from local dynamics along the memory states (**Methods**). The initial orientations of the low-dimensional model were set to be the orientations decoded from the full model at 1s into the delay. We compared the increase of bias from then on, i.e., at 1.2s, 1.5s, and 2s into the delay for the full model, but 0.2s, 0.5s, and 1s for the low-dimensional model. Low-dimensional projection captures characteristic patterns well despite relatively larger deviation in the SD compared to bias, and we found that projecting to the right eigenvector (**D-F**) generally yields better predictions than projecting to the left eigenvector (**A-C**). All parameters are the same as in **Figure 5** except for α. Shaded areas (too narrow to be seen) mark the $\pm$s.e.m. of 3000 realizations.

Figure 5 - Figure Supplement 2

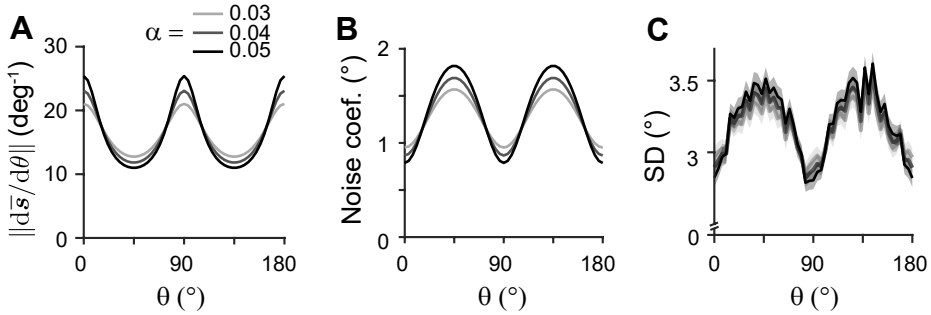

Figure 5 – Figure supplement 2. SD pattern remains consistent under different noise types. When Poisson noise is replaced by uniform Gaussian noise with $\boldsymbol{\xi} \sim N(0,\ \sigma_{\xi}^{2})$ in Eq. (3), the diffusion term and variability of errors are affected (**A-C**), but the drift and bias patterns remain unchanged (not shown). The distance between stimulus representation $\|\ \bar{\boldsymbol{s}}'(\theta)\ \|$ determining noise coefficients (**A**; Eq. (21)), noise coefficients (**B**), and SD (**C**) exhibit similar profiles and dependence on α as those under Poisson noise (see **Fig. 5C,F**). σ_{ξ} = 0.2 and all remaining parameters are the same as in **Figure 5**.

