Graduate Training Centre of Neuroscience
Neural Information Processing

University of Tübingen

Can you remember everything all at once?
Rate-distortion theory and human memory

Yirong Xiong

The study was supervised by

Charley Wu
Human and Machine Cognition Lab
Excellence Cluster Machine Learning
University of Tübingen

Duration of the essay rotation: 10.10.2022 - 24.11.2022

Abstract

Human memory has poor performance in everyday tasks. A framework based on efficient compression, rate-distortion theory, posits that memory systems are optimized to store information given limited resources. It addresses the trade-off between resource constraints and task performance. In this essay, I first illustrated that human memory is lossy compression. Then in the second part, I introduced framework of rate-distortion theory and its implementation based on artificial neural network. In the third part, I showed the principles of memory and how environmental and cognitive factors shape them using this framework. In the final part, I discussed about further adaptation of rate-distortion theory on neuroscience and other cognitive processes.

Contents

1 Introduction **3**

2 Theoretical framework **3**
2.1 RDT on memory 4
2.2 Variational autoencoders 5

3 Three principles of RDT and their implications on working memory **5**
3.1 Prior knowledge principle 6
3.1.1 Prior distribution 6
3.1.2 Context 7
3.2 Limited capacity principle 7
3.2.1 Capacity 9
3.2.2 Set-size 9
3.3 Task dependency principle 9
3.3.1 objective function 9

4 Discussion **10**
4.1 RDT in computational neuroscience 10
4.2 RDT in a broader picture of learning 11

Figure 1: Can you remember everything all at once?

1 Introduction

Memory is not a perfect reproduction of the past. When studying for an exam, your brain encodes a massive amount of information and prepares for memory retrieval. However, our memory system is capacity-limited: a finite physical storage medium with limited bits cannot remember everything all at once (Bates and Jacobs, 2020). Thus, information is often compressed during encoding by discarding redundancy and reducing the precision of unimportant parts (Nagy et al., 2020).

For instance, look at figure 1 for 10s. Now, here are two memory test questions: Q1. What is the text in the center of the figure? Q2. How many googly eyes are there? You might find Q1 relatively easy. While the absolute majority might find Q2 is an impossible mission, because our memory is lossy compression of reality (Nagy et al., 2020). Spending 2MB to encode a chaotic image in 10s is unreasonable for our brains. Therefore, our memory system summarized it as the title and filtered chaotic background, such as googly eyes.

Memory is optimized to store information precisely, subject to the constraints of limited mental resources (Gershman, 2021). This optimization problem can be formalized using rate-distortion theory (RDT), which focuses on the trade-off between storage cost and memory performance (Bates and Jacobs, 2020; Sims, 2016; Nagy et al., 2020). We will start with the theoretical framework of RDT.

2 Theoretical framework

As illustrated by Figure 2A, we can consider memory as a communication channel, including encoding and decoding (Lai and Gershman, 2021; Gershman, 2021). For encoding, sensory input is stored and transferred to memory. For decoding, the information is retrieved from memory and transferred to the output. The goal is to keep information from input during these two processes.

Before jumping to a solution, let's consider the definition of information and how to measure it. These questions are well addressed by information theory (Shannon, 1948). An information source generates an input signal x as a random variable. Given a specific value x_0, the larger the probability $p(x = x_0)$, the less surprising it is. Therefore, the 'surprise' of signal x_0 can be defined as $\log p(x = x_0)$. Considering all possible values, we get the expectation of surprise in this signal, which is entropy (Shannon, 1948). Entropy describes the average amount of 'surprising information' contained in a random variable.

