Cognition 152 (2016) 181–198

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Rate–distortion theory and human perception

Chris R. Sims

Department of Psychology, Drexel University, Philadelphia, PA, United States

ARTICLE INFO

Article history:
Received 9 August 2015
Revised 22 March 2016
Accepted 25 March 2016
Available online 20 April 2016

Keywords:
Rate–distortion theory
Information theory
Bayesian perception
Absolute identification
Visual working memory

ABSTRACT

The fundamental goal of perception is to aid in the achievement of behavioral objectives. This requires extracting and communicating useful information from noisy and uncertain sensory signals. At the same time, given the complexity of sensory information and the limitations of biological information processing, it is necessary that some information must be lost or discarded in the act of perception. Under these circumstances, what constitutes an 'optimal' perceptual system? This paper describes the mathematical framework of rate–distortion theory as the optimal solution to the problem of minimizing the costs of perceptual error subject to strong constraints on the ability to communicate or transmit information. Rate–distortion theory offers a general and principled theoretical framework for developing computational-level models of human perception (Marr, 1982). Models developed in this framework are capable of producing quantitatively precise explanations for human perceptual performance, while yielding new insights regarding the nature and goals of perception. This paper demonstrates the application of rate–distortion theory to two benchmark domains where capacity limits are especially salient in human perception: discrete categorization of stimuli (also known as absolute identification) and visual working memory. A software package written for the R statistical programming language is described that aids in the development of models based on rate–distortion theory.

1. Introduction

Perception is the act of extracting meaning from noisy and uncertain sensory signals, and in the process choosing what information to transmit and what to discard. Once perceived, perceptual memory is the act of sending a message to your future self. The fundamentally communicative nature of perception and memory suggests the relevance of *information theory* to the study of perceptual processing. However, for biological information processing systems, it is not enough to merely transmit information. Rather, the goal of perceptual processing must be to help the organism achieve goals. This suggests a utilitarian perspective on human perception. Rate–distortion theory (Berger, 1971; Shannon, 1959) represents the mathematical framework combining these two disciplines: information theory and decision theory.

This paper focuses on rate–distortion theory as a principled mathematical framework for understanding human perception and perceptual memory. The goal is to explain perception as a form of computational rationality (Gershman, Horvitz, & Tenenbaum, 2015)—the maximization of performance subject to constraints on information processing. When sensory signals are continuous rather than discrete, or when communication channels lack sufficient capacity, the loss of some information is inevitable. In this case, the goal of perception cannot be the perfect transmission, storage, or reproduction of afferent signals, but rather the minimization of some cost function subject to constraints on available capacity. Rate–distortion theory concerns the optimal solution to this difficult tradeoff.

With its focus on minimizing the costs of error, as well as optimally integrating prior beliefs and uncertain sensory evidence, rate–distortion theory shares much in common with the probabilistic inference approach to perception (Kersten, Mamassian, & Yuille, 2004; Knill & Richards, 1996) and in particular Bayesian decision theory (Körding, 2007; Maloney & Mamassian, 2009). Hence, rate–distortion theory has much to say about how biological organisms *should* behave in a particular environment, in keeping with ideal observer (Geisler, 2011) or rational analysis (Anderson, 1990) approaches to understanding human cognition. Such models can serve as a benchmark for comparing against human performance, or may inspire theories of the underlying neural mechanisms. Importantly, unlike fully rational Bayesian models of perception, rate–distortion theory offers a means of directly incorporating strong limits on the capabilities of the cognitive system (in terms of channel capacity limits) in a principled and theory-driven manner. In this regard, rate–distortion theory represents an important tool for those interested in studying the computational rationality of cognition.

E-mail address: chris.sims@drexel.edu
URL: http://www.pages.drexel.edu/~crs346/

http://dx.doi.org/10.1016/j.cognition.2016.03.020
0010-0277/

The theoretical approach advocated in this paper is an extension of much existing work in sensory neuroscience to higher-level perception. The efficient coding hypothesis (Barlow, 1961) suggests that the goal for neural information processing is to form efficient codes for sensory signals. In this context, 'efficient' refers to the reduction of redundancy. This idea has proven extremely useful for understanding the properties of early cortical processing of visual information, such as the nature of receptive fields in V1 (Olshausen & Field, 1997). However, reducing redundancy is but one possible goal for biological computation. For organisms acting in an environment, it may be more important for perceptual systems to be "good" than "efficient". Here a good perceptual system is one that accurately and reliably solves important perceptual problems. The distinction is that costs and constraints can be imposed not just by the internal neural architecture, but also by the goals of the organism and the structure of the external environment. Rate–distortion theory generalizes the idea of efficient coding to allow for a broader range of possible cost functions (Hino & Murata, 2009; Simoncelli & Olshausen, 2001).

The following section briefly introduces information theory and its core constructs. These constructs are then used to motivate the fundamental problem addressed by rate–distortion theory. The theory is then applied to two domains: absolute identification (the assignment of perceptual stimuli to ordinal categories) and perceptual working memory. In each case, rate–distortion theory contributes something fundamentally new to the understanding of human perception.

2. Information theory: a brief introduction

Information theory is a scientific field spanning the boundaries of mathematics and engineering. It was first codified by Claude Shannon in 1948 under the title "A Mathematical Theory of Communication", and the following year with an introductory essay by Warren Weaver, as "*The* Mathematical Theory of Communication" (Shannon & Weaver, 1949). The subtle change in definite article reflected the growing realization of the definitiveness of the theory—a Bell Labs engineer who followed the developments noted that Shannon's publication "came as a bomb, and something of a delayed action bomb" (Gleick, 2011, p. 221). In the decades that followed, information theory had a transformative effect on many fields, psychology and neuroscience included. Concise reviews of the history of information theory in psychology and neuroscience are given in Luce (2003) and Dimitrov, Lazar, and Victor (2011). More extensive introductions to information theory can be found in Gallistel and King (2009, chap. 1) and Cover and Thomas (2012).

Perhaps the most famous application of information theory within psychology is George Miller's "The Magical Number Seven, Plus or Minus Two" (Miller, 1956). This paper concerned two quite different topics: what Miller termed the span of immediate memory, and the span of absolute judgment. The former topic introduced the concept of a *chunk* to the lexicon of cognitive psychology. The latter proposed a limit on the number of *bits* available for the categorical identification of a perceptual signal. This latter topic offers the most direct approach to information theory.

Quite simply, a bit is a unit of measure for a quantity of information. It is important to emphasize that a unit of measure is not the same as the physical quantity that is being measured. For example, in the 19th century the meter was defined as the distance between two marks on a platinum bar. Clearly, objects can be measured in meters even if they are not constructed out of platinum. The distinction is important, because a bit is commonly understood to refer to a binary digit—a 1 or a 0—but this connection is often misleading. A binary digit *conveys* one bit of information, but information need not be transmitted via a binary code. A photoreceptor in the retina conveys information in the form of an analog and graded signal. Despite this, the signal conveyed by a photoreceptor is meaningfully measured and studied in terms of its information-theoretic content, measured in bits.

If a bit measures information, then what is information? Answering this question requires that a few elementary concepts first be introduced. The first such concept is that of a random variable, labeled x. Informally, a random variable is something that can take one of a set of different possible values, where each value has an associated probability. For example, x might refer to the roll of a 6-sided die, in which case the value of the random variable is defined by the set $\{1, 2, \ldots, 6\}$, and if the die is fair, the associated probabilities $P(x = x_i) = \frac{1}{6}$ for $x_i \in \{1 \ldots 6\}$. In information theory, the set of possible values that a random variable can take is also called its alphabet. Random variables can be defined over continuous alphabets as well. The height of a person is a random variable whose domain is (in principle) all possible positive values. In this case, the probability density function $p(x)$, describing the distribution of heights, might resemble a Gaussian or normal distribution.

For a discrete random variable taking a particular value $x = x_i$, it is possible to define the surprise of that event as $-\log p(x = x_i)$. Why should the logarithm be relevant for measuring surprise? If $p(x = x_i) = 0$ then $-\log 0 = \infty$. In other words, impossible events are infinitely surprising. On the other hand, if $p(x = x_i) = 1$ then $-\log 1 = 0$: outcomes that are certain to happen are not surprising at all. Another justification for a logarithmic measure of surprise relates to the additivity of information gained by independent outcomes. For example, if x and y are independent random variables, then the 'total surprise' of observing both should (intuitively) equal the sum of the surprise of each outcome individually. Using a logarithmic definition, $-\log p(x \wedge y) = (-\log p(x)) + (-\log p(y))$. Thus, the negative logarithm of probability provides an intuitively correct measure of the surprisingness of an event. With surprise formalized in this manner, the entropy of a random variable is simply its 'average surprise'[1]:

$$H(x) = -\sum_i P(x_i) \log P(x_i). \tag{1}$$

Note that this equation is simply the surprise of each outcome, weighted by its probability of occurrence. If a random variable has two equiprobable outcomes, the entropy of this binary random variable equals 1 when the logarithm is taken as base 2. This is defined to be 1 bit of information. Hence, the outcome of a fair coin flip conveys a single bit of information. When the natural logarithm is used, the corresponding unit of information is the nat (1 nat $\approx$ 1.44 bits).

Entropy describes the amount of information intrinsic to, or 'contained' in a random variable. Now consider a communication channel for conveying information from this source, as illustrated in Fig. 1. The input to this channel consists of samples from the random variable x. The output is also a random variable, labeled y. A communication channel relates a given input to the channel output via a conditional probability distribution, $P(y|x)$.

To give a concrete illustration of how the human perceptual system can be viewed as a communication channel as in Fig. 1, consider the task of visually judging (perceiving) the size of an object sitting on a table. The true size can be labeled x, and characterized by a probability distribution $p(x)$. The different possible values for x define the alphabet for the channel (if x is continuous, then the source alphabet is also infinite). The distribution $p(x)$ might reflect, for example, the fact that it would be unlikely to encounter extremely large objects sitting on a typical-sized table. Due to intrinsic noise in neural coding, it is physically impossible

[1] In Eq. (1) and throughout this paper, define $0 \times \log 0 = 0$.

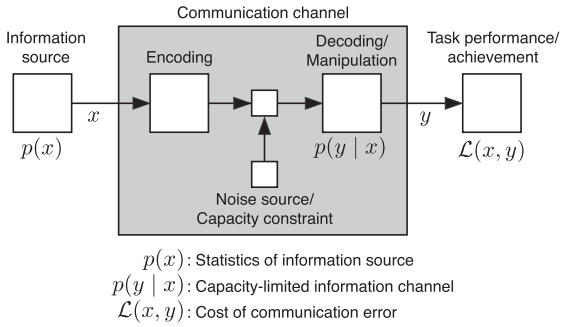

Fig. 1. The core constructs of rate–distortion theory. An information source is described by a probability distribution over its alphabet. Samples from this source are communicated over a noisy or capacity-limited channel, resulting in a conditional probability distribution over the channel output. A cost function defines the consequences of error in communication.

for the brain to represent the size *exactly*. Hence, the encoded representation is necessarily noisy and error-prone. Consequently, in decoding this neural representation to arrive at a "best guess", this guess y (the output of the channel) would also be error-prone. Hence, the channel itself can be described by a conditional probability distribution $p(y|x)$. Note that in the general case the alphabet for the channel input and output need not be the same. For example, the channel input might consist of a continuous signal, and the channel output a discrete category label or a binary response.

How do we measure the amount of information conveyed by such a channel? Intuitively, conveyed information can be no greater than the information contained in the source. But if the channel is noisy or error prone, the conveyed information of the output signal may be less than the entropy of the source. Thus, both the information source and the properties of the channel are relevant for measuring the communication of information. Mathematically, these two aspects are related by the mutual information of the channel input and output:

$$I(x,y) = H(x) - H(x|y) = \sum_i \sum_j P(y_j|x_i)P(x_i) \log \frac{P(y_j|x_i)}{P(y_j)}, \tag{2}$$

where $H(x|y)$ refers to the conditional entropy,

$$\begin{aligned} H(x|y) &= \sum_j P(y_j) H(x|y = y_j) \\ &= -\sum_j P(y_j) \sum_i P(x_i|y_j) \log P(x_i|y_j). \end{aligned} \tag{3}$$

As shown by the top line of Eq. (2), mutual information measures the reduction in entropy regarding the channel input, after observing the output of the channel. A channel that conveys a lot of information greatly reduces uncertainty regarding the input signal. When the channel input or output are defined by continuous signals, the summations above can be replaced by integrals. Like entropy, mutual information is also measured in units of bits (or nats). Mutual information defines the information rate of a given channel and information source. Here, *rate* does not intrinsically refer to time (bits per second), but rather the average number of bits communicated by the channel per symbol from the source alphabet (bits per symbol). For example, consider a sequence of outcomes generated from repeatedly rolling a 6-sided die. If these outcomes are communicated over a channel, the information rate of the channel—defined by the mutual information—is the number of bits conveyed per die roll (the reduction in uncertainty), averaged across the sequence of outcomes.

