REVIEW **Communicated by Laurence Abbott**

Quantifying Stimulus Discriminability: A Comparison of Information Theory and Ideal Observer Analysis

Eric E. Thomson
ethomson@ucsd.edu
William B. Kristan
wkristan@ucsd.edu
University of California, San Diego,
La Jolla, CA 92093-0357, U.S.A.

Performance in sensory discrimination tasks is commonly quantified using either information theory or ideal observer analysis. These two quantitative frameworks are often assumed to be equivalent. For example, higher mutual information is said to correspond to improved performance of an ideal observer in a stimulus estimation task. To the contrary, drawing on and extending previous results, we show that five information-theoretic quantities (entropy, response-conditional entropy, specific information, equivocation, and mutual information) violate this assumption. More positively, we show how these information measures can be used to calculate upper and lower bounds on ideal observer performance, and vice versa. The results show that the mathematical resources of ideal observer analysis are preferable to information theory for evaluating performance in a stimulus discrimination task. We also discuss the applicability of information theory to questions that ideal observer analysis cannot address.

1 Introduction

Many adaptive behaviors require that an organism respond differently to different stimuli, that is, discriminate among stimuli. Familiar examples include the frog's ability to discriminate prey location (Lettvin, Maturana, McCulloch, & Pitts, 1959) and the monkey's ability to indicate the net direction of motion in a field of randomly flickering dots (Britten, Shadlen, Newsome, & Movshon, 1992). What properties of sensory neurons mediate such abilities? Do different stimuli reliably produce spike trains with different temporal patterns, or is variability in spike timing merely noise that should be averaged away to extract the underlying signal? Such questions have generated heated and often productive discussions for at least 50 years (MacKay & McCulluch, 1952; for reviews, see Perkel & Bullock, 1968; Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997; Victor, 1999).

Before evaluating performance in a discrimination task, one must first decide what measure to use. While ideal observer performance has historically been used to measure stimulus discrimination (Green & Swets, 1966; Geisler, 2003), information measures are often used with the same goal (Alkasab et al., 1999; Arabzadeh, Panzeri, & Diamond, 2004; Buracas & Albright, 1999; Li, Piech, & Gilbert, 2004; Paz & Vaadia, 2004; Petersen, Panzeri, & Diamond, 2002; Pola, Thiele, Hoffmann, & Panzeri, 2003; Theunnissen & Miller, 1991). The main goal of this article is to analyze the relationship between these two approaches. The relationship between information theory and ideal observer analysis has been the subject of several studies by statisticians and engineers (Wagner, 1965; Tebbe & Dwyer, 1968; Kovalevsky, 1968; Golic, 1987; Feder & Merhav, 1994), providing an untapped vein of research relevant for neuroscience. We review and extend the results of this previous research, showing that information theory and ideal observer analysis are not equivalent and that ideal observer analysis is more appropriate for quantifying stimulus discriminability. We also describe how to derive upper and lower bounds on ideal observer performance as a function of the information measures.

Because one of our goals is to determine when it is more appropriate to use information theory or ideal observer analysis to answer a particular research question, we devote significant attention to the interpretation of the quantities typically encountered in the two approaches. Most of the article assumes only knowledge of basic probability theory (e.g., Yates & Goodman, 1999); we introduce and define all key terms from ideal observer analysis and information theory. We place proofs that would interrupt the flow of the paper in footnotes.

2 Preliminary Assumptions and Definitions

Initially, we limit our analysis to experiments in which on each trial an organism is presented with one of M stimuli from a discrete set $\mathbb{S} = \{s_1, \ldots, s_M\}$. The probability that a particular stimulus will be presented on a given trial is represented by the probability distribution $P(\mathbb{S}) = \{P(s_1), \ldots, P(s_M)\}$, which we assume does not change with time. We assume that each stimulus evokes a response from an N-element set $\mathfrak{R} = \{r_1, \ldots, r_N\}$. Typically, $N \gg M$. In section 5, we extend the results to the case in which $\mathbb{S}$ and $\mathfrak{R}$ are continuous variables. Also, note that none of our conclusions rests on the assumption that $\mathbb{S}$ describes a set of stimuli and $\mathfrak{R}$ describes a set of responses; the analysis applies to any pair of random variables, whether they describe sensory, neural, behavioral, or other states.

The dependence of $\mathfrak{R}$ on $\mathbb{S}$ is represented by a channel matrix $P(\mathfrak{R}|\mathbb{S})$, which is an M by N matrix in which row i contains $P(\mathfrak{R}|s_i)$ (Ash, 1965):

A. $P(\mathbb{S})=\{P(s_1),P(s_2)\}=\{\tfrac{1}{2},\ \tfrac{1}{2}\}$, and

$$P(\mathfrak{R} \mid \mathbb{S}) = \begin{pmatrix} P(r_1 \mid s_1) & P(r_2 \mid s_1) \\ P(r_1 \mid s_2) & P(r_2 \mid s_2) \end{pmatrix} = \begin{pmatrix} \tfrac{1}{2} & \tfrac{1}{2} \\ \tfrac{3}{4} & \tfrac{1}{4} \end{pmatrix}$$

B. $$P(\mathbb{S},\mathfrak{R}) = \begin{pmatrix} P(s_1,r_1) & P(s_1,r_2) \\ P(s_2,r_1) & P(s_2,r_2) \end{pmatrix} = \begin{pmatrix} \tfrac{1}{4} & \tfrac{1}{4} \\ \tfrac{3}{8} & \tfrac{1}{8} \end{pmatrix}$$

Figure 1: Examples of the quantities used to analyze performance in a sensory discrimination task. (A) A two-element stimulus distribution $P(\mathbb{S})$ (top) and a two-by-two channel matrix $P(\mathfrak{R} \mid \mathbb{S})$ (bottom). (B) The corresponding joint distribution $P(\mathbb{S},\mathfrak{R})$, calculated from $P(\mathbb{S})$ and $P(\mathfrak{R} \mid \mathbb{S})$ in A using the fact from probability theory that $P(s_i, r_j) = P(r_j|s_i)P(s_i)$.

$$P(\mathfrak{R} \mid \mathbb{S}) = \begin{pmatrix} P(r_1 \mid s_1) & \cdots & P(r_N \mid s_1) \\ \vdots & \ddots & \vdots \\ P(r_1 \mid s_M) & \cdots & P(r_N \mid s_M) \end{pmatrix}. \tag{2.1}$$

We assume that $P(\mathfrak{R}|\mathbb{S})$ does not change with time.[1] Note that $P(\mathfrak{R} \mid \mathbb{S})$ is not a probability distribution. Rather, each row of $P(\mathfrak{R} \mid \mathbb{S})$ is a conditional probability distribution $P(\mathfrak{R} \mid s_i)$ that sums to one. See Figure 1A for a numerical example.

Given $P(\mathbb{S})$ and $P(\mathfrak{R}|\mathbb{S})$, the joint distribution $P(\mathbb{S},\mathfrak{R})$ can easily be calculated (see Figure 1B). Hence, if $P(\mathbb{S})$ and $P(\mathfrak{R}|\mathbb{S})$ are given, then it is possible to calculate the values of all functions of $P(\mathbb{S},\mathfrak{R})$ such as mutual information and ideal observer performance in a stimulus-estimation task.

3 Ideal Observers: Minimum Error Classifiers

3.1 Defining Ideal Observer Performance: *P*(*c*). An ideal observer of the neural response $\mathfrak{R}$ is a minimum-error classifier. As shown in Figure 2, a classifier is a function **C** that maps the response variable $\mathfrak{R}$ into $\hat{\mathbb{S}}$, where $\hat{\mathbb{S}}$ is the M-element set $\{\hat{s}_1, \ldots, \hat{s}_M\}$ that contains all possible estimates of which stimulus was presented. A classifier is correct on a given trial if $\hat{s} = s$, that is, if the estimate is identical to the actual stimulus. Similarly, a classifier is in error on those trials in which $\hat{s} \neq s$. A useful and widespread measure of classifier performance is the probability that it will provide a correct

[1] That is, we assume the system implements a discrete memoryless channel (Cover & Thomas, 1991).

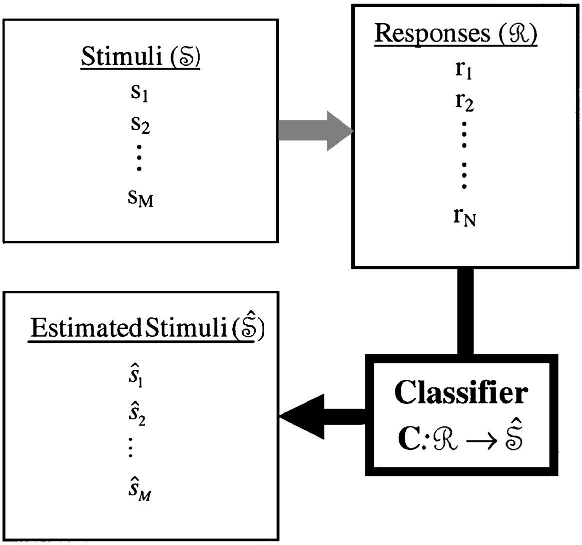
Stimuli ($\mathbb{S}$)
s_1
s_2
s_M
Responses ($\mathfrak{R}$)
r_1
r_2
r_N
Estimated Stimuli ($\hat{\mathbb{S}}$)
$\hat{s}_1$
$\hat{s}_2$
$\hat{s}_M$
Classifier
$C: \mathfrak{R} \to \hat{\mathbb{S}}$

Figure 2: Cartoon representation of the steps in a classification task. First, a stimulus from $\mathbb{S}$ is selected that evokes a response from $\mathfrak{R}$. Then the classifier C estimates which stimulus produced the given response. Formally, a classifier is a function that, given a response r_j as input, returns an estimate ($\hat{s}$) of the stimulus that produced that response.

estimate of $\mathbb{S}$ (Duda, Hart, & Stork, 2001). By definition, an ideal observer is a classifier that maximizes the probability of being correct (Liu, Knill, & Kersten, 1995; Geisler, 1989, 2003; Knill & Kersten, 1991).[2] This is equivalent to saying that an ideal observer minimizes the probability of error.

Ideal observers follow a simple decision rule (Duda et al., 2001): given response r_j, choose the estimate $\hat{s}_i$ that corresponds to the stimulus with the maximum probability in the conditional distribution $P(\mathbb{S} \mid r_j)$. More formally,

$$\text{Given } r_j \in \mathfrak{R}, \text{ choose } \hat{s}_i \text{ such that for all } s \in \mathbb{S},\ P(\mathbb{S}=s_i|r_j) \geq P(\mathbb{S}=s|r_j). \tag{3.1}$$

We denote the stimulus with the maximum probability, given r_j, as

[2] Albert Siegert, a physicist in the MIT Radiation Laboratory during World War II, introduced the term *ideal observer* to refer to minimum error classifiers. The term was first used in publication in a signal detection theory text in which the authors attribute the idea to Siegert (Lawson & Uhlenbeck, 1950).

A. $P(\mathbb{S}|r_1)=\left\{\frac{2}{5},\frac{3}{5}\right\}$, so $s_{\max(1)}=s_2$.
$P(\mathbb{S}|r_2)=\left\{\frac{2}{3},\frac{1}{3}\right\}$, so $s_{\max(2)}=s_1$.

B. *If r_1 is observed, then choose $\hat{s}_2$.*
If r_2 is observed, then choose $\hat{s}_1$.

C. $P(\boldsymbol{c}|r_1)= P(s_{max(1)}|r_1)=\frac{3}{5}$ $(P(\boldsymbol{e}|r_1)=\frac{2}{5})$.
$P(\boldsymbol{c}|r_2)= P(s_{max(2)}|r_2)=\frac{2}{3}$ $(P(\boldsymbol{e}|r_2)= \frac{1}{3})$.

D. $P(\boldsymbol{c}) = P(s_{\max(1)},r_1)+P(s_{\max(2)},r_2) = 3/8 + 1/4 = 5/8$.
Therefore, $P(\boldsymbol{e})=\frac{3}{8}$.

Figure 3: Example of the steps followed to calculate $P(c\,|\,r_j)$ and $P(c)$, a continuation of the example from Figure 1. (A) The conditional distributions $P(\mathbb{S}\,|\,r_j)$ are calculated using $P(\mathcal{R}\,|\,\mathbb{S})$ and $P(\mathbb{S})$ from Figure 1 and the fact that $P(s_i\,|\,r_j) = P(s_j, r_j)/P(r_j)$. (B) The corresponding ideal observer decision scheme, constructed using Rule 3.1. (C) $P(c\,|\,r_j)$ (and $P(e\,|\,r_j)$) are calculated using equation 3.2 and the fact that $P(e\,|\,r_j) = 1 - P(c\,|\,r_j)$. (D) $P(c)$ and $P(e)$ are calculated using equation 3.3 and the fact that $P(e) = 1 - P(c)$.

$s_{\max(j)}$. Hence, Rule 3.1 simplifies to "Choose $\hat{s}_{\max(j)}$" (see Figures 3A and 3B).[3] If multiple elements of $P(\mathbb{S}\,|\,r_j)$ have the same maximum probability, then the ideal observer makes an arbitrary choice among those elements (Duda et al., 2001). For instance, if $P(\mathbb{S}\,|\,r_j) = \{\frac{1}{5}, \frac{1}{5}, \frac{1}{10}, \frac{1}{5}, \frac{1}{10}, \frac{1}{5}\}$, then the ideal observer can choose $\hat{s}_1$, $\hat{s}_2$, $\hat{s}_4$, or $\hat{s}_6$.

Before describing how to calculate the probability that an ideal observer will correctly estimate $\mathbb{S}$, we introduce notation for use in the rest of the article. Let $\mathcal{C} = \{c, e\}$, where c is the event that a classifier is correct and e is the event that a classifier is in error. $P(\mathcal{C})$ is the probability distribution of $\mathcal{C}$ (i.e., $P(\mathcal{C}) = \{P(c), P(e) = 1 - P(c)\}$), and $P(\mathcal{C}\,|\,r_j)$ is the distribution of $\mathcal{C}$ when response r_j is observed.

The probability that an ideal observer will correctly estimate $\mathbb{S}$ if response r_j occurs is (Duda et al., 2001):

$$P(c\,|\,r_j) = P(\mathbb{S} = s_{\max(j)}\,|\,r_j) = \frac{P(s_{\max(j)}, r_j)}{P(r_j)}. \tag{3.2}$$

[3] In the classification literature, ideal observers are usually called *maximum a posteriori classifiers* or *Bayesian classifiers* (Duda et al., 2001). The reason for this name is that $s_{\max(j)}$ is the element with the maximum value in $P(\mathbb{S}\,|\,r_j)$, which in a Bayesian context is often called the a posteriori distribution of $\mathbb{S}$. We use the term *ideal observer* because it is more prevalent in psychophysics and neuroscience.

