Rate-distortion theory of neural coding and its implications for working memory

Anthony M.V. Jakob[1,2,*] and Samuel J. Gershman[3,4]
[1]Section of Life Sciences Engineering, École Polytechnique Fédérale de Lausanne
[2]Department of Neurobiology, Harvard Medical School
[3]Department of Psychology and Center for Brain Science, Harvard University
[4]Center for Brains, Minds, and Machines, MIT
[*]Correspondence: anthony_jakob@hms.harvard.edu

February 27, 2022

Abstract

Rate-distortion theory provides a powerful framework for understanding the nature of human memory by formalizing the relationship between information rate (the average number of bits per stimulus transmitted across the memory channel) and distortion (the cost of memory errors). Here we show how this abstract computational-level framework can be realized by a model of neural population coding. The model reproduces key regularities of visual working memory, including some that were not previously explained by population coding models. We verify a novel prediction of the model by reanalyzing recordings of monkey prefrontal neurons during an oculomotor delayed response task.

Introduction

All memory systems are capacity-limited in the sense that a finite amount of information about the past can be stored and retrieved without error. Most digital storage systems are designed to work without error. Memory in the brain, by contrast, is error-prone. In the domain of working memory, these errors follow well-behaved functions of set size, variability, attention, among other factors. An important insight into the nature of such regularities was the recognition that they may emerge from maximization of memory performance subject to a capacity limit or encoding cost [1, 2, 3, 4, 5, 6, 7].

Rate-distortion theory [8] provides a general formalization of the memory optimization problem (reviewed in [9]). The costs of memory errors are specified by a *distortion function*; the capacity of memory is specified by an upper bound on the mutual information between the inputs (memoranda) and outputs (reconstructions) of the memory system. Systems with higher capacity can achieve lower expected distortion, tracing out an optimal trade-off curve in the rate-distortion plane. The hypothesis that human memory operates near the optimal trade-off curve allows one to deduce several known regularities of working memory errors, some of which we describe below. Past work has studied rate-distortion trade-offs in human memory [1, 2, 10], as well as in

other domains such as category learning [4], perceptual identification [11], visual search [5], linguistic communication [12], and decision making [13, 14].

Our goal is to show how the abstract rate-distortion framework can be realized in a neural circuit using population coding. As exemplified by the work of Bays and his colleagues, population coding offers a systematic account of working memory performance [15, 16, 17, 18, 19, 20, 21], according to which errors arise from the readout of a noisy spiking population that encodes memoranda. We show that a modified version of the population coding model implements the celebrated Blahut-Arimoto algorithm for rate-distortion optimization [22, 23]. The modified version can explain a number of phenomena that were puzzling under previous population coding accounts, such as *serial dependence* (the influence of previous trials on performance [24]).

The Blahut-Arimoto algorithm is parametrized by a coefficient that specifies the trade-off between rate and distortion. In our circuit implementation, this coefficient controls the precision of the population code. We derive a homeostatic learning rule that adapts the coefficient to maintain performance at the capacity limit. This learning rule explains the dependence of memory performance on the intertrial and retention intervals [25, 26, 27]. It also makes the prediction that performance should adapt across trials to maintain a set point close to the channel capacity. We confirm these performance adjustments empirically. Finally, we show that variations in performance track changes in neural gain, consistent with our theory.

Results

The channel design problem

We begin with an abstract characterization of the channel design problem, before specializing it to the case of neural population coding. A communcation channel (Figure 1A) is a probabilistic mapping, $Q(\hat{\theta}|\theta)$, from input θ to a reconstruction $\hat{\theta}$. The input and output spaces are assumed to be discrete in our treatment (for continuous variables like color and orientation, we use discretization into a finite number of bins; see also [2]). We also assume that there is some capacity limit C on the amount of information that this channel can communicate about θ, as quantified by the mutual information $I(\theta;\hat{\theta})$ between θ and the stimulus estimate $\hat{\theta}$ decoded from the population activity. We will refer to $R \equiv I(\theta;\hat{\theta})$ as the channel's *information rate*. To derive the optimal channel design, we also need to specify what *distortion function* $d(\theta,\hat{\theta})$ the channel is optimizing—i.e., how errors are quantified. Details on our choice of distortion function can be found below.

With these elements in hand, we can define the channel design problem as finding the channel Q^* that minimizes expected distortion $D \equiv \mathbb{E}[d(\theta,\hat{\theta})]$ subject to the constraint that the information rate R cannot exceed the capacity limit C:

$$Q^* = \underset{Q:R\leq C}{\operatorname{argmin}}\, D. \tag{1}$$

For computational convenience, we can equivalently formulate this problem using a Lagrangian:

$$Q^* = \underset{Q}{\operatorname{argmin}}\, R + \beta D, \tag{2}$$

where β is a Lagrange multiplier equal to the slope of the rate-distortion function at the capacity limit:

$$\beta = -\frac{\partial R}{\partial D}. \tag{3}$$

The rate-distortion function is monotonically decreasing and convex; hence the value of β is always positive.

Using the Lagrangian formulation, one can show that the optimal channel for a discrete stimulus takes the following form:

$$Q^*(\hat{\theta}|\theta) \propto \exp[-\beta d(\theta, \hat{\theta}) + \log \bar{Q}(\hat{\theta})], \tag{4}$$

where the marginal probability $\bar{Q}(\hat{\theta})$ is defined by:

$$\bar{Q}(\hat{\theta}) = \sum_{\theta} P(\theta) Q^*(\hat{\theta}|\theta). \tag{5}$$

These two equations are coupled. One can obtain the optimal channel by initializing them to uniform distributions and iterating them until convergence. This is known as the Blahut-Arimoto algorithm [22, 23].

For a channel with a fixed capacity C but variable D across contexts, the Lagrange multiplier β will need to be adjusted for each context so that $R = C$. We can accomplish this by computing R for a range of β values and choosing the value that gets closest to the constraint C (later we will propose a more biologically plausible algorithm). Because the rate-distortion function is monotonically decreasing and convex, β will always be a decreasing function of D. Intuitively, β characterizes the sensitivity of the channel to the stimulus. When stimulus sensitivity is lower, the information rate is lower and hence the expected distortion is higher.

In general, we will be interested in communicating a collection of K stimuli, $\theta = \{\theta_1, \ldots, \theta_K\}$, with associated probing probabilities $\pi = \{\pi_1, \ldots, \pi_K\}$, where π_k is the probability that stimulus k will be probed [3]. The resulting distortion function is obtained by marginalizing over the probe stimulus:

$$d(\theta, \hat{\theta}) = \sum_{k} \pi_k d(\theta_k, \hat{\theta}_k). \tag{6}$$

Optimal population coding

We now consider how to realize the optimal channel with a population of spiking neurons, each tuned to a particular stimulus (Figure 1A). The firing rate of neuron i is determined by a simple Spike Response Model [28] in which the membrane potential is the difference between the excitatory input, u_i, and the inhibitory input, b, which we model as common across neurons (to keep notation simple, we will suppress the time index for all variables). Spiking is generated by a Poisson process, with firing rate modeled as an exponential function of the membrane potential [29]:

$$r_i = \exp[u_i - b]. \tag{7}$$

We assume that inhibition is strong enough such that only one neuron in the population will be active within an infinitesimally small time window. In particular, we assume that inhibition is given by $b = \log \sum_i \exp[u_i]$, in which case the firing rate is driven by the excitatory input with divisive normalization [30]:

$$r_i = \frac{\exp[u_i]}{\sum_j \exp[u_j]}. \tag{8}$$

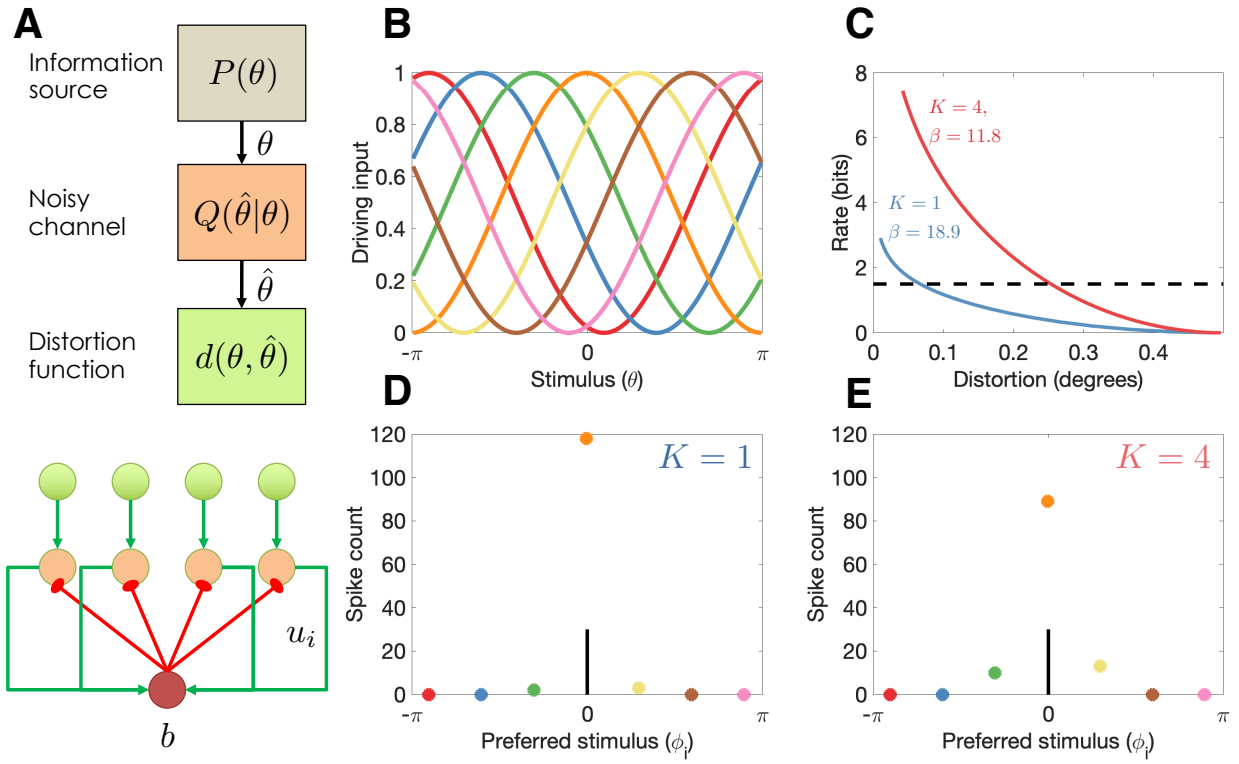

Figure 1: **Model illustration**. (A) Top: Abstract characterization of a communication channel. A stimulus θ is sampled from an information source $P(\theta)$ and passed through a noisy communication channel $Q(\hat{\theta}|\theta)$, which outputs a stimulus reconstruction $\hat{\theta}$. The reconstruction error is quantified by a distortion function, $d(\theta, \hat{\theta})$. Bottom: Circuit architecture implementing the communication channel. Input neurons encoding the negative distortion function provide the driving input to output neurons with membrane potential u_i and global feedback inhibition b. Each circuit codes a single stimulus at a fixed retinotopic location. When multiple stimuli are presented, the circuits operate in parallel, interacting only through a common gain parameter, β. (B) Tuning curves of input neurons encoding the negative cosine distortion function over a circular stimulus space. (C) Rate-distortion curves for two different set sizes ($K = 1$ and $K = 4$). The optimal gain parameter β is shown for each curve, corresponding to the point at which each curve intersects the channel capacity (horizontal dashed line). Expected distortion decreases with the information rate of the channel, but the channel capacity imposes a lower bound on expected distortion. (D) Example spike counts for output neurons in response to a stimulus ($\theta = 0$, vertical line). The output neurons are color-coded by their corresponding input neuron (arranged horizontally by their preferred stimulus, ϕ_i for neuron i; full tuning curves are shown in panel B). When only a single stimulus is presented ($K = 1$), the gain is high and the output neurons report the true stimulus with high precision. (E). When multiple stimuli are presented ($K = 4$), the gain is lower and the output has reduced precision (i.e., sometimes the wrong output neuron fires).

Note that allowing the feedback inhibition to be a function of the membrane potential b is biologically unrealistic, since the interneurons driving the inhibition do not have access to the membrane potential of other neurons. However, for our purposes it suffices to assume that the inhibitory population can approximate this function based on the spiking output of the excitatory neurons.

The resulting population dynamics is a form of "winner-take-all" circuit [31]. If each neuron has a preferred stimulus ϕ_i, then the winner can be understood as the momentary channel output, $\hat{\theta} = \phi_i$ whenever neuron i spikes (denoted $z_i = 1$). The probability that neuron i is the winner within a given infinitesimal time window is:

$$q(\hat{\theta} = \phi_i|\theta) = r_i. \tag{9}$$

Importantly, Equation 9 has the same functional form as Equation 4, and the two are equivalent if the excitatory input is given by:

$$u_i = -\beta d(\theta, \phi_i) + w_i, \tag{10}$$

where

$$w_i = \log \sum_{\theta} q(\hat{\theta} = \phi_i|\theta) P(\theta) \tag{11}$$

is the log marginal probability of neuron i being selected as the winner. We can see from this expression that the first term in Equation 10 corresponds to the neuron's stimulus-driven excitatory input and the second term corresponds to the neuron's excitability. The Lagrange multiplier β plays the role of a gain modulation factor.

The excitability term can be learned through a form of intrinsic plasticity [31], using the following spike-triggered update rule:

$$\Delta w_i = \eta \left(c \exp[-w_i] z_i - 1 \right), \tag{12}$$

where η is a learning rate and c a gain parameter. After a spike ($z_i = 1$), the excitability is increased proportionally to the inverse exponential of current excitability. In the absence of a spike, the excitability is decreased by a constant. This learning rule is broadly in agreement with experimental studies [32, 33].

We now address how to optimize β. We want the circuit to operate at the set point $R = C$, where the channel capacity C is understood as some fixed property of the circuit, whereas the information rate R can vary based on the parameters and input distribution, but cannot persistently exceed C. Assuming the total firing rate of the population is approximately constant across time, we can express the information rate as follows:

$$R = \mathbb{E}\left[\log \frac{Q(\hat{\theta}|\theta)}{\bar{Q}(\hat{\theta})} \right] = \sum_{i=1}^{N} \sum_{\theta} P(\theta) \mathbb{E}[\log r_i|\theta] - \mathbb{E}[\log r_i], \tag{13}$$

where N is the number of neurons. This expression reveals that channel capacity corresponds to a constraint on stimulus-driven deviations in firing rate from the marginal firing rate. When the stimulus-driven firing rate is persistently greater than the marginal firing rate, the population may incur an unsustainably large metabolic cost [34, 35]. When the stimulus-driven firing rate is lower than the marginal firing rate, the population is underutilizing its information transmission

resources. We can adapt the deviation through a form of homeostatic plasticity, by increasing β when the deviation is below the channel capacity, and decreasing β when the deviation is above the channel capacity. Concretely, a simple update rule implements this idea:

$$\Delta\beta = \alpha(C - R), \tag{14}$$

where α is a learning rate parameter. A similar adaptive gain modulation has been observed in neural circuits [36, 37, 38]. Mechanistically, this could be implemented by changes in background activity: when stimulus-driven excitation is high, the inhibition will also be high (the network is balanced), and the ensuing noise will effectively decrease the gain [39].

In the case where there are multiple stimuli, the same logic applies, but now we calculate the information rate over all the subpopulations of neurons (each coding a different stimulus). Specifically, the membrane potential becomes:

$$u_{ik} = -\beta\pi_k d(\theta_k, \phi_{ik}) + w_{ik}, \tag{15}$$

where k indexes both stimuli and separate subpopulations of neurons tuned to each stimulus location (or other stimulus feature that individuates the stimuli). As a consequence, β will tend to be smaller when more stimuli are encoded, because the same capacity constraint will be divided across more neurons.

Memory maintenance

In delayed response tasks, the stimulus is presented transiently, and then probed after a delay. The channel thus needs to maintain stimulus information across the delay. Our model assumes that the membrane potential u_i maintains a trace of the stimulus across the delay. The persistence of this trace is determined by the gain parameter β. Because persistently high levels of stimulus-evoked activity may, according to Equation 13, increase the information rate above the channel capacity, the learning rule in Equation 14 will reduce β and thereby functionally decay the memory trace.

The circuit model does not commit to a particular mechanism for maintaining the stimulus trace. A number of suitable mechanisms have been proposed [40]. One prominent model posits that recurrent connections between stimulus-tuned neurons can implement an attractor network that maintains the stimulus trace as a bump of activity [41, 42]. Other models propose cell-intrinsic mechanisms [43, 44] or short-term synaptic modifications [45, 46]. All of these model classes are potentially compatible with the theory that population codes are optimizing a rate-distortion trade-off, provided that the dynamics of the memory trace conform to the equations given above.