$$H(x) = -\sum_i P(x_i) \log P(x_i)$$

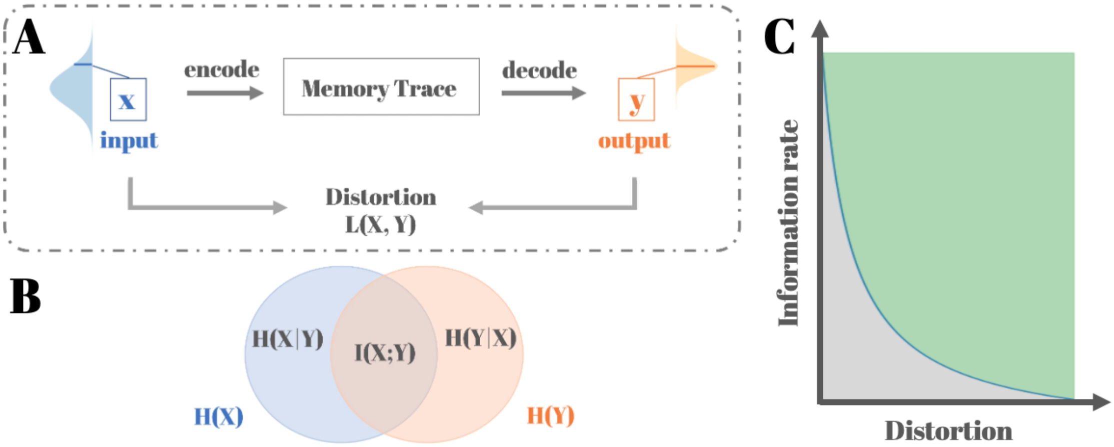

Figure 2: Theoretical framework of RDT. (A) Memory as a communication channel. An input stimulus sampled from a distribution is encoded into a memory trace by perception systems. During retrievial, a probability distribution is generated based on the memory trace, and output is sampled from the distribution. The distortion denotes the loss function of input and output. (B) Mutual information of input and output. $H(X)$ and $H(Y)$ indicate the entropy of input and output. $H(X|Y)$ and $H(Y|X)$ denote the information entropy carried by input and output exclusively. While $I(X;Y)$ shows the information shared by input and output. (C) The rate-distortion curve. The blue line indicates the rate-distortion curve. Channels on the curve are efficient, given their resource constraints. Channels in the green area above the curve are inefficient and have better performance by optimization. The grey area below the curve is unachievable.

How do we measure the information conveyed by a communication channel? Intuitively, we want the information from input and output to overlap as much as possible and reduce the loss of information in input and noise-induced uncertainty in output (Figure 2B). The overlap area, known as mutual information, can be defined as equation below:

$$I(x, y) = H(x) - H(x|y) = \sum_i \sum_j P(y_j|x_i)P(x_i) \log \frac{P(y_j|x_i)}{P(y_j)}$$

2.1 RDT on memory

It is easy to define a perfect memory channel where the output and input are the same. However, a channel in practical settings can rarely achieve perfect performance. There is always distortion between input and output. To this end, it's necessary to define a cost function, $\mathcal{L}(x, y)$. This function specifies the cost as the difference between an input signal value and its corresponding output, which means that the lower the cost, the better the performance. Therefore, the distortion of the channel, D, can be defined as the expectation of cost over all input and output:

$$D = E[\mathcal{L}(x, y)] = \sum_i \sum_j \mathcal{L}(x_i, y_j)P(y_j|x_i)P(x_i)$$

Rate-distortion curve: Given capacity constraints, when the distortion is minimized, the model achieves the best performance. Changing capacity constraints results in varying distortion. Therefore, we can draw the minimal distortion as a function of constraints, which is known as rate-distortion curve (Figure 2C). The area below the curve is impossible to achieve given the current setting, while the area above it is sub-optimal, which means it would be possible for models/human subjects in this area to improve their performance on reconstruction without increasing channel capacity.

The efficiency index: The efficiency of the channel is based on the capacity constraint and optimization level. We can evaluate the efficiency of a channel as Equation 4, where D_{emp} reflects the empirical distortion of this channel, D_{max} is the maximum distortion (the point where the rate-distortion curve intercepts the x-axis, or equivalently the optimal guessing performance), and D_{min} is the minimal distortion a channel can achieve given same information rate.

$$\epsilon = \frac{D_{emp} - D_{max}}{D_{min} - D_{max}}$$

2.2 Variational autoencoders

Can we build a model to implement RDT? Although there are already many algorithms for compression, contrary to perception and memory in biological systems, they are:

- not designed to adapt over time to new input distributions,
- not capable of capturing the uncertainty of encoding and decoding, and
- lacking varying loss function or the capacity.

Development in machine learning addresses our needs quite well. Autoencoder, a special artificial neural network, was invented to reconstruct high-dimensional data using a neural network model with a bottleneck layer in the middle (Hinton and Salakhutdinov, 2006). A recent development of it is variational autoencoders (VAE), which map input into a distribution instead of a fixed vector (Kingma and Welling, 2014).

A VAE consists of three components, encoder, sampler, and decoder (Figure 3). The first two parts map each input x_i to a probability distribution $p(z_i|x_i)$ over latent variables and sample a latent vector z_i from the distribution. The decoder takes z_i as input and reconstructs y_i based on it.

In order to reconstruct y that looks like a real input x from distribution of z, we need the probabilistic decoder $p_\theta(x|z)$ to maximise $p_\theta(x_i) = \int p_\theta(x_i|z)p_\theta(z)dz$. However, it's extremely computationally expensive to marginalize over latent variable z. We could find a solution the other way around by introducing a new approximation function $q_\phi(z|x)$, which should be very close to the real likelihood distribution $p_\theta(z|x)$. We can use Kullback-Leibler divergence to quantify the distance between these two distributions as $D_{KL}(q_\phi(z|x)||p_\theta(z|x))$. Using Variational Bayesian methods, the aim of minimizing D_{KL} can be realized by maximizing the evidence lower bound (ELBO).

$$E_{z\sim q_\phi(z|x)}(logp_\theta(x|z)) - D_{KL}(q_\phi(z|x)||p_\theta(z))$$

ELBO includes two terms. The first one penalizes low reconstruction quality and is referred to as the **distortion function**. The second one, **capacity constraint**, encourages high efficiency in latent encoding and constraints capacity of latent space. The negation of the ELBO defines the loss function of VAE.

$$L_{VAE}(\theta, \phi) = -E_{z\sim q_\phi(z|x)}(logp_\theta(x|z)) + D_{KL}(q_\phi(z|x)||p_\theta(z))$$

