ARTICLE **Communicated by Liam Paninski**

Indices for Testing Neural Codes

Jonathan D. Victor
jdvicto@med.cornell.edu
Department of Neurology and Neuroscience and Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY 10065, U.S.A.

Sheila Nirenberg
shn2010@med.cornell.edu
Department of Physiology and Biophysics and Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY 10065, U.S.A.

One of the most critical challenges in systems neuroscience is determining the neural code. A principled framework for addressing this can be found in information theory. With this approach, one can determine whether a proposed code can account for the stimulus-response relationship. Specifically, one can compare the transmitted information between the stimulus and the hypothesized neural code with the transmitted information between the stimulus and the behavioral response. If the former is smaller than the latter (i.e., if the code cannot account for the behavior), the code can be ruled out.

The information-theoretic index most widely used in this context is Shannon's mutual information. The Shannon test, however, is not ideal for this purpose: while the codes it will rule out are truly nonviable, there will be some nonviable codes that it will fail to rule out. Here we describe a wide range of alternative indices that can be used for ruling codes out. The range includes a continuum from Shannon information to measures of the performance of a Bayesian decoder. We analyze the relationship of these indices to each other and their complementary strengths and weaknesses for addressing this problem.

1 Introduction

Information-theoretic analysis is a powerful tool to illuminate how neurons represent the sensory world. A fundamental reason for this is that the classic information measure, Shannon's mutual information, places a limit on the number of possible stimuli that can be distinguished from the output of a neural channel. Thus, measuring mutual information can serve as a way to determine whether the activity of a channel can account for behavioral performance in a sensory discrimination task.

For example, suppose one wanted to determine which features of the activity (e.g., firing rate, interspike intervals) are critical for performing the task. One can measure the mutual information between a given feature and the stimulus set in the task. If the mutual information between the feature and the stimulus set is less than the mutual information between the stimulus set and the behavioral performance, then one can deduce that that feature is insufficient to account for the behavior. That is, one can rule out a neural code based solely on that feature.

Mutual information, though, is not perfect, that is, it is not a highly stringent test. While any code that fails this test is guaranteed to be nonviable, there are conditions in which nonviable codes can pass. The basic reason for this is that information is lost when neural activity is converted into a behavioral response (response discretization). By not taking this loss into account, mutual information can fail to exclude codes that are, in fact, nonviable for producing behavior.

Like mutual information, the best performance of a Bayesian decoder is an index that can be used to test the viability of neural codes in a rigorous fashion. This index takes into account the above information loss and might therefore appear to be a more universal test. However, this measure is highly sensitive to assumptions about the decision criterion used in performing the behavioral task. Consequently, it also can fail to exclude nonviable codes under some situations, but these situations are distinct from those in which mutual information fails.

Here we show that these indices are but two of many possible choices. In particular, they represent the extremes of a continuum of indices that provide rigorous tests of neural codes. However, these indices differ in their ability to test codes for a range of reasons, including differences in their sensitivity to decision criterion, differences in their sensitivity to response discretization, and differences in their bias and variance characteristics. In this article, we analyze the properties of these indices, their relationships to each other, and their complementary strengths and weaknesses.

In the main section of this article, we approach the problem of ruling codes out assuming no knowledge of what the subject is thinking, that is, what the subject's priors and decision rules are. The Data Processing Inequality (DPI) justifies this, since it places absolute limits on the behaviors a code can support. Thus, any code that is ruled out by these indices is truly ruled out. In the appendixes, we show how priors and decision rules can be taken into account to allow still more codes to be ruled out.

2 Results

This article has two main components: one that focuses on general indices of information and one that focuses on using these indices to test coding hypotheses.

In the first component, the ideas build on the central notion that information represents a reduction in uncertainty. Uncertainty can be formalized as the extent to which a probability distribution is dispersed. We therefore begin by recognizing that there are many ways of quantifying dispersion. We then show that for every way of quantifying dispersion, there is a corresponding generalized index of transmitted information. The corresponding index is the extent to which a distribution becomes less dispersed by making an observation. These general indices of information include the familiar Shannon mutual information, the performance of the best Bayesian decoder, and many others. All of these indices satisfy the DPI and therefore provide means to test coding hypotheses.

Then we show how these indices can be used to test coding hypotheses. We consider several examples that illustrate how the indices differ and identify their complementary strengths and weaknesses.

2.1 Indices of Concentration. Let P denote a probability distribution. To quantify its dispersion, we consider its opposite, namely concentration, as this will enable us to make use of the properties of convex functions. We define an index of concentration f to be a convex function on a probability distribution P and denote this by $f(P)$.[1]

The range of indices of concentration is captured by two well-known examples:

$$f_{\max}(P) = \max_i(p_i), \tag{2.1}$$

and

$$-H_1(P) = \sum_{i=1}^{L} p_i \log p_i, \tag{2.2}$$

where P is a probability distribution on L symbols and p_i is the probability associated with the ith symbol. The first index, equation 2.1, is associated with a Bayesian measure of information, as we show in the text following equation 2.4.[2] The second index, equation 2.2, is the negative of the Shannon entropy, the quantity associated with Shannon mutual information.

These two indices have quite different properties: the Bayes index $f_{\max}$, equation 2.1, depends on only the largest probability, while the Shannon index, equation 2.2, depends smoothly on all of the probabilities. These

[1]The convexity property (made explicit in equation 2.5) states that f never increases when distributions are mixed—that the concentration of a weighted average of distributions is no greater than the weighted average of the concentrations. This justifies the use of the term *index of concentration* for f.