The first equality in equation 3.2 asserts that $P(c\,|\,r_j)$ is the conditional probability of the stimulus chosen by the ideal observer rule (Duda et al., 2001). See Figure 3C. This is because on trials in which r_j occurs, the ideal observer will pick $\hat{s}_{\max(j)}$ as its estimate of $\mathbb{S}$, and by definition $s_{\max(j)}$ will be the stimulus that actually evoked r_j with frequency $P(\mathbb{S} = s_{\max(j)}\,|\,r_j)$. The second equality in equation 3.2 is an instance of Bayes' theorem (Yates & Goodman, 1999). Note that $P(s_{\max(j)}, r_j)$, the numerator on the right-hand side of equation 3.2, is the maximum element in the column that corresponds to response r_j in the joint distribution $P(\mathbb{S},\mathbb{R})$.[4]

$P(c)$, the probability that an ideal observer will correctly estimate $\mathbb{S}$, is the average (over $\mathbb{R}$) of $P(c\,|\,r_j)$ (Duda et al., 2001):

$$P(c) = \sum_{j=i}^{N} P(r_j)P(c\,|\,r_j) = \sum_{j=i}^{N} P(s_{\max(j)}, r_j). \tag{3.3}$$

The second equality in equation 3.3 follows from equation 3.2. Equation 3.3 shows that $P(c)$ is just the sum of the maximum from each column of the joint distribution $P(\mathbb{S},\mathbb{R})$. Further, since the N maximal elements from the columns of $P(\mathbb{S},\mathbb{R})$ sum to $P(c)$, the $N \cdot (M - 1)$ nonmaximal elements must sum to $P(e)$. See Figure 3D for an example.

In the rest of the article, when we use $P(c)$ or $P(c\,|\,r_j)$, we are specifically referring to the probability that an ideal observer will correctly estimate $\mathbb{S}$, probabilities that can be calculated using equations 3.2 and 3.3. Similarly, $P(e)$ and $P(e\,|\,r_j)$ will be used to refer to the probability that an ideal observer will make an error.

It follows from equation 3.3 that for a given value of M, $P(c)$ must lie between $\frac{1}{M}$ and 1 (inclusive). The upper bound ($P(c) = 1$) corresponds to the best performance possible by an ideal observer of $\mathbb{R}$, and equation 3.3 implies that this perfect performance can be achieved only when each column of $P(\mathbb{S},\mathbb{R})$ has a single nonzero element. The worst ideal observer performance possible is $P(c) = P(s_{\max})$, the probability of the most likely

[4] *Proof*: $P(s_i, r_j) = P(s_i|r_j)P(r_j)$, and since $P(r_j)$ is the same for all elements in the jth column of $P(\mathbb{S},\mathbb{R})$, the maximum conditional probability $P(s_{\max(j)}|r_j)$ must also pick out the row with the maximum joint probability in column j.

stimulus (Duda et al., 2001).[5] Since the smallest possible value of $P(s_{\max})$ is $\frac{1}{M}$,[6] the lowest possible $P(c)$ for any distribution over $\mathbb{S}$ is $\frac{1}{M}$. For example, consider the special case in which $\Re$ is independent of $\mathbb{S}$ (i.e., $P(\mathbb{S} \mid r_j) = P(\mathbb{S})$). In such a case, the ideal observer's estimate $\hat{s}$ is simply the most likely stimulus $\hat{s}_{\max}$. This stimulus occurs with probability$P(s_{\max}) = P(c)$, which can vary between $\frac{1}{M}$ when the stimuli are equiprobable (i.e.,$P(\mathbb{S}) = \{\frac{1}{M}, \ldots, \frac{1}{M}\}$) and 1.0 when only a single stimulus is presented (i.e., $P(\mathbb{S}) = \{0, \ldots, 1, \ldots, 0\}$).

3.2 Interpreting *P*(*c*). By definition, we say the response variable $\Re$ is a good discriminator of $\mathbb{S}$ if different stimuli tend to lead to different responses. In what sense does $P(c)$ quantify this notion? Figure 4 shows three arbitrary gaussian-shaped conditional distributions $P(\Re \mid s_i)$, each weighted by the probability of the corresponding stimulus $P(s_i)$. Note that each such response ensemble is a row of the joint distribution $P(\mathbb{S},\Re)$. As discussed in section 3.1, because $P(c)$ is the sum of the maxima from the N columns of $P(\mathbb{S},\Re)$, it follows that $P(e)$ is the sum of the nonmaximal elements from the columns. In Figure 4, these nonmaximal elements of $P(\mathbb{S},\Re)$ are shaded in gray, illustrating that $P(e)$ is the area of overlap among the M stimulus-dependent ensembles.

On the whole, then, ideal observer performance is a useful measure of how well a response $\Re$ discriminates among different stimuli. If the response ensembles corresponding to different stimuli are so segregated that there is very little overlap among them (i.e., low $P(e)$ and high $P(c)$), then different stimuli typically lead to different responses, the defining feature of discriminability. Conversely, if the ensembles show a good deal of overlap (i.e., high $P(e)$ and low $P(c)$), then different stimuli often evoke the same response and $\Re$ is a bad discriminator of $\mathbb{S}$.

4 Bounding Information-Theoretic Quantities with Ideal Observers

In this section, we quantitatively compare ideal observer analysis and information theory. In particular, we address the claim that they are

[5] In Duda et al. (2001) this property of classifiers is mentioned but not proven, so we prove it here. *Proof*: Let row k of $P(\mathbb{S},\Re)$ be the row whose marginal probability is $P(s_{\max})$, that is, $P(s_k) = P(s_{\max}) = \sum_{j=i}^{N} P(s_k, r_j)$. For each column j, we know that $P(s_{\max(j)}, r_j) \geq P(s_k, r_j)$. Summing the terms in this inequality over $\Re$ yields:

$$P(c) = \sum_{j=1}^{N} P(s_{\max(j)}, r_j) \geq \sum_{j=1}^{N} P(s_k, r_j) = P(s_{\max}).$$

This lower bound on $P(c)$ is attained when $P(s_{\max(j)}, r_j) = P(s_k, r_j)$ for each column of $P(\mathbb{S},\Re)$.

[6] *Proof*: If the maximum of an M element set were less than $\frac{1}{M}$, then the elements could not sum to 1.

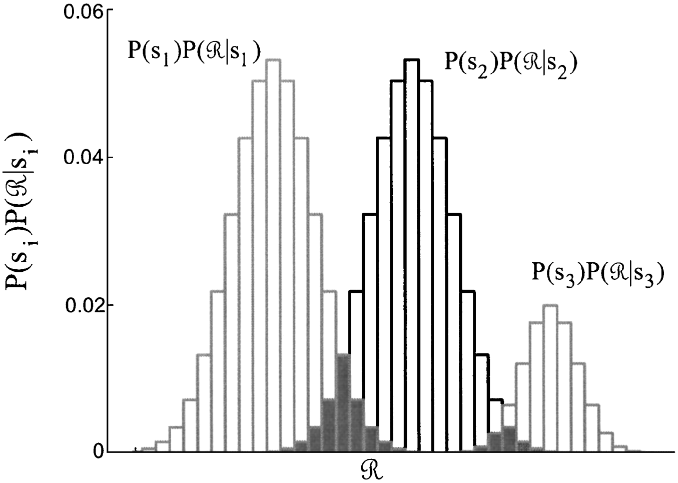

Figure 4: Graphical representation of the meaning of $P(c)$ and $P(e)$. The area of overlap among the weighted stimulus-conditional distributions (the gray-shaded bars) is identical to $P(e)$. The area of the unfilled bars sums to $P(c)$.

interchangeable theoretical methods that can be used to quantify stimulus discriminability. While most neuroscientists are interested in mutual information, we begin with entropy because it is useful for building intuitions for the more complicated cases and because most of the results generalize to the other information measures.

4.1 Entropy and the Ideal Observer

4.1.1 Three Interpretations of Entropy. If $P(\mathbb{S})$ is an M-element probability distribution, then the entropy of $\mathbb{S}$ is defined as (Cover & Thomas, 1991):

$$H(\mathbb{S}) = -\sum_{i=1}^{M} P(s_i) \log_2 P(s_i). \tag{4.1}$$

Because we take the logarithm to base two, all information measures are in units of bits, though in the rest of the review, we suppress the subscript. Qualitatively, $H(\mathbb{S})$ is usually described as a measure of our uncertainty about $\mathbb{S}$ (Ash, 1965), but how should this be interpreted? We discuss three interpretations of $H(\mathbb{S})$.

First, $H(\mathbb{S})$ measures how evenly the probability mass is spread among the M elements of $\mathbb{S}$ (Cover & Thomas, 1991). At one extreme, if all the probability mass in $P(\mathbb{S})$ is concentrated on one outcome ($P(\mathbb{S}) = \{0, \ldots, 1, \ldots, 0\}$), then $H(\mathbb{S})$ is minimized and is zero bits. At the other extreme, if each outcome

Codebook A

s_i	w_i	ℓ_i
s_1	11	2
s_2	10	2
s_3	01	2
s_4	00	2

$$\overline{L}_A = \tfrac{1}{4}2 + \tfrac{1}{4}2 + \tfrac{1}{4}2 + \tfrac{1}{4}2 = 2$$

Codebook B

s_i	w_i	ℓ_i
s_1	000	3
s_2	101	3
s_3	0101	4
s_4	1111	4

$$\overline{L}_B = \tfrac{1}{4}3 + \tfrac{1}{4}3 + \tfrac{1}{4}4 + \tfrac{1}{4}4 = 3\tfrac{1}{2}$$

Figure 5: Examples of two possible codebooks that encode the value of $\mathbb{S}\,(M=4)$. The stimulus value s_i is in the first column of each table, the corresponding codeword w_i is in the second column, and the third column shows ℓ_i, the corresponding codeword length. The average codeword length, $\overline{L}$, is calculated below each table using equation 4.2 under the assumption of equiprobable stimuli.

is equally probable ($P(\mathbb{S}) = \{\frac{1}{M}, \ldots, \frac{1}{M}\}$), then the mass is evenly spread ($P(\mathbb{S}) = \{\frac{1}{M}, \ldots, \frac{1}{M}\}$) and $H(\mathbb{S})$ is maximized at $\log(M)$ bits.

A second interpretation of $H(\mathbb{S})$ is that it provides lower bounds on the number of binary digits (bits) required to encode $\mathbb{S}$. This interpretation is based on the source coding theorem (see equation 4.3), a central result from coding theory. Before stating the theorem, we briefly review terminology from coding theory. To encode random variable $\mathbb{S}$ is to build a codebook, which assigns to each element s_i of $\mathbb{S}$ a unique codeword w_i. Each codeword is a sequence of symbols from a set of elementary symbols called the alphabet. The number of symbols in codeword w_i is the codeword length, which we denote ℓ_i. The codeword length ℓ_i can be considered a measure of the cost incurred by encoding stimulus s_i. The average codeword length, denoted $\overline{L}$, measures the average (over $\mathbb{S}$) cost when a particular codebook is used (Cover & Thomas, 1991):

$$\overline{L} = \sum_{i=1}^{M} P(s_i)\ell_i. \tag{4.2}$$

$\overline{L}$ is useful for comparing the cost of using different codebooks.

Figure 5 provides examples to illustrate these concepts. It shows two of the infinite number of possible codebooks if $\mathbb{S}$ is a set of four stimuli. The alphabet in this example is the set $\{0, 1\}$ of binary digits, so the codebooks assign to each element s_i of $\mathbb{S}$ a codeword w_i consisting of a sequence of binary digits. The corresponding codeword lengths ℓ_i are shown in the third column of the tables. If we assume the four stimuli are equiprobable, then $\overline{L}_A$, the average codeword length for codebook A, is 2 bits, and $\overline{L}_B$ is $3\frac{1}{2}$ bits. Hence, the average cost of using codebook A is less than the average cost of using codebook B.

The source coding theorem (also known as the noiseless coding theorem; Ash, 1965) provides the basis for the second interpretation of $H(\mathbb{S})$. The theorem is (Cover & Thomas, 1991)

$$H(\mathbb{S}) \leq \overline{L}_{\min} < H(\mathbb{S}) + 1, \tag{4.3}$$

where $\overline{L}_{\min}$ is the minimum possible average codeword length required to encode random variable $\mathbb{S}$. The theorem shows that as $H(\mathbb{S})$ increases, more bits are required to encode $\mathbb{S}$. The theorem also provides an absolute lower bound on $\overline{L}$ that can be used to evaluate any individual codebook. For example, it follows from equation 4.3 that codebook A in Figure 5 is in the set of best possible codebooks for $\mathbb{S}$ because it actually reaches the lower bound $\overline{L}_A = H(\mathbb{S})$. Also, because $\overline{L}_B > H(\mathbb{S}) + 1$, it follows that there exists a better codebook than codebook B. Note that in both codebooks, each stimulus is assigned a different codeword, so if there existed a channel in which the response to each stimulus was the corresponding codeword, $P(c)$ would be 1.0. Thus, the source coding theorem tells you the fewest number of bits required to encode a variable so that it can be decoded without error.

A third interpretation of $H(\mathbb{S})$, inspired by the description of $H(\mathbb{S})$ as a measure of uncertainty about $\mathbb{S}$, is that it is a measure of the difficulty an ideal observer would have estimating $\mathbb{S}$. The flaws in this interpretation are discussed in the next section.

4.1.2 Comparing Entropy and Ideal Observers. $H(\mathbb{S})$ is a function only of $P(\mathbb{S})$. Hence, to directly compare $H(\mathbb{S})$ and $P(c)$, we consider the behavior of an ideal observer faced with the task of estimating $\mathbb{S}$ given only $P(\mathbb{S})$. Given $P(\mathbb{S})$, the ideal observer picks the most likely stimulus $\hat{s}_{\max}$ as its estimate of $\mathbb{S}$, and $P(c) = P(s_{\max})$ (Duda et al., 2001).

$H(\mathbb{S})$ is sometimes interpreted as a measure of the difficulty in estimating $\mathbb{S}$ on a single trial: the higher the entropy, the less likely that an ideal observer will correctly estimate $\mathbb{S}$ (Alkasab et al., 1999). Formally:

$$H(\mathbb{S}_1) \geq H(\mathbb{S}_2) \Leftrightarrow P(c)_1 \leq P(c)_2, \tag{4.4}$$

where $\mathbb{S}_1$ and $\mathbb{S}_2$ are two random variables and $P(c)_i$ is ideal observer performance in estimating $\mathbb{S}_i$ on the basis of $P(\mathbb{S}_i)$. That equation 4.4 is incorrect can be shown by counterexample. Let $P(\mathbb{S}_1) = \{\frac{65}{100}, \frac{18}{100}, \frac{17}{100}\}$ and $P(\mathbb{S}_2) = \{\frac{50}{100}, \frac{49}{100}, \frac{1}{100}\}$. In this case, $H(\mathbb{S}_1) = 1.28 > 1.07 = H(\mathbb{S}_2)$ and $P(c)_1 = 0.65 > 0.50 = P(c)_2$.