β-VAE: The structure of the loss function allows us to manipulate the capacity of latent space by introducing a scalar multiplier β as a hyper-parameter, which scales the second term and can effectively trade-off between reconstruction quality and encoding efficiency (Higgins et al., 2022). When β is small, the bottleneck is relatively unconstrained, leading to large capacity and low encoding efficiency. When β is larger, the constraint on the latent bottleneck gets stronger, and its size tends to be small. The varying capacity of latent space provided by β-VAE is highly correlated with lossy compression. Therefore, β-VAE is an optimal model to implement RDT.

$$L_{\beta-VAE}(\theta, \phi) = -E_{z\sim q_\phi(z|x)}(logp_\theta(x|z)) + \beta D_{KL}(q_\phi(z|x)||p_\theta(z))$$

3 Three principles of RDT and their implications on working memory

Efficient data compression accounts for essential aspects of working memory. Several studies have examined three principles of RDT and their implications on working memory performance.

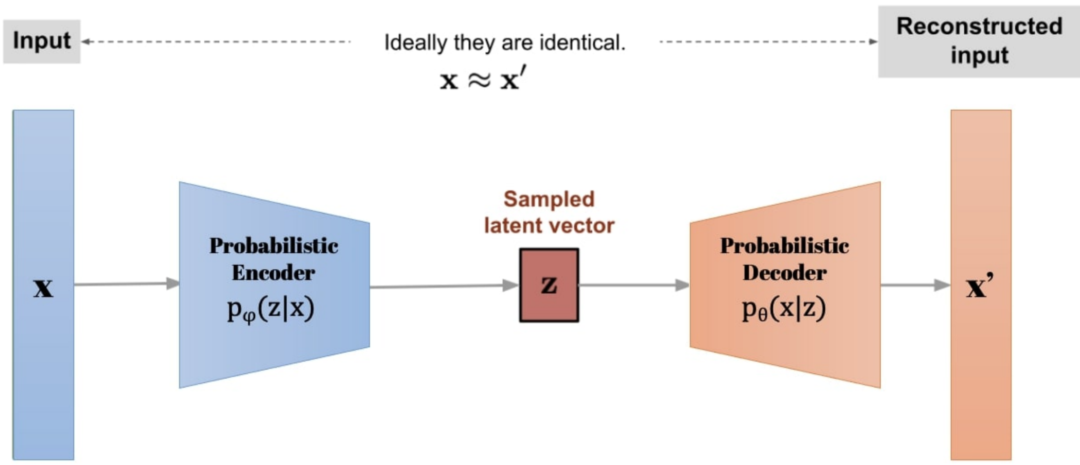

Figure 3: Conceptual framework of VAE. A VAE consists encoder, sampler, and decoder. Input is first mapped to a probability distribution $p(z_i|x_i)$ by the encoder. Then a latent vector sampled from the distribution is stored. During retrieval, a reconstructed input is generated based on the likelihood of the hidden vector $p(z_i|x_i)$.

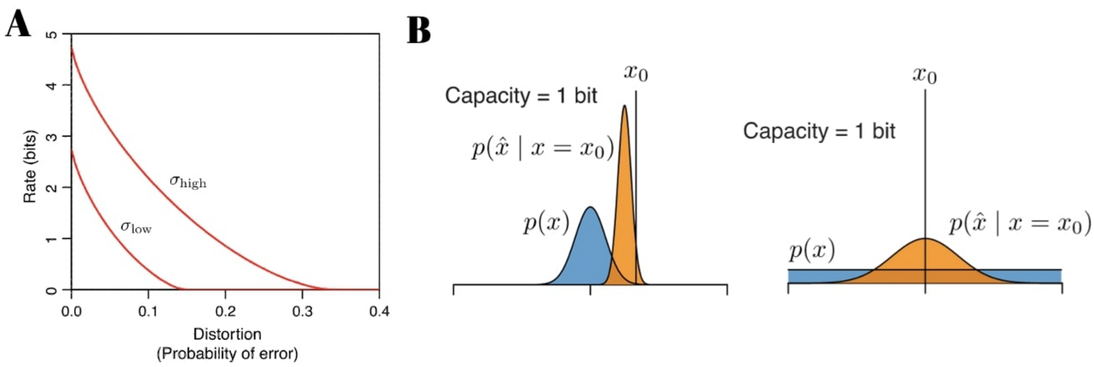

Figure 4: Prior knowledge principle. (A) Rate-distortion curves of two stimulus variance conditions. (B) Different prior distributions (uniform and Gaussian) and output distributions of x_0.

3.1 Prior knowledge principle

The prior knowledge principle states that efficient memory reconstruction must use knowledge of statistics of the to-be-remembered item. Thus, if people‘s memory system is similar to the rate-distortion model, their performance in memory tasks should be influenced by the prior knowledge in two ways:

- the performance should be worse with increasing entropy in the prior distribution
- the reconstruction should be biased toward the mean of the prior distribution

3.1.1 Prior distribution

Higher entropy leads to worse performance. Bates et al. (2019) presented stimuli of plant-like objects that varied by the width of their leaves, which were from distributions with different entropy. The participants were supposed to reconstruct the leave-width of stimulus. Bates et al. found that the number of correct responses with decreasing entropy was higher (Figure 5B). Stimuli with low probability in prior knowledge are more demanding to reconstruct than familiar ones.