Mutual information depends on the properties of the information source, $P(x)$ as well as the behavior of the channel, $P(y|x)$. If one were to take a fixed channel defined by $P(y|x)$, and use it to communicate signals from different source distributions, one would find that the mutual information would be higher for some information sources, and lower for others. The capacity of a channel is given by the maximum of the mutual information, over the space of all possible information sources $p(x)$.

3. Rate–distortion theory: information theory meets decision theory

It is deceptively easy to define a perfect communication channel. This is simply a channel where the output always equals the input. For example, in a perfect visual memory system, one would simply remember the world as it appears. However, in practical settings perfect performance is rarely achievable. Indeed, it is a somewhat counterintuitive fact that a perfect communication channel cannot exist when the information source is continuous. Even when signals are discrete rather than continuous, perfect communication might not be possible or feasible. The capacity of the channel might be lower than the entropy of the information source, the output alphabet of the channel might have fewer symbols than the input alphabet, or the channel might have known mechanistic limitations that limit its fidelity. In any of these cases, the goal of error-free communication must simply be abandoned.

In its place, the goal in such a setting must be to minimize the costs of communication error. To this end, it is necessary to define a cost function, $\mathcal{L}(x,y)$. This function specifies the cost associated with the event of an input signal x being conveyed as the value y. An intuitive cost function might be the squared error between the channel input and output: $\mathcal{L}(x,y) = (y-x)^2$. Channels where the output closely resembles the input will have low cost according to this measure. The distortion, or average cost associated with a particular information source and channel is given by

$$D = E[\mathcal{L}(x,y)] = \sum_i \sum_j \mathcal{L}(x_i, y_j) P(y_j|x_i) P(x_i) \tag{4}$$

If a particular cost function and information source are specified, then an optimal communication channel is one that minimizes distortion. Generally speaking, channels with higher capacity will be able to do a better job of minimizing costs than channels with lower capacity. But since any physical channel, labeled Q, must have a finite capacity, the problem becomes one of optimization under constraints. This is the fundamental problem addressed by rate–distortion theory (Berger, 1971; Shannon, 1959). We can state this problem more formally as follows:

$$Q^\star = \arg\min_Q D_{(Q)} \tag{5}$$

subject to $I_{(Q)}(x,y) \leqslant C$.

The equation states that an optimal information channel (labeled $Q^\star$) for a given information source is one that minimizes the channel distortion, subject to the constraint that the mutual information of the source and channel is at or below a given bound (indicated by C). The operator 'arg min' indicates the minimizing argument of the expression—the channel that achieves the lowest distortion. The distortion $D_{(Q)}$ and mutual information $I_{(Q)}$ are determined according to a given channel $Q = P(y|x)$, and the optimization is performed over the space of all possible channels (i.e., over all valid conditional probability distributions).

Shannon's noisy-channel coding theorem (Shannon & Weaver, 1949) states that when a channel has capacity greater than the information rate of the source (the average number of bits communicated per transmission), it is possible to achieve an arbitrarily small error rate. Most digital communication is concerned with this case of error-free performance. Conversely however, when

the channel capacity is lower than the information rate, the occurrence of error is unavoidable. In this case, the goal is not to avoid error, but rather minimize the expected cost of error. Eq. (5) seeks this lowest-possible cost.

The fundamental tradeoff described by rate–distortion theory is that decreasing distortion (average cost) requires a corresponding increase in the amount of information communicated by the channel. This tradeoff is captured by a rate–distortion curve, which indicates the minimum channel capacity necessary to achieve a particular level of performance. The important construct of a rate–distortion curve is illustrated next using a simple example.

3.1. Example rate–distortion problem: rolling a six-sided die

In this example, a simple but concrete demonstration is provided of the application of rate–distortion theory. The goal is to illustrate the mathematical concepts within a particularly simple example in order to facilitate the understanding of this same approach as applied to human perception. To illustrate the construct of a rate–distortion curve, consider again the example of rolling a six-sided die. The information source $P(x_i) = 1/6$ for $x_i \in \{1 \dots 6\}$. The task is to communicate the outcome of a roll via a capacity-limited channel, and the cost function for this example assumes that the goal is to maximize accuracy, or equivalently, minimize the probability of error. That is,

$$\mathcal{L}(x, y) = \begin{cases} 0 & x = y \\ 1 & x \neq y \end{cases}. \tag{6}$$

The goal is to find an optimal channel that minimizes expected cost according to this measure. The optimal rate–distortion curve for this example is illustrated in Fig. 2. This curve is expressed analytically as

$$R(D) = \log(6(1 - D)) + D \log\left(\frac{D}{5(1 - D)}\right). \tag{7}$$

Appendix A gives a complete derivation for this result. In brief, this solution is obtained by formulating the objective in Eq. (5) as an optimization problem, taking its derivative and analytically solving for the optimal information channel that achieves the minimum. The appendix also describes several other methods for solving rate–distortion problems, including both analytic and efficient numerical solution methods.

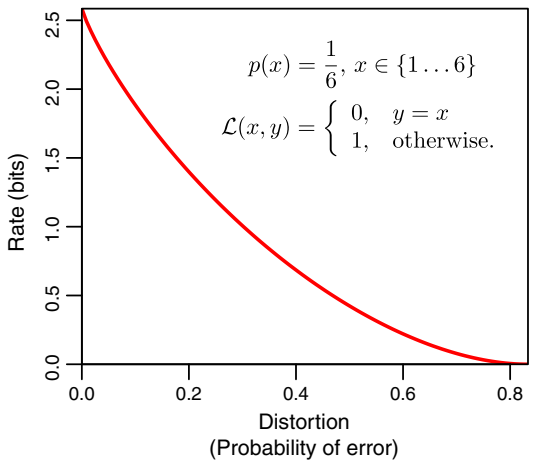

Fig. 2. Rate–distortion curve for communicating the outcome of rolling a 6-sided die, with a probability of error cost function.

There are several noteworthy features of the rate–distortion curve in Fig. 2. First, the curve reaches zero at the point $D = 5/6$ along the x-axis. This indicates the expected probability of error when no information is transmitted over the channel. Intuition confirms that this value is correct: By simply guessing the result of the die roll, one has a $1/6$ chance of guessing correctly, and an expected error rate of $5/6$. The curve intercepts the y-axis at an information rate of just over 2.5 bits. This is the minimum channel capacity in order to achieve zero communication error. This value also corresponds to the entropy of the source, $H(x) \approx 2.58$ bits. In between these two extremes, the curve shows that when some amount of error is tolerable, there is an absolute minimum channel capacity necessary to achieve the desired level of performance. Or in other words, no physical communication channel can exist at a point below this curve. Efficient communication channels are those that exist at points along, or close to, the theoretical bound predicted by a rate–distortion curve. If human perception or memory are efficient, one would predict their performance to lie close to a rate–distortion curve.

Note that the curve shown in Fig. 2 is specific to both the information source and the particular choice of cost function. Changing either of these elements would result in a different optimal channel and a different rate–distortion tradeoff. The cost function described by Eq. (6) is framed around minimizing the probability of communication error. However, it is easy to imagine that some types of errors may be more costly than others. In some gambling games the stakes depend on whether the roll of a die indicates an odd or even number. In such a scenario, $\mathcal{L}(3, 4)$ might be greater than $\mathcal{L}(3, 5)$, since in the former case an odd number is mistakenly reported as even. For every cost function there exists an information channel that is optimally efficient according to that measure.

3.2. Summary

Rate–distortion theory is concerned with finding optimal but capacity-limited communication channels. The approach defines a boundedly rational solution to the problem of information transmission. Rather than attempting to perfectly convey signals, the objective is to minimize the cost of error. This perspective very naturally addresses a wide range of biological information processing problems, as will be discussed below.

As a theoretical tool, a rate–distortion curve is highly useful. If one wishes to study the properties of a particular channel—be it a fiber optic cable or human memory—it is often difficult to directly compute the channel capacity, as this requires complete and accurate knowledge of the channel distribution $P(y|x)$. However, measuring performance or accuracy is often much easier. Having a rate–distortion curve allows one to directly map between these two quantities and infer a lower bound on channel capacity. Measuring a given level of performance enables a strong statement of the minimum capacity that the channel must possess. Further, this prediction is entirely independent of model parameters or assumptions.

The following two sections demonstrate the application of rate–distortion theory to two topics in human perception: discrete categorization of perceptual stimuli (absolute identification) and visual working memory. In each case the goal is an accessible introduction to the concepts and information-theoretic methods rather than a sophisticated and highly detailed model. For the moment, it is assumed that the optimization problem described by Eq. (5) can be solved to find an optimal channel $Q^{\star} = P(y|x)$ for a given information source and cost function. Efficient numerical and computational techniques for actually solving this problem are discussed in Appendix A.

4. Rate–distortion theory and absolute identification

Absolute identification involves the mapping of perceptual stimuli to ordinal categories. In a typical absolute identification experiment, a participant is trained to discriminate between a set of perceptual stimuli, such as a set of 10 different lines of varying length, or tones of varying frequency or loudness. Each stimulus item has an associated correct response, usually mapped to a different key on a keyboard. An absolute identification experiment consists of repeatedly presenting the subject with a randomly chosen stimulus, and asking the subject to respond with his or her best guess regarding the ordinal identity of the stimulus.

Rouder, Morey, Cowan, and Pealtz (2004) conducted an experiment examining absolute identification of line length.[2] On each trial, the subject was shown a line segment (displayed on a computer monitor) and was required to press a corresponding key on the keyboard. The line segment shown on each trial was randomly sampled from a set of different lengths. If shown the smallest line segment in the set, the correct response was to press '1'. If shown the second smallest, the correct response was '2', etc. Feedback was given to the participant after each response, indicating the correct response for that stimulus.

Across three conditions of the experiment the number of lines in the stimulus set was varied, using 13, 20, or 30 stimuli, and line lengths were spaced according to a power function with exponent 3.5. Two participants completed all three conditions, and a third participant completed just the 30-stimuli condition. Fig. 3a illustrates the most basic results from this experiment, plotting proportion of correct responses for each of the stimuli in each condition. The two most obvious features of the data are that performance decreases as the number of stimuli in the set is increased, and that performance is worse for stimuli in the middle of the range (termed the 'bow effect').

Absolute identification was a central component of George Miller's famous "magical number seven" paper (Miller, 1956). The reason is that performance in this type of experiment is intriguingly but systematically poor. When trained on a stimulus set of about seven items, participants can reach close to perfect performance. But as the number of stimuli in the set increases, identification accuracy decreases. Performance remains low even when neighboring stimuli are highly discriminable (termed the 'range effect').

How can information theory be applied to explain these data? For an experiment with a small discrete set of stimuli and responses, the distribution $P(y|x)$ can be approximated via empirical frequencies.[3] This enables directly estimating the entropy of the information source, $H(x)$ and the mutual information of the 'human channel', $I(x,y)$. These values are plotted in Fig. 3b, where each line indicates the data from one of the subjects. The dashed line in Fig. 3b indicates perfect performance in the absence of a limit on channel capacity. By contrast, transmitted information appears to reach an asymptote.[4] To Miller (1956) and others (Garner & Hake, 1951), this suggested that the fundamental limit in absolute identification was the channel capacity of the observer in processing perceptual information; as Miller termed it, the 'span of absolute judgment'. Note however that this assessment is not without challenge. The data illustrated in Fig. 3 are obtained by collapsing across many blocks of the experiment; Rouder et al. (2004) demonstrated that performance improved substantially with training (see also Dodds, Donkin, Brown, & Heathcote, 2011); thus it remains unclear the extent to which channel capacity is invariably fixed. MacRae (1970) argued that biases in the calculation of mutual information using small datasets may give the illusion of a constant channel capacity. When correcting for possible bias, mutual information may peak, and then decline with increased source entropy, suggesting that channel capacity is "limited, but not fixed" (MacRae, 1970).