The rest of this section provides a more general analysis of the relationship between $P(c)$ and $H(\mathbb{S})$. We first review previous results that show how to calculate the range of $H(\mathbb{S})$ values that are consistent with a given value of $P(c)$ (Tebbe & Dwyer, 1968; Kovalevsky, 1968). Toward this end, we define $M_{P(c)}$ as the set of all M-element probability distributions for which ideal observer performance is $P(c)$. Since entropy is invariant under permutations

$$\mathrm{M}_{\mathrm{P}(c)} = 4_{0.5} = \left\{ \begin{array}{l} \quad \{\frac{1}{2},\frac{3}{8},\frac{1}{8},0\}, \\ \circ\ \{\frac{1}{2},\frac{1}{2},0,0\}, \\ \quad \{\frac{1}{2},\frac{1}{3},\frac{1}{12},\frac{1}{12}\}, \\ \qquad \vdots \\ \bullet\ \{\frac{1}{2},\frac{1}{6},\frac{1}{6},\frac{1}{6}\} \end{array} \right\}$$

Figure 6: Some members of the set $M_{P(c)} = 4_{0.5}$. The empty circle indicates $P_{\min}(4_{0.5})$, the distribution in $4_{0.5}$ with the minimum entropy ($H = 1$ bit). The filled circle indicates $P_{\max}(4_{0.5})$, the distribution with the maximum entropy ($H = 1.8$ bits).

of elements of $P(\mathbb{S})$ we can, without loss of generality, sort the probability values in each distribution in $M_{P(c)}$ in descending order (Feder & Merhav, 1994). That is, by construction, all distributions in $M_{P(c)}$ satisfy

$$P(c) = P(s_1) \geq P(s_2) \geq \cdots \geq P(s_M). \tag{4.5}$$

Figure 6 shows an example of $M_{P(c)}$ for the case in which $P(c) = 0.5$ and $M = 4$. Each distribution in $M_{P(c)} = 4_{0.5}$ has the same maximal element $P(s_1) = P(s_{\max}) = 0.5$, but the probability mass in the remaining three elements can arbitrarily vary as long as the constraints provided by equation 4.5 are satisfied. $H(\mathbb{S})$ is sensitive to this variability, but $P(c)$ is not. In fact, the distributions in $M_{P(c)}$ take on a range of $H(\mathbb{S})$ values. We denote the distribution in $M_{P(c)}$ that has the minimum entropy $P_{\min}(M_{P(c)})$ and the distribution with the maximum entropy $P_{\max}(M_{P(c)})$. We denote the corresponding entropy values $h_{\min}(M_{P(c)})$ and $h_{\max}(M_{P(c)})$, respectively.

Previous papers (Tebbe & Dwyer, 1968; Kovalevsky, 1968) show that $P_{\max}(M_{P(c)})$ and $h_{\max}(M_{P(c)})$ are

$$P_{\max}(M_{P(c)}) = \left\{ P(c), \frac{P(e)}{M-1}, \ldots, \frac{P(e)}{M-1} \right\} \tag{4.6}$$

$$h_{\max}(M_{P(c)}) = H(\mathbb{C}) + P(\mathrm{e})\log(M-1), \tag{4.7}$$

where $\mathbb{C}$ is the random variable with outcomes $\{c, e\}$ defined in section 3.1. Equation 4.6 holds because once $P(c)$ is fixed, the way to maximize entropy is to distribute the residual $P(e) = [1 - P(c)]$ probability mass evenly to the remaining $M - 1$ elements of the distribution. Equation 4.7 follows when equation 4.6 is substituted into equation 4.2. In Figure 6, $P_{\max}(4_{0.5})$ is indicated with a filled circle, and this distribution has entropy $h_{\max}(4_{0.5}) = 1.8$ bits. In Figure 7A, $h_{\max}(M_{P(c)})$ is plotted as a function of $P(c)$ for $M = 2$,

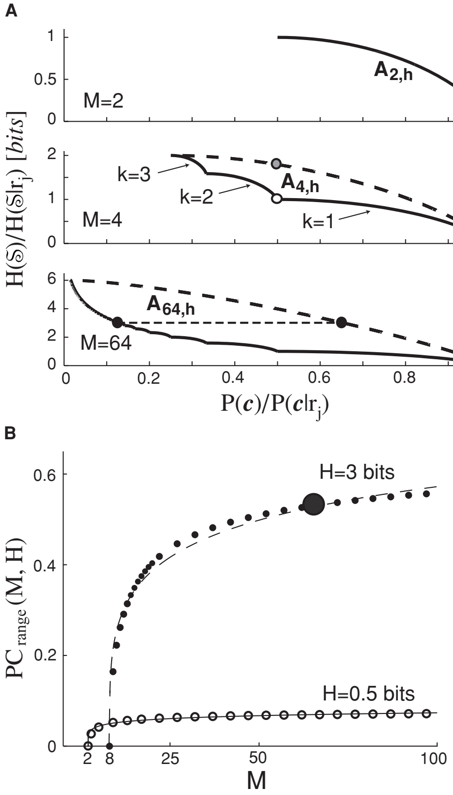

Figure 7: Comparison of $H(\mathbb{S})$ and $P(c)$. (A) Plot of the upper and lower bounds on $H(\mathbb{S})$ as a function of $P(c)$ for $M = 2$, $M = 4$, and $M = 64$ (panels 1, 2, and 3, respectively). The upper bound, $h_{\max}(M_{P(c)})$, is indicated by a dashed line, and the lower bound, $h_{\min}(M_{P(c)})$, is indicated by a solid line. The sets of allowable $\{P(c), H(\mathbb{S})\}$ points are labeled $\mathrm{A}_{M,h}$. Note that $\mathrm{A}_{2,h}$ (panel 1) is a line. In panel 2, the arch-shaped line segments corresponding to different values of k are indicated, as are the points corresponding to $h_{\max}(4_{0.5})$ and $h_{\min}(4_{0.5})$ from Figure 6 (filled and empty circles, respectively). (Recall that the integer k describes the minimum number of stimuli that must have probability $P(c)$ when the goal is to minimize $H(\mathbb{S})$; see the text). In panel 3, the lower (0.14) and upper (0.69) bounds on $P(c)$ when $H(\mathbb{S}) = 3$ bits are indicated with filled circles and connected with a dashed line. As discussed in section 4.2, the graphs also plot the relationship between $P(c|r_j)$ and $H(\mathbb{S}\,|\,r_j)$, so these quantities are included on the x- and y-axes, respectively. (B) Plot of $PC_{\text{range}}(M,H)$ as a function of M shows that $PC_{\text{range}}(M,H)$ increases exponentially with M. The filled circles show $PC_{\text{range}}(M,3)$, and the unfilled circles plot $PC_{\text{range}}(M,0.5)$. The lines are the best saturating exponential fit to the points (see text). The large filled circle on the $PC_{\text{range}}(M,3)$ line indicates $PC_{\text{range}}(64,3)$, the range delineated by the dashed line in panel 3 in Figure 7A.

4, and 64 (dotted lines). It can be seen that at a given value of $H(\mathbb{S})$, the maximum possible $P(c)$ value is specified by $h_{\max}(M_{P(c)})$, so $h_{\max}(M_{P(c)})$ provides an upper bound on $P(c)$.

Note that for a given $P(c)$ value, $h_{\max}(M_{P(c)})$ increases with $\log(M - 1)$, the only term in equation 4.7 that depends on M. $h_{\max}(M_{P(c)})$ increases with M because for a larger M, there exist a greater number of elements across which the residual probability mass $P(e)$ can be spread. At the other extreme, to minimize $H(\mathbb{S})$, the residual probability mass $P(e)$ must be concentrated as much as possible while satisfying the constraints in equation 4.5. It is as if there are M cups, each of which can hold $P(c)$ liters of water, and the goal is to distribute a single liter of water to the cups while filling as few of the cups as possible. For a given $P(c)$ value, there exists an integer k that indicates the minimum number of cups that must be completely filled when following such a strategy. For the case in which $M=4$ and $P(c)=0.5$ (see Figure 6), $P_{\min}(4_{0.5})$ is $\{\frac{1}{2}, \frac{1}{2}, 0, 0\}$, so $k=2$ and no water is distributed to the third or fourth cups. Consider also the case in which $M=4$ and $P(c)=0.4$. In this case, $P_{\min}(4_{0.4})=\{0.4, 0.4, 0.2, 0\}$, k is again two, but there remain 0.2 liters of water once the second cup is full, and this water must be distributed to the third cup (i.e., cup $k + 1$).

Mathematically, this procedure of concentrating probability mass leads to the following equations for $P_{\min}(M_{P(c)})$ and $h_{\min}(M_{P(c)})$ (Tebbe & Dwyer, 1968; Kovalevsky, 1968). For each $P(c)$ value between $\frac{1}{M}$ and 1, there exists an integer k between 1 and $M-1$ such that $\frac{1}{k+1} \leq P(c) \leq \frac{1}{k}$, and

$$P_{\min}(M_{P(c)}) = \{P(c)_1, \ldots, P(c)_k, \ 1 - kP(c), \ 0, \ldots, 0\} \tag{4.8}$$

$$h_{\min}(M_{P(c)}) = -\,[kP(c)\log(P(c)) + (1 - kP(c))\log(1 - kP(c))]. \tag{4.9}$$

As can be seen in equation 4.8, k corresponds to the case in which k out of the M stimuli are assigned a probability of $P(c)$. The remaining $[1 - kP(c)]$ probability mass is assigned to stimulus $k + 1$, and stimuli $k + 2$ through M are assigned probabilities of zero. Figure 7A plots $h_{\min}(M_{P(c)})$ as a function of $P(c)$ for $M=2$, 4, and 64 (solid lines). Each arch-shaped line segment in Figure 7A corresponds to a different value of k (in Figure 7A.2 the line segments are labeled with their corresponding k values). It can be seen in the figure that $h_{\min}(M_{P(c)})$ provides a lower bound on $P(c)$. That is, for a given value of $H(\mathbb{S})$, the lowest possible value of $P(c)$ is given by equation 4.9.

Note that $h_{\min}(M_{P(c)})$ does not depend on M. Increasing M, the number of possible stimuli, does not affect the outcome when the goal is to concentrate the probability mass to those stimuli as much as possible. For instance, if $P(c)$ is 0.5, then no matter how large M is, only the first two stimuli are assigned nonzero probabilities (e.g., $P(6_{0.5})=\{\frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0\}$).

The upper and lower bounds $h_{\min}(M_{P(c)})$ and $h_{\max}(M_{P(c)})$ circumscribe the set $\mathrm{A}_{M,h}$ of all possible points $\{P(c), H(\mathbb{S})\}$ (see Figure 7). We note four general features of $\mathrm{A}_{M,h}$. First, its upper and lower bounds, $h_{\max}(M_{P(c)})$ and

$h_{\min}(M_{P(c)})$, are tight (Feder & Merhav, 1994). That is, for a given M and $P(c)$, there exists a distribution $P(\mathbb{S})$ that actually equals $h_{\max}(M_{P(c)})$ and $h_{\min}(M_{P(c)})$. These distributions are given by equations 4.6 and 4.8, respectively.

Second, for a given value of $P(c)$, the range of possible $H(\mathbb{S})$ values increases with $\log(M - 1)$. For the case in which there are only two stimuli ($M = 2$), $H(\mathbb{S})$ is uniquely specified by $P(c)$. This is because once $P(s_1)$ is fixed, there is no freedom to vary the remaining probability mass: $P(s_2)$ must equal $1 - P(s_1)$. As M increases, $h_{\max}(M_{P(c)})$ increases with $\log(M - 1)$ without bound. This is because the only term in equations 4.7 and 4.9 that depends on M is $\log(M - 1)$, which is in the equation for $h_{\max}(M_{P(c)})$.

Third, while equation 4.4 is incorrect (i.e., higher entropy does not imply a decrement in ideal observer performance), it is possible to infer the range of $P(c)$ values consistent with a given value of $H(\mathbb{S})$. That is, given $H(\mathbb{S})$, it is possible to calculate upper and lower bounds on $P(c)$. This is because both $h_{\max}(M_{P(c)})$ and $h_{\min}(M_{P(c)})$ are one-to-one, strictly monotonic functions of $P(c)$ (Feder & Merhav, 1994). Hence, both $h_{\max}(M_{P(c)})$ and $h_{\min}(M_{P(c)})$ have inverses that provide upper and lower bounds of $P(c)$, respectively. While closed-form analytical solutions for the inverses do not exist, the bounds can be numerically estimated with accuracy limited only by machine precision.[7] For example, numerical methods applied with $M = 64$ and $H(\mathbb{S}) = 3$ bits show that $P(c)$ can vary between 0.14 and 0.69, bounds marked with filled circles in Figure 7A, panel 3.[8] In this case the range of $P(c)$ values, defined as the difference between the maximum and minimum $P(c)$ value, is 0.55.

Fourth, the range of $P(c)$ values consistent with a given $H(\mathbb{S})$ is a saturating exponential function of M. If we define the maximum and minimum $P(c)$ values consistent with a given $H(\mathbb{S})$ as $PC_{\max}(M, H)$ and $PC_{\min}(M, H)$, respectively, then the range of $P(c)$ values at a given entropy is given by $PC_{\text{range}}(M, H) = PC_{\max}(M, H) - PC_{\min}(M, H)$. Figure 7B plots $PC_{\text{range}}(M, 0.5)$ and $PC_{\text{range}}(M, 3.0)$ as functions of M (filled and open circles, respectively). As can be seen in the figure, $PC_{\text{range}}(M, H)$ is a saturating exponential function of M that saturates at $1 - PC_{\min}(M, H)$. We expected this saturating exponential based on the following two considerations. First, the range of $H(\mathbb{S})$ values at a given $P(c)$ increases logarithmically with M (see above), and this is due to the fact that $h_{\max}(M_{P(c)})$ increases logarithmically

[7] Equations 4.7 and 4.9 are of the form $y = x + \log(x)$, for which there exists no analytical solution for the inverse.

[8] We carried out the calculation of the lower bound on $P(c)$ as follows. A set of $h_{\max}(M_{P(c)})$ values was obtained over the range $P(c) = \frac{1}{4}$ to 1 using equations 4.7 and 4.9. Then a cubic spline was fit to these numbers using $P(c)$ as the dependent variable (de Boor, 1978). Then, given any value $H(\mathbb{S})$, $P(c)_{\max}$ can be estimated using the spline. A similar method was used to calculate the upper bound on $P(c)$.

with M. Because the inverse of the log is an exponential function, and $PC_{\max}(M, H)$ grows with the inverse of $h_{\min}(M_{P(c)})$, $PC_{\text{range}}(M, H)$ should grow exponentially with M. However, since $P(c)$ can never be greater than 1.0, $PC_{\text{range}}(M, H)$ must also be bounded from above by $1 - PC_{\min}(M, H)$, so we also expected the curve to saturate at that value. While there do not exist analytical solutions for $PC_{\text{range}}(M, H)$ (see the previous paragraph), the points are indeed well fit by a saturating exponential function constrained to have a maximum of $1 - PC_{\min}(M, H)$(see Figure 7B).[9]

4.1.3 Multiple Question Ideal Observers and Entropy. If an ideal observer makes an error on the first guess and is given another chance to guess $\mathbb{S}$, then it will pick the stimulus with the maximum probability in the conditional distribution $P(\mathbb{S}\,|\,r_j, \mathbb{S} \neq s_{\max(j)})$. In that case, the analysis from the previous section applies, though with the number of stimuli effectively reduced to $M - 1$. More generally, after G guesses, the stimulus set is effectively reduced to $M - G$ stimuli, where G can vary between zero and $M - 1$.