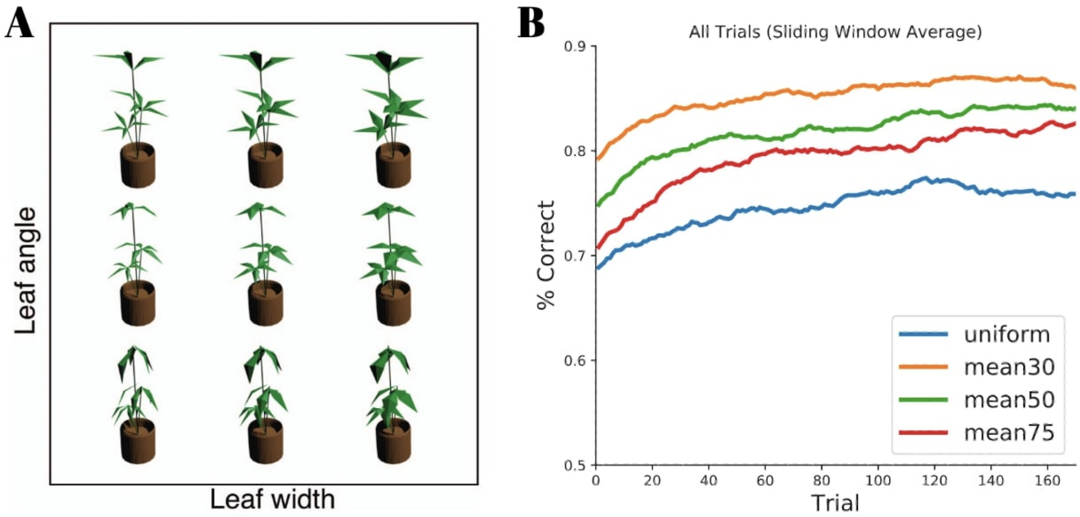

Figure 5: Entropy of prior distribution affects recall performance. (A) Stimuli of artificial potted plants used by Bates et al. (2019) Leaf angle varies along the vertical axis, while leaf width varies along the horizontal axis. (B) Average percent correct for memory task as a function of trial number for different prior stimulus distributions (uniform distribution and Gaussian distribution with various means and variances) (Bates and Jacobs, 2020).

Nagy et al. (2020) studied the effect of prior distribution in a real-life chessboard setting. They presented game- and random-configurations to subjects with different chess skills and β-VAE models trained on a different number of games (Figure 6). They found the increased expertise contributed explicitly to the enhancement of recall performance in real-game configurations but had no significant influence on the performance of random configurations. A potential explanation is that prior knowledge of real chess games affects the efficient compression of data.

3.1.2 Context

Recall can be biased towards prior knowledge. Nagy et al. (2020) tested the effect of contextual information on memory recall (Figure 7). In the human experiment, subjects were shown ambiguous hand-drawn sketches of objects (e.g., inputs shown in Figure 7A) and were asked to reproduce the images after a delay. Before recalling, they saw a category name, which provided one possible interpretation of the image as context. The authors found that the reconstruction of images was biased toward the contextual category. The simulation was conducted on conditional models on sketches belonging to each label. Similar to results on human subjects, reconstructions from the conditional posterior resulted in systematic distortions of the original image consistent with the provided label.

3.2 Limited capacity principle

The limited capacity principle states that all finite systems have limits in processing and storage capacities. Thus, the performance of human memory depends on the system's capacity. The limited capacity principle suggests at least two predictions on human memory. First, subjects with lower working memory capacity should have worse recall performance. Second, the performance should decrease with increasing stimuli due to allocating storage.