Figure 7 - Figure Supplement 1

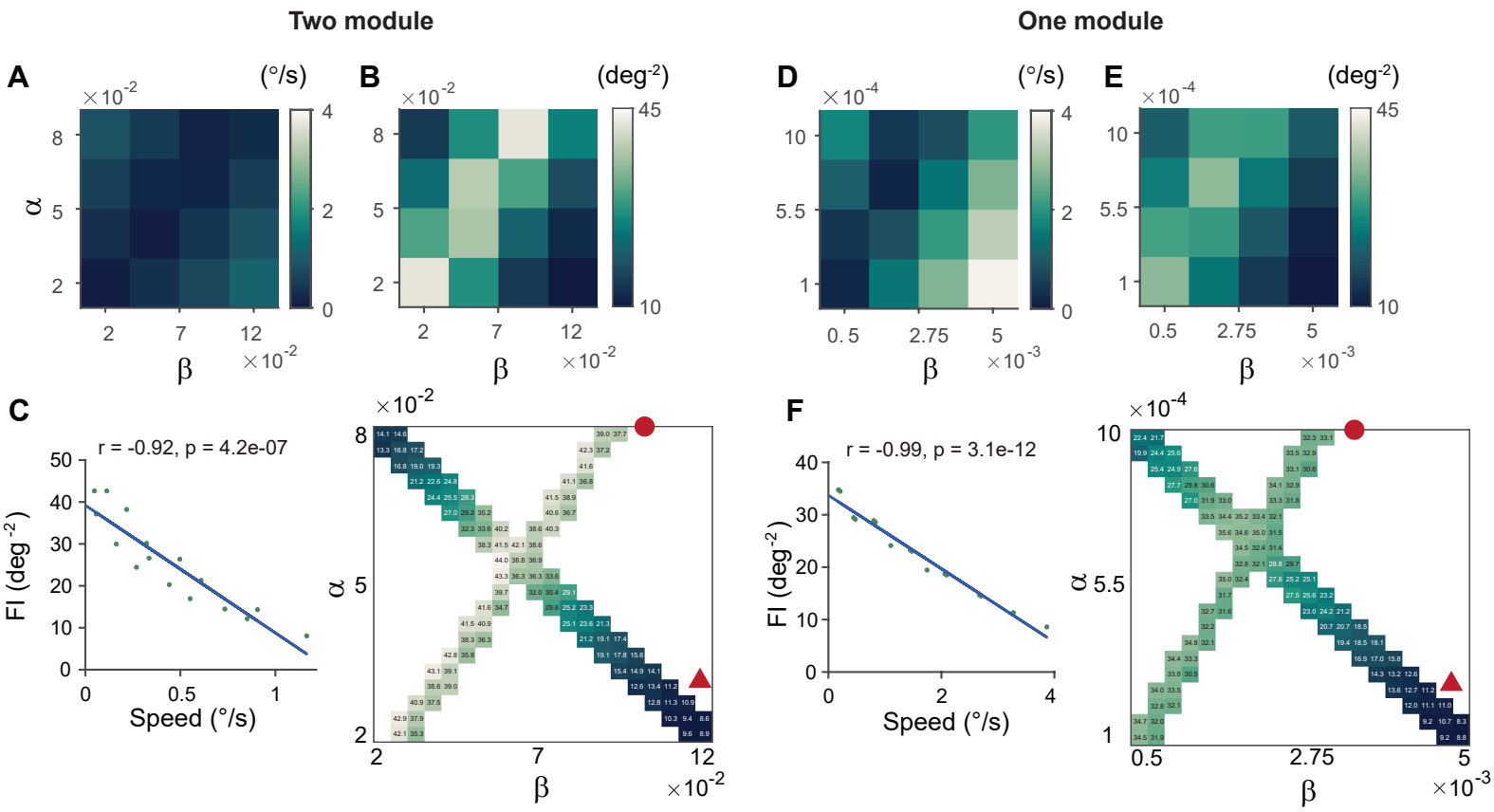

Figure 7 – Figure supplement 1. Relationship between drift speed and memory loss in two-module (**A-C**) and one-module (**D-F**) networks. (**A,D**) Drift speed for different heterogeneity degrees, α and β. (**B,E**) Minimum Fisher information (FI). The upper panels were obtained with a coarse parameter grid, and lower panels were obtained with a fine grid but only along parameters for the smallest bias increase (red circle) and the orthogonal direction (red triangle). When bias speed is large with unbalanced excitation and inhibition strengths (triangle), the minimum FI decreases quickly in both two-module and one-module networks, suggesting memory loss. On the other hand, along the direction with the smallest bias increase (circle), the minimum FI is relatively high. The FI was estimated at 4s into the delay epoch using 1000 realizations. (**C,F**) Negative correlation between minimum FI and drift speed.

Figure 7 - Figure Supplement 2

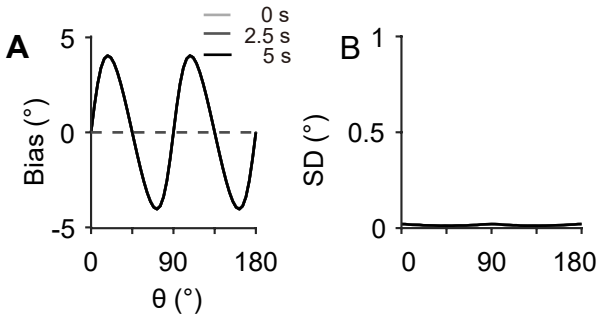

Figure 7 – Figure supplement 2. Error patterns in sensory networks with long intrinsic time constants. In the sensory module alone, with an intrinsic time constant τ=1s in Eq. (3), the bias (**A**) and SD (**B**) remain nearly unchanged during a 5-second delay period. Preferred orientations are estimated 10s after stimulus onset, and all other parameters are the same with those in **Fig. 7F**.