During time periods when no memory trace needs to be maintained, such as the intertrial interval (ITI) in delayed response tasks, we assume that the information rate is 0. Because the information rate is the *average* number of bits communicated across the channel, these "silent" periods effectively increase the achievable information rate during "active" periods (which we denote by R_A). Specifically, if T_A is the active time (delay period length), and T_S is the silent time (ITI length), then the channel's rate is given by:

$$R = \frac{T_A}{T_A + T_S} R_A. \tag{16}$$

Equivalently, we can ignore the intervals in our model and simply rescale the channel capacity by $(T_A + T_S)/T_A$. This will allow us to model the effects of delay and ITI on performance in working memory tasks.

Implications for working memory

Continuous report with circular stimuli

We apply the framework described above to the setting in which each stimulus is drawn from a circular space (e.g., color or orientation), $\theta_k \in (-\pi, \pi)$, which we discretize. Reconstruction errors are evaluated using a cosine distortion function:

$$d(\theta, \hat{\theta}) = -\omega \cos(\theta - \hat{\theta}), \tag{17}$$

where $\omega > 0$ is a scaling parameter. This implies that the input neurons have cosine tuning curves (Figure 1B). All of our subsequent simulations use the same tuning curves.

As an illustration of the model behavior in the continuous report task, we compare performance for set sizes 1 and 4. The optimal trade-off curves are shown in Figure 1C. For every point on the curve, the same information rate achieves a lower distortion for set size 1, due to the fact that all of the channel capacity can be devoted to a single stimulus (a hypothetical capacity limit is shown by the dashed horizontal line). In the circuit model, this higher performance is achieved by a narrow bump of population activity around the true stimulus (Figure 1D), compared to a broader bump when multiple stimuli are presented (Figure 1E).

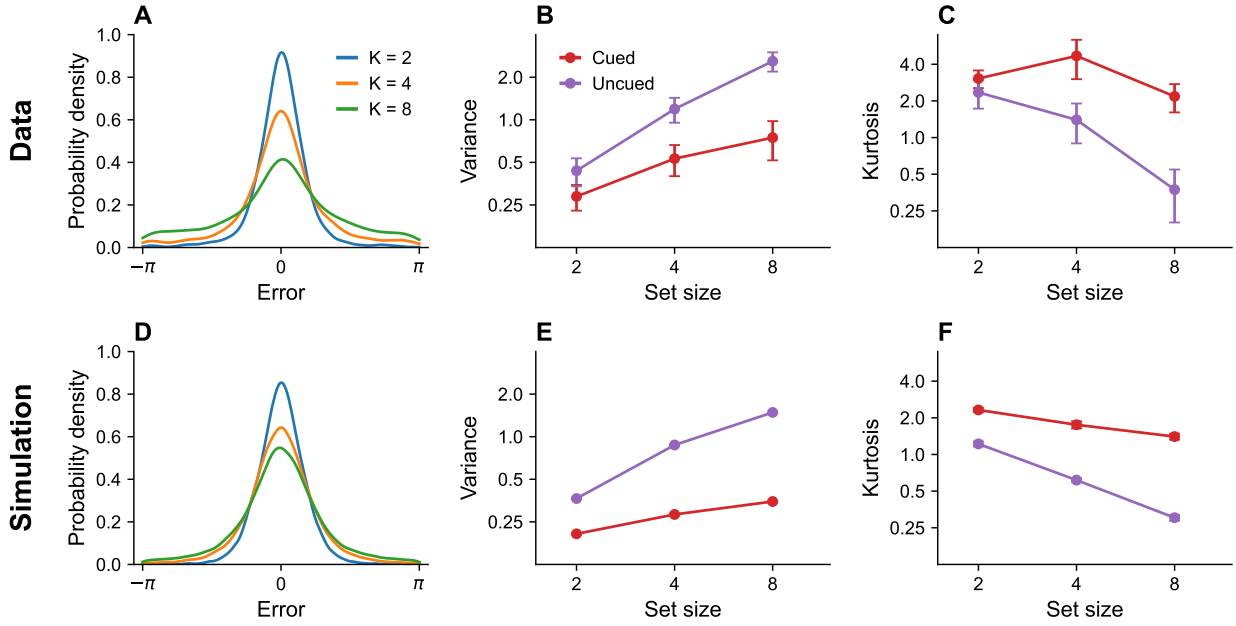

Figure 2: **Set size effects and prioritization**. (A) Error distributions for different set sizes, as reported in [15]. Error variability increases with set size. (B) Error variance as a function of set size for cued and uncued stimuli. Reports for cued stimuli have lower error variance. (C) Kurtosis as a function of set size for cued and uncued stimuli. (D, E, F) Simulation results replicating the observed effects. Error bars represent standard error of the mean.

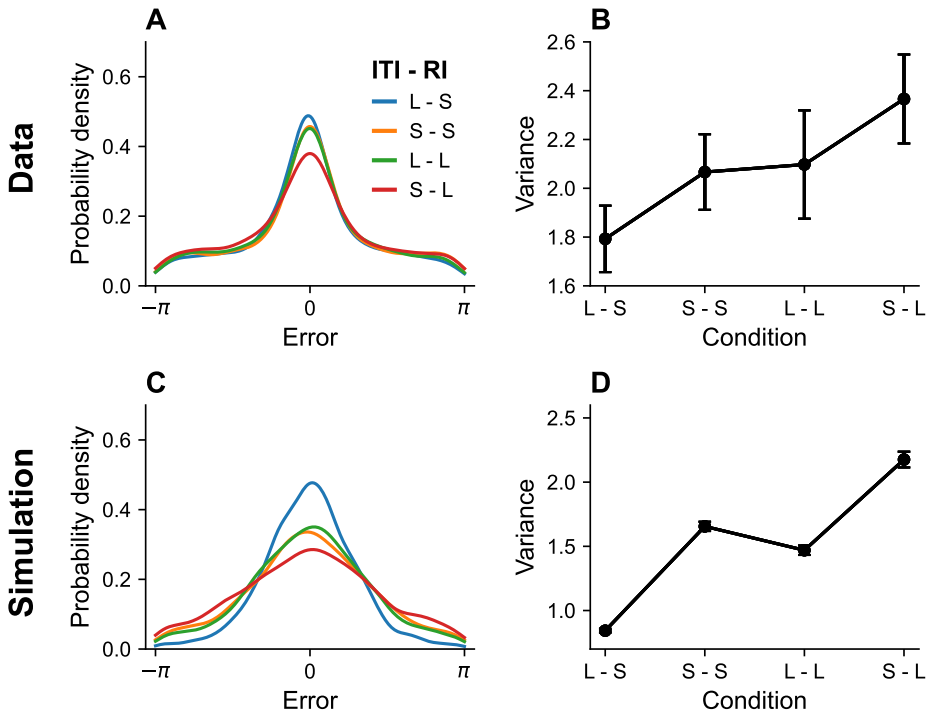

Figure 3: **Timing effects**. (A) Error distributions for different intertrial intervals (ITI) and retention intervals (RI), as reported in [26]. "S" denotes a short interval, and "L" denotes a long interval. (B) Error variance as a function of timing parameters. Longer ITIs are associated with lower error variance, whereas longer RIs are associated with larger error variance. (C, D) Simulation results replicating the observed effects. Error bars represent standard error of the mean.

Set size

One of the most fundamental findings in the visual working memory literature is that memory precision decreases with set size [47, 15, 48]. Our model asserts that this is the case because the capacity constraint of the system is divided across more neurons as the number of stimuli to be remembered increases, thus reducing the recall accuracy for any one stimulus. Figure 2A shows the distribution of recall error for different set sizes as published in previous work [15]. Figure 2D shows simulation results replicating these findings.

Prioritization

Stimuli that are attentionally prioritized are recalled more accurately. For example, error variance is reduced by a cue that probabilistically predicts the location of the probed stimulus [15, 49]. In our model, the cue is encoded by the probing probability π_k, which alters the expected distortion. This results in greater allocation of the capacity budget to cued stimuli than to uncued stimuli. Figure 2B and C show empirical findings, which are reproduced by our simulations shown in Figure 2E and F.

