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]:

$$H(x) = -\sum_i P(x_i) \log P(x_i). \quad (1)$$

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 $0 \times \log 0 = 0$.