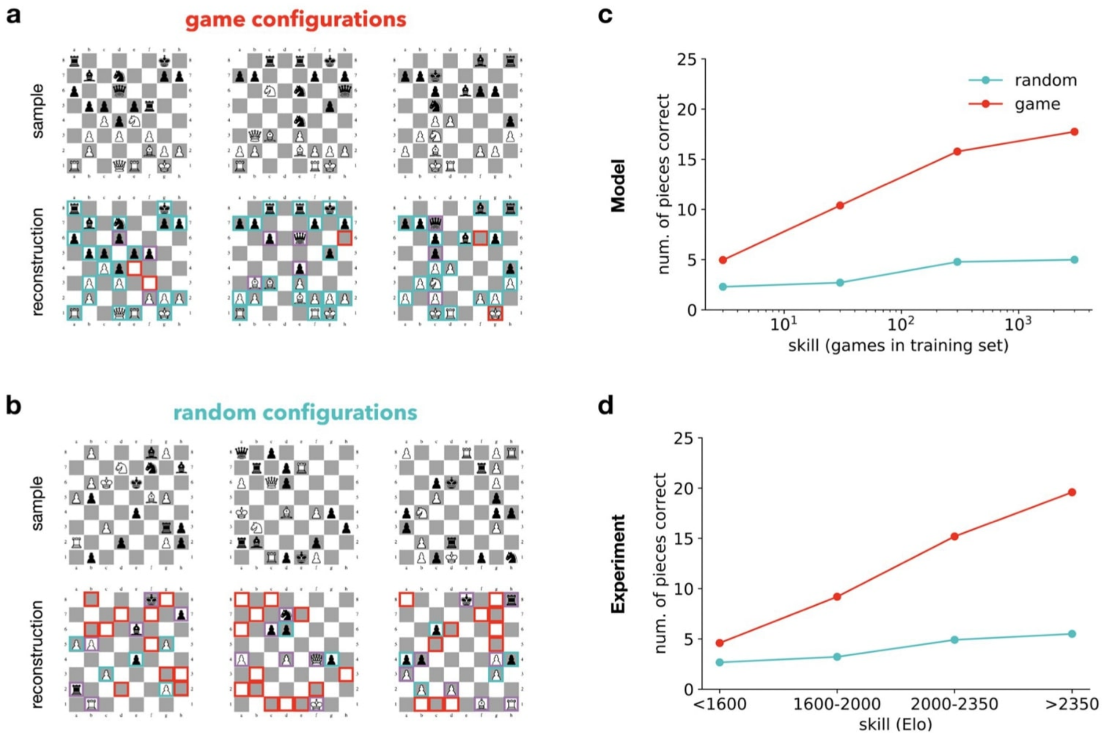

Figure 6: Prior knowledge affects recall performance (Nagy et al., 2020). (A) *Top*, Chessboard configurations from real-game settings. *Bottom*, reconstructions of the chessboards. Blue, purple, and red indicate correct reconstruction, switched identity, and missed/false reconstruction. (B) Same as A but from randomized settings. (C) Reconstruction accuracy by VAE models as a function of the training set. (D) Reconstruction accuracy by human participants as a function of the chess skill.

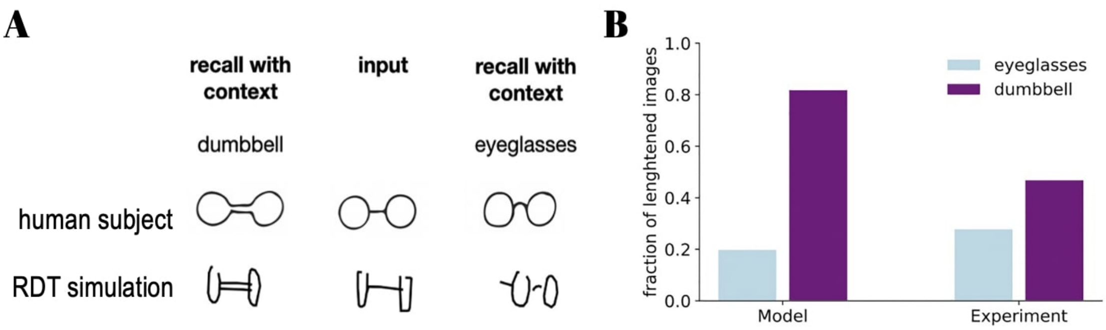

Figure 7: Context affects recall performance (Nagy et al., 2020). (A) Recall input with different contexts. *Top*, The left one is reconstructed by a human subject in the 'dumbbell' context, while the right one is reconstructed in the 'eyeglasses' context. *Bottom*, Same as top, but reconstructed by model. (B) Quantitative changes in visual features (fraction of the length of the line in reconstruction) with changing context.

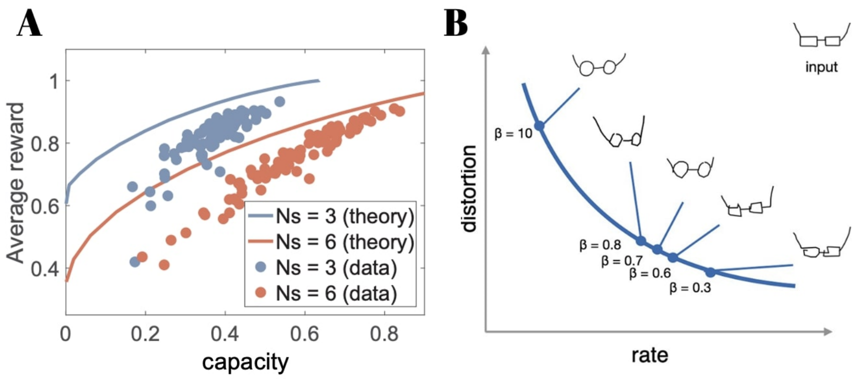

Figure 8: Capacity affects recall performance. (A) Average reward as a function of capacity (Gershman, 2020), applied to data from Collins (2018). Each solid line shows the optimized models for a particular set size (Ns = 3 or 6). The circles show data from human participants. (B) Changes in β result in different reconstruction points on the rate-distortion curve (Nagy et al., 2020).