If absolute identification is subject to a limit on channel capacity in any form, then rate–distortion theory suggests that the goal for perception should not be perfect identification, but rather the minimization of error according to some cost function. This constitutes a 'boundedly rational' view of perceptual categorization. But what is the cost that should be minimized? Despite the long history of absolute identification in psychology, this question has not previously been addressed. An obvious choice, matching the instructions given to participants in the typical absolute identification experiment, is to maximize the percent of trials answered correctly. This implies a cost function identical to that given in Eq. (6). Note that this cost function is discrete and discontinuous: the magnitude of the cost does not increase gradually with the magnitude of the error.

It is also worth noting that this cost function can be stated directly without introducing any free parameters into the model, under the assumption that participants in the experiment rationally seek to maximize their performance on the task. The predicted behavior according to this model can be compared against observed performance, and hence the hypothesis of computational rationality is fully falsifiable.

Fig. 4 illustrates the resulting optimal rate–distortion curves using this cost function and an information source with $N = 13$, 20, or 30 equally-likely values. Appendix B describes a software package implemented in the R programming language that was used to compute this curve. Overlaid on the plot are marker points corresponding to the empirical performance from each subject. Empirical information rate and distortion were estimated via Eqs. (2) and (4), respectively, with the cost function defined by Eq. (6). This figure illustrates that human performance is not only limited in capacity, but also rather *inefficient* at absolute identification, at least according to the explicit instructions given to participants. Inefficiency is defined in the following sense: without increasing channel capacity, it would have been possible for participants to substantially reduce their error rate. Graphically, this is indicated by the fact that the plot markers in the graph could be translated to the left without crossing the rate–distortion curve. A more formal index of efficiency, ϵ, can be defined as follows:

$$\epsilon = \frac{D_{\text{emp}} - D_{\max}}{D^{\star} - D_{\max}}, \tag{8}$$

where D_{emp} reflects the empirical distortion according to a given cost function, $D_{\max}$ is the maximum distortion (the point where the rate–distortion curve intercepts the x-axis, or equivalently the optimal 'guessing' performance), and $D^{\star}$ is the minimal distortion for a channel with the same information rate as empirical performance. Mean efficiency was $\epsilon = 0.94, 0.88, 0.69$ for the three set size conditions. Thus, human efficiency at absolute identification appears to decrease as the number of stimuli to be discriminated increases. Notably, this finding is undetectable in the absence of rate–distortion theory.

Given this result, there are two obvious questions: How, and why is human absolute identification inefficient?

To answer these questions it is useful to compare three different models of absolute identification (Fig. 5a–c). Each model assumes an information capacity limit constrained by observed human performance, determined by computing the mutual information from

[2] At the time of writing, the complete data set for this experiment is available from the website of Jeff Rouder, http://pcl.missouri.edu/.

[3] Note however, that using empirical frequencies with small datasets may significantly bias the calculation of mutual information. More sophisticated techniques for computing mutual information from empirical data have been developed (MacRae, 1970).

[4] Diagnosing an asymptote in performance is questionable with three data points, however the pattern reported here is consistent with a much larger body of evidence (Miller, 1956).

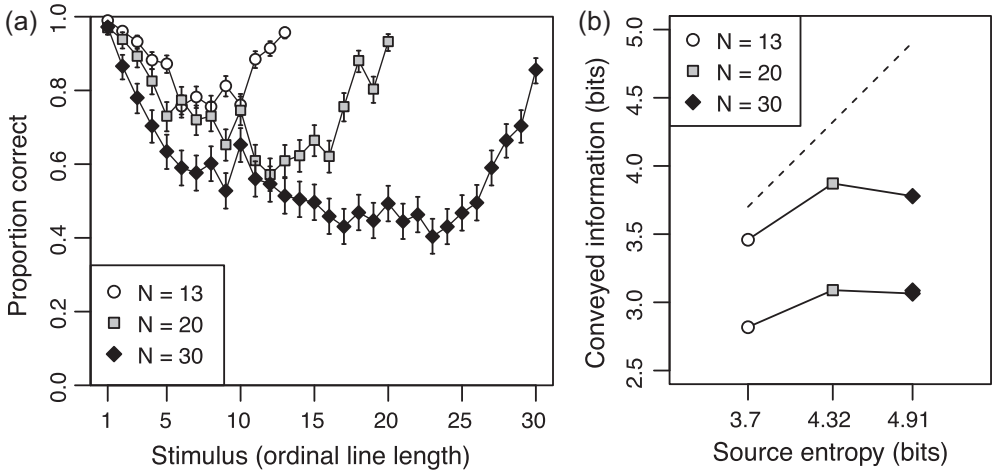

Fig. 3. (a) Results from an experiment on absolute identification of line length conducted by Rouder et al. (2004). Error bars indicate 95% confidence intervals, computed using a binomial test. (b) Human performance analyzed in terms of transmitted information (y-axis) as a function of the entropy of the information source (x-axis). Mutual information was computed according to Eq. (2) by approximating conditional response distributions with empirical response frequencies. The dashed line indicates perfect performance of a channel with no capacity limit.

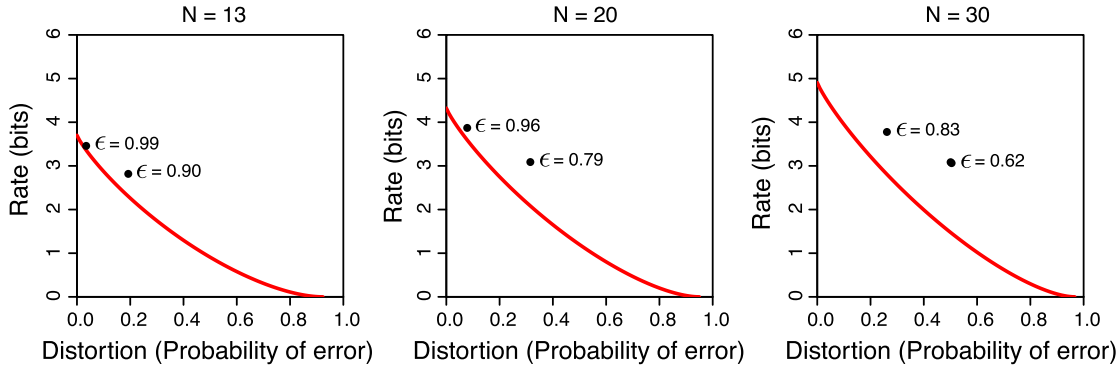

Fig. 4. Optimal rate–distortion curves for an absolute identification experiment with $N = 13$, 20, or 30 equiprobable stimuli (left, middle, and right panels), and a maximum accuracy cost function. Marker points indicate empirical performance for each subject. Information-theoretic efficiency, ϵ, is reported next to each point.

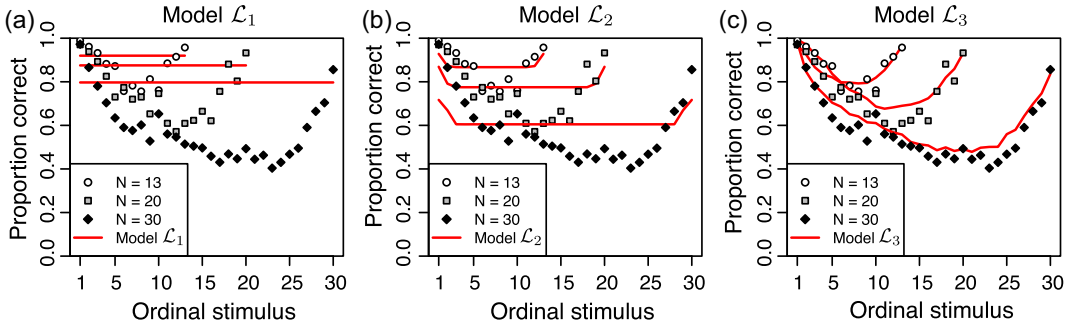

Fig. 5. Comparison of three information-theoretic models of absolute identification. Each panel plots the proportion correct according to the model (solid lines) and empirical performance from the Rouder et al. dataset (plot markers). The models differ only in the cost function that is minimized. (a) Probability of error cost function. (b) Absolute error cost function. (c) Empirically-derived cost function.

the empirical response distributions. For simplicity, the models are based on the aggregated data from each condition of the experiment rather than fitting models separately to each subject. This results in capacity limits of 3.01, 3.25, and 3.19 bits for the N = 13, 20, and 30 conditions, respectively. Each of the three models is an optimal information channel according to Eq. (5), each has the same limit on information transmission, and the models differ only in the cost function that is minimized. The predictions of the models follow directly from the optimal channel distributions $Q^\star$.

Model $\mathcal{L}_1$ uses the probability of error cost function, as in Eq. (6). This cost function also matches the presumed task given to subjects: Maximize the percent of trials answered correctly. Hence, model $\mathcal{L}_1$ represents a normative model of human perceptual performance. Given a constraint on capacity, the model performs in a mathematically optimal fashion.

The performance of this model is illustrated in Fig. 5a. Although this model transmits the same amount of information as human participants, it substantially exceeds human performance.

In addition, the model fails to exhibit the bow effect that is characteristic of human performance. The reason for this difference lies in the *type* of errors that human subjects and the model commit. When subjects responded incorrectly in the experiment, their errors were not random but rather tended to be close to the correct value. By contrast, model $\mathcal{L}_1$ has no conception of 'close': a response that is off by 5 stimulus levels is no worse than a response that is off by 1. As a result, the error distribution for model $\mathcal{L}_1$ is flat across all response categories. Near-misses convey some amount of information about the correct stimulus, but do not contribute to maximizing percent correct. This discrepancy also explains why human performance was inefficient according to the objective definition of the task.

The results from model $\mathcal{L}_1$ suggest that participants' implicit goals in the experiment were not to answer each trial correctly, but rather to select a response category that is as close as possible to the correct response. This distinction is subtle but important. The latter objective implies a metric representation of stimuli and responses. Model $\mathcal{L}_2$ (Fig. 5b) embodies this by using the cost function $\mathcal{L}_2 = |y - x|$, where y and x are the ordinal response and stimulus categories, respectively. Hence, according to model $\mathcal{L}_2$, the best outcome is to generate the correct response, as this would achieve zero cost. Failing that, the cost function suggests producing a response that is as close as possible, where distance is linear in the ordinal range of responses. Fig. 5b shows that this model exhibits hints of the bow effect (higher performance for stimuli at the extremes of the stimulus range). Accuracy for the model is also lower than model $\mathcal{L}_1$, and closer to human performance, even though both models convey the same amount of information. However, model $\mathcal{L}_2$ still fails to capture empirical performance at a detailed quantitative level. Model $\mathcal{L}_2$ used a cost function defined over the ordinal stimulus and response categories, but similar predictions are obtained if the stimulus values (the actual line lengths) are used instead.

In changing the cost function between model $\mathcal{L}_1$ and $\mathcal{L}_2$, the theoretical status of the model has also changed in a subtle, but important manner. The first model represents a quantitative, formal, and falsifiable prediction for human performance. Indeed, based on the quantitative inaccuracy of this model, one can confidently reject it as a sufficient explanation for human perceptual categorization. Importantly, model $\mathcal{L}_2$ is equally falsifiable, in the sense that it also makes quantitative predictions that can be compared against actual data. However, its status as a normative model is now less clear-cut. There is no reason to believe that participants *should* seek to minimize the absolute error in perceptual categorization. The goal is to instead assess the descriptive adequacy of a particular hypothesized cost function, rather than ascribe to it normative status.

Model $\mathcal{L}_3$ also assumes a metric representation of the stimulus and response space. But unlike $\mathcal{L}_2$, it is not assumed that this space is linear with respect to either ordinal rank or physical line length. In particular, it may be the case that $\mathcal{L}(1,2)$ is higher or lower than $\mathcal{L}(4,5)$. To allow for this possibility, model $\mathcal{L}_3$ assumes that each ordinal stimulus or response category has an associated 'anchor' in psychological space. An anchor is essentially the mental representation of a particular stimulus. The resulting cost of an error is the distance between two anchors in this space, or $\mathcal{L}_3 = |\alpha_{(y)} - \alpha_{(x)}|$, where $\alpha_{(j)}$ is the location of the anchor corresponding to ordinal stimulus or response category j. The assumption that physical stimuli exist in a psychological metric space is in fact common to many models of absolute identification (reviews of several different modeling approaches are given in Shiffrin & Nosofsky, 1994; Stewart, Brown, & Chater, 2005).