Figure 9 - Figure Supplement 1

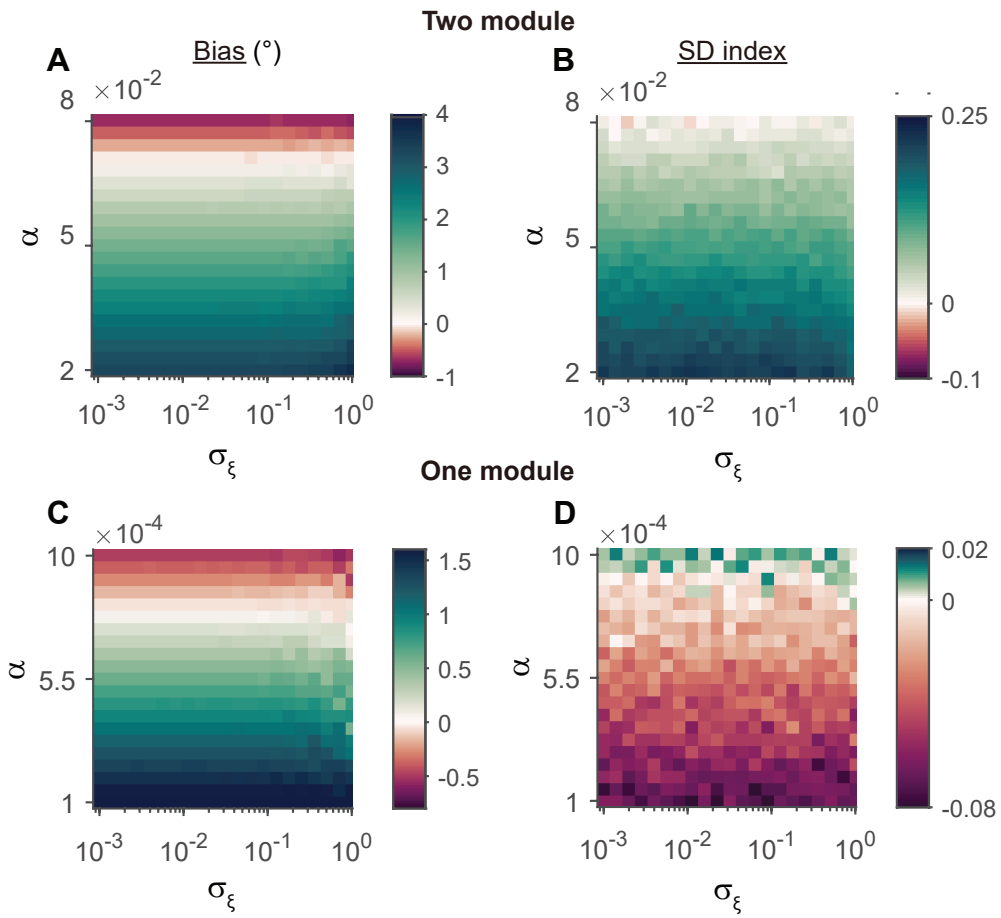

Figure 9 – Figure supplement 1. Error patterns remain unchanged under different levels of noise. Bias (left) and SD (right) in two-module (**A,B**) and one-module (**C,D**) networks with uniform Gaussian noise $\xi \sim N(0, \sigma_\xi{}^2)$ at varying σ_ξ levels. For each heterogeneity level α, the error patterns are unchanged across different σ_ξ values. Bias and SD are estimated 1 second into a delay period from 6000 realizations. All other parameter are the same as in **Figure 7** and **8**.