[2]In the ecology literature, this is the Berger-Parker (1970) diversity index.

dependencies put them at two ends of a continuum. Indices along the continuum have intermediate dependence on the nonpeak probabilities. This viewpoint is useful not only conceptually but also practically: these intermediate indices can, under some circumstances, have advantages over either the Bayes or Shannon indices (as we will show in the second half of section 2).

To display formally that the Bayes and Shannon indices constitute the ends of a continuum, we consider the function $f_\alpha(P)$ defined by

$$f_\alpha(P) = \left(\sum_i p_i^\alpha \right)^{\frac{1}{\alpha}}. \tag{2.3}$$

This function is convex and thus is an index of concentration.[3] As $\alpha \to \infty$, f_α is progressively dominated by the single largest probability and approaches the Bayes index $f_{\max}$, equation 2.1. As $\alpha \to 1$, $(f_\alpha - 1)/(\alpha - 1)$ approaches the negative of the Shannon entropy, equation 2.2.[4]

2.2 For Every Index of Concentration, There Is a Generalized Index of Transmitted Information. To construct an index of transmitted information, I_f, from any index of concentration f, we generalize the relationship between Shannon mutual information and Shannon entropy. As is well known, Shannon mutual information is the difference between an a priori entropy and an a posteriori entropy, for example, the difference between the entropy of a stimulus distribution and the entropy of that distribution conditional on observing a response. By generalizing this relationship, we show that any index of concentration f can be turned into an information-theoretic quantity and can therefore be used to test codes. As we will also show, the new indices have distinct properties that make them useful for analyzing different kinds of behavioral experiments.

To generalize the Shannon construction, consider two random variables X (with distribution P_X) and Y (with distribution P_Y). We will think of X as

[3] Convexity of equation 2.3 is a straightforward consequence of the Minkowski inequality (Abramowitz & Stegun, 1970). Although equation 2.3 might suggest otherwise, we cannot create indices of concentration $f_g(P) = g^{-1}(\sum_i g(p_i))$ from any convex function g. For example, $g(p) = -\log(p)$ leads to $f_g(P) = \prod p_i$, which is not convex. Necessary and sufficient conditions of g for convexity of f_g are given in Zhao, Fang, and Li (2005). We thank Liam Paninksi for pointing us to this literature.

[4] The indices f_α are monotonically related to the generalized entropies of Rényi (1970) and Tsallis (1988), but there are important distinctions that justify our focus on f_α. The negative of the Rényi entropy is not convex (Wehrl, 1978) and is therefore not an index of concentration. The negative of the Tsallis entropy, while convex (as shown in appendix B), is limited because of its bias properties. Specifically, since the negative of the Tsallis entropy is a linear transformation of $(f_\alpha)^\alpha$ (a concave-up function), its estimators will have greater upward bias than corresponding estimators of f_α.

the stimulus set and Y as the response set. X assumes values $i \in \{1, \ldots, M\}$ with probability $p_{X;i} = x_i$, and Y assumes values $j \in \{1, \ldots, N\}$ with probability $p_{Y;j} = y_j$. (Here and below, we use upper-case letters such as X to denote a random variable, P_X to denote its associated probability distribution, and $p_{X;i}$, to denote the probability that X assumes a specific value i.)

We construct an index $I_f(X, Y)$ that tells us to what extent observation of response variable Y narrows the possibilities for X. That is, to what extent does the a priori concentration $f(X)$ increase when a response is observed? The index $I_f(X, Y)$ is given by

$$I_f(X, Y) = \sum_{j=1}^{N} p_{Y;j} f(P_{X|Y=j}) - f(P_X), \tag{2.4}$$

where $P_{X|Y=j}$ is the conditional distribution of X, given the observation $Y = j$.

Note that for $f = -H_1$, equation 2.4 is the familiar Shannon mutual information, which we will denote $I_{Shannon}$. The first term is the expected concentration of the stimulus distribution X, given the benefit of an observation in Y. The second term is the a priori concentration of the stimulus distribution. For $f(P) = \max(p_i)$, equation 2.4 is a corresponding Bayesian measure. In this case, the first term of equation 2.4 is the expected fraction correct of the Bayesian decoder that has the benefit of an observation in Y. The second term is the fraction correct that could be achieved by choosing the a priori most likely symbol. So, in both cases, equation 2.4 is the increase in concentration that is achieved on the basis of the observations Y. Since the two indices of concentration f represent the ends of a continuum, so do their corresponding indices of transmitted information I_f, with $\alpha \to \infty$ yielding a Bayes measure and $\alpha \to 1$ yielding a Shannon information.[5]

2.3 Properties of Indices of Transmitted Information. Here we show that several key properties of Shannon information extend to all of the indices $I_f(X, Y)$ as defined by equation 2.4. These properties are important because they show the indices $I_f(X, Y)$ behave in a manner that merits the designation "transmitted information" and because they will allow us to use these indices to test coding hypotheses. The properties are (1) nonnegativity: $I_f(X, Y) \geq 0$, with strict inequality implying that X and Y are dependent, and $I_f(X, Y) = 0$ whenever X and Y are independent; (2)

[5] Although the intermediate indices of concentration f_α are monotonically related to the generalized entropies of Renyi (via an exponential transformation) and Tsallis (via a power law transformation), the corresponding indices of transmitted information I_f are not monotonically related to these entropies, because values of f_α are combined by addition in equation 2.4.

refinement (Rényi, 1961): $I_f(X, Y)$ should behave in a lawful manner if the response variable Y is refined into a more detailed representation; and (3) the DPI: stimulus-independent processing of the response variable Y should not increase $I_f(X, Y)$. In this section, we show that all of the $I_f(X, Y)$ have these properties.