3.2.1 Capacity

Experimental evidence indicates that with more information rate, the memory trace can represent finer details of the stimuli. Thus the reconstruction can be less distorted. Gershman (2020) examined this using data from Collins (2018), which consists of 91 subjects performing a reinforcement learning task. They found reward generally increases monotonically with the capacity, with values close to the optimal trade-off curve. In Nagy et al. (2020) experiment, they manipulated the capacity of sketch-VAE models via changing beta values. The distortion between input and output is getting less with increasing rate.

3.2.2 Set-size

Set-size effects suggest decreases in memory performance with increases in the number of to-be-remembered items, which has already been reported in many working memory experiments. In Pratte (2020) visual experiment, subjects were supposed to remember square colors with varying set sizes from two to ten items. They found that precision declined monotonically with set size (Figure 9A). RDT predicts working memory distribution as a function of the number of items stored, assuming that the total capacity is evenly divided among items. For instance, as illustrated in Figure 9B, every item in the set-size six conditions shares the memory distribution for the set-size six conditions (Bates et al., 2019). Bates and Jacobs (2020) used a β-VAE model to reconstruct data set consisting of images with varying numbers of stimuli. The result is the same as on human subjects. When less capacity is available, a model tries to "spread" its resources across the to-be-remembered stimuli.

3.3 Task dependency principle

To efficiently compress the stimuli, memory systems must consider behavioral goals. The task dependency principle implies that the allocation of mental resources should consider the information's value. For example, task-irrelevant features should be allocated minimal resources. Therefore reconstruction of them should be more distorted than task-relevant features.

3.3.1 objective function

Bates and Jacobs (2020) used an extension decision layer to train β-VAE model in two conditions. In leaf-width relevant condition, the model was penalized solely for errors in leaf width; in the other condition, it was penalized solely for mistakes in leaf angle. As illustrated in Figure 10A, models ignored task-irrelevant features in reconstruction as expected. Similarly, in Gazzaley et al.

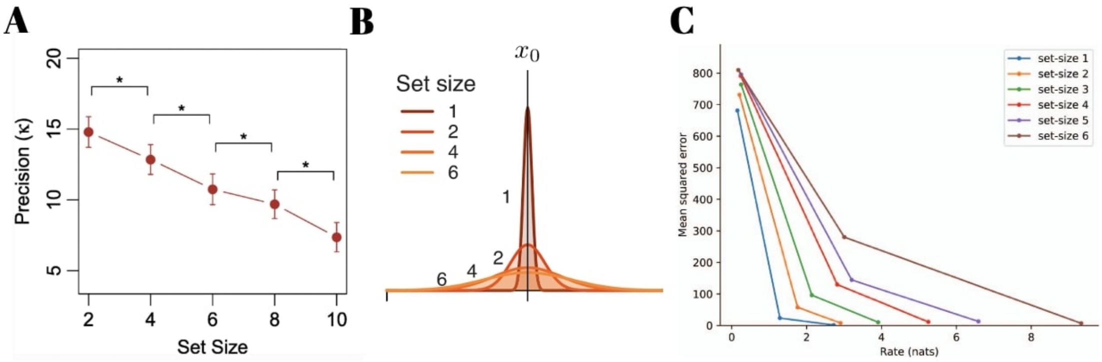

Figure 9: Set size affects recall performance. (A) Memory precision as a function of set size in a visual memory experiment (Pratte, 2020). (B) Reconstruction distribution of one stimulus in different set-size settings (Bates et al., 2019). (C) Estimated rate-distortion curves for each set-size setting (Bates and Jacobs, 2020).

(2008), the human subjects' recall performance on task-relevant stimuli is significantly better than on task-irrelevant stimuli (Figure 10B).

4 Discussion

So far, I have focused on the framework of RDT and examined its application to understanding human memory characteristics by comparing human subjects' performance and implementation models β-VAE.

What is the role of RDT in learning the nature and limits of human memory? According to Marr (1982), there are three levels of cognitive processing: the computational level, the algorithmic level, and the implementation level. These levels represent a gradient from the internal (neural mechanisms) to the external layer (the principles of the system's behavior). RDT serves as a theoretical framework for understanding human memory at a computational level (Sims, 2016). It regards human memory as a principled, capacity-limited system and provides explanations at a behavioral level. Moreover, it also gives insight into the neural implementation levels.

4.1 RDT in computational neuroscience

How does the brain realize RDT in neural activity? An impressive amount of experimental evidence has suggested that neurons encode sensory information in an efficient coding strategy (Borst and Theunissen, 1999; Lewicki, 2002; Olshausen and Field, 2004; Smith and Lewicki, 2006; Deneve and Chalk, 2016; Chalk et al., 2018). This strategy indicates that sensory systems, given internal constraints, encode maximal information from inputs by adapting neural coding to the statistical structure of the environment. The cost of the neural code to represent a stimulus is inversely proportional to the frequency of the stimulus in the environment.

Another critical question is, what leads to the constraint in neural systems? Zénon et al. (2019) has proposed two resources: information capacity and energetic demands. The informational constraint posits that increasing information demand decreases the capacity for other parallel processes. The energetic constraint postulates total brain energetic consumption is constant (Sokoloff et al., 1955; Raichle and Gusnard, 2002; Lennie, 2003), so the metabolic cost of neural activity in one brain area should influence total capacity in other brain areas. These two hypotheses suggest a local and global perspective of understanding the constraints in neural coding and require further studies.