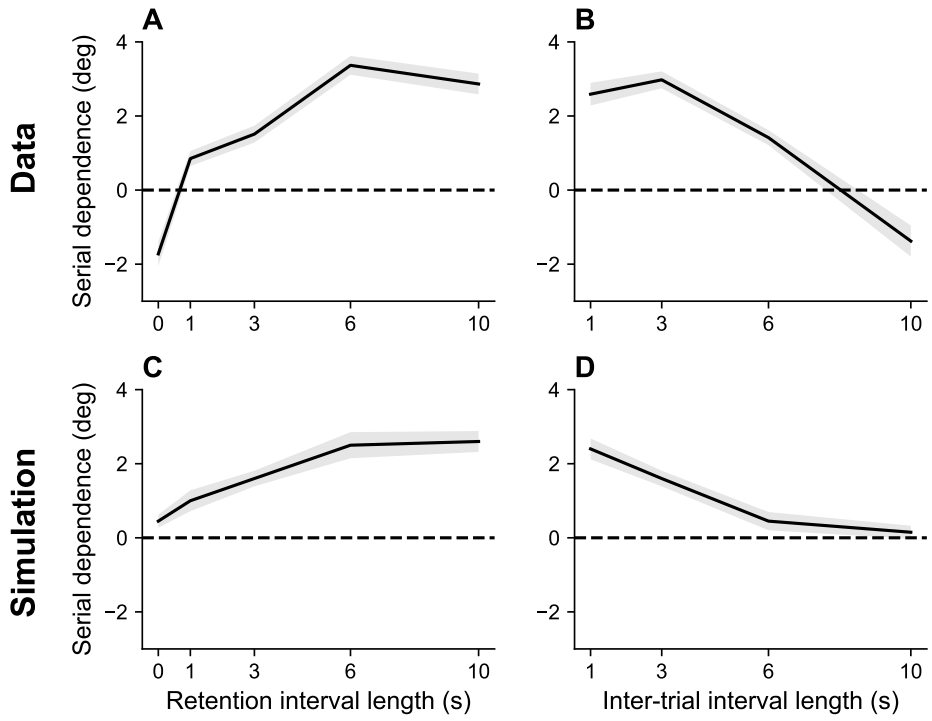

Figure 4: **Serial dependence as a function of retention interval and intertrial interval**. (A) Serial dependence increases with the retention interval until eventually reaching an asymptote, as reported in [27]. Serial dependence is quantified as the peak-to-peak amplitude of a derivative of Gaussian (DoG) tuning function fitted to the data using least squares (see Methods). (B) Serial dependence decreases with intertrial interval. (C,D) Simulation results. Shaded area corresponds to standard error of the mean.

Timing

It is well-established that memory performance typically degrades with the retention interval [50, 51, 18, 52], although the causes of this degradation are controversial [53], and in some cases the effect is unreliable [54]. According to our model, this occurs because long retention intervals tax the information rate of the neural circuit. In order to stay within the channel capacity, the circuit reduces the gain parameter β for long retention intervals, thereby reducing the information rate and degrading memory performance.

Memory performance also depends on the intertrial interval, but in the opposite direction: longer intertrial intervals improve performance [26, 25]. The critical determinant of performance is in fact the ratio between the intertrial and retention intervals. Souza and Oberauer [26] found that performance in a color working memory task was similar when both intervals were short or both intervals were long. They also reported that a *longer* retention interval could produce *better* memory performance when it is paired with a longer intertrial interval. Figure 3 shows a simulation of the same experimental paradigm, reproducing the key results. This timescale invariance, which is also seen in studies of associative learning [55], arises as a direct consequence of Equation 16. Increasing the intertrial interval reduces the information rate, since no stimuli are

being communicated during that time period, and can therefore compensate for longer retention intervals.

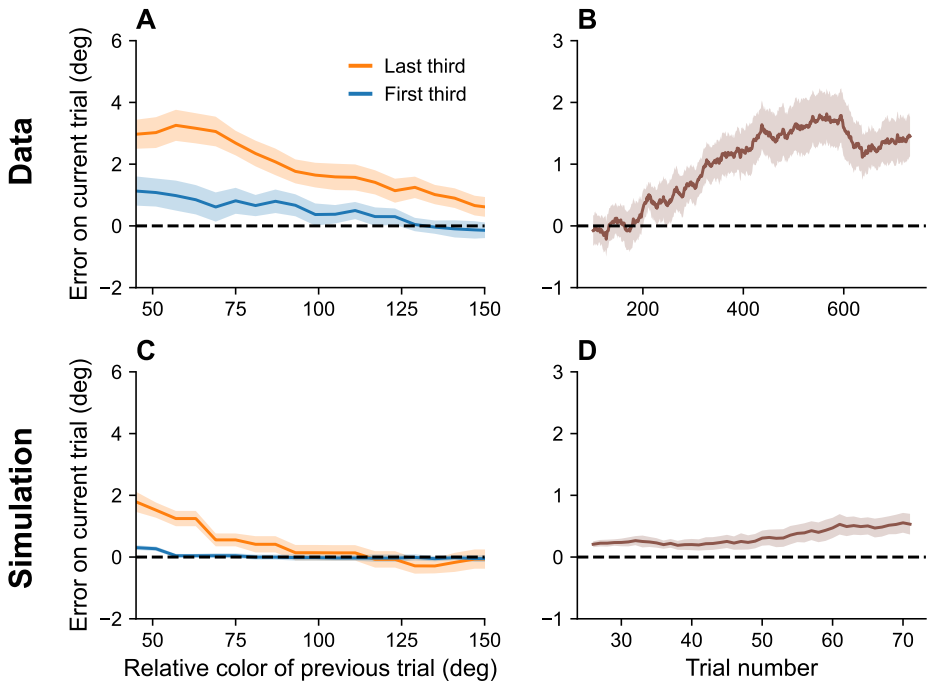

Figure 5: **Serial dependence builds up during an experiment**. (A) Serial dependence computed using first third (blue) and last third (orange) of the trials within a session, as reported in [56]. Data shown here were originally reported in [57]. To obtain a trial-by-trial measure of serial dependence, we calculated the folded error as described in [56] (see Methods). Positive values indicate attraction to the last stimulus, while negative values indicate repulsion. Serial dependence is stronger in the last third of the trials in the experiment compared to the first third. (B) Serial dependence increases over the course of the experimental session, computed here with a sliding window of 200 trials. (C, D) Simulation results. Shaded area corresponds to standard error of the mean.

Serial dependence

Working memory recall is biased by recent stimuli, a phenomenon known as *serial dependence* [58, 59, 27, 60]. Recall is generally attracted toward recent stimuli, though some studies have reported repulsive effects when the most recent and current stimulus differ by a large amount [61, 27]. Our theory explains serial dependence as a consequence of the marginal firing rate of the output cells, which biases the excitatory input u_i (see Equation 10). Because the marginal firing rate is updated incrementally, it will reflect recent stimulus history.

An important benchmark for theories of serial dependence is the finding that it increases with the retention interval and decreases with intertrial interval [27]. These twin dependencies are reproduced by our model (Figure 4). Our explanation of serial dependence is closely related to our explanation of timing effects on recall error: the strength of serial dependence varies inversely

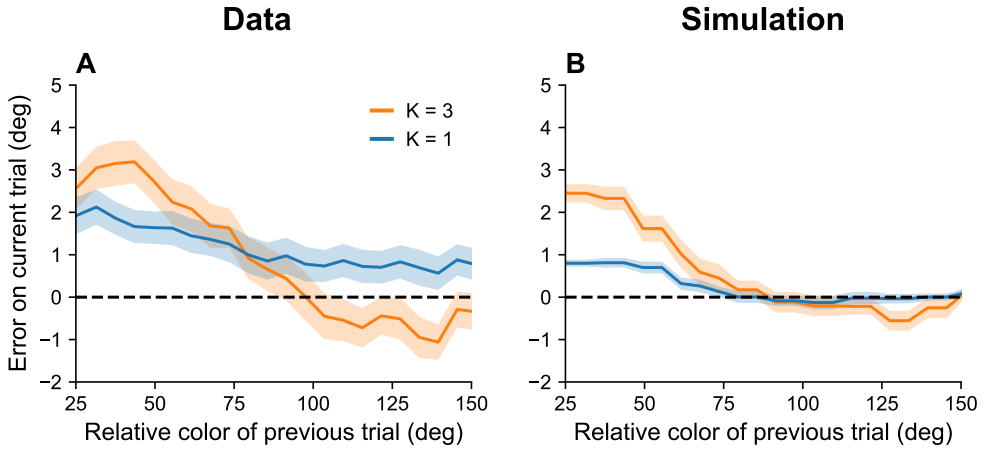

Figure 6: **Serial dependence increases with set size**. (A) Serial dependence (quantified using folded error) for set sizes $K = 1$ (blue) and $K = 3$ (orange), using data originally reported in [51]. Serial dependence computed as the peak amplitude of a DoG tuning function fitted to the data using least squares is stronger for larger set sizes (see Methods). (B) Simulation results. Shaded area corresponds to standard error of the mean.

with the information rate, which in turn increases with the intertrial interval and decreases with the retention interval. Mechanistically, this effect is mediated by adjustments of the gain parameter β in order to keep the information rate near the channel capacity.