In addition, statistical properties of estimates of $I_f(X, Y)$ are important to consider, because in laboratory applications, the indices must be estimated from a finite amount of data. As is well known, naive estimates of the Shannon mutual information are upwardly biased. Naive estimates of the general indices $I_f(X, Y)$ are also biased. However, the nature of this bias depends on f, as does the variance of the estimates, as we discuss at the end of this section.

Finally, we mention that there are several properties of Shannon mutual information that do not generalize, but these properties are not essential to our purpose, ruling out codes. First, Shannon information is symmetric in X and Y, but this is not true of I_f in general. For this reason, we use the term *transmitted information* for I_f rather than *mutual information*. Second, the channel coding theorem and Sanov's theorem (large deviation approximation) (Cover & Thomas, 1991; Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997) do not apply to the other indices. While these properties of Shannon information are at the foundation of classical information theory and formalize its privileged place as a tool for characterizing codes and analyzing information flow, they are not required for eliminating codes.

2.3.1 Nonnegativity. Convexity of f guarantees that $I_f(X, Y) \geq 0$. To see this, we first observe that the unconditional distribution P_X is a mixture of the conditional distributions $P_{X|Y=j}$. That is, $\sum_{j=1}^{N} y_j P_{X|Y=j} = P_X$, where y_j is the probability of the jth symbol in Y. The convexity property states that the concentration of their mixture is no greater than the weighted sum of the concentrations of the individual components. That is, for any set of N distributions P_n and any set of mixing weights (real numbers $\lambda_n \in [0, 1]$ that sum to 1),

$$\sum_{n=1}^{N} \lambda_n f(P_n) \geq f\left(\sum_{n=1}^{N} \lambda_n P_n\right). \tag{2.5}$$

The conclusion that $I_f(X, Y) \geq 0$ follows from equation 2.5 by choosing $\lambda_j = y_j$. The minimum value $I_f(X, Y) = 0$ is achieved when X and Y are independent (so $P_{X|Y=j} = P_X$).

2.3.2 Refinement. The generalized transmitted information I_f obeys a refinement rule (Rényi, 1961) that governs its behavior when the response variable Y is "refined" into a more detailed representation Z. Suppose that we initially analyze the relationship between a stimulus set X and a response

set Y and only later realize that the final value $Y = N$ actually represents R distinguishable responses. The refinement rule dictates how the unrefined information $I_f(X, Y)$ and the refined information $I_f(X, Z)$ are related:

$$I_f(X, Z) = I_f(X, Y) + p_{Y;N} I_f(X \mid Y = N, Z \mid Y = N). \tag{2.6}$$

This relationship follows from the definition 2.4 of I_f, as we now show. When Y takes on one of its first $N-1$ values, then $Z = Y$. That is, for the unrefined symbols $j \leq N-1$, $p_{Z|y=k;\,j} = \delta(j,k)$ and $P_{X|Y=j} = P_{X|Z=j}$. When Y takes on the final value $Y = N$, Z can take any value in $\{N, N+1, \ldots, N+R-1\}$. For the symbols $r \in \{N, N+1, \ldots, N+R-1\}$ generated by this refinement, $p_{Z|y=N;r} = p_{Z;r}/p_{Y;N}$, and $p_{Z|y=k;r}$ is zero for $k < N$. Therefore,

$$\begin{aligned}
I_f(X, Z) &= \sum_{j=1}^{N-1} p_{Z;\,j} f(X \mid Z = j) - f(X) + \sum_{r=N}^{N+R-1} p_{Z;r} f(X \mid Z = r) \\
&= \left(\sum_{j=1}^{N-1} p_{Z;\,j} f(X \mid Z = j) + p_{Y;N} f(X \mid Y = N) - f(X) \right) \\
&\quad + \sum_{r=N}^{N+R-1} p_{Z;r} f(X \mid Z = r) - p_{Y;N} f(X \mid Y = N) \\
&= I_f(X, Y) + p_{Y;N} \left(\sum_{r=N}^{N+R-1} \frac{p_{Z;r}}{p_{Y;N}} f(X \mid Z = r) - f(X \mid Y = N) \right) \\
&= I_f(X, Y) + p_{Y;N} \left(\sum_{r=N}^{N+R-1} p_{Z|Y=N;r} f(X \mid Z = r) - f(X \mid Y = N) \right) \\
&= I_f(X, Y) + p_{Y;N} I_f(X \mid Y = N, Z \mid Y = N).
\end{aligned}$$

2.3.3 Data Processing Inequality. The DPI, an important property of Shannon mutual information, states that stimulus-independent transformation of the response variable Y into another variable Z cannot increase the amount of information. This property is shared by the generalized transmitted information I_f. More formally, if $X \to Y \to Z$ form a Markov chain, then

$$I_f(X, Z) \leq I_f(X, Y). \tag{2.7}$$

We remark that if both $X \to Y \to Z$ and $X \to Z \to Y$ are Markov chains, then the DPI implies that $I_f(X, Z) = I_f(X, Y)$, since, in addition to equation 2.7, $I_f(X, Y) \leq I_f(X, Z)$ must hold.

To demonstrate the DPI, we first observe that any transformation from Y to Z can be achieved by a sequence of simple steps $Y = Y_0$, $Y_1, \ldots, Y_k, \ldots Y_s = Z$, where each step consists of one of the following kinds of transformations: (1) symbols in Y_k are relabeled in Y_{k+1}; (2) a symbol a in Y_k splits into one of two new symbols b or c in Y_{k+1}; or (3) two distinct symbols in Y_k, say g and h, are merged into a new symbol u in Y_{k+1}.