The assumption of this model is that both perceptual stimuli and ordinal responses are mapped into a common metric space. This formulation constrains the cost function to be symmetric such that $\mathcal{L}_3(x,y) = \mathcal{L}_3(y,x)$, but this is not a necessity of the mathematical framework. While models $\mathcal{L}_1$ and $\mathcal{L}_2$ have no free parameters, in $\mathcal{L}_3$ the locations of the anchors are estimated from the empirical data via maximum likelihood estimation (Myung, 2003). The likelihood function for the model is given directly by the conditional distribution $Q^{\star} = P(y|x)$. The anchors for the smallest and largest stimuli were constrained to lie at 0 and 1, respectively. This results in $N - 2$ free parameters in the model, where N is the number of stimuli in the training set (N = 13, 20, or 30).

The relatively large number of parameters in this models warrants brief discussion. Models $\mathcal{L}_2$ and $\mathcal{L}_3$ represent a progressive shift away from normative models of performance, and towards descriptive accounts of how the perceptual system actually performs. By fitting the cost function to human performance, one can understand performance in terms of the implicit cost function that it seeks to minimize. This approach is also known as *inverse decision theory* (Körding, Fukunaga, Howard, Ingram, & Wolpert, 2004), where the goal is to infer or estimate a utility function starting from the observation of behavioral preferences. As the number of free parameters in a model increases, the falsifiability of the model decreases assuming that parameters are fit to data, rather than specified in advance as a specific point hypothesis. A completely descriptive account of the perceptual cost function could be obtained by estimating $\mathcal{L}(x,y)$ separately for each possible (x,y) combination. Importantly however, this does not negate the utility of a descriptive model of human perceptual processing.

The fit of this model is illustrated in Fig. 5c. The model accurately captures both the average proportion correct, as well as the distribution of errors (not shown). The estimated anchor positions for each of the ordinal stimuli are shown in Fig. 6a. The spacing between anchors is expansive at the edges of the stimulus range and compressive in the middle. When the anchor locations are replotted as a function of the fraction of the physical stimulus range (such that 0 and 1 correspond to the smallest and largest physical line lengths), the anchors from all three set size conditions essentially lie on a common curve (Fig. 6b). Consequently, nearly all free parameters in the model could feasibly be replaced by a parametric function involving a small number of parameters. Further, these parameters appear invariant to the number of stimuli in the training set. Fig. 7 illustrates the resulting cost function for the $N = 30$ condition. The costs of error are flattened in the middle of the stimulus range, and steepest at the two extremes. Future research will be necessary to determine the extent to which this same cost function generalizes to other datasets and stimuli (including nonvisual dimensions).

Nearly all properties of model $\mathcal{L}_3$ are constant across the three conditions of the experiment. The primary factor driving the change in performance is the fixed limit on the information rate of the channel. As the entropy of the information source increases, a capacity-limited channel necessitates that some amount of information is lost. Rate–distortion theory describes the optimal performance that can be achieved in this circumstance. Human performance is parsimoniously explained as minimizing the cost of identification error, according to a particular (and previously unobserved) cost function, while subject to a constraint on channel capacity.

Absolute identification has long been argued to be a task where human performance faces a strong limit on information-theoretic capacity. If capacity is limited, and if the ability to discretely categorize stimuli is important in our evolutionary history, then it is reasonable to suppose that the brain might make efficient use of the capacity that is available. Efficiency might be defined as reducing the redundancy of neural signals (Barlow, 1961; Olshausen & Field, 1997),

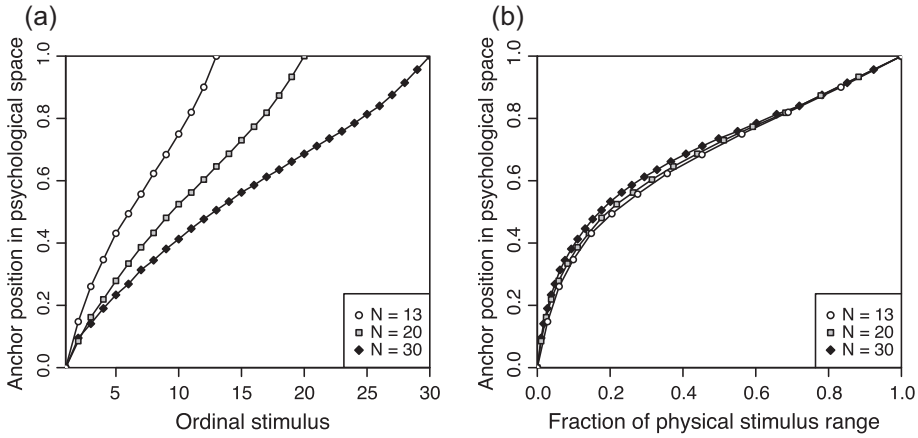

Fig. 6. Estimated anchor positions from model $\mathcal{L}_3$. (a) Anchor positions plotted as a function of the ordinal rank of the stimuli in each condition. (b) The same data, with anchor position replotted against fraction of the total physical stimulus range.

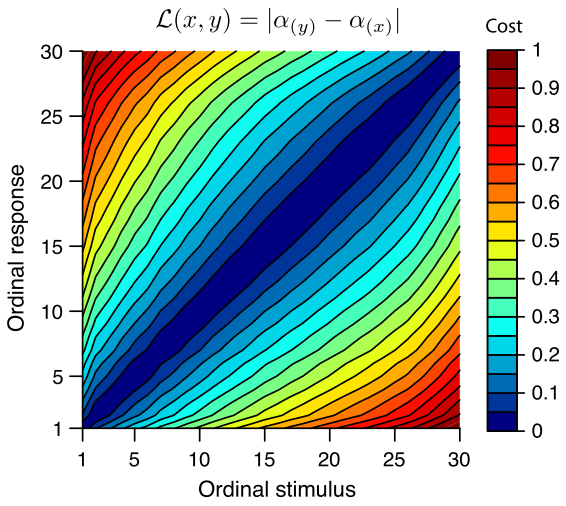

Fig. 7. Resulting cost function for the N = 30 condition, according to the estimated anchor locations.

but this perspective overlooks the behavioral cost of error as a driving factor in perception. Rate–distortion theory offers a means of recovering (via inverse decision theory) the implicit cost function based on empirical performance. The results in this section have shown that this cost function is highly conserved across different set size conditions. Future work will be necessary to determine if this same cost function generalizes across different stimulus dimensions.

The model presented in this section was developed with the goal of theoretical and conceptual clarity, at the expense of explaining all aspects of human performance. Notably, human absolute identification is characterized by strong sequential dependencies between responses (Stewart et al., 2005). The current model could plausibly be extended to account for these effects by assuming that the implicit model of the statistics of the stimuli is biased by recent experience; an approach that has been successful in explaining sequential dependencies in other binary choice tasks (Jones, Curran, Mozer, & Wilder, 2013). The current framework represents a principled and theory-grounded means of studying how implicit statistical learning interacts with perceptual identification.

5. Rate–distortion theory and perceptual working memory

Visual working memory (VWM) is a highly limited storage system used for the temporary maintenance and manipulation of visual information. But *how* is it limited? In recent years there has been extensive and ongoing debate regarding the nature of this limit.

An early view, advanced most famously by Miller's "magic number" paper (Miller, 1956), suggests that working memory is limited to maintaining a small number of discrete representations. Miller's original limit of seven items has subsequently been revised downwards to 3–4 items (Cowan, 2000) or chunks (Mathy & Feldman, 2012). These studies primarily examined memory performance for discrete, categorical items such as letters or digits. As applied to visual working memory, the analogous view posits that visual working memory consists of a small number of discrete 'slots', each of which is capable of maintaining a single visual object (Luck & Vogel, 1997). The key feature of this basic model is that an item is either remembered, subject to some fixed perceptual noise, or it is not (Luck & Vogel, 2013).

More recently, an alternative view has emerged suggesting that the fundamental limit in visual working memory is not the number of discrete slots, but rather a continuous resource (for a recent review, see Ma, Husain, & Bays, 2014). The distinguishing feature of this model family is that there is a continuum in the fidelity of memory representations. Items that are encoded with a large amount of resource will exhibit high-fidelity or high-precision memory representations, whereas memory items with limited resource allocation will have poor memory precision or fidelity. An important question for this model family is the nature of the continuous resource that is limited. One intuitively plausible explanation suggests that the gain of neural activity maintaining memory representations is the limiting resource (Bays, 2014, 2015; van den Berg, Awh, & Ma, 2014).

In this section, a computational level (Marr, 1982) account of perceptual working memory is developed. Here memory resource is identified with channel capacity in the formal sense of information theory. Notably, this account does not contradict neural or mechanistic explanations for a limit in channel capacity. Information theoretic transmission is closely related to the limits on the gain in neural populations (Dayan & Abbott, 2001; Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1999). However, by adopting a computational-level theory it is possible to study and define optimal memory performance in the context of a cost function defined by an external task or environment—a perspective

that is obscured by focusing solely on internal constraints or mechanisms.

Unlike categorization tasks, perceptual working memory often requires the maintenance of a continuous-valued signal rather than a discrete categorical representation. For example, when visually comparing two apples to determine which is larger, visual working memory must maintain a metric rather than categorical representation of size. Despite this difference, the application of rate–distortion theory follows essentially the same pattern as in the previous section. Perceptual working memory is defined as a capacity-limited channel described by a conditional distribution $p(y|x)$. The goal is for this channel to minimize the expected cost defined by some function $\mathcal{L}(x,y)$. For a given information source (for example, a probability distribution over visual features), cost function, and limit on memory capacity, it is possible to derive the boundedly rational memory system. This can be used either as a model of human performance, or as a benchmark to compare performance against, following in the tradition of ideal observer analyses (Geisler, 2011).

For some combinations of information source and cost function, it is possible to analytically derive a rate–distortion curve and corresponding optimal channel. One such example is a Gaussian information source with a squared error cost function, $\mathcal{L}(x,y) = (y-x)^2$. In this case, the optimal channel takes the form of an additive Gaussian noise channel, with the magnitude of the noise determined by the constraint on information rate. Details of this case are given in Appendix A, and a model of visual working memory based on this approach has previously been described (Sims, Jacobs, & Knill, 2012). However, the empirical cost function for perceptual memory is largely unknown, and existing data suggests that it may deviate substantially from a quadratic function (Sims, 2015). Consequently, strong assumptions regarding the parametric form of the cost function may be unwarranted. In the current section, the data from Sims et al. (2012) are re-analyzed in the more general case where both the distribution of visual stimuli and cost function are unconstrained.

In particular, Sims et al. (2012) conducted an experiment on visual working memory for line length. Participants viewed displays containing 1, 2, 4, or 8 line segments of varying length. After a brief memory retention interval, a new 'probe' stimulus was displayed at the location formerly occupied by one of the memory items. The participant was asked to report whether the new stimulus was shorter or longer than the remembered item (a 2-AFC procedure). The probe stimulus length was determined according to $x+\Delta$, where x indicates the study line length, and Δ was randomly chosen from the set $\{-1,-0.5,-0.25,-0.075,+0.075,+0.25,+0.5,+1\}$ cm. The line lengths were sampled from a log-normal distribution,

$$p(x) = \log \mathcal{N}(\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right) \tag{9}$$

with parameter $\mu = 1.1757$. In two conditions the variance of this distribution was manipulated. In the low variance condition, $\sigma_{\text{low}} = 0.0748$ and in the high variance condition $\sigma_{\text{high}} = 0.299$. Complete methodological details can be found in Sims et al. (2012).