Serial dependence has also been shown to build up over the course of an experimental session [56]. This is hard to explain in terms of theories based on purely short-term effects, but it is consistent with our account in terms of the bias induced by the marginal firing rate. Because this bias reflects continuous incremental adjustments, it integrates over the entire stimulus history, thereby building up over the course of an experimental session.

If, as we hypothesize, serial dependence reflects a capacity limit, then we should expect it to increase with set size, since β must decrease to stay within the capacity limit. To the best of our knowledge, this prediction has not been tested. We confirmed this prediction for color working memory using a large dataset reported in [51]. Figure 6 shows that the attractive bias for similar stimuli on consecutive trials is stronger when the set size is larger ($p < 0.05$, group permutation test).

Systematic biases

Working memory exhibits systematic biases towards stimuli that are shown more frequently than others [51]. Moreover, these biases increase with the retention interval, and build up over the course of an experimental session. Our interpretation of serial dependence, which also builds up over the course of a session, suggests that these two phenomena may be linked (see also [62]).

Our theory posits that, over the course of the experiment, the marginal firing rate asymptotically approaches the distribution of presented stimuli (assuming there are no inhomogeneities in the distortion function). Thus, the neurons corresponding to high-frequency stimuli become more excitable than others and bias recall towards their preferred stimuli. This bias is amplified by

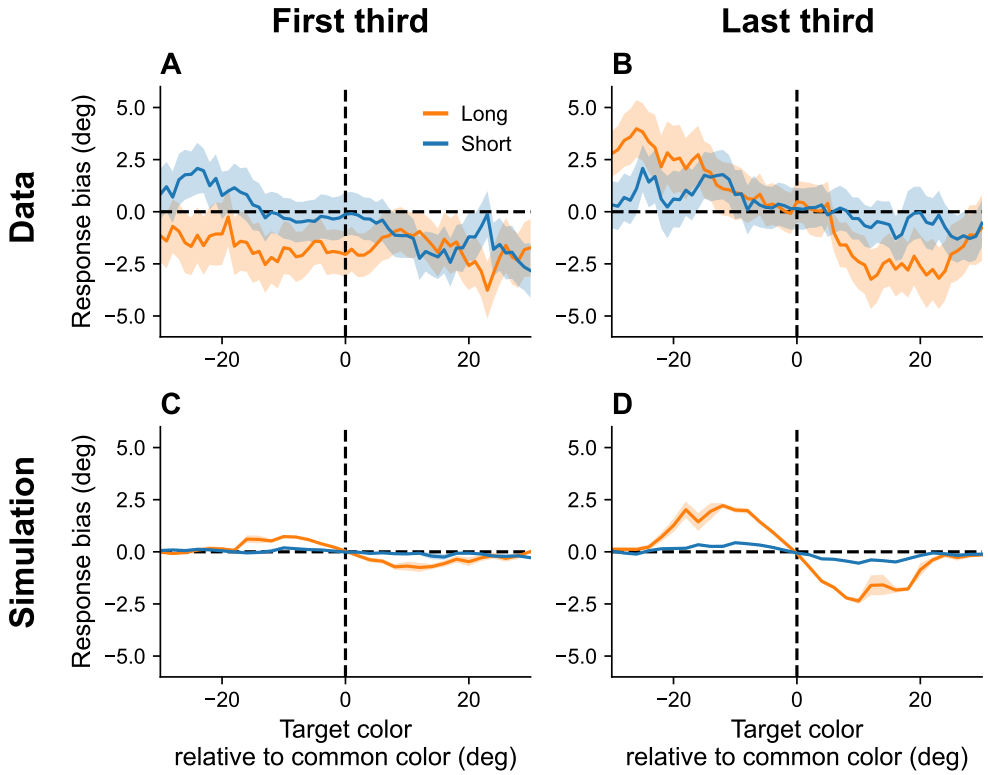

Figure 7: **Continuous reports are biased towards high frequency colors**. (A, B) Bias for targets around common colors during the first (Panel A) and last (Panel B) third of the session, as reported in [51]. Bias refers to the difference between the stimulus and the mean reported color. X axis is centered around high-frequency colors. Bias increases with RI length (blue = short RI, orange = long RI). Bias also increases as the experiment progresses. (C, D) Simulation results. Shaded area corresponds to standard error of the mean.

lower effective capacities brought about by longer retention intervals. Figure 7 shows simulation results replicating these effects.

Variations in gain

Equation 14 predicts that operating below the channel capacity will lead to an increase in the gain term β, which, in turn, leads to a higher information rate and better memory performance. Therefore, our model predicts that recall accuracy should improve after a period of poor memory performance, and degrade after a period of good memory performance. At the neural level, the model predicts that error will tend to be lower when gain (β) is higher.

We tested these predictions by re-analyzing the monkey neural and behavioral data reported in [61] ($N = 2$). Squared error was significantly lower following higher-than-average error than following lower-than-average error (linear mixed model, $p < 0.001$; Figure 8A).

In order to estimate the neural gain, we first inferred the preferred stimulus of each neuron by fitting a bell-shaped tuning function to its spiking behavior (Equation 21, Figure 8B). We then per-

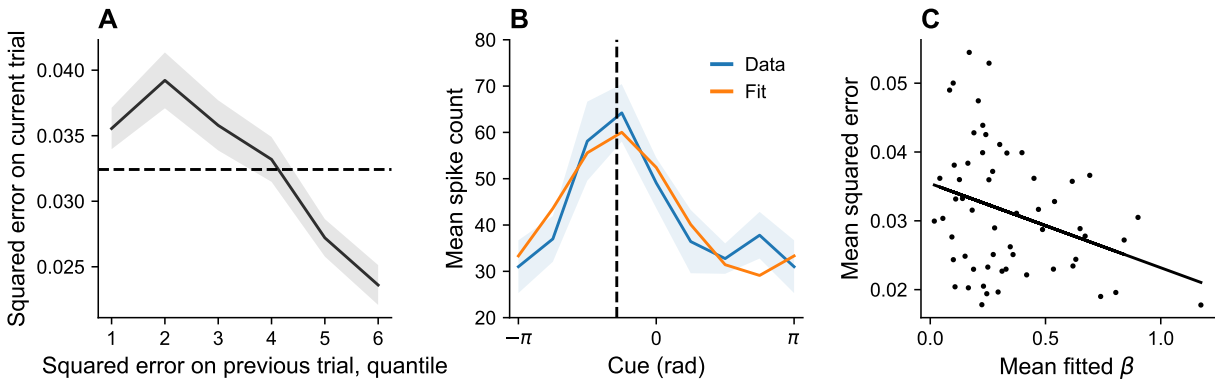

Figure 8: **Dynamic variation in memory precision and neural gain**. (A) Mean squared error on current trial, classified by quantiles of squared error on previous trial. Squared error tends to be above average (dashed black line) following low squared error on the previous trial, and tends to be below average following large squared error on the previous trial. (B) Orientation tuning curve (orange) fitted to mean spike count (blue) during the retention interval, shown for one example neuron. The neuron's preferred stimulus (dashed black line) corresponds to the peak of the tuning curve. Shaded region corresponds to standard error of the mean. (C) Mean squared error for different sessions plotted against mean fitted β. According to our theory, β plays the role of a gain control on the stimulus. Consistent with this hypothesis, memory error decreases with β.

formed Poisson regression to fit a β for each neuron (Equation 22). Model comparison using the Bayesian information criterion (BIC) established that both the distortion function (which captures driving input) and spiking history were significant predictors of spiking behavior (full model: $54,545$; no history: $59,163$; neither distortion nor history: $67,903$). We then examined the relationship between neural gain and memory precision across sessions, finding that session-specific mean squared error was negatively correlated with the average β estimate ($r = -0.32$, $p < 0.02$; Figure 8C).