We now proceed to show that none of these transformations can increase $I_f(X, Y_k)$. The relabeling transformation (transformation 1) does not affect I_f, since it merely reorders the terms in the sum of equation 2.4. The splitting transformation (transformation 2) does not affect I_f for the following reason. The assumption that X, Y, and Z form a Markov chain implies that the choice of b versus c is independent of X. Therefore, the a posteriori distributions $P_{X|Y_k=a}$, $P_{X|Y_{k+1}=b}$ and $P_{X|Y_{k+1}=c}$ are identical. Moreover, since b and c arise only from a, the marginal probabilities $p_{Y_{k+1};b}$ and $p_{Y_{k+1};c}$ must sum to $p_{Y_k;a}$. Together, these observations imply that transformation 2 does not affect I_f (see equation 2.4). Finally, the merging transformation, transformation 3, may affect I_f, but the convexity property, equation 2.5, implies that I_f cannot increase, as the following calculation shows. Since u can arise only from g or h,

$$p_{Y_{k+1};u} = p_{Y_k;g} + p_{Y_k;h} \tag{2.8}$$

and

$$\begin{aligned} p_{Y_{k+1};u} P_{X|Y_{k+1}=u} &= P_{X,Y_{k+1}=u} = P_{X,Y_k=g} + P_{X,Y_k=h} \\ &= p_{Y_k;g} P_{X|Y_k=g} + p_{Y_k;h} P_{X|Y_k=h}. \end{aligned} \tag{2.9}$$

Thus, the conditional probability $P_{X|Y_{k+1}=u}$ is a mixture of the conditional probabilities $P_{X|Y_k=g}$ and $P_{X|Y_k=h}$, with mixing weights $\lambda_1 = p_{Y_k;g}/(p_{Y_k;g} + p_{Y_k;h})$ and $\lambda_2 = 1 - \lambda_1$ in equation 2.5. The convexity property shows that the contribution of the terms of I_f that involve g and h cannot increase when they are merged into the new symbol u.

In sum, the above paragraph emphasizes the importance of convexity for the DPI, that is, it *implies* the DPI. Conversely, the DPI implies convexity. More formally, if (for an arbitrary f) an expression of the form of equation 2.4 satisfies the DPI, then the function f must be convex. To demonstrate equation 2.5, create random variables X and Y for which $P_{Y=y_n} = \lambda_n$ and $P_{X|Y=y_n} = P_n$. "Processing" Y by merging all of its symbols into a single symbol Z sets up a situation in which $I_f(X, Z) = 0$, and $I_f(X, Y) \geq 0$ is equivalent to equation 2.5.

2.3.4 Bias and Variance of Naive Estimates of the Indices. To apply the indices $I_f(X, Y)$ to an experiment, they must be estimated from a finite quantity of laboratory data. Therefore, we consider the bias and variance properties

of simple estimates of the indices $I_f(X, Y)$. These are the naive (plug-in) estimates, which consist of inserting the observed frequencies of joint observations of X and Y into the definition of $I_f(X, Y)$, equation 2.4. As we show, these naive estimates of $I_f(X, Y)$ tend to be overestimates; they are upwardly biased. However, depending on the choice of the concentration index f, bias affects some kinds of data sets more than others. For Bayes-like indices, bias is most severe when performance is at threshold (chance performance) and minimal when performance is far from threshold (close to perfect). For Shannon-like indices, bias is independent of the level of performance to a first approximation.

To see how the bias properties of the indices $I_f(X, Y)$ emerge, we make use of the fact that they are sums of indices of concentration, f. We therefore begin by discussing the statistical behavior of estimates of f.

It is well known that the naive estimate of the Shannon entropy is downwardly biased (Carlton, 1969; Miller, 1955; Treves & Panzeri, 1995; Victor, 2000). Consequently, naive estimates of its negative, the Shannon index of concentration $f = -H_1$, are upwardly biased. As we show in appendix C, naive estimates of all indices of concentration are upwardly biased, and, moreover, as the sample size increases, the bias decreases monotonically. The proof hinges on the convexity of the index of concentration, along with the following observation: a probability distribution estimated from the frequencies encountered in a data set containing N observations is a mixture of probability distributions estimated from the N data sets, each of size $N - 1$, in which one of the observations is omitted.

The specific bias behavior of the naive estimate of an index of concentration depends in a systematic fashion on the form of the index. To illustrate this, we consider the bias behavior of f_α for a range of values of the parameter α. Figure 1A considers the case of two symbols, with probability p_1 and $p_2 = 1 - p_1$. The left panel shows how the true value of f_α depends on p_1. For all values of α, the minimum concentration occurs when the probabilities are equal ($p_1 = p_2 = \frac{1}{2}$). As α increases, this minimum becomes progressively sharper, approaching a singularity as $\alpha \to \infty$ (here, $\alpha = 16$).

Column 2 of Figure 1A shows how the expected bias in the naive estimate of f_α depends on p_1. As shown in appendix C, this expected bias is asymptotically proportional to $1/N$; we plot the proportionality constant in the figure. In the Shannon limit ($\alpha = 1$), bias is independent of p_1, as is well known (Carlton, 1969; Miller, 1955; Treves & Panzeri, 1995; Victor, 2000). For large α, there is a large bias when the probabilities are nearly equal (near $p_1 = \frac{1}{2}$) and little bias when the probabilities are very unequal.