In this experiment, participants were instructed to answer each trial as accurately as possible. If a line length x is mis-remembered as a different length y, what is the task-defined cost of this error? The experiment used a fixed set of perturbations, Δ. If the remembered line length y is less than the probe stimulus, $x+\Delta$, then the probe stimulus will appear longer and the participant should respond as such. This response would be correct if on that particular trial $\Delta > 0$. Following this reasoning, the true cost function for the task can be written as

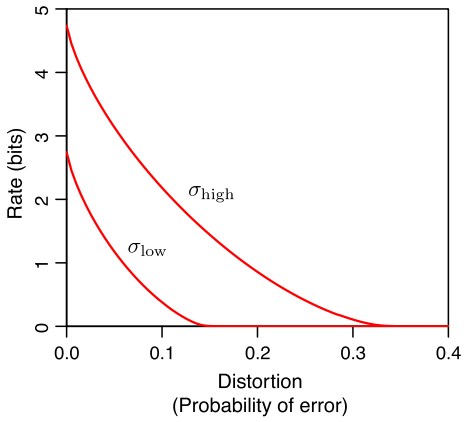

Fig. 8. Rate–distortion curves for a visual working memory experiment using a 2-AFC procedure and a probability of error cost function. The two curves correspond to the two stimulus variance conditions of the experiment.

$$\mathcal{L}(x,y|\Delta) = \begin{cases} 0 & \text{if } (y < x+\Delta) \wedge (\Delta > 0), \\ 0 & \text{if } (y > x+\Delta) \wedge (\Delta < 0), \\ 1 & \text{otherwise.} \end{cases} \tag{10}$$

$$\mathcal{L}(x,y) = \sum_{\Delta} \mathcal{L}(x,y|\Delta) P(\Delta) \tag{11}$$

Fig. 8 illustrates the optimal rate–distortion curves for this experiment, computed using the software package described in Appendix B. This figure demonstrates a strong prediction of the theory, namely that for an efficient memory system, increasing the variance of the information source will necessarily lead to a decrease in performance. This is depicted visually by the fact that at a fixed information rate (a horizontal line through the graph) the minimum achievable distortion depends on the variance of the information source. The rate–distortion curves can also be used to estimate an assumption-free lower bound on human memory capacity in each condition of the experiment. However, the tightness of this bound depends on whether the cost function used to generate the rate–distortion curve matches the goals of the individual performing the task.

In Section 4, it was possible to directly estimate the mutual information between the information source and participants' responses, and use this to compute information-theoretic efficiency via Eq. (8). This was possible because the experiment used a small number of discrete stimuli and responses. In the current case, the stimuli are drawn from a continuous probability distribution. Estimating $H(x,y)$ in this case would require either an infeasibly larger number of trials or else making strong assumptions regarding the nature of the conditional distribution $p(y|x)$. However, it is still possible to use rate–distortion theory to develop efficient models of visual working memory by estimating memory capacity from the observed level of human performance. If $Q^\star = p(y|x)$ represents an optimal capacity-limited channel, then for a given stimulus x, the likelihood of the participant responding that the probe stimulus is larger is given by the probability that $y < x+\Delta$.

For an ideal memory system, the corresponding memory channel will be optimized for the statistics of the information source as well as the cost function. It may be the case however that a person's implicit knowledge of sensory statistics may not match the true stimulus statistics. It may also be the case that the implicit cost function deviates from the task-defined cost of memory error (as in the case of absolute identification). To examine this possibility, both the cost function as well as the distribution of the stimuli can be treated as parameters estimated from the data. For the

current experiment, a parametric family of cost functions is considered with the form

$$\mathcal{L}(x,y) = \frac{|y-x|^{\beta}}{|y-x|^{\beta} + \alpha^{\beta}}. \quad (12)$$

The parameters α and β determine the shape of the cost function. All curves in this family are monotonically increasing and reach an asymptote at 1. The parameter α determines the memory error at which the cost reaches half its maximum, $\mathcal{L} = 0.5$. The parameter β determines the slope or steepness of the cost function. A similar family of cost functions was used to model visual working memory for other features such as color or line orientation (Sims, 2015).

The implicit stimulus distribution is assumed to be a log-normal distribution with parameters μ and σ. This is also the distribution actually used to generate stimuli in the experiment (separate analyses show that a log-normal distribution offers a better model of the implicit statistics than a Gaussian distribution). By estimating the parameters μ and σ rather than assuming the veridical values it is possible to examine the extent to which visual working memory adapts to the statistics of the information source, and the extent to which performance is explainable by a mismatch between assumed and veridical statistics.

Lastly, different trials of the experiment varied the set size (the number of stimuli that were presented simultaneously). It is to be expected that as more items are held in memory, less capacity will be available to encode or represent each item (Sims, 2015; Sims et al., 2012). Hence, the constraint on memory capacity is also estimated separately for each set size condition. Capacity is assumed to be independent of the distribution of stimuli (low versus high variance condition), and the cost function and statistics of the information source as assumed to be independent of the set size. All parameters were fit to the combined data from all participants via maximum likelihood estimation.

Fig. 9 illustrates the main features of human performance in this experiment, as well as the resulting model fit (red curves). Each panel plots the proportion of trials where the subject reported that the probe item was larger than the corresponding memory item, for each value of the perturbation magnitude Δ. Psychometric curves are plotted separately for each quartile of the stimulus distribution (shown as the four separate curves in each panel).

The most notable feature is that performance decreases with increasing set size, indicated by the flattening of the psychometric curves. However, there are other important properties of the data that are captured by the model. At a given set size, performance is worse in the high variance condition of the experiment (comparing top versus bottom rows). In addition, the response for a fixed value of Δ is biased by the initial stimulus that was presented. Unusually short line segments are less likely to be reported as being longer when probed, and vice versa for the longest line segments. Lastly, both the effects if the stimulus variance, as well as the dependence on the remembered stimulus value strongly interact with set size (both effects are largest at high set sizes). All of these effects are explained by rate–distortion theory without additional mechanistic assumptions.

Fig. 10 shows the estimated memory capacity per item for each set size condition of the experiment. Note that this is the estimated capacity limit for the implicit statistics of the information source, and not the amount of information conveyed for the true distribution of stimuli. If a fixed memory capacity were perfectly shared among all encoded items, estimated memory capacity would be inversely proportional to the set size, $R \propto 1/N$. This prediction is shown for comparison by the dashed curve in Fig. 10. The match between the estimated and predicted capacity is reasonably close

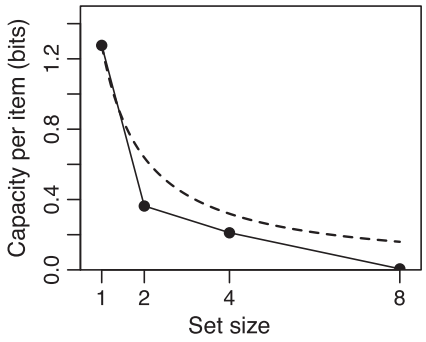

Fig. 10. Estimated memory capacity from each set size condition is shown by the black line and marker points. Perfect memory sharing would predict that capacity should be inversely proportional to the set size, as shown by the dashed line.

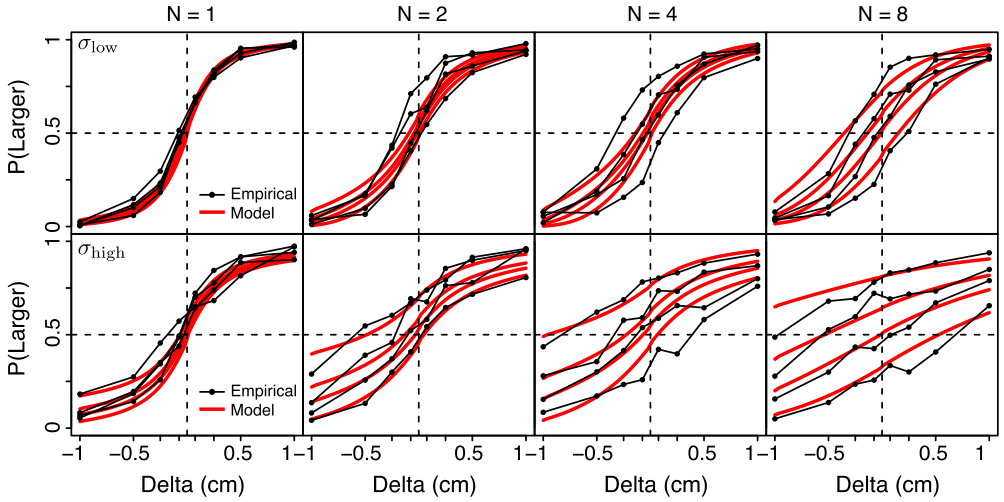

Fig. 9. Psychometric performance on a visual working memory task examining memory for line length, data from Sims et al. (2012). Participants were tasked with detecting whether a probe item was larger or smaller than a remembered stimulus. Each column shows memory performance for a different set size. Stimuli were drawn from two different source distributions, a low variance (top row) or high variance condition (bottom row). Empirical data is indicated by plot markers and black lines. Model fit is indicated by smooth red curves. Psychometric curves are plotted for each quartile of the stimulus range (shown as the four separate curves in each panel). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

but not exact. This suggests that human visual working memory is reasonably, but not perfectly efficient at dividing available memory resources among multiple items, a finding that has been replicated across a large number of datasets (Sims, 2015). The capacity estimates reported here are substantially lower than those previously reported in Sims et al. (2012). Two factors account for this difference. One is that the current model does not attempt to account for sensory noise, but rather folds this into the estimate of memory capacity. The other factor is that the model reported in Sims et al. (2012) allowed for variability in the number of items encoded. In the current model, capacity estimates reflect a mixture of memory-based responses, and trials where the tested item was not encoded in memory. Including these latter responses will necessarily bring down total memory capacity estimates. There are clearly many possible extensions to the model. The current implementation was chosen to highlight the formal elements of rate–distortion theory: the statistics of the information source, the cost function, and the constraint on capacity. However, formal model comparisons (Sims, 2015) have shown that even this simple model family offers a superior quantitative account of human memory performance compared to currently available alternative models.

The empirically estimated cost function from the dataset is shown in Fig. 11 (red curve), based on the parametric family described by Eq. (12). The estimated cost function is compared to the task-defined cost function (step-like function, black curve) from Eq. (11), as well as a simple quadratic cost function (blue curve) that was previously assumed by (Sims et al., 2012). The empirical cost of error rises more steeply for small errors compared to a quadratic function, but saturates for larger errors. This pattern has also been replicated across a number of datasets (Sims, 2015), as well as in the case of estimating the implicit cost function for sensorimotor errors (Körding & Wolpert, 2004). Note however, that by assumption all of the cost functions considered are stimulus invariant. That is, they depend on the difference $y - x$, but not on the particular stimulus value x that is encoded. This assumption was shown to be inappropriate in the case of absolute identification (Fig. 7). Hence it is likely that a more complicated family of cost functions could improve the accuracy of the model.

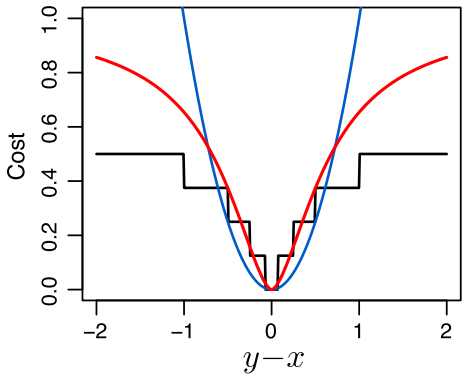

Fig. 11. Comparison of the task-defined cost function (step-like function) to a quadratic cost function (blue curve) and the implicit cost function estimated from human performance (red curve). The estimated cost function is based on the parametric family described by Eq. (12). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 12a plots the error distribution predicted by the model, that is the probability distribution for the quantity $y - x$ (shown by the blue shaded distribution). For comparison, the error distribution predicted by a quadratic cost function is also shown (shaded red). Note that the information constraint is the same for both cases, and only the cost function differs. The predicted error distribution exhibits a sharper peak, and heavier tails compared to that predicted by a quadratic cost function. This difference is highlighted in Fig. 12b, which shows the residual, defined as the error distribution for the estimated cost function, minus the error distribution assuming a quadratic cost function. A highly similar pattern of residual also remains between the empirical error distribution and that predicted by a mixture of Gaussian and uniform responses (Zhang & Luck, 2008). This pattern of residual has also been observed in other datasets and used as evidence against a slot-based memory representation (van den Berg, Shin, Chou, George, & Ma, 2012).