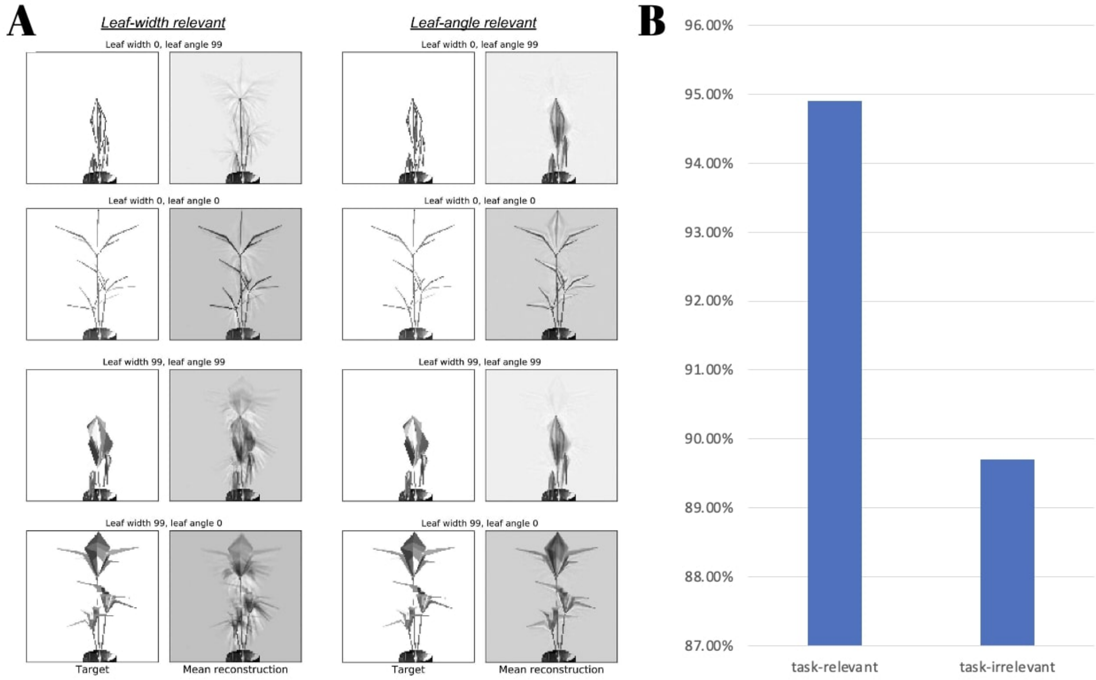

Figure 10: The goal of tasks affects memory representation. (A) Image reconstructions of target images when models were penalized for error in leaf-width (left column) or leaf-angle (right column) (Bates and Jacobs, 2020). (B) Recognition accuracy for task-relevant and task-irrelevant stimuli, applied to data from Gazzaley et al. (2008).

4.2 RDT in a broader picture of learning

Memory serves a central role in learning. To make adaptive decisions, one must refer to past experiences and evaluate the outcomes of different actions (Shohamy and Daw, 2015). These two processes can affect each other remarkably. For example, different learning procedures (e.g., free recall v.s. forced choice recognition) might change stored information (Breen, 1993). One possible explanation for this is the complexity of learning tasks (the difficulty of learning materials, ways of memory retrieval, time limits, etc.) might change subjects' usage of the distortion function, which represents their understanding of mistakes (Sims, 2016; Hamidi et al., 2022).

These results suggested there are multiple constraints in learning. However, how these constraints interplay with each other remains to be discovered. A possible way to solve it is by combining RDT with reinforcement learning (Gershman, 2020; Lai and Gershman, 2021), therefore inducing different constraints in the framework.

References

Christopher J. Bates and Robert A. Jacobs. Efficient data compression in perception and perceptual memory. *Psychological Review*, 127(5):891–917, 2020. ISSN 1939-1471. doi: 10.1037/rev0000197. Place: US Publisher: American Psychological Association.

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(2):11, February 2019. ISSN 1534-7362. doi: 10.1167/19.2.11. URL `https://doi.org/10.1167/19.2.11`.

Alexander Borst and Frédéric E. Theunissen. Information theory and neural coding. *Nature Neuroscience*, 2(11):947–957, November 1999. ISSN 1546-1726. doi: 10.1038/14731. URL `https://www.nature.com/articles/nn1199_947`. Number: 11 Publisher: Nature Publishing Group.

E. Kristin Breen. Recall and Recognition Memory in Parkinson's Disease. *Cortex*, 29(1):91–102, March 1993. ISSN 0010-9452. doi: 10.1016/S0010-9452(13)80214-6. URL `https://www.sciencedirect.com/science/article/pii/S0010945213802146`.

Matthew Chalk, Olivier Marre, and Gašper Tkačik. Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences*, 115(1):186–191, January 2018. doi: 10.1073/pnas.1711114115. URL `https://www.pnas.org/doi/10.1073/pnas.1711114115`. Publisher: Proceedings of the National Academy of Sciences.