The behavior of bias of estimates of f_α can be understood intuitively as follows. Two factors need to be considered: first, the variability in the estimate of p, and second, how the naive estimate of f_α depends on the estimate of p. The variability in the estimate of p is determined by multinomial statistics and is independent of α (i.e., it is the same for all of the indices), so we focus on the second factor, the shape of f_α. We consider

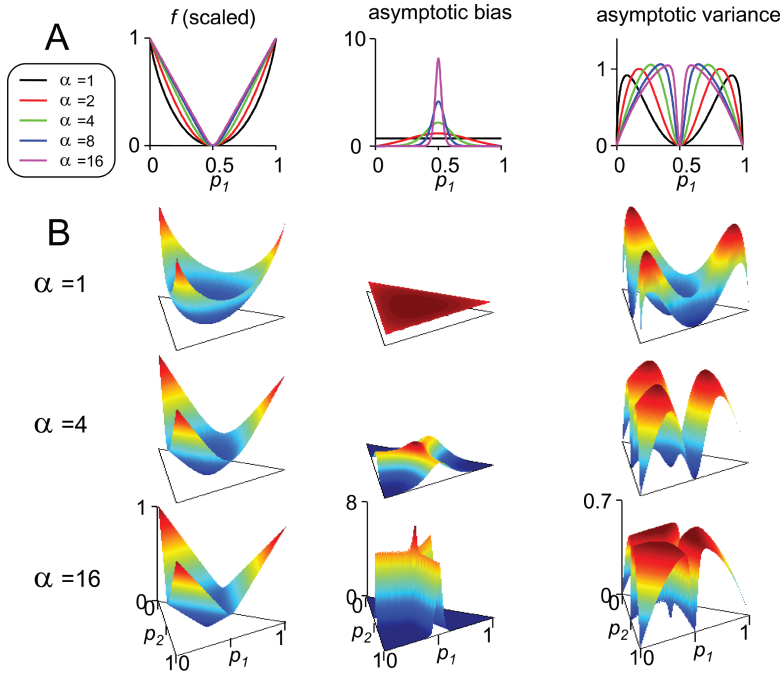

Figure 1: Behavior of bias and variance of naive estimates of indices of concentration f_α. (A) Two symbols, with probabilities p_1 and $p_2 = 1 - p_1$. (B) Three symbols, with probabilities p_1, p_2, and $p_3 = 1 - p_1 - p_2$. Column 1: f_α, linearly rescaled into the range $[0, 1]$. Column 2: the coefficient of $1/N$ in the asymptotic bias (same normalization as column 1). Column 3: the coefficient of $1/N$ in the asymptotic variance (same normalization as column 1).

first the peak in the bias near $p_1 = p_2 = \frac{1}{2}$ for large α. The origin of this peak is that f_α is sharply curved at this point. In particular, as $\alpha \to \infty$, f_α approaches $f_{\max}$. In this limit, $f_{\max}$ is V-shaped at this point, because it is determined by whichever symbol happens to have the highest frequency in the data sample. Thus, any fluctuation away from equal occurrences of the two symbols will produce a naive estimate that is higher than the correct value. Consequently, the bias is large. However, if p_1 is far from $\frac{1}{2}$, fluctuations in the number of occurrences of the two symbols are just as likely to lead to an overestimate of $f_{\max}$ as an underestimate. Consequently, the bias in the estimate of $f_{\max}$ is minimal. For small α, the sharpness of the trough of f_α is progressively smaller, and the peak in the bias is progressively smaller and broader.[6] Note that although the expected bias of naive estimators of

[6]It is generally stated that the bias properties of the Renyi entropies are more favorable than that of the Shannon entropy. For example, Strong, Koberle, de Ruyter van Steveninck,

the indices f_α is generally less than that of the Shannon entropy ($\alpha = 1$) when the probabilities are far from equality, the naive estimate of the Shannon entropy is asymptotically less biased than that of the other indices near $p_1 = p_2 = \frac{1}{2}$.

The third column of Figure 1A shows how the variance of the naive estimate of f_α depends on p_1. For all indices f_α, variance is smallest when f_α has its extreme values: at $p_1 = \frac{1}{2}$, where f_α is minimum, and at $p_1 = 0$ and $p_1 = 1$, where f_α is maximum. However, the range of values of p_1 at which the variance in the estimate of f_α is largest depends strongly on α, with p_1 near $\frac{1}{2}$ associated with the largest variances for large α, and p_1 near 0 or 1 associated with the largest variances for small α.

Intuitively, the behavior of the variance can be understood as the net result of two factors. One factor is the local slope of the index f_α (see Figure 1A, column 1). A large slope implies a high sensitivity to errors in the estimates of p_1, while a small slope implies insensitivity to errors in the estimates of p_1. The slope is 0 at $p_1 = \frac{1}{2}$, and this accounts for the trough of the variance at this point. The second factor is the behavior of the variance of binomial distributions, which determines how accurately one can estimate p_1 from a finite sample. For N samples, the variance is given by $Np_1(1 - p_1)$, indicating that estimates of p_1 are least variable at the extremes of its range. This accounts for the minima of the variance at $p_1 = 0$ and $p_1 = 1$.

The case of three symbols (see Figure 1B) shows how the observations of Figure 1A (two symbols) extend to the general case. For large α, f_α has a trough where the two largest probabilities are equal (e.g., $p_1 = p_2 = 2/5$, $p_3 = 1/5$), and a sharp minimum where these troughs converge ($p_1 = p_2 = p_3 = 1/3$). This leads to local maxima in the bias of the naive estimator along these troughs and a global maximum at their convergence. As α decreases, the sharpness of the bias distribution is blunted, attaining uniformity in the Shannon ($\alpha = 1$) case. As in the two-symbol case, the variance of the naive estimator has a minimum at the point of equal probabilities and at the extremes, and the largest variances occur near the point of equal probabilities when α is large.