In contrast, if the goal is to guess the value of a random variable using the fewest number of yes-or-no questions (i.e., the game of 20 Questions; Cover & Thomas, 1991), then the questions that receive yes and no answers can be represented by ones and zeros, respectively. Then, by the source coding theorem, $H(\mathbb{S})$ describes the best possible performance $\overline{L}_{\min}$ (Cover & Thomas, 1991). In 20 Questions, the optimal question sequence will cut down the range of possible outcomes to quickly specify the stimulus. For instance, if there are eight equiprobable stimuli, then the initial question should be of the form, "Is it above four?" The ideal observer, on the other hand, must always guess a single outcome, that is, the ideal observer must always ask questions of the form, "Was it an 8?"

4.2 Response-Conditional Entropy and the Ideal Observer. The entropy of $\mathbb{S}$, given that r_j is observed, is known as the response-conditional entropy $H(\mathbb{S}\,|\,r_j)$, which is defined as (Cover & Thomas, 1991)

$$H(\mathbb{S}\,|\,r_j) = -\sum_{i=1}^{M} P(s_i\,|\,r_j)\log(P(s_i\,|\,r_j)). \tag{4.10}$$

Note that equation 4.10 is simply equation 4.2 with the conditional distribution $P(\mathbb{S}\,|\,r_j)$ substituted for $P(\mathbb{S})$. Hence, all the properties of $H(\mathbb{S})$ (see section 4.1.1) extend to $H(\mathbb{S}\,|\,r_j)$. For instance, in addition to measuring how evenly spread the conditional distribution $P(\mathbb{S}\,|\,r_j)$ is, $H(\mathbb{S}\,|\,r_j)$ provides bounds on the minimum average number of binary digits required to encode $\mathbb{S}$ once r_j is known.

[9] The lines are the least-squares fits to saturating exponentials of the form $(1 - PC_{\min}(M))(1 - e^{-(\frac{M-T}{\alpha})^{\beta}}) + PC_{range}(T)$, where T is the smallest value of M for which the given $H(\mathbb{S})$ is defined and α/β are free parameters.

Since $H(\mathbb{S}\,|\,r_j)$ is a function of $P(\mathbb{S}\,|\,r_j)$, we compare $H(\mathbb{S}\,|\,r_j)$ to the performance of an ideal observer of r_j (i.e., $P(c\,|\,r_j)$), which is also a function of $P(\mathbb{S}\,|\,r_j)$. Recall that an ideal observer of r_j will pick the most likely stimulus from the conditional distribution $P(\mathbb{S}\,|\,r_j)$ (Rule 1). Because $H(\mathbb{S}\,|\,r_j)$ is mathematically equivalent to $H(\mathbb{S})$ (i.e., $H(\mathbb{S}\,|\,r_j)$ is a function of an ordinary M-element probability distribution, and the ideal observer picks the stimulus with the maximum probability from this distribution), $H(\mathbb{S}\,|\,r_j)$ has the exact same properties with respect to $P(c\,|\,r_j)$ that $H(\mathbb{S})$ has with respect to $P(c)$ (see section 4.1.2). For example, for a given $P(c\,|\,r_j)$, there is a range of possible $H(\mathbb{S}\,|\,r_j)$ values bounded by

$$h_{\max}(M_{P(c\,|\,r_j)}) = H(\mathcal{C}\,|\,r_j) + P(\mathrm{e}\,|\,r_j)\log(M-1) \tag{4.7$'$}$$

$$\begin{aligned} h_{\min}(M_{P(c\,|\,r_j)}) &= -[kP(c\,|\,r_j)\log P(c\,|\,r_j) \\ &\quad + (1-kP(c\,|\,r_j))\log(1-kP(c\,|\,r_j)). \end{aligned} \tag{4.9$'$}$$

We label equations 4.7′ and 4.9′ because they are simply equations 4.7 and 4.9 with $P(c)$ replaced by $P(c\,|\,r_j)$ and $M_{P(c)}$ replaced by $M_{P(c\,|\,r_j)}$ (see Figure 7). Also note that the set of possible $\{P(c\,|\,r_j),\ H(\mathbb{S}\,|\,r_j)\}$ points is identical to the set $A_{M,h}$ of possible points $\{P(c),\ H(\mathbb{S})\}$ discussed in section 4.1 (see Figure 7). We use the label $A_{M,h}$ for both sets of points, letting the context make clear whether we are referring to entropy or response-conditional entropy.

4.3 Specific Information and the Ideal Observer. How much information does a specific response r_j provide about the stimulus set $\mathbb{S}$? In information theory, this is quantified by the specific information between $\mathbb{S}$ and r_j (DeWeese & Meister, 1999). Specific information is formally defined as

$$\begin{aligned} I(\mathbb{S}, r_j) = H(\mathbb{S}) - H(\mathbb{S}\,|\,r_j) &= -\sum_{i=1}^{M} P(s_i)\log(P(s_i)) \\ &\quad + \sum_{i=1}^{M} P(s_i\,|\,r_j)\log(P(s_i\,|\,r_j)). \end{aligned} \tag{4.11}$$

Since $H(\mathbb{S})$ is the original entropy of $\mathbb{S}$, and $H(\mathbb{S}\,|\,r_j)$ is the entropy of $\mathbb{S}$ remaining after r_j is observed, it follows that $I(\mathbb{S}, r_j)$ is a measure of how much entropy about $\mathbb{S}$ is removed upon observation of response r_j. (For examples, see Figure 8A). Interestingly, $I(\mathbb{S}, r_j)$ can be negative (DeWeese & Meister, 1999), which occurs when $P(\mathbb{S}\,|\,r_j)$ is more evenly spread than $P(\mathbb{S})$ (e.g., rows 1 and 2 in Figure 8A). Conversely, $I(\mathbb{S}, r_j)$ is positive when $P(\mathbb{S}\,|\,r_j)$ is more concentrated than $P(\mathbb{S})$ (e.g., rows 3 and 4 in Figure 8A). Interpreted in terms of coding theory (see section 4.1.1), $I(\mathbb{S}, r_j)$ indicates how many fewer binary digits, on average, are needed to encode $\mathbb{S}$ once r_j is known. This number is negative when more digits are required to encode $\mathbb{S}$ once r_j is observed.

A. Let $P(\mathbb{S}) = \{\frac{2}{3}, \frac{1}{6}, \frac{1}{6}\}$, so $H(\mathbb{S}) = 1.25$

j	P(𝕊|r_j)	P(c|r_j)	H(𝕊|r_j)	I(𝕊,r_j)
1	$\{\frac{3}{5}, \frac{1}{5}, \frac{1}{5}\}$	.6	1.4	-0.15
2	$\{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\}$	.5	1.5	-0.25
3	$\{\frac{1}{2}, \frac{1}{2}, 0\}$	.5	1.0	0.25
4	{.77,.12,.11}	.77	1.0	0.25

B.

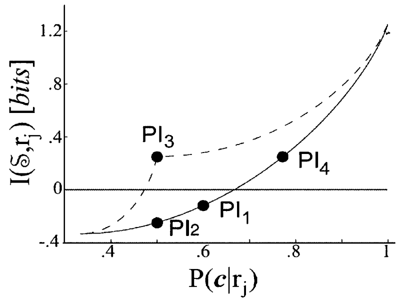

Figure 8: Examples that compare $P(c)$ and specific information. (A) Each row of the table contains a conditional distribution $P(\mathbb{S}\,|\,r_j)$. Columns 2–5 contain the corresponding ideal observer performance $P(c\,|\,r_j)$, response-conditional entropy $H(\mathbb{S}\,|\,r_j)$, and specific information $I(\mathbb{S}, r_j)$, respectively. The calculation of $I(\mathbb{S}, r_j)$ assumes that $P(\mathbb{S})$ is as given with an entropy $H(\mathbb{S})$ of 1.25 bits. (B) Scatter plot of the four $\{P(c\,|\,r_j), I(\mathbb{S}, r_j)\}$ pairs from the rows in the table in A. The point from row j of the table in A is denoted PI_j. The absolute upper and lower bounds on $I(\mathbb{S}, r_j)$ as a function of $P(c\,|\,r_j)$ are superimposed for reference (dashed and solid lines, respectively).

While specific information has not received much attention from neuroscientists (but see DeWeese & Meister, 1999), it is a potentially useful measure that can be used to determine whether certain classes of neuronal responses transmit more information than others. For instance, is $I(\mathbb{S}, r_j)$ greater for spike trains with higher firing rates?

How is $I(\mathbb{S}, r_j)$ related to ideal observer performance? While this question has not been discussed in the literature, we extend the previous results to address it. Since $I(\mathbb{S}, r_j)$ measures the amount of information carried by a particular response r_j about the stimulus set $\mathbb{S}$, we compare $I(\mathbb{S}, r_j)$ to the performance of an ideal observer that has observed a particular response r_j, that is, $P(c\,|\,r_j)$.

Before examining the general relationship between $I(\mathbb{S}, r_j)$ and $P(c\,|\,r_j)$, we use the examples in Figure 8A to highlight three features of their relationship. First, surprisingly, a response can carry negative information about $\mathbb{S}$ and be a better predictor of $\mathbb{S}$ than a response that carries positive information about $\mathbb{S}$ (e.g., compare rows 1 and 3). Second, $P(c\,|\,r_j)$ does not uniquely specify $I(\mathbb{S}, r_j)$ (compare rows 2 and 3). The converse also holds (rows 3 and 4). Third, associated with each conditional distribution $P(\mathbb{S}|r_j)$ is a point $\{P(c\,|\,r_j),\, I(\mathbb{S}, r_j)\}$ that indicates the ideal observer performance and specific information for that distribution. Figure 8B plots the points that correspond to the four conditional distributions in Figure 8A. The bounds on the set of allowable such points, $A_{M,i}$, are superimposed on Figure 8B for comparison.

The derivation of $A_{M,i}$, the set of allowable $\{P(c\,|\,r_j),\, I(\mathbb{S}, r_j)\}$ points, is a natural extension of the results from sections 4.1 and 4.2. Assume that we know $P(\mathbb{S})$ (a reasonable assumption, as in most discrimination tasks $P(\mathbb{S})$ is controlled by the experimenter). Then $H(\mathbb{S})$, the first term in equation 4.11 is fixed. Hence, $I(\mathbb{S}, r_j)$ varies only with the response-specific entropy $H(\mathbb{S}\,|\,r_j)$ (see section 4.2), the second term in equation 4.11. As shown in section 4.2 $\{P(c\,|\,r_j),\, H(\mathbb{S}\,|\,r_j)\}$ must lie in the set $A_{M,h}$, circumscribed by $h_{\max}(M_{P(c\,|\,r)})$ and $h_{\min}(M_{P(c\,|\,r)})$, bounds that are defined in equations 4.7′ and 4.9′, respectively. It follows from basic properties of inequalities that for a given $P(c\,|\,r_j)$ the lower and upper bounds on specific information are:

$$i_{\min}(P(\mathbb{S})_{P(c\,|\,r)}) = H(\mathbb{S}) - h_{\max}(M_{P(c\,|\,r)}) \tag{4.12}$$

$$i_{\max}(P(\mathbb{S})_{P(c\,|\,r)}) = H(\mathbb{S}) - h_{\min}(M_{P(c\,|\,r)}). \tag{4.13}$$

Figure 9 shows examples of $i_{\min}(P(\mathbb{S})_{P(c\,|\,r)})$ and $i_{\max}(P(\mathbb{S})_{P(c\,|\,r)})$ for different values of M and $H(\mathbb{S})$.

Four consequences of equations 4.12 and 4.13 deserve mention. First, if M (the number of stimuli) is fixed and $H(\mathbb{S})$ varies, the shape of $A_{M,i}$ does not change, but is merely shifted vertically with $H(\mathbb{S})$ (see Figure 9).

Second, the greater $H(\mathbb{S})$ is, the fewer possible negative $I(\mathbb{S}, r_j)$ values there are. If $H(\mathbb{S})$ is maximized and equals $\log(M)$, then $I(\mathbb{S}, r_j)$ is always greater than or equal to zero (see Figure 9). This is because in such a case, $P(\mathbb{S}\,|\,r_j)$ cannot be more evenly spread than $P(\mathbb{S})$, which is the necessary condition for the existence of a negative $I(\mathbb{S}, r_j)$. At the other extreme, as $H(\mathbb{S}) \to 0$, most distributions $P(\mathbb{S}\,|\,r_j)$ have greater entropy than the highly concentrated $P(\mathbb{S})$, so there exists a greater number of possible negative values of $I(\mathbb{S}, r_j)$.

Third, the range of permissible $I(\mathbb{S}, r_j)$ values at a particular $P(c\,|\,r_j)$ increases with $\log(M - 1)$. This is because $i_{\max}(P(\mathbb{S})_{P(c\,|\,r)})$ does not depend on M, and for a given $P(c\,|\,r_j)$, $i_{\min}(P(\mathbb{S})_{P(c\,|\,r)})$ decreases with $\log(M - 1)$. As with $H(\mathbb{S})$, at a given value of $H(\mathbb{S}\,|\,r_j)$, the range of permissible $P(c\,|\,r_j)$ values is a saturating exponential function of M (see section 4.1.2).

Fourth, because $i_{\min}(P(\mathbb{S})_{P(c\,|\,r)})$ and $i_{\max}(P(\mathbb{S})_{P(c\,|\,r)})$ are invertible functions, it is possible to use numerical methods to calculate the range of $P(c\,|\,r_j)$—values consistent with a given $I(\mathbb{S}, r_j)$ (see section 4.1.2).