Anne G. E. Collins. The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. *Journal of Cognitive Neuroscience*, 30(10):1422–1432, October 2018. ISSN 0898-929X. doi: 10.1162/jocn_a_01238. URL `https://doi.org/10.1162/jocn_a_01238`.

Sophie Deneve and Matthew Chalk. Efficiency turns the table on neural encoding, decoding and noise. *Current Opinion in Neurobiology*, 37:141–148, April 2016. ISSN 0959-4388. doi: 10.1016/j.conb.2016.03.002. URL `https://www.sciencedirect.com/science/article/pii/S0959438816300113`.

Adam Gazzaley, Wesley Clapp, Jon Kelley, Kevin McEvoy, Robert T. Knight, and Mark D'Esposito. Age-related top-down suppression deficit in the early stages of cortical visual memory processing. *Proceedings of the National Academy of Sciences*, 105(35):13122–13126, September 2008. doi: 10.1073/pnas.0806074105. URL `https://www.pnas.org/doi/full/10.1073/pnas.0806074105`. Publisher: Proceedings of the National Academy of Sciences.

Samuel J. Gershman. Origin of perseveration in the trade-off between reward and complexity. *Cognition*, 204:104394, November 2020. ISSN 0010-0277. doi: 10.1016/j.cognition.2020.104394. URL `https://www.sciencedirect.com/science/article/pii/S0010027720302134`.

Samuel J Gershman. The rational analysis of memory. page 19, 2021.

Mani Hamidi, Mihaly Banyai, and Charley Wu. Learning dynamics are shaped by the interplay of constraints on representation and policy complexity. *Conference on Neural Information Processing Systems*, 2022.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. July 2022. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, July 2006. doi: 10.1126/science.1127647. URL `https://www.science.org/doi/10.1126/science.1127647`. Publisher: American Association for the Advancement of Science.

Diederik Kingma and Max Welling. Auto-Encoding Variational Bayes. December 2014.

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. ISBN 978-0-12-824586-6. doi: 10.1016/bs.plm.2021.02.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0079742121000049.

Peter Lennie. The Cost of Cortical Computation. *Current Biology*, 13(6):493–497, March 2003. ISSN 0960-9822. doi: 10.1016/S0960-9822(03)00135-0. URL https://www.sciencedirect.com/science/article/pii/S0960982203001350.

Michael S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, April 2002. ISSN 1546-1726. doi: 10.1038/nn831. URL https://www.nature.com/articles/nn831. Number: 4 Publisher: Nature Publishing Group.

David Marr. *Vision: a computational investigation into the human representation and processing of visual information.* San Francisco, CA: W.H. Freeman, Cambridge, Mass, 1982. ISBN 978-0-262-51462-0. OCLC: ocn472791457.

David G. Nagy, Balázs Török, and Gergő Orbán. Optimal forgetting: Semantic compression of episodic memories. *PLOS Computational Biology*, 16(10):e1008367, October 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008367. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008367. Publisher: Public Library of Science.

Bruno A Olshausen and David J Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, August 2004. ISSN 0959-4388. doi: 10.1016/j.conb.2004.07.007. URL https://www.sciencedirect.com/science/article/pii/S0959438804001035.

Michael S. Pratte. Set size effects on working memory precision are not due to an averaging of slots. *Attention, Perception, & Psychophysics*, 82:1–13, 2020. ISSN 1943-393X. doi: 10.3758/s13414-019-01902-5. Place: Germany Publisher: Springer.

Marcus E. Raichle and Debra A. Gusnard. Appraising the brain's energy budget. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10237–10239, August 2002. ISSN 0027-8424. doi: 10.1073/pnas.172399499. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC124895/.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x.

Daphna Shohamy and Nathaniel D Daw. Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences*, 5:85–90, October 2015. ISSN 23521546. doi: 10.1016/j.cobeha.2015.08.010. URL https://linkinghub.elsevier.com/retrieve/pii/S2352154615001163.

Chris R. Sims. Rate–distortion theory and human perception. *Cognition*, 152:181–198, July 2016. ISSN 0010-0277. doi: 10.1016/j.cognition.2016.03.020. URL https://www.sciencedirect.com/science/article/pii/S0010027716300750.

Evan C. Smith and Michael S. Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–982, February 2006. ISSN 1476-4687. doi: 10.1038/nature04485. URL https://www.nature.com/articles/nature04485. Number: 7079 Publisher: Nature Publishing Group.

Louis Sokoloff, Renward Mangold, Richard L. Wechsler, Charles Kennedy, and Seymour S. Kety. THE EFFECT OF MENTAL ARITHMETIC ON CEREBRAL CIRCULATION AND METABOLISM 1. *Journal of Clinical Investigation*, 34(7 Pt 1):1101–1108, July 1955. ISSN 0021-9738. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC438861/.

Alexandre Zénon, Oleg Solopchuk, and Giovanni Pezzulo. An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, 123:5–18, February 2019. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2018.09.013. URL https://www.sciencedirect.com/science/article/pii/S0028393218306328.