Discussion

We have shown that a simple population coding model with spiking neurons can solve the channel design problem: signals passed through the spiking network are transmitted with close to the minimum achievable distortion under the network's capacity limit. We focused on applying this general model to the domain of working memory, unifying several seemingly disparate aspects of working memory performance: set size effects, stimulus prioritization, serial dependence, approximate timescale invariance, and systematic bias. Our approach builds a bridge between biologically plausible population coding and prior applications of rate-distortion theory to human memory [1, 2, 9, 4, 5, 10].

Relationship to other models

The hypothesis that neural systems are designed to optimize a rate-distortion trade-off has been previously studied through the lens of the information bottleneck method [63, 64, 65, 66], a special case of rate-distortion theory in which the distortion function is derived from a compression principle. Specifically, the distortion function is defined as the Kullback-Leibler divergence between $P(\theta'|\theta)$ and $P(\theta'|\hat{\theta})$, where θ' denotes the probed stimulus. This distortion function applies a "soft" penalty to errors based on how much probability mass the channel places on each stimulus. The expected distortion is equal to the mutual information between θ' and $\hat{\theta}$. Thus, the information bottleneck method seeks a channel that maps the input θ into a compressed representation $\hat{\theta}$ satisfying the capacity limit, while preserving information necessary to predict the probe θ'.

As pointed out by Leibfried and Braun [67], using the Kullback-Leibler divergence as the distortion function leads to a harder optimization compared to classical rate-distortion theory because $P(\theta'|\hat{\theta})$ depends on the channel distribution, which is the thing being optimized. One consequence of this dependency is that minimizing the rate-distortion objective using alternating optimization (in the style of the Blahut-Arimoto algorithm) is not guaranteed to find the globally optimal channel. It is possible to break the dependency by replacing $P(\theta'|\hat{\theta})$ with a reference distribution that does not depend on the channel. This turns out to strictly generalize rate-distortion theory, because an arbitrary choice of the reference distribution allows one to recover any lower-bounded distortion function up to a constant offset [67]. However, existing spiking neuron implementations of the information bottleneck mthod [64, 65] do not make use of such a reference distribution, and hence do not attain the same level of generality.

Leibfried and Braun [67] propose a spiking neuron model that explicitly optimizes the rate-distortion objective function for arbitrary distortion functions. Their approach differs from ours in several ways. First, they model a single neuron, rather than a population. Second, they posit that the channel optimization is realized through synaptic plasticity, in contrast to the intrinsic plasticity rule that we study here. Third, they treat the gain parameter β as fixed, whereas we propose an algorithm for optimizing β.

Open questions

A cornerstone of our approach is the assumption that the neural circuit responsible for working memory dynamically modifies its output to stay within a capacity limit. What, at a biological level, is the nature of this capacity limit? Spiking activity accounts for a large fraction of cortical energy expenditure [68, 69]. Thus, a limit on the overall firing rate of a neural population is a natural transmission bottleneck. Previous work on energy-efficient coding has similarly used the cost of spiking as a constraint [34, 70, 71]. One subtlety is that the capacity limit in our framework is an upper bound on the stimulus-driven firing rate *relative* to the average firing rate (on a log scale). This means that the average firing rate can be high provided the stimulus-evoked transients are small, consistent with the observation that firing rate tends to be maintained around a set point rather than minimized [36, 37, 38]. The set point should correspond to the capacity limit.

The next question is how a neural circuit can control its sensitivity to inputs in such a way that the information rate is maintained around the capacity limit. At the single neuron level, this might be realized by adaptation of voltage conductances [70]. At the population level, neuromodulators could act as a global gain control. Catecholamines (e.g., dopamine and norepinephrine), in

particular, have been thought to play this role [72, 73]. Directly relevant to this hypothesis are experiments showing that local injection of dopamine receptor antagonists into the prefrontal cortex impaired performance in an oculomotor delayed response task [74].

Our model can be extended in several ways. One, as already mentioned, is to develop a biologically plausible implementation of gain adaptation, either through intrinsic or neuromodulatory mechanisms. A second direction is to consider channels that transmit a compressed representation of the input. Previous work has suggested that working memory representations are efficient codes that encode some stimuli with higher precision than others [75, 20]. Finally, an important direction is to enable the model to handle more complex memoranda, such as natural images. Recent applications of large-scale neural networks, such as the variational autoencoder, to modeling human memory hold promise [5, 76], though linking these to more realistic neural circuits remains a challenge.

Methods

We re-analyzed 6 datasets of monkey and human subjects performing a delayed response task. The detailed experimental procedures can be found in the original reports [15, 26, 61, 56, 51, 46]. In 3 of the 6 datasets, one or multiple colors were presented on a screen at equally spaced locations. After a retention interval, during which the cues were no longer visible, subjects had to report the color at a particular cued location, measured as angles on a color wheel. In one dataset, angled color bars were presented, and the angle of the bar associated with a cued color had to be reported [15]. In the two last datasets, only the location of a black cue on a circle had to be remembered and reported [61, 46].

Set size and stimulus prioritization

Human subjects ($N = 7$) were presented with 2, 4 or 8 color stimuli at the same time. On each trial, one of the locations was cued before the appearance of the stimuli. Cued locations were 3 times as likely to be probed [15].

We computed trial-wise error as the circular distance between the reported angle and the target angle, separately for each set size and cuing condition. We then calculated circular variance and kurtosis as presented in the original paper, using the following equations:

$$\sigma^2 = -2 \log |\bar{m}_1|, \tag{18}$$

and

$$k = (|\bar{m}_2| \cos(\text{Arg}(\bar{m}_2) - 2\text{Arg}(\bar{m}_1)) - |\bar{m}_1|^4)/(1 - |\bar{m}_1|)^2, \tag{19}$$

where $\bar{m}_n$ is the nth uncentered trigonometric moment.

Timing effects

Human subjects ($N = 36$) were presented with 6 simultaneous color stimuli and had to report the color at a probed location as an angle on a color wheel. The RI and ITI lengths varied across sessions (RI: 1 or 3 seconds, ITI: 1 or 7.5 seconds) [26].

Serial dependence increases with retention interval and decreases with intertrial interval

Human subjects ($N = 55$) were presented with a black square at a random position on a circle and had to report the location of the cue [46]. The RI and ITI were varied across blocks of trials (RI: 0, 1, 3, 6, or 10 seconds, ITI: 1, 3, 6 or 10 seconds). For each block and subject, we computed serial dependence as the peak-to-peak amplitude of a derivative of Gaussian (DoG) function fit to the data. The DoG function is defined as follows:

$$y = xawc \exp(-(wx)^2), \tag{20}$$

where y is the trial-wise error, x is the relative circular distance to the target angle of the previous trial, a is the amplitude of the DoG peak, w is the width of the curve, and c is the constant $\sqrt{2e}$, chosen such that the peak-to-peak amplitude of the DoG fit—the measure of serial dependence in [46]—is exactly $2a$.

Build-up of serial dependence

Human subjects ($N = 12$) performed a delayed continuous report task with one item [57]. Following [56], we obtained a trial-by-trial measure of serial dependence using their definition of folded error.

Let θ_d denote the circular distance between the angle reported on the previous trial and the target angle on the current trial. In order to aggregate trials with negative θ_d (preceding target is located clockwise to current target) and trials with positive θ_d (preceding target is located counterclockwise to current target), we computed the folded error as $\theta'_e = \theta_e \times \text{sign}(\theta_d)$, where θ_e is the circular distance between the reported angle and the target angle. Positive θ'_e corresponds to attraction to the previous stimulus, whereas negative θ'_e corresponds to repulsion.

We excluded trials with absolute errors larger than $\pi/4$. We then computed serial bias as the average folded error in sliding windows of width $\pi/2$ rad and steps of $\pi/30$ rad. We repeated this procedure separately for the trials contained in the first and last third of all sessions. Finally, we computed the increase in serial dependence over the course of a session using a sliding window of 200 trials on the folded error.

Serial dependence increases with set size

We re-analyzed the dataset collected by [51], experiment 1a, in which human subjects ($N = 90$) performed a delayed response task with 1 or 3 items.