4.4 Equivocation and the Ideal Observer. If we average the response-conditional entropy $H(\mathbb{S}\,|\,r_j)$ (see section 4.2) over all $\Re$, we obtain the

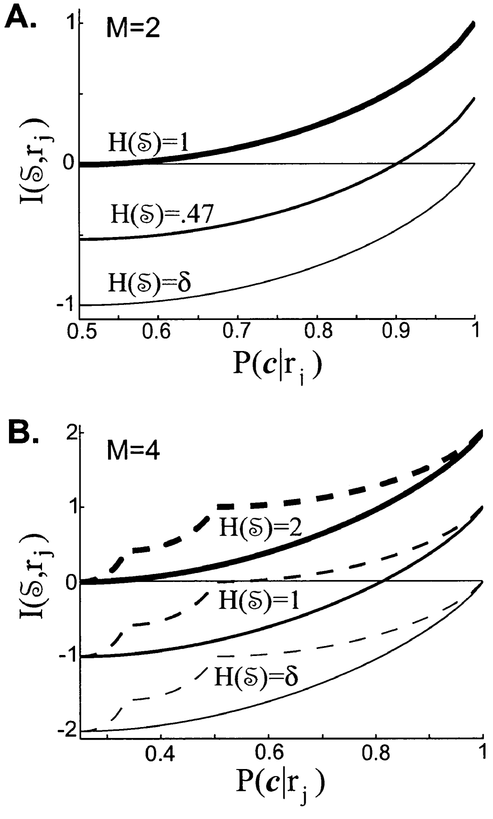

Figure 9: Comparison of $I(\mathbb{S}, r_j)$ and $P(c)$. (A) Plot of upper and lower bounds on $I(\mathbb{S}, r_j)$ as a function of $P(c)$ (dashed and solid lines, respectively). In this example, $M=2$, so the upper and lower bounds are the same and the dashed lines are not visible. The three sets of bounds shown correspond to cases in which stimulus distributions with different $H(\mathbb{S})$ values are chosen, and these entropy values are indicated next to the corresponding three lines. δ stands for an arbitrarily small real number. (B) Same as A, but with $M=4$.

equivocation (Cover & Thomas, 1991)

$$H(\mathbb{S} \mid \mathfrak{R}) = \sum_{j=1}^{N} P(r_j) H(\mathbb{S} \mid r_j) = -\sum_{j=1}^{N} \sum_{i=1}^{M} P(s_i, r_j) \log(P(s_i \mid r_j)). \quad (4.14)$$

$H(\mathbb{S} \mid \mathfrak{R})$ is often described as the average uncertainty remaining in $\mathbb{S}$ once $\mathfrak{R}$ is given (Ash, 1965). It describes the average minimum number of binary digits required to encode the value of $\mathbb{S}$ once $\mathfrak{R}$ is specified.

Before describing the general relationship between $H(\mathbb{S} \mid \mathfrak{R})$ and $P(c)$, we consider the examples in Figure 10A, which show four different joint

A.

i	P($\mathbb{S}$,$\mathfrak{R}$)	P(c)	H($\mathbb{S}$ | $\mathfrak{R}$)	I($\mathbb{S}$,$\mathfrak{R}$)
1	$\begin{pmatrix} .25 & 0 & .25 \\ 0 & .25 & .25 \end{pmatrix}$	.75	.5	.5
2	$\begin{pmatrix} .375 & .0625 & .0625 \\ .125 & .1875 & .1875 \end{pmatrix}$	.75	.81	.19
3	$\begin{pmatrix} .4125 & 0 & .0875 \\ .0875 & 0 & .4125 \end{pmatrix}$	.825	.67	.33
4	$\begin{pmatrix} .2225 & .055 & .2225 \\ .0275 & .445 & .0275 \end{pmatrix}$	.89	.5	.5

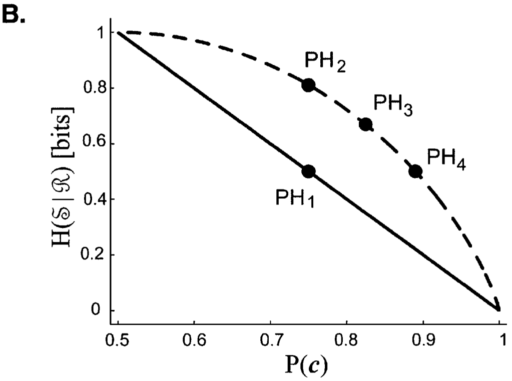

Figure 10: Examples comparing $P(c)$ with $H(\mathbb{S} \mid \mathfrak{R})$ and $I(\mathbb{S},\mathfrak{R})$. (A) The second column of the table contains four joint probability distributions. Columns 3–5 contain the corresponding values of $P(c)$ (see equation 3.3), $H(\mathbb{S} \mid \mathfrak{R})$ (see equation 4.14), and $I(\mathbb{S},\mathfrak{R})$ (see equation 4.17), respectively. (B) Scatter plot of the points $\{P(c), H(\mathbb{S} \mid \mathfrak{R})\}$ corresponding to the rows from the table: the point labeled PH_i corresponds to row i. The absolute upper and lower bounds on $H(\mathbb{S} \mid \mathfrak{R})$ are overlaid for comparison.

distributions ($M = 2$, $N = 3$). For each distribution, $P(\mathbb{S}) = \{\frac{1}{2}, \frac{1}{2}\}$, so $H(\mathbb{S}) = 1$ bit. Given the joint distribution $P(\mathbb{S},\mathfrak{R})$, $H(\mathbb{S} \mid \mathfrak{R})$ and $P(c)$ can be calculated by substituting the appropriate terms into equations 4.14 and 3.3, respectively. The examples highlight two features of the relationship between $H(\mathbb{S} \mid \mathfrak{R})$ and $P(c)$.

First, $H(\mathbb{S} \mid \mathfrak{R})$ is not uniquely specified by $P(c)$ (rows 1 and 2). The converse also holds (rows 1 and 4). These examples illustrate that $H(\mathbb{S} \mid \mathfrak{R})$ and $P(c)$ measure quite different features of $P(\mathbb{S},\mathfrak{R})$. $P(c)$ remains unchanged as long as the sum of the maximal elements in the columns remains the same, and the nonmaximal elements of $P(\mathbb{S},\mathfrak{R})$ can arbitrarily vary within this constraint. Equivocation, on the other hand, increases as the entropy of this nonmaximal probability mass is increased. For example, the gray area in

Figure 4 shows the probability mass that contributes to $P(e)$. Provided that the maximal elements remain unchanged, $H(\mathbb{S} \mid \mathfrak{R})$ will increase as this gray area is spread out and decrease as the gray area becomes more concentrated.

Second, higher equivocation does not imply lower $P(c)$ (rows 1 and 3). That is, just because a variable $\mathfrak{R}_1$ removes more uncertainty about $\mathbb{S}$ than another variable $\mathfrak{R}_2$ (i.e., $H(\mathbb{S} \mid \mathfrak{R}_1) < H(\mathbb{S} \mid \mathfrak{R}_2)$), this does not imply that an ideal observer can better estimate $\mathbb{S}$ on the basis of $\mathfrak{R}_1$.

More generally, just as with $H(\mathbb{S})$, there exist upper and lower bounds on $H(\mathbb{S} \mid \mathfrak{R})$ as a function of $P(c)$. For a given M and $P(c)$, we denote the upper and lower bounds on $H(\mathbb{S} \mid \mathfrak{R})$ as $H_{\max}(M_{P(c)})$ and $H_{\min}(M_{P(c)})$, respectively. Previous papers (Tebbe & Dwyer, 1968; Kovalevsky, 1968) show that

$$H_{\max}(M_{P(c)}) = h_{\max}(M_{P(c)}). \tag{4.15}$$

Equation 4.15 shows that the upper bound on equivocation is the same as the upper bound on entropy. Figure 11A plots $H_{\max}(M_{P(c)})$ for the cases $M = 2$, 4, and 64 (dashed lines). Obviously, $H_{\max}(M_{P(c)})$ has the exact same properties as $h_{\max}(M_{P(c)})$ discussed in section 4.1.2. Equation 4.15 is equivalent to Fano's inequality (Cover & Thomas, 1991), and provides an upper bound on $P(c)$. That is, it delineates the best $P(c)$ consistent with a given equivocation.

On the other hand, it is possible that the actual $P(c)$ of a distribution with a given value of $H(\mathbb{S} \mid \mathfrak{R})$ is much lower than the upper bound provided by equation 4.15. Previous papers derive $H_{\min}(M_{P(c)})$ (Tebbe & Dwyer, 1968; Kovalevsky, 1968), which provides the lowest $P(c)$ consistent with a given equivocation. We present their result without proof. For $P(c)$ between $\frac{1}{M}$ and 1, there exists an integer k (the same k that was introduced in equation 4.9) such that $\frac{1}{k+1} \leq P(c) \leq \frac{1}{k}$, and

$$H_{\min}(M_{P(c)}) = \log(k) + k(k+1)\log\left(\frac{k+1}{k}\right)\left(P(e) + \frac{1-k}{k}\right). \tag{4.16}$$

Just as was the case with $h_{\min}(M_{P(c)})$, $H_{\min}(M_{P(c)})$ is broken up into $M - 1$ line segments. Each segment is the straight line that connects the end points of the $M - 1$ arch-shaped segments that delineate $h_{\min}(M_{P(c)})$, the lower bounds on entropy (see section 4.1.2). The lower bounds on equivocation are shown in Figure 11A for the $M = 2$, 4, and 64 cases (solid lines), and we include the lower bounds on entropy $[h_{\min}(M_{P(c)})]$ in the same figure for comparison (light gray lines). Note that $H_{\min}(M_{P(c)})$ is not a function of M, so at a given value of $P(c)$, increasing M does not change the lower bound on equivocation.

$H_{\max}(M_{P(c)})$ and $H_{\min}(M_{P(c)})$ circumscribe the set $A_{M,H}$ of allowable $\{P(c), H(\mathbb{S} \mid \mathfrak{R})\}$ points, which are indicated in Figure 11A for $M = 2$, 4, and 64.[10] We highlight four features of $A_{M,H}$. First, Feder and Merhav (1994) give

[10] Formally, $A_{M,H}$ is the convex hull of $A_{M,h}$ (Tebbe & Dwyer, 1968; Kovalevsky, 1968).

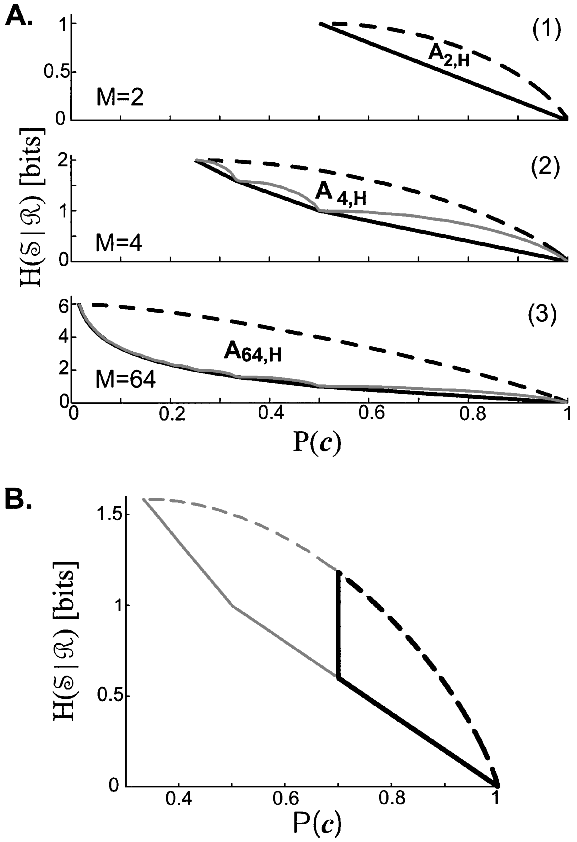

Figure 11: Comparison of $P(c)$ and $H(\mathbb{S}\mid\mathfrak{R})$. (A) Plot of the upper and lower bounds on $H(\mathbb{S}\mid\mathfrak{R})$ as a function of $P(c)$ for $M=2$, $M=4$, and $M=64$ (panels 1, 2, and 3, respectively). The upper bound ($H_{\max}(M_{P(c)})$) is indicated by a dashed line and the lower bound ($H_{\min}(M_{P(c)})$) by a solid line. The corresponding lower bounds on entropy ($h_{\min}(M_{P(c)})$) are overlaid in gray for comparison. The sets of allowable $\{P(c), H(\mathbb{S}\mid\mathfrak{R})\}$ pairs are labeled $A_{M,H}$. (B) Example of how $P(\mathbb{S})$ further constrains what values of $P(c)$ are consistent with a given $H(\mathbb{S}\mid\mathfrak{R})$ ($M=3$ in this example). In this case, we assume that $P(s_{\max})=0.7$, so $P(c)$ cannot be less than 0.7 (see section 3.1). The bounds on $H(\mathbb{S}\mid\mathfrak{R})$ with this additional constraint included are shown in black, and the bounds provided by equations 4.15 and 4.16 alone are shown in gray for comparison.

algorithms for generating joint distributions that actually attain the bounds, so the bounds are tight.[11]

[11] As a caveat, note that for a fixed input distribution $P(\mathbb{S})$, these bounds are not necessarily tight (see section 4.5.2).

Second, just as was the case with $H(\mathbb{S})$, for a given $P(c)$ the range of possible $H(\mathbb{S}\,|\,\mathbb{R})$ values increases with $\log(M-1)$. This is because the only term from equations 4.15 and 4.16 that varies with M is the $\log(M-1)$ term that equation 4.15 inherits from equation 4.7′. As a corollary, the range of $P(c)$ values consistent with a given $H(\mathbb{S}\,|\,\mathbb{R})$ increases exponentially with M.

Third, although $H(\mathbb{S}\,|\,\mathbb{R})$ does not uniquely map onto $P(c)$, the fact that $H_{\max}(M_{P(c)})$ and $H_{\min}(M_{P(c)})$ are invertible (Feder & Merhav, 1994) implies that it is possible to calculate upper and lower bounds on $P(c)$ for a given $H(\mathbb{S}\,|\,\mathbb{R})$ (see section 4.1 for details). Also, if $P(\mathbb{S})$ is known, then an additional constraint can be used to further narrow the range of $P(c)$ values consistent with a given equivocation. Namely, since $P(c)$ must be greater than or equal to $P(s_{\max})$ (see section 3.1), we can eliminate all $P(c)$ values below $P(s_{\max})$. As illustrated in Figure 11B, this constraint can considerably tighten the range of $P(c)$ values consistent with a given $H(\mathbb{S}\,|\,\mathbb{R})$.

Fourth, $A_{M,H}$ does not depend on the response distribution $P(\mathbb{R})$ or the number of possible responses N. If the goal is to make inferences between $H(\mathbb{S}\,|\,\mathbb{R})$ and $P(c)$, this is a very useful property, as it allows us to avoid two practical problems. First, it is in general a very difficult problem to obtain unbiased estimates of N and $P(\mathbb{R})$. This is partly because when $\mathbb{R}$ is a variable describing a neuronal response, N is usually quite large and most outcomes have a very low probability of occurring. In such cases, estimates of N and $P(\mathbb{R})$ suffer from biases due to undersampling (Paninski, 2004; Orlitsky, Santhanam, & Zhang, 2003). Second, if $H_{\max}(M_{P(c)})$ and $H_{\min}(M_{P(c)})$ depended on N, then for each representation of the neural response, we would have to recalculate the upper and lower bounds on $H(\mathbb{S}\,|\,\mathbb{R})$ in order to calculate the corresponding bounds on $P(c)$. The independence of $A_{M,H}$ from N and $P(\mathbb{R})$ circumvents both of these problems.