The properties of estimates of generalized transmitted information I_f are consequences of the bias properties of estimates of the underlying index of concentration f. In the typical laboratory situation, the input symbols X can be chosen by the experimenter, so that the frequencies encountered after N trials exactly match those of the true distribution P_X. In this case, only the conditional probabilities $P_{X|Y=j}$ in equation 2.4 must be estimated. Naive estimates of I_f thus decrease monotonically with N, since they inherit the

---

and Bialek (1998) used the Ma bound ($\alpha = 2$) because of this property. However, the indices f_α are exponential functions of the Renyi entropies, which makes their bias behavior less favorable. Additionally, as Figure 1 shows, the bias behavior depends strongly on the range of probabilities, with relatively greater bias for the Shannon entropy when the probabilities are unequal.

monotonically decreasing behavior of estimates of f (see appendix C). For large α, this bias will be large when the conditional distributions have nearly equal probabilities, while for $\alpha = 1$, the bias will depend only on the number of symbols in X and Y (Treves & Panzeri, 1995).

We caution that the above analysis hinges on the assumption that the input distribution P_X is known. If P_X must also be estimated, then estimates of I_f may approach the true value from below rather than from above. For example, consider estimates of $I_{f_{\max}}$ in a scenario in which each of the symbols of X occurs with equal probability and each is signaled reliably by a corresponding symbol in Y. That is, the probabilities in the a posteriori distributions $P_{X|Y=j}$ are unequal, but the probabilities in the a priori distribution P_X are equal. Consequently, naive estimates of $f_{\max}(P_{X|Y=j})$ are nearly unbiased, but naive estimates of $f_{\max}(P_X)$ have a large upward bias. Therefore, their difference, the naive estimates of the Bayesian index $I_{f_{\max}}$, has a downward bias. This bias behavior for $I_{f_{\max}}$ contrasts with the bias behavior naive estimates of $I_{Shannon}$, which is always upward if all of the co-occurrence probabilities of symbols in X and Y are nonzero (Treves & Panzeri, 1995).

2.4 Using Indices of Concentration to Test Coding Hypotheses. Now that we have constructed a variety of indices of transmitted information and shown that they share many of the properties of Shannon information, we describe how they can be used to test hypotheses about neural codes. We consider a generic behavioral task consisting of presentation of input symbols that elicit observable behavioral responses. We assume that we have recorded from all of the neurons that provide the sensory signals relevant to this task and that we have hypothesized a "neural code," that is, a way to represent these neural signals by instances of some random variable. The question to be asked is, can the neural activity as represented in this fashion account for the behavioral performance? If the answer to the question is no, then (since we have assumed that we have recorded from all the relevant neurons), we have rigorously ruled out a neural code (Nirenberg et al., 2006). This strategy, which seeks to determine whether specific statistical features of neuronal activity are insufficient to support a behavior, stands in contrast to the more standard approach (Bialek, Rieke, de Ruyter van Steveninck, & Warland, 1991; McClurkin, Optican, Richmond, & Gawne, 1991; Victor & Purpura, 1996) of identifying statistical features that might support a behavior.

2.5 Setup. The question of whether a neural code can support a stimulus-behavior linkage is equivalent to the question of whether the stimulus set, the neural representation, and the behavior constitute a Markov chain. In the analysis below, the stimulus is a random variable X assuming one of L discrete values, characterized by a distribution P_X. The neural code is represented by a random variable Y, whose relationship to X is

characterized by the conditional distributions $P_{X|Y=y}$. Like the stimulus X, the behavioral response Z is a discrete random variable; its relationship to X is characterized by $P_{X|Z=z}$. Our question is whether the neural code Y can account for the observed dependence of the behavior Z on the stimulus X. That is, we want to determine whether $P_{X|Y=y}$ and $P_{X|Z=z}$ are consistent with a Markov chain $X \to Y \to Z$, and to do so from two sets of joint measurements: stimulus and code ($P_{X|Y=y}$) and stimulus and behavior ($P_{X|Z=z}$). According to the DPI, a necessary condition is that $X \to Y \to Z$ constitute a Markov chain that

$$I_f(X, Z) \leq I_f(X, Y). \tag{2.10}$$

for every generalized transmitted information I_f. However, for different choices of the index of concentration f, the above inequality places different conditions on the neural activity Y. Our goal is to analyze this situation, comparing the utility of different choices of I_f. The main points are evident even in the simplest scenario: two stimuli and two behavioral responses. We thus analyze this scenario first and then describe how the analysis generalizes to scenarios in which there are multiple stimuli or behavioral responses.

2.6 Two Stimuli, Two Responses: The Indices Are Inequivalent. We focus on the two-stimulus, two-response case in which both stimuli are equally probable. This case illustrates the complementary properties of the constraints (see equation 2.10) for different choices of the indices I_f. In our first example, the Shannon index is generally stronger. In the second example, the Bayes index is generally stronger. But no index is guaranteed to be the stronger in either case. The subsequent examples examine the basis for this complementarity.

2.6.1 Example 1: Shannon Index Is Superior. In our first example (see Figure 2) we analyze a scenario in which there are only two words in the neural code Y, each of which signals the two stimuli with moderate reliability. For one of the code words (say, y_1) the a posteriori probabilities of the stimuli x_1 and x_2 are given by 0.75 and 0.25; for the other code word (say, y_2), the a posteriori probabilities of the stimuli are given by 0.25 and 0.75.

The inset in Figure 2 illustrates the relationship between the stimulus and the code. Each code word y is associated with an a posteriori probability of the two stimuli, $q_1 = p(x_1 \mid y)$ and $q_2 = p(x_2 \mid y)$. One can describe the relationship between stimuli and code words by tabulating how often a code word has a particular pair (q_1, q_2) of a posteriori probabilities. Since $q_1 + q_2 = 1$ (i.e., the a posteriori probabilities for all of the stimuli must sum to 1), for this tabulation we need to consider only q_1. We call this density $\rho(q_1)$, and it is the ordinate of the inset.