In the current model, and unlike previous analysis of this dataset (Sims et al., 2012), the possibility was considered that the subjects' implicit understanding of the stimulus statistics might differ from the true generative model. Fig. 13 compares the distribution used to generate the stimuli with the estimated implicit distribution, for each variance condition of the experiment. The inferred distributions are substantially broader than the true stimulus distribution. These differences render human participants strictly suboptimal at the task: given the same memory capacity, higher performance could be achieved by adopting an implicit source distribution that matched the task-defined statistics exactly. This discrepancy may reflect incomplete statistical learning on the part of the experimental participants. In this case, participants completed

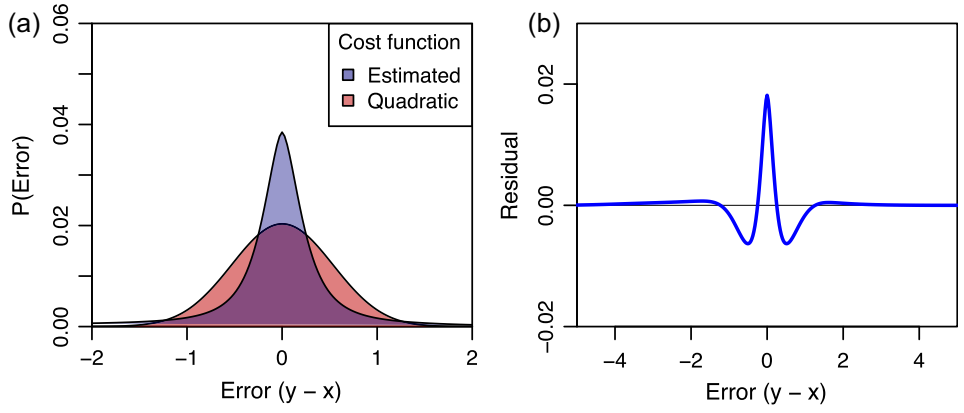

Fig. 12. (a) Error distribution, or the probability distribution of the quantity $y - x$, under either a quadratic cost function (red) or the cost function estimated from the empirical data (blue). (b) Residual, defined as the difference between the error distribution predicted by the empirical and quadratic cost functions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

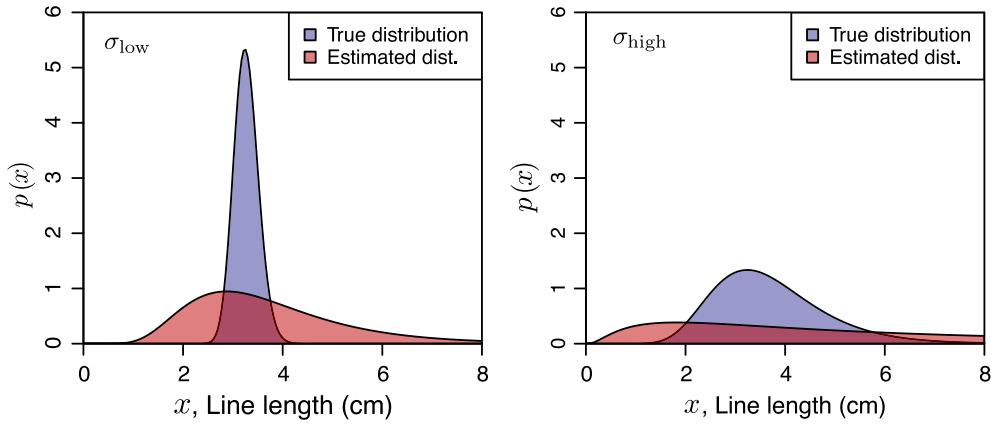

Fig. 13. Comparison of the distribution used to generate stimuli in the low and high variance conditions (blue shaded distributions, left and right panels) to the estimated implicit distribution used by visual working memory (red shaded distributions). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

only one experimental session in each variance condition. Part of the nature of perceptual expertise may involve fine tuning implicit knowledge of the statistics of task-relevant information.

Despite the difference between the true and estimated distributions, there is evidence that human memory performance adapts to the statistics of the task. The variance of the implicit distribution is higher in the high-variance condition of the experiment, as would be expected by an adaptive memory system. The means of the inferred distributions are somewhat lower than the true distribution (estimated $\mu = \{1.06, 0.60\}$ for the low and high variance conditions, true $\mu = 1.18$).

Another possibility is that the broader distributions shown in Fig. 13 may reflect conservatism on the part of visual memory. Optimizing memory performance for a narrow range of stimuli would preclude the accurate memory of stimuli that unexpectedly fall outside of this range. Future research would be necessary to untangle these possibilities. The most important point illustrated by Fig. 13 is that human visual memory cannot meaningfully be separated from the study of implicit statistical learning (Orhan, Sims, Jacobs, & Knill, 2014).

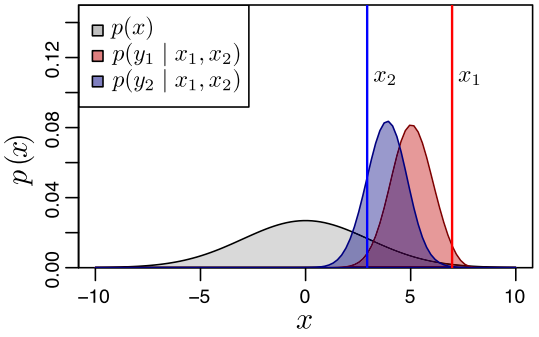

Fig. 14. Illustration of dependencies among items encoded in visual working memory. Stimuli x_1 and x_2 are drawn independently from a Gaussian distribution. However, visual working memory assumes a correlation between features. As a result, the recalled values y reflect a bias towards the across-trial mean of the stimuli, as well as the within-trial mean of the two stimuli. This is illustrated by the fact that the channel distributions, $p(y_1 | x_1, x_2)$ and $p(y_2 | x_1, x_2)$, indicating the distribution of recalled values, are shifted towards the mean of the two stimuli x_1 and x_2.

5.1. Inter-item dependencies in visual working memory

In the preceding section, it was assumed that each stimulus item is encoded independently in visual working memory. The recalled memory representation is in general biased towards the mean of the stimulus distribution (with the exact magnitude of the bias depending on memory capacity and the particular cost function). However, conditioned on the stimulus distribution, each memory representation is statistically independent. As a number of researchers have demonstrated, assuming complete independence among memory items neglects empirically-observed dependencies (Brady & Alvarez, 2011; Jiang, Olson, & Chun, 2000; Orhan & Jacobs, 2013). For example, when shown two different stimuli, memory recall might be biased towards the mean of the two stimuli (a perceptual magnet effect). In this section, a toy example is given to illustrate this effect and show how it can be explained within the rate–distortion theory framework.

To account for inter-item dependencies, Orhan and Jacobs (2013) developed the *probabilistic clustering theory* of visual working memory. According to this model, visual memory assumes an incorrect generative model for visual information, according to which individual stimuli are drawn from clusters (using a nonparametric Bayesian prior over cluster assignments). The consequence of this assumption is akin to a form of visual chunking. Stimulus items that are nearby exhibit a magnet effect, such that memory representations are biased towards the mean of each cluster.

The idea that the brain may adopt an implicit understanding of visual statistics that differs from the true statistics is not incompatible with the current theoretical framework (as illustrated by Fig. 13). Indeed, probabilistic clustering theory can be incorporated in its entirety within rate–distortion theory as an additional assumption on the implicit prior over visual stimuli. Fig. 14 illustrates this approach using a simplified example. In this example, rather than independently encoding features, a single memory channel is used to simultaneously encode and recall two features, x_1 and x_2. For example, x_1 and x_2 might correspond to two line segments of different length presented to the observer to store in visual working memory.

In the example, these features are generated independently from a Gaussian distribution with mean = 0 and SD = 3. However, the memory channel assumes that these features are correlated (in this example using a correlation coefficient $\rho = 0.5$). The cost function is assumed to minimize the summed squared error for x_1 and x_2. Fig. 14 demonstrates that the distribution of recalled values, y_1 and y_2 exhibit both a bias towards the prior distribution of the stimuli, as well as inter-item dependencies (a bias towards the mean of x_1 and x_2). The figure demonstrates that visual working memory is biased towards the mean of the two presented stimuli,

as represented by the shift of the conditional distributions toward the midpoint of x_1 and x_2. Without assumed correlations between stimuli, the recall distributions would be biased towards the overall mean, $p(x)$, but not towards the within-trial mean of x_1 and x_2.

Note that this example is somewhat exaggerated for the sake of visualizing the inter-item dependencies. In addition, the simple assumption of correlated features could be replaced by a more complicated statistical model such as the clustering theory proposed by Orhan and Jacobs (2013). The only major limitation is that of computational complexity of determining an optimal information channel.[5]

6. Discussion

What is the place of rate–distortion theory in understanding the nature and limits on human perception? Clearly the answer to this question depends on one's goals. Marr (1982) suggested a three-level hierarchy for understanding human vision: the implementation, algorithmic, and computational level. These three levels represent a successive shift from explanations focused on the internal (neural architecture) to explanations focused on the external (the structure of the environment and the high level goals of the system). The motivation behind the present paper has been largely to argue on behalf of rate–distortion theory (Berger, 1971; Shannon, 1959) as a useful and productive theoretical framework for studying human perception at the computational level. Rate–distortion theory combines the central elements of both information theory and decision theory, and is uniquely situated for explaining biological computation as a principled, but capacity-limited system. As a computational-level theory, the goal is not to contradict explanations formulated at the neural or algorithmic level, but rather provide an explanation for the 'why' of behavior, and provide inspiration for the development of mechanistic theories.

The application of information theory to human perception is an idea nearly as old as information theory itself (Attneave, 1954; Barlow, 1961; Garner & Hake, 1951; Miller, 1956). However, rate–distortion theory contributes something qualitatively different to this area. In particular, a fixed information transmission limit is not sufficient as an explanation for behavior. If human perception is to be described as an information channel, it is also necessary to understand the goal of communication. In this regard, rate–distortion theory contributes an important perspective to the study of human perception. The objective for biological information processing is not (merely) the communication of information, but rather the minimization of relevant costs. Information is simply a means to an end. The incorporation of costs into the framework of information theory defines rate–distortion theory.

This framework was applied to two benchmark domains in the study of human perception: absolute identification (the assignment of perceptual stimuli to ordinal categories) and perceptual working memory. The study of absolute identification has long been concerned with apparent capacity limits, without regard for the goal of information processing. The most intuitive goal—that of maximizing performance on a typical absolute identification experiment—offers a poor account of human performance. It is only by extending information theory to account for the goals of information processing (as captured by a cost function) that a quantitatively accurate model of human performance emerges. The current paper represents the first time the implicit cost function for human absolute identification has been investigated. At the least, this represents a new means of characterizing human performance. However, this result also represents a starting point for a much broader set of research questions, for example, examining the extent to which task instructions or extensive practice might alter the implicit cost function for perceptual identification.

The goal in developing this model was not to propose a highly detailed model of human absolute identification. In fact there are important behavioral phenomena that are not explained by the current model. Human responses show strong sequential dependencies, including assimilation of responses towards recent stimuli (reviewed in Stewart et al., 2005). Although not explained by the current model, these findings are not fundamentally incompatible with rate–distortion theory. Sequential dependencies can arise in the current framework through stimulus-dependent changes in either the cost function or assumed statistics of the information source. Indeed, it is an important element of the framework that one's implicit understanding of the statistics of the environment may not match the ecological statistics. The current approach allows for quantifying and studying mismatch in statistical learning in a direct and principled manner.

The second half of the paper examined the use of rate–distortion theory in modeling visual working memory. Much of the history of research in visual working memory is characterized by debates between alternative implementation and algorithmic-level models of performance (reviewed in Luck & Vogel, 2013). The present goal was the development of a computational-level exploration of visual working memory performance. The most important element that accompanies this shift in perspective is the emphasis on the behavioral or biological costs of memory error, something that is lost in focusing purely on mechanistic explanations. This approach has been employed previously (Orhan et al., 2014; Sims, 2015; Sims et al., 2012). What is new in the current paper is the simultaneous estimation of the implicit cost function for behavior, as well as the implicit model or understanding of the statistics of the visual environment. Thus, rate–distortion theory defines three critical components that limit human perceptual memory: information-theoretic limits on memory capacity, mismatches between ecological statistics and implicit statistical learning (see also Orhan & Jacobs, 2014), and mismatches between task-defined and implicit cost functions. As demonstrated in the current paper, each of these components can shed light on the nature of human visual working memory. Beyond the scope of simple laboratory tasks, these three components can be used as a novel and principled means of characterizing or understanding the relationship between perceptual memory and perceptual expertise (Curby, Glazek, & Gauthier, 2009; Herzmann & Curran, 2011). Intuitively, perceptual experts such as radiologists or satellite image analysts should have implicit cost functions that more closely align with the costs of error in their domain of expertise, as well as statistical learning that reflects the veridical structure of the environment. This area repersents a promising direction for future research.