4.5 Mutual Information and the Ideal Observer. Typically, researchers calculate equivocation as an intermediate step in the estimation of mutual information, the information-theoretic quantity most often used by neuroscientists. The mutual information (also called transinformation and transmitted information) between random variables $\mathbb{S}$ and $\mathbb{R}$ is the average (over $\mathbb{R}$) specific information:

$$I(\mathbb{S},\mathbb{R}) = \sum_{j=1}^{N} P(r_j)\, I(\mathbb{S},r_j) = H(\mathbb{S}) - H(\mathbb{S}\,|\,\mathbb{R}) = \sum_{j=1}^{N}\sum_{i=1}^{M} P(s_i,r_j) \times \log\left(\frac{P(s_i,r_j)}{P(s_i)P(r_j)}\right). \quad (4.17)$$

$I(\mathbb{S},\mathbb{R})$ is the standard measure of how much information a variable $\mathbb{R}$ transmits about variable $\mathbb{S}$ (Ash, 1965; Cover & Thomas, 1991).

4.5.1 Three Interpretations of I($\mathbb{S}$,$\Re$). Almost universally, $I(\mathbb{S},\Re)$ is interpreted as a measure of the average amount of uncertainty removed by $\Re$ about $\mathbb{S}$ (Ash, 1965; Cover & Thomas, 1991). Operationally, what does this mean? In this section, we discuss three interpretations of $I(\mathbb{S},\Re)$ that often motivate the use of $I(\mathbb{S},\Re)$ by neuroscientists.

First, $I(\mathbb{S},\Re)$ indicates how many fewer binary digits are required, on average, to encode $\mathbb{S}$ once $\Re$ is specified. This follows from the fact that the specific information $I(\mathbb{S}, \mathrm{r}_j)$ quantifies how many fewer digits are required to encode S once r_j is known (see section 4.3), and $I(\mathbb{S},\Re)$ is the average of $I(\mathbb{S},r_j)$ over all $\Re$.

Second, $I(\mathbb{S},\Re)$ is a direct measure of the degree of statistical dependence between $\mathbb{S}$ and $\Re$ (Schneidman, Bialek, & Berry, 2003). If $\mathbb{S}$ and $\Re$ are independent variables, then for all stimulus-response pairs, $P(s_i,r_j) = P(s_i)P(r_j)$ (Yates & Goodman, 1999). This equality implies that all arguments of the log in equation 4.17 are one, so $I(\mathbb{S},\Re)$ is zero. Interestingly, if observed frequencies are used to estimate the probabilities in equation 4.7, then $I(\mathbb{S},\Re)$ is one-half the log-likelihood test statistic (G^2) when the null hypothesis is that $\mathbb{S}$ and $\Re$ are independent (Forbes, 1995).

A third interpretation of $I(\mathbb{S},\Re)$ is that it measures stimulus discriminability, or how well the stimulus can be predicted given the neuronal response (Alkasab et al., 1999; Arabzadeh et al., 2004; Buracas & Albright, 1999; Li et al., 2004; Paz & Vaadia, 2004; Petersen et al., 2002; Pola et al., 2003; Theunnissen & Miller, 1991; for exceptions, see Oram, Foldiak, Perret, & Sengpiel, 1998; Treves, 1997). That is, researchers often assume that $I(\mathbb{S},\Re)$ can be used as a surrogate for ideal observer performance, $P(c)$. For example, one paper claims, "Mutual information quantifies how well an ideal observer of neuronal responses can discriminate between all the different stimuli based on a single trial" (Pola et al., 2003, p. 37). Even in time-series analysis, *predictive information* is defined as the mutual information between past and future events (Bialek, Nemenman, & Tishby, 2001), again suggesting that higher mutual information implies improved predictability.

When examined quantitatively, this third interpretation of $I(\mathbb{S},\Re)$ is shown to be incorrect. Let us assume $P(\mathbb{S})$ is fixed, $\Re_1$ and $\Re_2$ are different representations of the neural response (e.g., PSTHs at different bin widths), and $P(c)_i$ is ideal observer performance when observing response variable $\Re_i$. The third interpretation is equivalent to

$$I(\mathbb{S},\Re_1) \geq I(\mathbb{S},\Re_2) \Leftrightarrow P(c)_1 \geq P(c)_2. \tag{4.18}$$

That is, if $\Re_1$ carries more information about $\mathbb{S}$ than $\Re_2$, then an ideal observer would be better at estimating $\mathbb{S}$ on the basis of $\Re_1$ than $\Re_2$. That equation 4.18 is false has been known since 1965 when a single counterexample was published (Wagner, 1965). In the next section we examine the general relationship between $I(\mathbb{S},\Re)$ and $P(c)$.

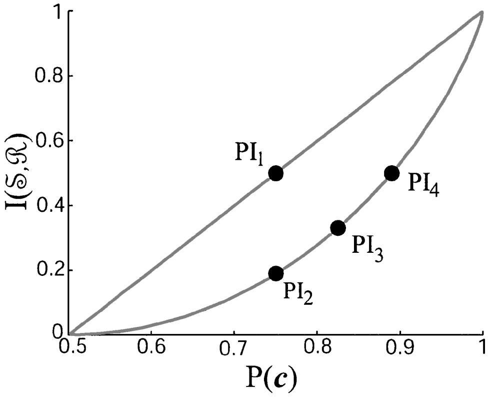

Figure 12: Examples comparing $I(\mathbb{S},\mathfrak{R})$ to $P(c)$. Each point labeled PI_i is the $\{P(c), I(\mathbb{S},\mathfrak{R})\}$ point corresponding to row i of the table in Figure 10A. The upper and lower bounds on $I(\mathbb{S},\mathfrak{R})$ as a function of $P(c)$ are overlaid in gray for comparison. In this example, $M = 2$.

4.5.2 Comparing $I(\mathbb{S},\mathfrak{R})$ and P(c). We highlight certain features of the relationship between $I(\mathbb{S},\mathfrak{R})$ and $P(c)$ using previous examples: the mutual information corresponding to the four joint distributions in Figure 10A is shown in column 5 of the table in Figure 10A. Figure 12 illustrates the $\{P(c), I(\mathbb{S},\mathfrak{R})\}$ points associated with each of these joint distributions. We note two features of the relationship between $P(c)$ and $I(\mathbb{S},\mathfrak{R})$ from these examples. First, equation 4.18 is incorrect, as can be seen by comparing rows 1 and 3. Second, $P(c)$ is not associated with a unique $I(\mathbb{S},\mathfrak{R})$ (rows 1 and 2). The converse is also true (rows 1 and 4). These examples suggest that there is a range of $I(\mathbb{S},\mathfrak{R})$ values consistent with a given $P(c)$, and vice versa.

To address their relationship more generally, we derive upper and lower bounds on $I(\mathbb{S},\mathfrak{R})$ as a function of $P(c)$.[12] Assume $P(\mathbb{S})$, and hence $H(\mathbb{S})$, is fixed by the experimenter. For a given $P(c)$, there exist the following upper and lower bounds on $I(\mathbb{S},\mathfrak{R})$:

$$I_{\min}\left(P(\mathbb{S})_{P(c)}\right) = H(\mathbb{S}) - H_{\min}(M_{p(c)}) \tag{4.19}$$

$$I_{\max}\left(P(\mathbb{S})_{P(c)}\right) = H(\mathbb{S}) - H_{\max}(M_{p(c)}), \tag{4.20}$$

where $H_{\max}(M_{P(c)})$ and $H_{\min}(M_{P(c)})$ are as defined in equations 4.12 and 4.13 (see section 4.4). The proof is as follows. Since the first term in equation 4.17 $[H(\mathbb{S})]$ is constant, $I(\mathbb{S},\mathfrak{R})$ varies only with $H(\mathbb{S}\,|\,\mathfrak{R})$. However, as

[12] Treves (1997) addresses a similar question under the assumption that $M = N$, that is, the number of responses equals the number of stimuli. Our results relax that assumption and produce tighter upper bounds.

described in section 4.4, we know that all points $\{P(c), H(\mathbb{S}\,|\,\mathfrak{R})\}$ must lie in $A_{M,H}$, the region whose lower and upper bounds are given by equations 4.12 and 4.13, respectively. Equations 4.19 and 4.20 follow from this fact and basic properties of inequalities.

Clearly the bounds imposed on $I(\mathbb{S},\mathfrak{R})$ by equations 4.19 and 4.20 can be tightened further. For one, $I(\mathbb{S},\mathfrak{R})$ is always nonnegative (Cover & Thomas, 1991). Second, $P(c)$ cannot be less than $P(s_{\max})$ (see section 3.1). Plots of the bounds on $I(\mathbb{S},\mathfrak{R})$ with these additional two constraints added are shown in Figure 13 for different stimulus distributions $P(\mathbb{S})$ (solid bold lines). The bounds provided by equations 4.19 and 4.20 alone are shown in light gray.

We mention three facts about the relationship between $P(c)$ and $I(\mathbb{S},\mathfrak{R})$. First, since $I_{\min}(P(\mathbb{S})_{P(c)})$ and $I_{\max}(P(\mathbb{S})_{P(c)})$ are one-to-one, monotonically increasing functions of $P(c)$, both functions have inverses that can be estimated using the methods discussed in section 4.1. Hence, given an estimate of mutual information $I(\mathbb{S},\mathfrak{R})$, it is possible to calculate the range of $P(c)$ values consistent with that estimate. As with all other quantities discussed so far, the range of $P(c)$ values consistent with a given $I(\mathbb{S},\mathfrak{R})$ increases exponentially with M, the number of stimuli.

Second, neither $I_{\min}(P(\mathbb{S})_{P(c|r)})$ nor $I_{\max}(P(\mathbb{S})_{P(c|r)})$ depends on $P(\mathfrak{R})$ or N. We discussed the benefits of this fact in section 4.4.

Finally, using numerical optimization, we have generated many joint distributions $P(\mathbb{S},\mathfrak{R})$ that reach the bounds $I_{\min}(P(\mathbb{S})_{P(c)})$ and $I_{\max}(P(\mathbb{S})_{P(c)})$. Four such distributions are provided in Figure 10A. However, we have not proved in general that the bounds are tight. That is, for an arbitrary $P(\mathbb{S})$ and $P(c)$, it is not guaranteed that there exists a distribution $P(\mathfrak{R},\mathbb{S})$ such that $I(\mathbb{S},\mathfrak{R}) = I_{\min}(P(\mathbb{S})_{P(c)})$ or $I_{\max}(P(\mathbb{S})_{P(c)})$. We conjecture that the bounds are not generally tight. Aside from numerical optimization procedures we have implemented in which the bounds were not reached (data not shown), this conjecture is based on the following reasoning. Given $P(\mathbb{S})$ and $P(c)$, there must exist channel matrices $P(\mathfrak{R}\,|\,\mathbb{S})$ and response distributions $P(\mathfrak{R})$ such that $P(\mathfrak{R}\,|\,\mathbb{S})^T P(\mathbb{S}) = P(\mathfrak{R})$. For this equation to hold and the bounds on $I(\mathbb{S},\mathfrak{R})$ to be tight, four constraints must be satisfied: the rows of $P(\mathfrak{R}\,|\,\mathbb{S})$ must sum to 1, as must the elements of $P(\mathfrak{R})$, the resultant joint distribution $P(\mathfrak{R},\mathbb{S})$ must satisfy equation 3.3, and $P(\mathbb{S},\mathfrak{R})$ must reach the bound on $I(\mathbb{S},\mathfrak{R})$ given in equation 4.19 or 4.20. Even if $N > M > 1$, we think it is unlikely that these constraints can be satisfied for an arbitrary $P(\mathbb{S})$, $P(c)$ pair. An interesting area for future research would be to use these constraints to derive even tighter bounds on $I(\mathbb{S},\mathfrak{R})$.

4.6 Channel Capacity and the Ideal Observer. The capacity of a channel $P(\mathfrak{R}\,|\,\mathbb{S})$ is defined as (Cover & Thomas, 1991)

$$C(P(\mathfrak{R}\,|\,\mathbb{S})) = \max_{P(\mathbb{S})} I(\mathbb{S},\mathfrak{R}). \tag{4.21}$$

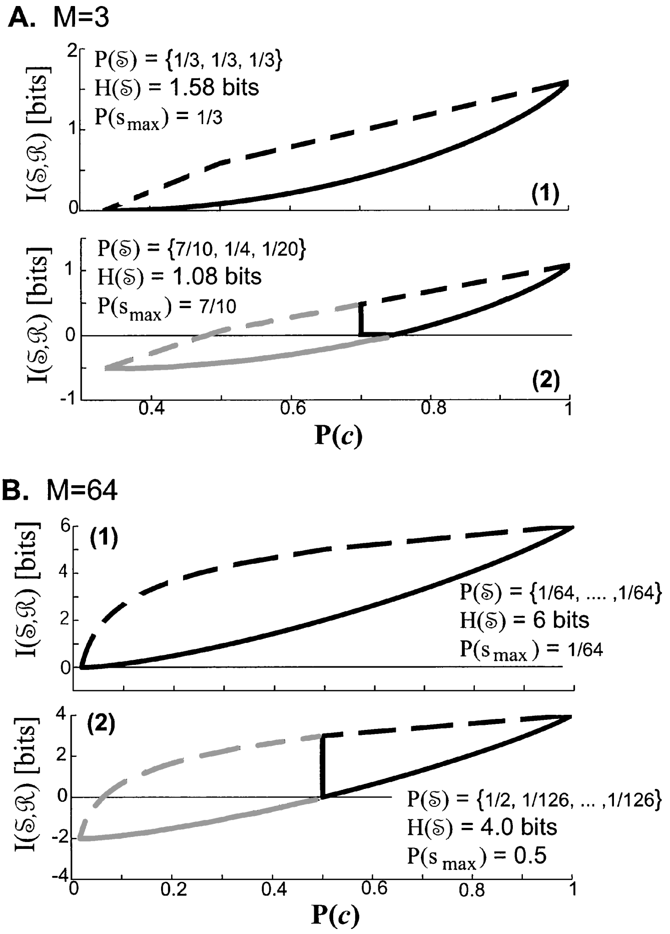

Figure 13: Comparison of $I(\mathbb{S},\mathfrak{R})$ and $P(c)$. (A) Plots of $I_{\max}(P(\mathbb{S})_{P(c)})$ and $I_{\min}(P(\mathbb{S})_{P(c)})$ as functions of $P(c)$ (dashed and solid gray lines, respectively). In these examples, $M=3$. Panels 1 and 2 plot bounds on mutual information under two different assumptions about the value of the stimulus distribution $P(\mathbb{S})$, and these $P(\mathbb{S})$ values are shown in each panel. The corresponding entropy ($H(\mathbb{S})$) and maximum stimulus probability ($P(s_{\max})$) values are also shown in each panel. The black lines show the constraints that are added to the bounds due to the nonnegativity of $I(\mathbb{S},\mathfrak{R})$ and the fact that $P(c)$ cannot be less than $P(s_{\max})$ (see the text). (B) Same as in A, except $M=64$.