We calculated folded error using the procedure mentioned above. We excluded trials with absolute errors larger than $\pi/4$. We then computed serial bias as the average folded error in sliding windows of width $\pi/4$ rad and steps of $\pi/30$ rad. We repeated this procedure separately for the trials with $K = 1$ or $K = 3$ items. In order to test whether serial dependence was stronger for one of the set size conditions, we performed a permutation test: We shuffled the entire dataset and partitioned it into two groups of size $S_{K=1}$ and $S_{K=3}$, where $S_{K=k}$ denotes the number of trials recorded for the set size condition $K = k$. We fitted a DoG curve (Equation 20) to each partition using least squares and computed the difference between the peak amplitude of the two fits. We repeated this process $20,000$ times. We then calculated the p-value as the proportion of shuffles for which the difference between the peak amplitudes was equal to or larger than the one computed using the unshuffled dataset.

Continuous reports are biased towards high frequency colors

Human subjects ($N = 120$) performed a delayed continuous report task with a set size of 2 [51]. On each trial, the RI was either 0.5 or 4 seconds. The stimuli were either drawn from a uniform distribution or from a set of 4 equally-spaced bumps of width $\pi/9$ rad with equal probability. The centers of each bump were held constant for each subject.

We defined systematic bias as mean error versus distance to the closest bump center and computed it in sliding windows of width $\pi/45$ rad and steps of $\pi/90$ rad, as done in the original study. We repeated this procedure separately for the trials with $RI = 0.5s$ or $RI = 4s$, and for the first and last third of trials within a session.

Simulation parameters

We used the following parameters for all simulations, unless stated otherwise: $N = 1000$, $K = 1$, $C = 0.9$, $\omega = 1$, $\eta = 3 \times 10^{-6}$, $c = 0.5$, $\alpha = 10^{-3}$, $\bar{r} = 20$ Hz, $\Delta t = 1$ ms. Spikes contributed to intrinsic synaptic plasticity for 10 timesteps. Weights w were clipped to be in the range $[-12, 0]$. β was initialized at $\beta_0 = 20$ and clipped to be in the range $[0, 200]$. C was the only free parameter we allowed to vary across simulations, and we chose $C = 0.1$ to fit the data in [26].

In order to account for the higher probing probability of the cued stimulus in [15], we used $\pi_k = \frac{\alpha_k}{\sum_{k'} \alpha_{k'}}$ with $\alpha_{\text{priority}} = \{1.5_{K=2}, 2.2_{K=4}, 3.0_{K=8}\}$ and $\alpha_k = 1$ otherwise, as reported in [15].

Dynamics of memory precision and neural gain

We re-analyzed the behavioral and neural dataset collected in [61]. Since neural recordings were not available for all trials within a session, we ignored sessions in which only a subset of the 8 potential cues were displayed.

We sorted the squared error on trial t (denoted by e_t^2) based on 6 quantiles of the squared error on the previous trial. We then defined the indicator variable $i_t = \mathcal{I}(e_{t-1}^2 > \bar{e^2})$, taking the value $+1$ if the squared error on the previous trial was larger than the mean squared error, and -1 otherwise. We then fit the linear mixed model $e_t^2 \sim 1 + i_t + (1|\text{session})$.

In order to infer the preferred stimulus of each recorded neuron, we used a least squares approach to fit the mean spike count for each presented stimulus and neuron to a bell-shaped tuning function:

$$f_i(\theta) = A_i \exp(w_i^{-1}(\cos(\theta - \phi_i) - 1)), \tag{21}$$

where θ is the presented stimulus, A_i and w_i control the amplitude and width of the tuning function, respectively, and ϕ_i is the preferred stimulus of neuron i [15].

We then fitted the neural data by performing Poisson regression for each neuron using the following model:

$$\log(s_j) \sim 1 + D_j + \bar{s}_j, \tag{22}$$

where s_j is the number of spikes emitted by the neuron on trial j, D_j is the expected distortion between the stimulus θ_j and the neuron's preferred stimulus, and $\bar{s}_j$ is an exponential moving average of the neuron's spike history with decay rate 0.8. We discarded 3 neurons for which the fitted β was negative and 1 neuron for which the fitted β was larger than 5 standard deviations

above the mean of the fitted values. In order to ascertain the utility of the different regressors, we fitted another model without the history term, and another without both the distortion and history terms, and compared them based on their Bayesian Information Criterion (BIC).

Source code

All simulations and analyses were performed using Julia, version 1.6.2. Source code can be found at `https://github.com/amvjakob/wm-rate-distortion`.

Acknowledgments

Johannes Bill and Chris Bates generously provided constructive feedback and discussion. We are also grateful to Dan Bliss and Matt Panichello for sharing data. This research was supported by a Bertarelli Fellowship and by the Center for Brains, Minds, and Machines (funded by NSF STC award CCF-1231216).

References

[1] Chris Sims, Robert Jacobs, and David Knill. An ideal observer analysis of visual working memory. *Psychological Review*, 119:807–830, 2012.

[2] Chris R Sims. The cost of misremembering: Inferring the loss function in visual working memory. *Journal of Vision*, 15:2–2, 2015.

[3] Ronald Van den Berg and Wei Ji Ma. A resource-rational theory of set size effects in human visual working memory. *ELife*, 7:e34963, 2018.

[4] Christopher J Bates, Rachel A Lerch, Chris R Sims, and Robert A Jacobs. Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, 19:11–11, 2019.

[5] Christopher J Bates and Robert A Jacobs. Efficient data compression in perception and perceptual memory. *Psychological Review*, 127:891–917, 2020.

[6] Timothy F Brady, Talia Konkle, and George A Alvarez. Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, 138:487–502, 2009.

[7] MR Nassar, JC Helmers, and MJ Frank. Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review*, 125:486–511, 2018.

[8] Claude E Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 7:325–350, 1959.

[9] Chris R Sims. Rate-distortion theory and human perception. *Cognition*, 152:181–198, 2016.

[10] David G Nagy, Balázs Török, and Gergő Orbán. Optimal forgetting: Semantic compression of episodic memories. *PLoS Computational Biology*, 16:e1008367, 2020.

[11] Chris R Sims. Efficient coding explains the universal law of generalization in human perception. *Science*, 360:652–656, 2018.

[12] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115:7937–7942, 2018.

[13] Samuel J Gershman. Origin of perseveration in the trade-off between reward and complexity. *Cognition*, 204:104394, 2020.

[14] Lucy Lai and Samuel J Gershman. Policy compression: An information bottleneck in action selection. In *Psychology of Learning and Motivation*, volume 74, pages 195–232. Elsevier, 2021.

[15] Paul M Bays. Noise in neural populations accounts for errors in working memory. *Journal of Neuroscience*, 34:3632–3645, 2014.

[16] Paul M Bays. Spikes not slots: noise in neural populations limits working memory. *Trends in Cognitive Sciences*, 19:431–438, 2015.

[17] Paul M Bays. A signature of neural coding at human perceptual limits. *Journal of Vision*, 16:4–4, 2016.

[18] Sebastian Schneegans and Paul M Bays. Drift in neural population activity causes working memory to deteriorate over time. *Journal of Neuroscience*, 38:4859–4869, 2018.

[19] Sebastian Schneegans, Robert Taylor, and Paul M Bays. Stochastic sampling provides a unifying account of visual working memory limits. *Proceedings of the National Academy of Sciences*, 117:20959–20968, 2020.

[20] Robert Taylor and Paul M Bays. Efficient coding in visual working memory accounts for stimulus-specific variations in recall. *Journal of Neuroscience*, 38:7132–7142, 2018.

[21] Ivan Tomić and Paul M Bays. Internal but not external noise frees working memory resources. *PLoS Computational Biology*, 14:e1006488, 2018.

[22] Richard Blahut. Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, 18:460–473, 1972.

[23] Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18:14–20, 1972.

[24] Anastasia Kiyonaga, Jason M Scimeca, Daniel P Bliss, and David Whitney. Serial dependence across perception, attention, and memory. *Trends in Cognitive Sciences*, 21:493–497, 2017.

[25] Zach Shipstead and Randall Engle. Interference within the focus of attention: Working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39:277–289, 2013.

[26] Alessandra S Souza and Klaus Oberauer. Time-based forgetting in visual working memory reflects temporal distinctiveness, not decay. *Psychonomic Bulletin & Review*, 22:156–162, 2015.

[27] Daniel P Bliss, Jerome J Sun, and Mark D'Esposito. Serial dependence is absent at the time of perception but increases in visual working memory. *Scientific Reports*, 7:1–13, 2017.

[28] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge University Press, 2002.