6.1. Relationship to Bayesian models of perception

The most obvious alternative to the current approach is the framework of Bayesian inference (Knill & Richards, 1996; Körding, 2007; Maloney & Mamassian, 2009). However, it would be misleading to characterize the two approaches as incompatible. Both Bayesian inference and rate-distortion theory are concerned with optimal inferences and minimizing cost. The primary difference lies in the origin of the constraints on performance. Bayesian models of perception often incorporate two qualitatively different types of uncertainty. The first is fundamental uncertainty due to ambiguous information available to the observer. For example,

[5] In this example, the distribution $p(x_1, x_2)$ is treated as discrete with 100 bins per feature dimension. As the number of features increases, the size of $p(x)$ scales exponentially. This suggests the utility of developing Monte Carlo techniques that can be applied to rate–distortion problems (e.g., Jalali & Weissman, 2008).

the 3D shape of an object cannot unambiguously be reconstructed from its 2D projection on the retina (Kersten et al., 2004); this uncertainty is inherently external to the organism. A second type of uncertainty frequently encountered in Bayesian models of perception is internal noise. For example, one might model the sensory encoding of a stimulus as a Gaussian noise-corrupted version of the afferent signal (Orhan & Jacobs, 2013; Stocker & Simoncelli, 2006). In this latter case, the framework of Bayesian inference is agnostic regarding *why* signals should be noisy, or what form that noise should take. In particular, what is 'rational' about Gaussian noise (as opposed to some other noise distribution)? In this regard, rate–distortion theory takes a small, but important step beyond Bayesian models of perception.

Specifically, rate–distortion theory assumes that physical communication is necessarily capacity-limited, and that the goal of information transmission is to minimize some particular cost function. Neither of these assumptions is particularly controversial in psychology or neuroscience. However, this formulation generates specific predictions for the nature of encoding noise, rather than treating noise in an atheoretical manner. For example, it has been shown that rate–distortion theory offers a parsimonious explanation as to why the empirical error distribution in visual memory is non-Gaussian (Sims, 2015). By optimizing the channel encoder to the task, rate–distortion theory is, in one sense, more optimal than Bayesian models that may make arbitrary assumptions about the nature of noise (such as the common Gaussian noise assumption) and only perform optimal inference (decoding).

6.2. Falsifiability and normative status of the framework

Rate–distortion theory is not a specific model of human perception, but rather a general theoretical framework in which hypotheses can be formulated and tested. This property is also shared by the Bayesian approach to understanding perception or cognition. In recent years, Bayesian models of cognition have been critiqued in terms of their falsifiability (Bowers & Davis, 2012a, 2012b) due to the potential for arbitrary choices regarding the prior, likelihood, and/or utility function specified in a given model. Given their similarity, models developed in an information-theoretic framework are of course also subject to such critiques. However, as has been pointed out elsewhere (Hahn, 2014), charges of arbitrary assumptions may be leveled against *any* mathematical modeling approach. Hence, it is more appropriate to talk about the falsifiability of specific models, rather than of an entire theoretical framework. Indeed, the central element of rate–distortion theory—that physical communication channels must possess only finite capacity—is closer in status to a scientific law than a hypothesis.

In the current paper, rate–distortion theory has been used to develop two qualitatively different kinds of models of perception. The first represents a strong prediction for boundedly rational behavior. In this case, the cost function is fixed by the environment or the task that is being performed, and the only assumption is that the individual seeks to maximize task performance subject to a constraint on channel capacity. Such models are clearly falsifiable, as demonstrated in the current paper (Fig. 5). In the second class of models, the goal is not a normative justification for behavior, but rather a descriptive explanation for behavior at Marr's computational level. Models in this category do not seek to test or reject specific hypotheses regarding the perceptual cost function, but rather intend to accurately describe or infer the cost function. This approach is essentially the application of inverse decision theory (Körding et al., 2004) to human perception. Both classes of models can exist comfortably within the same theoretical framework. However, it is essential that the theoretical status for a particular model is clear.

An explanation for behavior must include an understanding of the goals of the organism, the environment in which it must behave, and the constraints (both internal and external) that limit performance. The greatest strength of rate–distortion theory as applied to human perception is that it enables the specification of 'ideal observer' models of performance (Geisler, 2011) that formally incorporate each of these elements, as well as new tools for the quantitative analysis and description of perception. While framed at the computational level, the ultimate goal of this approach is to constrain and inspire both the algorithmic and mechanistic understanding of human perception.

Acknowledgements

CRS was supported by a Drexel University Career Development Award and a Drexel ExCITe Center Seed Grant. This paper is dedicated to the memory of David Knill.

Appendix A. Solving rate–distortion problems

This section briefly describes four approaches to solving rate–distortion problems. A rigorous treatment of the first three approaches can be found in Berger (1971). The final approach described in this section, based on the Blahut algorithm (Blahut, 1972), is also the favored method due to its simplicity, generality, and computational efficiency. A software package implementing this algorithm, written in the R programming language, is made available. Details on installing and using this package are given in Appendix B.

A.1. Numerical solution via convex optimization

For discrete signals, the fundamental optimization problem can be restated as follows:

$$\begin{aligned} \text{Minimize} \quad & \sum_i \sum_j Q(y_j|x_i) P(x_i) \log \frac{Q(y_j|x_i)}{Q(y_j)} \\ \text{subject to} \quad & Q(y_j|x_i) \geqslant 0 \quad \text{for all } x_i, y_j, \\ & \sum_j Q(y_j|x_i) = 1 \quad \text{for all } x_i, \\ & \sum_i \sum_j \mathcal{L}(x_i, y_j) Q(y_j|x_i) P(x_i) = D. \end{aligned} \tag{A.1}$$

This problem corresponds to finding the minimum channel capacity necessary to achieve a desired level of performance (the mutual information of the channel is being minimized). Note that this differs from the statement given in the main text, where the goal instead was to minimize cost subject to a capacity constraint. In fact, the current formulation defines a distortion–rate function, the mathematical inverse of the rate–distortion curve considered in the main text. The only difference is the exchange of the independent and dependent variables. Since rate–distortion functions are monotonically decreasing, the same curve would result from minimizing distortion subject to a capacity constraint, as minimizing capcity subject to a distortion constraint (Berger, 1971).

The first two constraints in Eq. (A.1) arise from the fact that Q must represent a valid conditional probability distribution. This problem falls into the general class of convex optimization problems, since mutual information is a convex function of Q. Hence, there are many general purpose solution methods available. Readers may consult (Boyd & Vandenberghe, 2004) for a general introduction to convex optimization.

A.2. Analytical solution via Lagrange multipliers

The major difficulty in obtaining an analytical solution to Eq. (A.1) is the inequality constraint requiring that all probabilities be non-negative. An alternative solution method is to (temporarily) ignore this constraint, leaving a minimization problem with only equality constraints. The solution to this modified problem can then be checked to ensure that all probabilities are non-negative. To solve this problem, we construct an augmented function:

$$J_{(Q)} = I_{(Q)} - \sum_i \mu_i \sum_j Q(y_j|x_i) - s \sum_i \sum_j \mathcal{L}(x_i, y_j) Q(y_j|x_i) P(x_i), \quad \text{(A.2)}$$

where μ_i and s are Lagrange multipliers associated with the equality constraints. This augmented function can be solved by setting its derivative equal to zero, and solving for $Q(y_j|x_i)$, μ_i, and s. The result is a set of j simultaneous equations involving the unknowns q_j:

$$\sum_i \frac{P(x_i) e^{s\mathcal{L}(x_i,y_j)}}{\sum_k q_k e^{s\mathcal{L}(x_i,y_k)}} = 1 \quad \text{(A.3)}$$

For any chosen value of s ($s < 0$) the above can be solved (in principle) for the values q_j. If $q_j > 0$ for all j, the optimal channel can be obtained according to

$$Q(y_j|x_i) = \frac{q_j e^{s\mathcal{L}(x_i,y_j)}}{\sum_k q_k e^{s\mathcal{L}(x_i,y_k)}}. \quad \text{(A.4)}$$

The optimal rate–distortion curve can also be computed as a parametric function of s according to

$$D(s) = \sum_i \sum_j \lambda_i P(x_i) q_j e^{s\mathcal{L}(x_i,y_j)} \mathcal{L}(x_i, y_j), \quad \text{(A.5)}$$

$$R(s) = sD + \sum_i P(x_i) \ln \lambda_i,$$

$$\text{with } \lambda_i = \left(\sum_k q_k e^{s\mathcal{L}(x_i,y_k)} \right)^{-1}.$$

The parameter s turns out to be the slope of the rate–distortion curve at the coordinates $(D(s), R(s))$. Since rate–distortion curves are monotonically decreasing, the slope of the curve, and hence s must be negative. Note that the above computes the information rate in nats. To convert to bits it is necessary to multiply $R(s)$ by $\log_2(e) \approx 1.442695$.

Example: We wish to design an optimal, but capacity-limited channel for communicating the outcome of rolling a six-sided die. The cost function to be minimized is the probability of error (also known as the Hamming distance function), as given by Eq. (6). In this case, $P(x_i) = 1/6$ for $i \in 1 \ldots 6$. Applying Eq. (A.3) generates a system of 6 simultaneous equations. This system is most easily be solved with a computer algebra system such as Mathematica. In the present case, this yields the particularly simple solution $q_j = 1/6$ for all j. Substituting this result in Eq. (A.5) gives:

$$D(s) = \frac{5e^s}{6\left(\frac{1}{6} + \frac{5e^s}{6}\right)},$$

$$R(s) = s\frac{5e^s}{6\left(\frac{1}{6} + \frac{5e^s}{6}\right)} + \log\left(\frac{1}{\frac{1}{6} + \frac{5e^s}{6}}\right).$$

Solving to eliminate the implicit parameter s results in the solution given in the main text, Eq. (7).

A.3. Analytical solution for continuous information sources

Rate–distortion theory is not limited to discrete input and output alphabets. The same framework applies to continuous signals, although unfortunately in many cases no analytical solution can be obtained. The general approach is to follow the pattern for the discrete case. The results already presented in the previous section are applicable merely by replacing discrete summations with integrals. One exception however is the interpretation of the entropy of a continuous random variable, $H(x)$. For such a random variable described by a probability density function $p(x)$, the *differential entropy* is given by

$$H(x) = \mathrm{E}[-\log p(x)] = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx. \quad \text{(A.6)}$$

Importantly, this quantity does not indicate the absolute amount of information contained in a source, but rather the relative entropy compared to the coordinate frame (the differential entropy is zero for a uniform distribution over the unit interval). This is highlighted by the fact that differential entropy, unlike its discrete counterpart, can be negative.

With the above caveat, rate–distortion problems can be solved for continuous information sources by forming the continuous analog of the Lagrangian in Eq. (A.2), replacing summations by integrals. The optimal information channel is determined by the stationary point of this equation. Solving this requires techniques from the calculus of variations. For an introduction to this topic, see Sagan (1992). One important result in the continuous domain is stated in the following example.

Example: Rate–distortion curve for a Gaussian source and quadratic cost function. If the information source is described by a Gaussian distribution with mean μ and variance σ^2, and the cost to be minimized is the quadratic function $\mathcal{L}(x, y) = (y - x)^2$, then it is possible to analytically solve for the optimal communication channel, as well as the optimal rate–distortion curve. A full derivation is provided in Berger (1971). For this problem, the optimal communication channel is given by

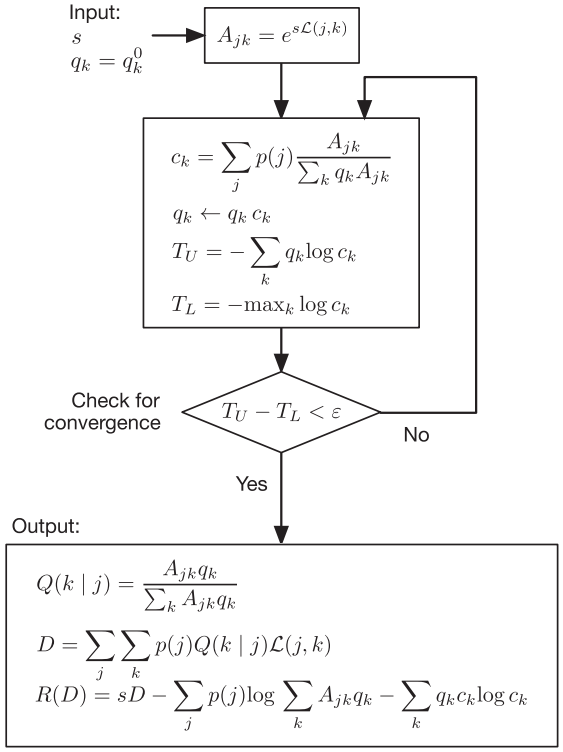

Fig. A.1. Blahut algorithm (Blahut, 1972) for efficiently computing the rate–distortion curve and optimal channel for an arbitrary discrete source and cost function.

$$Q^{\star}(y|x) = \text{Normal}(\mu_m, \sigma_m^2) \tag{A.7}$$
$$\mu_m = x + e^{-2R}(\mu - x)$$
$$\sigma_m^2 = e^{-4R}(e^{2R} - 1)\sigma^2$$

where R is the information rate of the channel, specified in nats. The corresponding rate–distortion curve is given by

$$R = \begin{cases} \frac{1}{2}\log\left(\frac{\sigma^2}{D}\right), & 0 \leqslant D \leqslant \sigma^2, \\ 0, & D \geqslant \sigma^2. \end{cases} \tag{A.8}$$

This result captures several properties that are true of human visual working memory. First, the channel output will be biased towards the mean of the prior distribution, and the magnitude of this bias will increase as the information rate decreases. Second, for a channel with fixed capacity, increasing the variance of the source (indicated by σ^2) will result in an increase in recall error as measured by squared error. Both of these predictions have been shown to hold for human visual working memory (Sims et al., 2012).