The main portion of Figure 2 illustrates the range of possible behaviors that is supported by code diagrammed in the inset. Since there are two stimuli and two behavioral responses, the stimulus-response behavior is completely specified by the probabilities that each of the two stimuli (x_1 and x_2) will elicit the behavior z_1. These probabilities, $p(z_1 \mid x_1)$ and $p(z_1 \mid x_2)$, are the two axes of the main plot. There are four ways that the subject could use these two neural responses; we refer to them as decision rules:

1. For any code word y, always respond with z_1. That is, $p(z_1 \mid x_1) = p(z_1 \mid x_2) = 1$. This results in the behavior plotted at point B_1.
2. For any neural code word y, always respond with z_2. That is, $p(z_1 \mid x_1) = p(z_1 \mid x_2) = 0$. This results in the behavior plotted at point B_2.
3. When (and only when) y_1 is present, respond with z_1. That is, $p(z_1 \mid x_1) = 0.75$ and $p(z_1 \mid x_2) = 0.25$. This results in the behavior plotted at point B_3.
4. When (and only when) y_2 is present , respond with z_1. That is, $p(z_1 \mid x_1) = 0.25$ and $p(z_1 \mid x_2) = 0.75$. This results in the behavior plotted at point B_4.

The general calculation of the above probabilities from the density $\rho(q_1)$ is detailed in appendix D.

These four behaviors ($B_1, \ldots, B_4$) are the corners of a diamond-shaped region. The interior of this diamond represents all of the behaviors that can be supported by the code Y. The reason that the interior points are included is that they correspond to behaviors that can be achieved by mixing the behaviors at the corners, that is, by applying one decision rule on one trial and another decision rule on another trial. Appendix D shows that no other behaviors are possible.

Now that we have delineated all possible behaviors that can be supported by the code Y, we can show how the different indices allow us to test whether this code is viable. For each index I_f, we can calculate the transmitted information available in the code Y, $I_f(X, Y)$. The data processing inequality (see equation 2.10) requires that for any behavior derived from this code, $I_f(X, Z) \leq I_f(X, Y)$. Therefore, for each index I_f, we plot the locus of behaviors for which $I_f(X, Z) = I_f(X, Y)$. If we observe a behavior outside this locus, it would allow us to rule out that code. In other words, that particular code Y did not carry enough information to account for that particular behavior Z.

Note that the bounds are different for the different indices I_{f_α}. We show the Shannon bound ($\alpha = 1$) in red, the Bayes bound ($\alpha = \infty$) in blue, and the bound determined by an intermediate index ($\alpha = 2$) in green. Importantly, these bounds are not only distinct but also differ in strength.

Here is an example that illustrates this point that the bounds provided by the different indices are both distinct and differ in strength. Suppose that we

made the assumption that only spike count matters, but in reality, behavior makes use of spike timing. That is, our assumed code Y is insufficient, and the observed behavior B' is outside the diamond of supportable behaviors, as in Figure 2. But although B' cannot be supported by Y, it is inside the bounds of the Bayes index ($\alpha = \infty$, blue). Hence, had we analyzed the experiment using the Bayes index, the code Y would appear to be a viable code. Importantly, though, B' is outside the bounds of the Shannon index ($\alpha = 1$, red). In other words, in this example, the Shannon index does rule out the code Y, even though the Bayes index does not. Indices that are intermediate between the Shannon and Bayes indices (e.g., $\alpha = 2$) have intermediate behavior: they succeed in ruling out the code Y for some of the behaviors, but not for all of them.

2.6.2 Example 2: Bayes Index Is Superior. In the first example (see Figure 2), the Shannon index provided a stronger test of the neural code Y than the Bayes index. That is, the region of behaviors within the bounds for $\alpha = 1$ (red) was smaller than the region of behaviors within the bounds for $\alpha = \infty$ (blue). Here we set up a situation in which the opposite is true.

In this example, the neural code Y has many code words y (see Figure 3, inset). Specifically, code words with all a posteriori probabilities $q_1 = P(x_1 \mid y)$ are represented, and all are equally likely (i.e., the density $\rho(q_1)$ is constant). Any a posteriori probability q_1 can serve as the cutoff for a decision rule. For example, a typical decision rule chooses behavior z_1 if the a posteriori probability of x_1 is sufficiently high. This decision rule is characterized by a cutoff criterion q_{cut}, along with the policy of choosing behavior z_1 if $q_1 \geq q_{cut}$. These decision rules correspond to the behaviors that form the curved black trajectory below the diagonal in Figure 3. There are also behaviors that form a curved trajectory above the diagonal. These correspond to decision rules that choose the behavior z_1 if the a posteriori probability of x_1 is sufficiently low, that is, $q_1 \leq q_{cut}$. All possible behaviors that can be supported by Y are mixtures of these behaviors and correspond to points that lie within the lens-shaped region bounded by these two curves (proof is in appendix D).

In contrast to the situation of Figure 2, the bounds associated with the Bayes index (blue) are tighter than the bounds associated with the Shannon index (red).

Note that from the insets of Figures 2 and 3, it might appear that the critical difference between these two examples is that the former has a discrete set of code words and the latter has a continuum of code words. However, the critical factor is not the discreteness of the code itself. Rather, the critical difference is that the range of certainties associated with the code words is wider in Figure 3 than Figure 2, combined with the fact that the behavior is discrete. We show two more examples that illustrate this point and then turn to why it is the case.