That is, the capacity is the maximum possible mutual information between $\mathbb{S}$ and $\mathfrak{R}$ using the specified channel, where the maximum is calculated over the set of all possible input distributions. We use $P_C(\mathbb{S})$ to denote the stimulus distribution that solves this maximization problem. In general,

Channel 1

$$P_1(\mathfrak{R} \mid \mathbb{S}) = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

$$P_{C_1}(\mathbb{S}) = \{\tfrac{1}{2}, \tfrac{1}{2}\}$$

$$P_1(c) = \tfrac{2}{3} \text{ and } C_1 = .08 \text{ bits}$$

Channel 2

$$P_2(\mathfrak{R} \mid \mathbb{S}) = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

$$P_{C_2}(\mathbb{S}) = \{\tfrac{1}{2}, \tfrac{1}{2}\}$$

$$P_2(c) = \tfrac{2}{3} \text{ and } C_2 = .33 \text{ bits}$$

Figure 14: Examples comparing $C(P(\mathfrak{R} \mid \mathbb{S}))$ to $P(c)$. As discussed in the text, although $C(P_2(\mathfrak{R} \mid \mathbb{S}))$ is greater than $C(P_1(\mathfrak{R} \mid \mathbb{S}))$ (i.e., $C_2 > C_1$), an ideal observer of both channels achieves the same performance ($P(c) = 2/3$).

calculating the channel capacity is a difficult problem that is often solved using numerical optimization techniques (Cover & Thomas, 1991).

As with the other information-theoretic quantities, we would like to know whether the fact that a channel has a higher capacity than another implies that an ideal observer of $\mathfrak{R}$ would better be able to predict $\mathbb{S}$ using that channel. More precisely, it would be interesting to know whether the following is true:

$$C(P_1(\mathfrak{R} \mid \mathbb{S})) > C(P_2(\mathfrak{R} \mid \mathbb{S})) \Leftrightarrow P_1(c) > P_2(c), \tag{4.22}$$

where $P_i(\mathfrak{R} \mid \mathbb{S})$ is channel i and $P_i(c)$ is ideal observer performance using channel i. Also, we stipulate that the $P(c)$ values are calculated when the input distribution $P(\mathbb{S})$ is set to $P_{Ci}(\mathbb{S})$, the stimulus distribution that satisfies equation 4.21 for channel i. To our knowledge, nobody has published analytical results that bear on equation 4.22. Also, one must be cautious in interpreting counterexamples to equation 4.22 because different channels will likely have different input distributions that cause the channel to reach capacity, and it is not clear that a direct comparison of $P(c)$ in such cases is appropriate.

The channels shown in Figure 14 provide a counterexample to equation 4.22, and we have constructed the example so that in both channels,

$P_C(\mathbb{S})$ is the same. The first channel is a symmetrical channel, so $C(P(\Re \mid \mathbb{S}))$ and $P_C(\mathbb{S})$ can be calculated with known formulas (Cover & Thomas, 1991).[13] By substituting the second channel's values into equation 4.17, the mutual information of the second channel can be simplified to $I(\mathbb{S}, \Re) = \frac{1}{3} H(\mathbb{S})$, which is maximized when $P(\mathbb{S}) = \{1/2, 1/2\}$. While the results of the previous sections do not suggest an obvious way to calculate upper and lower bounds on $C(P(\Re \mid \mathbb{S}))$ as a function of $P(c)$, the example in Figure 14 suggests that such bounds likely exist.

5 Extension to Continuous Distributions

5.1 Background and Definitions. The results in section 4 depend on the assumption that $\mathbb{S}$ and $\Re$ are discrete sets. In practice, this is not a significant limitation because continuous distributions are always effectively binned and discretized since we can never represent them with infinite precision. However, for conceptual completeness, we extend the previous analysis to continuous distributions. This extension requires that we modify both the definitions of the information measures and the measure of estimation error. The upshot of the analysis is that if $\mathbb{S}$ is continuous, the information measures provide even fewer constraints on the error measure than when $\mathbb{S}$ is discrete.

The entropy of a continuous distribution $\mathbb{S}$, also known as the *differential entropy*, is defined analogously to entropy for the discrete case, with the sum replaced by an integral (Cover & Thomas, 1991):

$$h(\mathbb{S}) = -\int_{-\infty}^{\infty} P(s) \log(P(s)) ds. \tag{5.1}$$

The remaining information measures such as equivocation are also defined analogously to the discrete case (Cover & Thomas, 1991). The term *differential entropy* is used instead of *entropy* because the differential entropy does not have the same mathematical properties as entropy. For example, the differential entropy changes when $\mathbb{S}$ is multiplied by a scaling factor a (i.e., $h(a\mathbb{S}) = h(\mathbb{S}) + \log(|a|)$) (Shannon & Weaver, 1949). Also, $h(\mathbb{S})$ can be negative (Shannon & Weaver, 1949). For example, the differential entropy of a uniform distribution defined over the interval $[-w/2, w/2]$ is $\log(w)$, which is negative for $w < 1$ (Cover & Thomas, 1991).

A second change we make to accommodate continuous variables is in our evaluation of estimation error. This is because when $\hat{s}$ is a point estimate of the stimulus, it follows from equation 3.2 that $P(c|r_j) = 0$, so $P(c) = 0$. Hence,

[13] For a symmetrical channel, $P_C(\mathbb{S})$ is the uniform distribution and the capacity is $\log(M) - H(P(\Re \mid si))$, where $H(P(\Re \mid s_i))$ is the entropy of an arbitrary row of the channel (Cover & Thomas, 1991).

$P(c)$ is an inappropriate measure of the success in estimating the value of a continuous variable $\mathbb{S}$ and needs to be replaced by an error term that increases with the distance between s and $\hat{s}$. The most common such measure is the squared error between s and $\hat{s}$, $(s - \hat{s})^2$ (Yates & Goodman, 1999). The average (over $\mathbb{S}$) squared error, or mean squared error, is a useful overall measure of the quality of $\hat{s}$ as an estimate of $\mathbb{S}$ (Yates & Goodman, 1999):

$$MSE(\mathbb{S}) = \int_{-\infty}^{\infty} P(s)(s - \hat{s})^2\, ds. \tag{5.2}$$

When using $MSE(\mathbb{S})$ to evaluate an estimate of $\mathbb{S}$, the goal is to use an estimate $\hat{s}$ that will minimize $MSE(\mathbb{S})$, or the minimum mean squared error estimator of $\mathbb{S}$. The minimum mean square estimator of any random variable $\mathbb{S}$ (discrete or continuous) is the mean of $\mathbb{S}$, which we denote as $\mu_{\mathbb{S}}$ (Yates & Goodman, 1999). As a corollary, the variance of $\mathbb{S}$, $\sigma^2_{\mathbb{S}}$, is the actual value of the minimum mean squared error, which we denote $E(\mathbb{S})$:

$$E(\mathbb{S}) = \int_{-\infty}^{\infty} p(s)(s - \mu_{\mathbb{S}})^2\, ds = \sigma^2_{\mathbb{S}}. \tag{5.3}$$

The first equality is just equation 5.2 with the minimum mean squared error estimator $\mu_{\mathbb{S}}$ substituted for $\hat{s}$, and the second equality is the definition of the variance of $\mathbb{S}$ (Yates & Goodman, 1999). While not technically an ideal observer, the minimum mean-squared error estimator is a close relative because it minimizes an error function.

In sum, to extend the results from section 4 to continuous distributions, we must first replace the entropy-based information measures with the differential entropy-based measures and then replace $P(c)$ with $E(\mathbb{S})$. In the following section, we briefly show how to calculate the bounds on each information measure as a function of $E(\mathbb{S})$. Note that this idea that such bounds could be derived was suggested but not carried out in Feder & Merhav (1994).

5.2 Comparing Information Measures and Minimum Mean Squared Error. The results from section 4 assume that $\mathbb{S}$ and $\mathbb{R}$ are discrete sets. The extension to the case where $\mathbb{R}$ is continuous and $\mathbb{S}$ is discrete is trivial, as none of the results depends on the assumption that $\mathbb{R}$ is discrete. However, for reasons described in the previous section, the extension to the case of continuous $\mathbb{S}$ requires a significant extension of the analysis, to which we now turn.

5.2.1 Differential Entropy and $E(\mathbb{S})$*.* For a given value of $E(\mathbb{S})$, there is an upper bound on $h(\mathbb{S})$ given by Cover and Thomas (in press):

$$h_{\max}(\mathbb{S}) = \frac{1}{2}\log(2\pi e E(\mathbb{S})). \tag{5.4}$$

Figure 11A plots this relationship.[14]

Equation 5.4 places a lower bound on $E(\mathbb{S})$ for a given value of $h(\mathbb{S})$ (Cover & Thomas, in press), as can be seen in Figure 15A. We prove that there exists no upper bound on $E(\mathbb{S})$ for a given value of $h(\mathbb{S})$. Assume that there exists such an upper bound $\mathfrak{U}$ on $E(\mathbb{S})$ when $h(\mathbb{S})$ takes on some arbitrary value $\mathfrak{h}$. To generate a counterexample, let $[-w/2, w/2]$ be the interval of a uniform distribution that is constructed to satisfy $\mathfrak{h} = \log(w)$, as depicted in Figure 15B.1. We then split this uniform distribution into two segments of width $w/2$, such that there is a distance D between these two segments (see Figure 15B.2). Because $E(\mathbb{S}) = \frac{w}{12} + \frac{1}{4}(D^2 + Dw)$, which can be verified by substituting the equation for the split distribution into equation 5.3, there exists a number D such that $E(\mathbb{S}) > \mathfrak{U}$. By indirect proof, we have shown that there exists no upper bound $\mathfrak{U}$ on the error for a given entropy.

The set **B** of possible $\{h(\mathbb{S}), E(\mathbb{S})\}$ points for continuous distributions includes the curve delineated by equation 5.4 and all points to the right of the curve (see Figure 15A). It follows that for a given $E(\mathbb{S})$, there is no lower bound on $h(\mathbb{S})$, or:

$$h_{\min}(\mathbb{S}) = -\infty \tag{5.5}$$

In sum, $h(\mathbb{S})$ can vary between $-\infty$ and $h_{\max}(\mathbb{S})$ at a given $E(\mathbb{S})$.

5.2.2 Response-Conditional Entropy and $E(\mathbb{S} \mid r_j)$. If we know that response r_j occurred and the resulting conditional distribution of $\mathbb{S}$ is $P(\mathbb{S} \mid r_j)$, then the same analysis as in section 5.2.1 applies. That is, the minimum MSE estimator of $\mathbb{S}$ given r_j is $\mu_{\mathbb{S}|r_j}$, the mean of $P(\mathbb{S} \mid r_j)$, and the error $E(\mathbb{S} \mid r_j)$ in this case is $\sigma^2_{\mathbb{S}|r_j}$, the variance of $P(\mathbb{S} \mid r_j)$ (Yates & Goodman, 1999). In this case, equations 5.4 and 5.5 apply, with $\mathbb{S}$ replaced by $\mathbb{S}|r_j$.

5.2.3 Specific Information and $E(\mathbb{S}|r_j)$. Following a proof strategy exactly analogous to that in section 4.3, it can be shown that

$$I(\mathbb{S}, r_j)_{\max} = \infty \tag{5.6}$$

$$I(\mathbb{S}, r_j)_{\min} = -h_{\max}(\mathbb{S}), \tag{5.7}$$

bounds, which are illustrated in Figure 15C.

[14] Note that equation 5.3 is stated in Cover and Thomas (in press) without proof. *Proof of equation* 5.4: In the set of all continuous distributions with fixed variance σ^2, the gaussian has the maximum differential entropy, which is $h_{\text{gauss}}(\mathbb{S}) = \frac{1}{2}\log\left(2\pi e\sigma^2\right)$ (Cover & Thomas, 1991). Also, since the minimum error estimator has error $E(\mathbb{S}) = \sigma^2$(see section 5.1), this implies that the maximum differential entropy consistent with a certain value of $E(\mathbb{S})$ is given by equation 5.3.

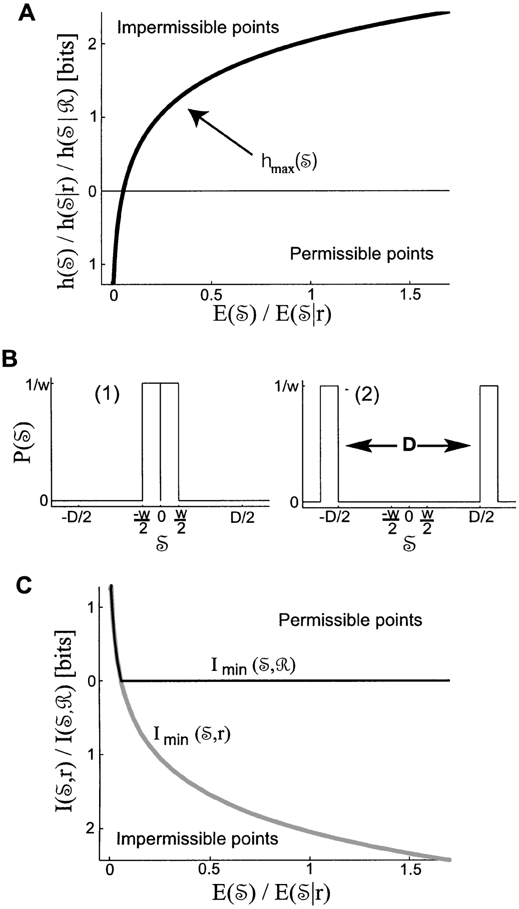

Figure 15: Comparison of continuous information measures with minimum mean squared error estimators. (A) Plot of the maximum differential entropy ($h_{\max}(\mathbb{S})$) as a function of the minimum mean squared error $E(\mathbb{S})$ (see equation 5.7). The set of points to the right of, and including, the curve is the set of permissible $\{E(\mathbb{S}), h(\mathbb{S})\}$ points. The same equation also describes the maximum response-conditional entropy and equivocation as a function of minimum mean-squared error (see the text). (B) Geometrical representation of the proof that there is no upper bound on error for a given entropy. Panel 1 shows a uniform distribution with width w. When this probability mass is split into two sections of width $w/2$, and these sections are separated by distance D (panel 2), $h(\mathbb{S})$ remains the same but the error increases with D^2 (see the text). (C) Plot of the lower bound on specific information (gray line) and mutual information (black line) as a function of $E(\mathbb{S}\,|r)$ and $E(\mathbb{S})$, respectively. There is no upper bound on the information measures: the points to right of, and including, the lines are the set of permissible $\{E(\mathbb{S}), I(\mathbb{S},\mathfrak{R})\}$ points (see the text).