A.4. Efficient numerical solution for discrete sources via the Blahut algorithm

In many cases, including problems of practical relevance and interest, analytical solutions to rate–distortion problems are either unobtainable or computationally intractable. Fortunately, Blahut (1972) developed an efficient algorithm for numerically solving rate–distortion problems. The algorithm requires discrete information sources, but can be applied to continuous signals by first discretizing the source distribution $p(x)$ at a suitable resolution.

The steps of this algorithm are illustrated in Fig. A.1. The input consists of the parameter s (this corresponds to the slope of the rate–distortion curve at a particular point), and initial values for the vector q_k. It is sufficient to initialize $q_k = 1/N$ where N is the size of the output alphabet. The algorithm is iterative, and is guaranteed to converge to a point on the optimal rate–distortion curve. The parameter ε determines the tolerance for convergence, and in most cases can be set to 0.001.

It is often the case that the appropriate value for s is not known. For instance, one may wish to find an optimal channel for a given source and cost function, such that the information rate does not exceed a constraint given by the value C. What value of s will yield a channel with this information rate? One simple means of solving this problem is to implement an outer-loop optimization: search for the value of s such that the information rate is as close as possible to the desired constraint given by C. The result of this outer-loop optimization is the optimal information channel subject to the capacity constraint. Since this is a unidimensional optimization problem it can often be performed efficiently.

Appendix B. An R package for rate–distortion theory

This section briefly describes a custom software package written to accompany this paper. The package is written for the

```
library(RateDistortion) # Load the library

# Define a discretized Gaussian information source
x <- seq(from = -10, to = 10, length.out = 100)
Px <- dnorm(x, mean = 0, sd = 3)
Px <- Px / sum(Px) # Ensure that probability sums to 1
y <- x # The destination alphabet is the same as the source

# Define a quadratic cost function
cost.function <- function(x, y) {
    (y - x)^2
}

R.max <- 2 # Assume a constraint on information rate (bits)
# Find an optimal information channel
Q <- FindOptimalChannel(x, Px, y, cost.function, R.max)

# Compute the conditional distribution for a given value of x
cpd <- ConditionalDistribution(Q, index = 25)

# Plot the results
plot(c(-10, 10), c(0, 0.15), type = "n",
     xlab = "x", ylab = "Probability",
     yaxs = "i")
lines(x, Px, col = "black")
lines(cpd$y, cpd$p, col = "red")
abline(v = x[25], col = "red")
legend("topright",
       lty = 1, col = c("black", "red"),
       legend = c("P(x)", "P(y | x)"))
```

Listing 1. Using the Blahut algorithm to find an optimal information rate limited channel.

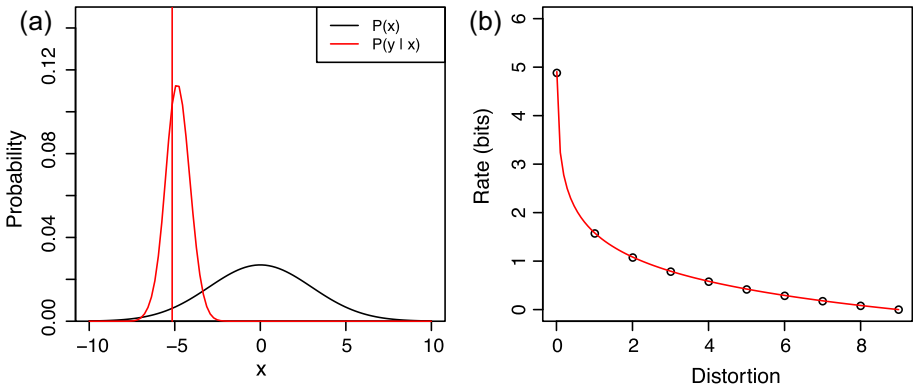

Fig. B.2. Demonstration of the R software package for solving rate–distortion problems. This example assumes a Gaussian information source with mean = 0 and SD = 3. The cost function is the quadratic cost function. (a) Information source $p(x)$, and the conditional probability distribution $p(y|x)$. (b) The rate–distortion curve for this example. The plot markers show the output of the Blahut algorithm, the smooth curve shows the analytical solution.

```
# Vector of points at which to compute the rate-distortion curve
D <- seq(from = 1e-2, to = 9, length.out = 10)
R <- vector(mode = "numeric", length = 10)
for(i in 1:10) {
    Q <- FindRate(x, Px, y, cost.function, D[i])
    R[i] <- Q$R
}

## Plot the rate-distortion curve
plot(c(0, 9), c(0, 6), type = "n",
     xlab = "Distortion",
     ylab = "Rate (bits)")
points(D, R)

# Compare to the analytical solution
d <- seq(from = 1e-2, to = 9, length.out = 100)
r <- (1/2) * log2(9 / d)
lines(d, r, col = "red")
```

Listing 2. Computing and plotting the rate–distortion curve.

R statistical programming language and implements the basic Blahut algorithm described in the previous section. In addition, the package contains a number of helper or convenience routines that simplify the development of models based on rate–distortion theory. All of the models and examples described in this paper were implemented with the aid of this package. Complete documentation accompanies the package. The full source code for all of the models described in this paper can be obtained by contacting the author.

B.1. Installation

Instructions for installing the software package can be found at http://www.pages.drexel.edu/∼crs346/code.html. This webpage also contains additional examples of its usage.

B.2. Usage

As a minimal example, consider a Gaussian information source with mean = 0 and standard deviation = 3. The goal is to find an optimally efficient communication channel for conveying samples from this distribution, subject to a constraint on information rate. The cost function for this example is assumed to be the quadratic error, $\mathcal{L}(x,y) = (y-x)^2$. This problem can be solved using the Blahut algorithm by using a discretized version of the information source. The example code in Listing 1 demonstrates the use of the R package for accomplishing this, and Fig. B.2a shows the resulting output.

In addition to computing an optimal information channel for a given constraint on information rate, it is also possible to plot the full rate–distortion curve. The example in Listing 2 assumes the same Gaussian information source and cost function, and compares the rate–distortion curve computed via the Blahut algorithm to the analytical solution described in Appendix A. The output of this code is shown in Fig. B.2b.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.
Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*(3), 183–193.
Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In *Sensory communication* (pp. 217–234). MIT Press.
Bays, P. M. (2014). Noise in neural populations accounts for errors in working memory. *The Journal of Neuroscience, 34*(10), 3632–3645.

Bays, P. M. (2015). Spikes not slots: Noise in neural populations limits working memory. *Trends in Cognitive Sciences, 19*(8), 431–438.
Berger, T. (1971). *Rate distortion theory: A mathematical basis for data compression*. Prentice-Hall.
Blahut, R. E. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory, 18*(4), 460–473.
Bowers, J. S., & Davis, C. J. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*(3), 389–414.
Bowers, J. S., & Davis, C. J. (2012b). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin, 138*(3), 423–426.
Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.
Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory ensemble statistics bias memory for individual items. *Psychological Science, 22*(3), 384–392.
Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.
Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87–185.
Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance, 35*(1), 94–107.
Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience*.Cambridge, MA: MIT Press.
Dimitrov, A. G., Lazar, A. A., & Victor, J. D. (2011). Information theory in neuroscience. *Journal of Computational Neuroscience, 30*(1), 1–5.
Dodds, P., Donkin, C., Brown, S. D., & Heathcote, A. (2011). Increasing capacity: Practice effects in absolute identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 477.
Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience*. Wiley-Blackwell.
Garner, W. R., & Hake, H. W. (1951). The amount of information in absolute judgments. *Psychological Review, 58*(6), 446.
Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research, 51*(7), 771–781.
Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349*(6245), 273–278.
Gleick, J. (2011). *The information: A history, a theory, a flood*. Pantheon Books.
Hahn, U. (2014). The bayesian boom: Good thing or bad? *Frontiers in Psychology, 5*, 765.
Herzmann, G., & Curran, T. (2011). Experts memory: An ERP study of perceptual expertise effects on encoding and recognition. *Memory & Cognition, 39*(3), 412–432.
Hino, H., & Murata, N. (2009). An information theoretic perspective of the sparse coding. In *Advances in neural networks – ISNN 2009* (pp. 84–93). Springer.
Jalali, S., & Weissman, T. (2008). Rate-distortion via markov chain monte carlo. In *IEEE international symposium on information theory, 2008. ISIT 2008* (pp. 852–856). IEEE.
Jiang, Y., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 683.
Jones, M., Curran, T., Mozer, M. C., & Wilder, M. H. (2013). Sequential effects in response time reveal learning mechanisms and event representations. *Psychological Review, 120*(3), 628.
Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology, 55*, 271–304.
Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.
Körding, K. (2007). Decision theory: What "should" the nervous system do? *Science, 318*(5850), 606–610.
Körding, K. P., Fukunaga, I., Howard, I. S., Ingram, J. N., & Wolpert, D. M. (2004). A neuroeconomics approach to inferring utility functions in sensorimotor control. *PLoS Biology, 2*(10), e330.
Körding, K. P., & Wolpert, D. M. (2004). The loss function of sensorimotor learning. *Proceedings of the National Academy of Sciences of the United States of America, 101*(26), 9839–9842.
Luce, R. D. (2003). Whatever happened to information theory in psychology? *Review of General Psychology, 7*(2), 183.
Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279–281.
Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences, 17*(8), 391–400.
Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17*(3), 347–356.
MacRae, A. W. (1970). Channel capacity in absolute judgment tasks: An artifact of information bias? *Psychological Bulletin, 73*(2), 112–121.
Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing bayesian transfer. *Visual Neuroscience, 26*(01), 147–155.
Marr, D. (1982). *Vision: A computational approach*.San Francisco: Freeman & Co..
Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition, 122*(3), 346–362.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81.
Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47*(1), 90–100.
Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*(23), 3311–3325.
Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review, 120*(2), 297.
Orhan, A. E., & Jacobs, R. A. (2014). Toward ecologically realistic theories in visual short-term memory research. *Attention, Perception, & Psychophysics, 76*(7), 2158–2170.
Orhan, A. E., Sims, C. R., Jacobs, R. A., & Knill, D. C. (2014). The adaptive nature of visual working memory. *Current Directions in Psychological Science, 23*(3), 164–170.
Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1999). *Spikes: Exploring the neural code*. MIT Press.
Rouder, J. N., Morey, R. D., Cowan, N., & Pealtz, M. (2004). Learning in a unidimensional absolute identification task. *Psychonomic Bulletin & Review, 11*(5), 938–944.
Sagan, H. (1992). *Introduction to the calculus of variations*. Dover.
Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. *Institute of Radio Engineers, National Convention Record, 4*, 142–163.
Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of information*. University of Illinois Press.
Shiffrin, R. M., & Nosofsky, R. M. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review, 101*(2), 357–361.
Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*(1), 1193–1216.
Sims, C. R. (2015). The cost of misremembering: Inferring the loss function in visual working memory. *Journal of Vision, 15*(3), 1–27.
Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review, 119*(4), 807–830.
Stewart, N., Brown, G. D., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*(4), 881.
Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9*(4), 578–585.
van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121*(1), 124.
van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences, 109*(22), 8780–8785.
Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233–235.