[29] Renaud Jolivet, Alexander Rauch, Hans-Rudolf Lüscher, and Wulfram Gerstner. Predicting spike timing of neocortical pyramidal neurons by simple threshold models. *Journal of Computational Neuroscience*, 21:35–49, 2006.

[30] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13:51–62, 2012.

[31] Bernhard Nessler, Michael Pfeiffer, Lars Buesing, and Wolfgang Maass. Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *PLoS Computational Biology*, 9:e1003037, 2013.

[32] Gaël Daoudal and Dominique Debanne. Long-term plasticity of intrinsic excitability: learning rules and mechanisms. *Learning & Memory*, 10:456–465, 2003.

[33] Robert H Cudmore and Gina G Turrigiano. Long-term potentiation of intrinsic excitability in lv visual cortical neurons. *Journal of Neurophysiology*, 92:341–348, 2004.

[34] William B Levy and Robert A Baxter. Energy efficient neural codes. *Neural Computation*, 8:531–543, 1996.

[35] Simon B Laughlin, Rob R de Ruyter van Steveninck, and John C Anderson. The metabolic cost of neural information. *Nature Neuroscience*, 1:36–41, 1998.

[36] Niraj S Desai, Lana C Rutherford, and Gina G Turrigiano. Plasticity in the intrinsic excitability of cortical pyramidal neurons. *Nature Neuroscience*, 2:515–520, 1999.

[37] Keith B Hengen, Mary E Lambo, Stephen D Van Hooser, Donald B Katz, and Gina G Turrigiano. Firing rate homeostasis in visual cortex of freely behaving rodents. *Neuron*, 80:335–342, 2013.

[38] Keith B Hengen, Alejandro Torrado Pacheco, James N McGregor, Stephen D Van Hooser, and Gina G Turrigiano. Neuronal firing rate homeostasis is inhibited by sleep and promoted by wake. *Cell*, 165:180–191, 2016.

[39] Frances S Chance, Larry F Abbott, and Alex D Reyes. Gain modulation from background synaptic input. *Neuron*, 35:773–782, 2002.

[40] Joel Zylberberg and Ben W Strowbridge. Mechanisms of persistent activity in cortical circuits: possible neural substrates for working memory. *Annual Review of Neuroscience*, 40:603–627, 2017.

[41] Xiao-Jing Wang. Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences*, 24:455–463, 2001.

[42] Daniel J Amit and Nicolas Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral cortex*, 7:237–252, 1997.

[43] Alexei V Egorov, Bassam N Hamam, Erik Fransén, Michael E Hasselmo, and Angel A Alonso. Graded persistent activity in entorhinal cortex neurons. *Nature*, 420:173–178, 2002.

[44] D Durstewitz and JK Seamans. Beyond bistability: biophysics and temporal dynamics of working memory. *Neuroscience*, 139:119–133, 2006.

[45] Gianluigi Mongillo, Omri Barak, and Misha Tsodyks. Synaptic theory of working memory. *Science*, 319:1543–1546, 2008.

[46] Daniel P Bliss and Mark D'Esposito. Synaptic augmentation in a cortical circuit model reproduces serial dependence in visual working memory. *PLoS One*, 12:e0188927, 2017.

[47] Paul M Bays, Raquel FG Catalao, and Masud Husain. The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, 9:7–7, 2009.

[48] Patrick Wilken and Wei Ji Ma. A detection theory account of change detection. *Journal of Vision*, 4:11–11, 2004.

[49] Aspen H Yoo, Zuzanna Klyszejko, Clayton E Curtis, and Wei Ji Ma. Strategic allocation of working memory resource. *Scientific Reports*, 8:1–8, 2018.

[50] Yoni Pertzov, Sanjay Manohar, and Masud Husain. Rapid forgetting results from competition over time between items in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43:528–536, 2017.

[51] Matthew F Panichello, Brian DePasquale, Jonathan W Pillow, and Timothy J Buschman. Error-correcting dynamics in visual working memory. *Nature Communications*, 10:1–11, 2019.

[52] Weiwei Zhang and Steven J Luck. Sudden death and gradual decay in visual working memory. *Psychological Science*, 20:423–428, 2009.

[53] Klaus Oberauer, Simon Farrell, Christopher Jarrold, and Stephan Lewandowsky. What limits working memory capacity? *Psychological Bulletin*, 142:758–799, 2016.

[54] Hongsup Shin, Qijia Zou, and Wei Ji Ma. The effects of delay duration on visual working memory for orientation. *Journal of Vision*, 17:10–10, 2017.

[55] Peter D Balsam and C Randy Gallistel. Temporal maps and informativeness in associative learning. *Trends in Neurosciences*, 32:73–78, 2009.

[56] João Barbosa and Albert Compte. Build-up of serial dependence in color working memory. *Scientific reports*, 10(1):1–7, 2020.

[57] Joshua J Foster, Emma M Bsales, Russell J Jaffe, and Edward Awh. Alpha-band activity reveals spontaneous representations of spatial position in visual working memory. *Current Biology*, 27(20):3216–3223, 2017.

[58] Jason Fischer and David Whitney. Serial dependence in visual perception. *Nature Neuroscience*, 17:738–743, 2014.

[59] Matthias Fritsche, Pim Mostert, and Floris P de Lange. Opposite effects of recent history on perception and decision. *Current Biology*, 27:590–595, 2017.

[60] Charalampos Papadimitriou, Afreen Ferdoash, and Lawrence H Snyder. Ghosts in the machine: memory interference from the previous trial. *Journal of Neurophysiology*, 113:567–577, 2015.

[61] Joao Barbosa, Heike Stein, Rebecca L Martinez, Adrià Galan-Gadea, Sihai Li, Josep Dalmau, Kirsten CS Adam, Josep Valls-Solé, Christos Constantinidis, and Albert Compte. Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nature neuroscience*, 23(8):1016–1024, 2020.

[62] Ke Tong and Chad Dubé. A tale of two literatures: A fidelity-based integration account of central tendency bias and serial dependency. *Computational Brain & Behavior*, pages 1–21, 2022.

[63] William Bialek, Rob R De Ruyter Van Steveninck, and Naftali Tishby. Efficient representation as a design principle for neural coding and computation. In *2006 IEEE International Symposium on Information Theory*, pages 659–663. IEEE, 2006.

[64] Stefan Klampfl, Robert Legenstein, and Wolfgang Maass. Spiking neurons can learn to solve information bottleneck problems and extract independent components. *Neural Computation*, 21:911–959, 2009.

[65] Lars Buesing and Wolfgang Maass. A spiking neuron as information bottleneck. *Neural Computation*, 22:1961–1992, 2010.

[66] Stephanie E Palmer, Olivier Marre, Michael J Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112:6908–6913, 2015.

[67] Felix Leibfried and Daniel A Braun. A reward-maximizing spiking neuron as a bounded rational decision maker. *Neural Computation*, 27:1686–1720, 2015.

[68] David Attwell and Simon B Laughlin. An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, 21:1133–1145, 2001.

[69] Peter Lennie. The cost of cortical computation. *Current Biology*, 13:493–497, 2003.

[70] Martin Stemmler and Christof Koch. How voltage-dependent conductances can adapt to maximize the information encoded by neuronal firing rate. *Nature Neuroscience*, 2:521–527, 1999.

[71] Vijay Balasubramanian, Don Kimber, and Michael J Berry II. Metabolically efficient information processing. *Neural Computation*, 13:799–815, 2001.

[72] David Servan-Schreiber, Harry Printz, and Jonathan D Cohen. A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. *Science*, 249:892–895, 1990.

[73] Daniel Durstewitz, Marian Kelc, and Onur Güntürkün. A neurocomputational theory of the dopaminergic modulation of working memory functions. *Journal of Neuroscience*, 19:2807–2822, 1999.

[74] Toshiyuki Sawaguchi and Patricia S Goldman-Rakic. D1 dopamine receptors in prefrontal cortex: involvement in working memory. *Science*, 251:947–950, 1991.

[75] Onur Ozan Koyluoglu, Yoni Pertzov, Sanjay Manohar, Masud Husain, and Ila R Fiete. Fundamental bound on the persistence and capacity of short-term memory stored as graded persistent activity. *Elife*, 6:e22225, 2017.

[76] Nicholas T Franklin, Kenneth A Norman, Charan Ranganath, Jeffrey M Zacks, and Samuel J Gershman. Structured event memory: A neuro-symbolic model of event cognition. *Psychological Review*, 127:327—361, 2020.