5.2.4 Equivocation and $E(\mathbb{S})$. The average minimum mean squared error $E(\mathbb{S})$ when response variable $\Re$ is available to help estimate $\mathbb{S}$ is the mean (over $\Re$) of $E(\mathbb{S}\,|\,r_j)$:

$$E(\mathbb{S}) = \sum_{j=1}^{N} p(r_j) E(\mathbb{S}\,|\,r_j). \tag{5.8}$$

If $\Re$ is continuous, the sum can be replaced by the appropriate integral.

The bounds on $h(\mathbb{S}\,|\,\Re)$ are identical to the bounds on $h(\mathbb{S})$. To prove this, we use a mathematical technique borrowed from the papers that derived the bounds on equivocation in the case of discrete $\mathbb{S}$ (Tebbe & Dwyer, 1968; & Kovalevsky, 1968).[15] First, recall from section 5.2.2 that **B**, the set of permissible $\{E(\mathbb{S}\,|\,r_j), h(\mathbb{S}\,|\,r_j)\}$ points, is fixed by equations 5.4 and 5.5. Second, note that

$$\{E(\mathbb{S}), h(\mathbb{S}\,|\,\Re)\} = \sum_{j=1}^{N} P(r_j)\{E(\mathbb{S}\,|\,r_j), h(\mathbb{S}\,|\,r_j)\}, \tag{5.9}$$

where the sum can be replaced by an integral if $\Re$ is continuous. The right-hand side of equation 5.9 is simply a convex combination of points of the form $\{E(\mathbb{S}\,|\,r_j), h(\mathbb{S}\,|\,r_j)\}$, which from section 5.2.2 we know must be in the set **B**. Hence, the set $\mathbf{B}^*$ of permissible $\{E(\mathbb{S}), h(\mathbb{S}\,|\,\Re)\}$, pairs must lie in the convex hull of **B**. However, **B** is already a convex set, so $\mathbf{B} = \mathbf{B}^*$ (Boyd & Vandenberghe, 2004). In other words, the bounds on $h(\mathbb{S}\,|\,\Re)$ as a function of $E(\mathbb{S})$ are identical to the bounds on $h(\mathbb{S}\,|\,r_j)$ as a function of $E(\mathbb{S}\,|\,r_j)$, that is,

$$h_{\max}(\mathbb{S}\,|\,\Re) = h_{\max}(\mathbb{S}) \tag{5.10}$$

$$h_{\min}(\mathbb{S}\,|\,\Re) = h_{\min}(\mathbb{S}) = -\infty. \tag{5.11}$$

Bounds are shown in Figure 15A.

5.2.5 Mutual Information and E($\mathbb{S}$). Using the results from section 5.2.4, it is possible to use an argument analogous to that in section 4.5 to show

$$I_{\max}(\mathbb{S},\Re) = h(\mathbb{S}) - h_{\min}(\mathbb{S}) = \infty \tag{5.12}$$

$$I_{\min}(\mathbb{S},\Re) = h(\mathbb{S}) - h_{\max}(\mathbb{S}). \tag{5.13}$$

The fact that $I(\mathbb{S},\Re) \leq h(\mathbb{S})$ (Cover & Thomas, 1991) provides an additional constraint and the set of permissible $\{E(\mathbb{S}), I(\mathbb{S},\Re)\}$ points with all of these constraints included is shown in black in Figure 15C.

[15] For an introduction to the concepts from convex analysis used in the following proof, see Boyd & Vandenberghe (2004).

6 Discussion

6.1 Measuring Stimulus Discriminability. An animal that cannot discriminate successfully on single trials will not survive long in natural conditions. A frog, for example, will not get a second chance to catch a fly if it misses on its first try. As discussed in section 3.2, ideal observer performance, $P(c)$, is a useful and natural measure of single-trial stimulus discriminability. Researchers also often use information measures with the goal of quantifying discrimination performance. In section 4, we showed that this is typically not justified. In particular, rather than there being a one-to-one relationship between information-theoretic quantities and $P(c)$, there is typically a range of permissible $P(c)$ values associated with a given information measure. In section 5 we showed that when the analysis is extended to continuous stimulus distributions, the problems with the information measures are only exacerbated, as they provide no upper bounds on estimation error.

If the goal is to make inferences from information measures to $P(c)$, a caveat should be noted: as the number of stimuli (M) increases, the range of $P(c)$ values associated with a given information-theoretic quantity increases exponentially (see section 4). Hence, it would be prudent to pick a stimulus set that is small enough to significantly narrow the range of permissible $P(c)$ values. While the most conservative option is to use only two stimuli ($M = 2$), in some cases this will not be desirable because the stimulus space will not be adequately sampled.

6.2 Information Theory in Neuroscience. We have shown that to quantify stimulus discriminability, the tools of ideal observer analysis are preferable to those of information theory. Information theory, however, is helpful for answering questions that ideal observer analysis cannot address. We list three. We do not intend the list to be exhaustive, but are simply mentioning those applications that follow most naturally from the results in section 4.

First, as we discussed in section 4.5.1, mutual information $I(\mathbb{S},\Re)$ measures the degree of statistical dependence between random variables $\mathbb{S}$ and $\Re$. It is a useful measure because it is completely general: it will detect any deviation from independence whether it is due to linear correlation or some nonlinear dependency.[16] Ideal observer analysis does not directly quantify such dependencies. But even if two random variables are dependent, the tools of ideal observer analysis are required to determine how well one can be predicted from the other.

Second, there clearly exist psychophysical tasks that should be evaluated using information measures rather than $P(c)$. For instance, if the goal is to

[16] Of course, there exist other tests for statistical dependence (e.g., the χ^2 test), and it is up to the researcher to decide which statistic is appropriate for his or her data.

evaluate a subject's performance in the game of 20 Questions, then $H(\mathbb{S})$ indicates the best possible performance $\overline{L}_{\min}$ (see section 4.1.3).

Third, natural selection may have discovered a coding strategy that uses the fewest number of elementary symbols required to encode certain classes of stimuli. That is, natural selection may have solved the optimization problem of reaching $\overline{L}_{\min}$ (see section 4.1.1). Such a hypothesis would be challenging to test empirically (e.g., it requires determining the set of elementary symbols used in the neural code, a problem discussed extensively in Brenner, Strong, Koberle, Bialek, & de Ruyter van Steveninck, 2000). Regardless of such practical difficulties, information theory contains the quantitative resources to address such questions about neural coding, resources that ideal observer analysis does not provide.

6.3 Open Questions and Future Directions. We finish by discussing five outstanding questions that are theoretically interesting and, to our knowledge, have not been addressed by previous work. First, it would be interesting to extend the discussion in section 4.6 by generating general bounds on channel capacity as a function of ideal observer performance. Also, an extension of the result to capacity in continuous channels would be useful.

Second, we have analyzed quantities that depend on entropy rather than entropy rates (Cover & Thomas, 1991; Strong, Koberle, de Ruyter van Steveninck, & Bialek, 1998). However, Feder and Merhav (1994) show that for stationary processes, the results for $H(\mathbb{S})$ described in section 4.1 also apply to entropy rates, so we expect our results to generalize to information rates (Strong et al., 1998). That is, we expect that a neuron that transmits information at a higher rate than another neuron is not necessarily the better predictor of a time-varying stimulus. However, this idea should be formally developed.

Third, since spike trains are nonstationary (Berry & Meister, 1998), it would be interesting to determine how neurally inspired nonstationarities in $P(\mathbb{S})$ and $P(\mathfrak{R} \mid \mathbb{S})$ would affect the results in sections 4 and 5.

Fourth, by picking as our point of comparison the ideal observer, we have treated all errors equally.[17] This would be inappropriate in some instances. For example, if the goal is to estimate the spatial location of the stimulus, an estimate close to the actual location is better than an estimate that is completely off. In such cases, instead of judging an estimate as categorically right or wrong, the estimate should be evaluated by an error term that increases with the distance between $\hat{s}$ and s, such as the mean squared error (see section 5.1). We implicitly addressed this concern in section 5.2, in which we used the minimum mean squared error as a cost

[17] Technically, we have used a 0/1 loss function, also known as the Hamming distance between s and $\hat{s}$.

function. We showed that using the mean squared error as the cost function only amplifies the discrepancies between ideal observers and the information measures. However, a more general consideration of the relationship between arbitrary cost functions (Schneidman et al., 2003) and the information measures deserves analysis.

Fifth, one of the virtues often stressed of the information-theoretic approach to neural coding is that the information measures are nonparametric (Paz & Vaadia, 2004; Peterson et al., 2002). Ironically, it is partly because we freed the proofs in sections 4 and 5 from assumptions about $P(\mathbb{S})$ and $P(\Re \mid \mathbb{S})$ that it was possible to demonstrate the differences between the information measures and ideal observers. On the other hand, under certain assumptions (e.g., the responses to different stimuli are univariate gaussians with identical variances), the relationship between $P(c)$ and the information-theoretic quantities is one-to-one. We would like to know the minimal assumptions required about $P(\mathbb{S})$ and $P(\Re \mid \mathbb{S})$ for ideal observer performance to be uniquely specified by the information measures. We hope that the work presented here will act as an impetus for such an analysis so that we can know under what conditions it is possible to make unambiguous inferences between encoding and decoding measures in the nervous system.

Acknowledgments

Thanks to Kevin Briggman, E. J. Chichilnisky, Philip Gill, Dan Kersten, Samar Mehta, Joy Thomas, Jonathan Victor, the Sejnowski lab, Diane Whitmer, and an anonymous reviewer. Thanks especially to Barbara Leary.

References

Alkasab, T. K., Bozza, T. C., Cleland T. A., Dorries, K. M., Pearce, T. C., White, J., & Kauer, J. S. (1999). Characterizing complex chemosensors: Information-theoretic analysis of olfactory systems. *TINS*, *22*, 102–108.

Arabzadeh, E., Panzeri, S., & Diamond M. E. (2004). Whisker vibration information carried by rat barrel cortex neurons. *J. Neurosci.*, *24*, 6011–6020.

Ash, R. B. (1965). *Information theory*. New York: Dover.

Berry, M. J., & Meister, M. (1998). Refractoriness and neural precision. *J. Neurosci.*, *18*, 2200–2211.

Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computation*, *13*, 2409–2463.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Brenner, N., Strong, S. P., Koberle, R., Bialek, W., & de Ruyter van Steveninck, R. R. (2000). Synergy in a neural code. *Neural Comp.*, *12*, 1531–1552.

Britten, K., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *J. Neurosci.*, *12*, 4745–4767.

Buracas, G. T., & Albright, T. D. (1999). Gauging sensory representations in the brain. *TINS*, *22*, 303–309.

Cover, T. J., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

Cover T. J., & Thomas J. (In press). *Elements of information theory* (2nd ed.). New York: Wiley.

de Boor, C. (1978). *A practical guide to splines*. New York: Springer-Verlag.

DeWeese, M. R., & Meister, M. (1999). How to measure the information gained from one symbol. *Network: Comput. Neural Syst.*, *10*, 325–340.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley.

Feder, M., & Merhav, M. (1994). Relations between entropy and error probability. *IEEE Transactions on Information Theory*, *40*, 259–266.

Forbes, D. A. (1995). Classification-algorithm evaluation: Five performance measures based on confusion matrices. *J. Clin. Monit.*, *11*, 189–206.

Geisler, W. S. (1989). Ideal observer theory in psychophysics and physiology. *Physica Scripta*, *39*, 153–160.

Geisler, W. S. (2003). Ideal observer analysis. In L. M. Chalupa & J. S. Werner (Eds.), *Visual neurosciences*. Cambridge, MA: MIT Press.

Golic, J. (1987). On the relationship between the information measures and the Bayes probability of error. *IEEE Transactions on Information Theory*, *33*, 681–693.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Knill, D. C., & Kersten, D. (1991). Ideal perceptual observers for computation, psychophysics, and neural networks. In R. J. Watt (Ed.), *Pattern recognition by man and machine* (pp. 83–97). New York: Macmillan Press.

Kovalevsky, V. A. (1968). *Character readers and pattern recognition*. New York: Spartan Press.

Lawson, J. L., & Uhlenbeck, G. E. (1950). *Threshold signals*. New York: McGraw-Hill.

Lettvin, J., Maturana, H., McCulloch, W. S., & Pitts, W. (1959). What the frog's eye tells the frog's brain. *Proc. IRE*, *47*, 1940–1951.

Li, W., Piech, V., & Gilbert, C. D. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nat. Neurosci.*, *7*, 651–657.

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research*, *35*, 549–568.

MacKay, D., & McCulluch, D. S. (1952). The limiting information capacity of a neuronal link. *Bull. Math. Biophys.*, *14*, 127–135.

Oram, M. W., Foldiak, P., Perret, D. I., & Sengpiel, F. (1998). The "ideal homunculus": Decoding neural population signals. *TINS*, *21*, 259–265.

Orlitsky, A., Santhanam, N. P., & Zhang, J. (2003). Always good Turing: Asymptotically optimal probability estimation. *Science*, *302*, 427–431.

Paninski, L. (2004). Estimating entropy on *m* bins given fewer than *m* samples. *IEEE Transactions on Information Theory*, *50*, 2200–2203.

Paz, R., & Vaadia, E. (2004). Learning-induced improvement in encoding and decoding of specific movement directions by neurons in the primary motor cortex. *PLoS Biol*, *2*(2), e45. Available online: http://www.plosbiology.org/plosonline/?request=get-document&doi=10.1371%2F journal.pbio.0020045.

Perkel, D. H., & Bullock, T. H. (1968). Neural coding. *Neurosciences Research Bulletin*, *6*, 221–348.

Petersen, R. S., Panzeri, S., & Diamond, M. E. (2002). Population coding in somatosensory cortex. *Curr. Opin. Neurobiol., 12*, 441–447.

Pola, G., Thiele, A., Hoffmann, K. P., & Panzeri, S. (2003). An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network: Comput. Neural Syst., 14*, 35–60.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Schneidman, E., Bialek, W., & Berry, M. J. (2003). Synergy, redundancy, and independence in population codes. *J. Neurosci., 23*, 11539–11553.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Phys. Rev. Lett., 80*, 197–200.

Tebbe, D. L., & Dwyer, S. J. (1968). Uncertainty and the probability of error. *IEEE Transactions on Information Theory, 14*, 516–518.

Theunissen, F. E., & Miller, J. P. (1991). Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. *J. Neurophys., 66*, 1690–1703.

Treves, A. (1997). On the perceptual structure of face space. *BioSystems, 40*, 189–196.

Victor, J. D. (1999). Temporal aspects of neural coding in the retina and lateral geniculate: A review. *Network, 10*, R1–66.

Wagner, T. J. (1965). Some remarks concerning uncertainty and the probability of error. *IEEE Transactions on Information Theory, 11*, 144–145.

Yates, R. D., & Goodman, D. J. (1999). *Probability and stochastic processes: A friendly introduction for electrical and computer engineers.* New York: Wiley.

Received May 13, 2004; accepted September 4, 2004.