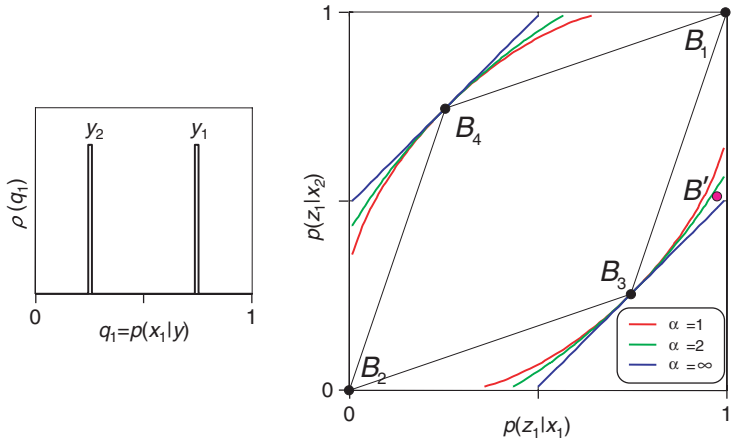

Figure 2: Testing a neural code in a scenario in which the Shannon index has the advantage. The neural code has two words: y_1, for which the a posteriori probabilities of the stimuli x_1 and x_2 are given by 0.75 and 0.25, and y_2, for which the a posteriori probabilities of the stimuli are given by 0.25 and 0.75 (inset). The diamond-shaped region in the main graph shows the range of behaviors that can be supported by the code and the bounds provided by several different indices of transmitted information. The Shannon bound ($\alpha = 1$) is tighter than the Bayes bound ($\alpha = \infty$). For further details, see the text.

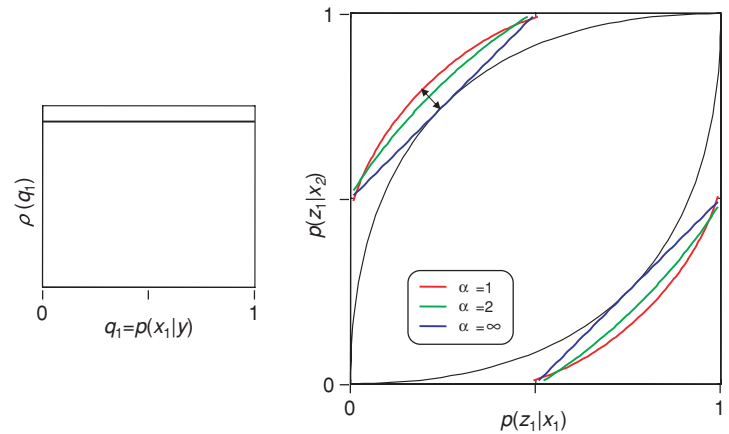

Figure 3: Testing a neural code in a scenario in which the Bayes index has the advantage. The neural code words cover the entire range of a posteriori probabilities (see inset). The lens-shaped region in the main graph shows the range of behaviors that can be supported by the code, and the bounds provided by several different indices of transmitted information. The Bayes bound ($\alpha = \infty$) is tighter than the Shannon bound ($\alpha = 1$). For further details, see the text.

2.6.3 Further Examples: The Range of Certainty Is Crucial. To see that the critical factor is the range of certainty (rather than whether the code has a continuum of words), we consider a code with four code words, covering a range of certainties. Specifically, in the code illustrated in the inset of Figure 4, Y consists of four code words y_1, y_2, y_3, and y_4, with a posteriori likelihoods $q_1 = P(x_1 \mid y)$ of $1, 2/3, 1/3$, and 0. That is, y_1 and y_2 indicate that stimulus x_1 was probably present, while y_3 and y_4 indicate that stimulus x_2 was probably present. However, y_1 and y_4 are reliable, while y_2 and y_3 are ambiguous.

As in Figure 2, the set of supported behaviors forms a polygonal region. The corners of this region correspond to decision rules in which one set of code words $\{y_1, \ldots, y_c\}$ elicits one behavior, and the complementary set $\{y_{c+1}, \ldots, y_4\}$ elicits the alternative behavior. However, unlike Figure 2, the bounds associated with the Bayes index (blue) are tighter than the bounds associated with the Shannon index (red).

In the final example, Figure 5, the code words form a continuum (as in Figure 3), but the certainties are tightly clustered. This happens because the density $\rho(q_1)$ is bimodal, with modes at 0.25 and 0.75.

As in the uniformly distributed example in Figure 3, the region of supportable behaviors forms a lens-shaped region. However, in contrast to the uniformly distributed example, the Bayes index does not always provide the strongest test. Rather, there are behaviors (e.g., B') that will result in excluding the code only if the Shannon index is used.

Comparing the above examples suggests that the difference in performance of the indices depends primarily on the range of certainty of the code words. We now discuss why this is the case by focusing on the geometry of the bounds corresponding to the Bayes and Shannon indices.

There are two interacting factors: the relative positions of the Bayes and Shannon bounds and their shapes. We first consider their positions. The Bayes bound will always make contact with the region of behaviors that the code can support (e.g., at points B_3 and B_4 in Figure 2 and at the inner arrowhead in Figure 3.) This is because the Bayes bound can always be attained by a decision rule that maximizes the fraction of correct responses (this decision rule chooses the stimulus with the maximum a posteriori probability). In contrast, the Shannon bound need not make contact with the region of supportable behaviors (see Figures 3–5). This is because converting the neural code to a binary behavioral response causes a loss of information about the level of certainty. In other words, the code has a greater Shannon information than the subject can possibly transmit with a binary behavioral decision. The difference between the Shannon information transmitted by the code and the maximum Shannon information that can be transmitted by processing the code into a binary behavior results in a gap between the Shannon bound and the region of supportable behaviors (see the arrows in Figures 3–5). This gives an advantage to the Bayes index, whose bound is always in contact with the region of supportable behaviors.

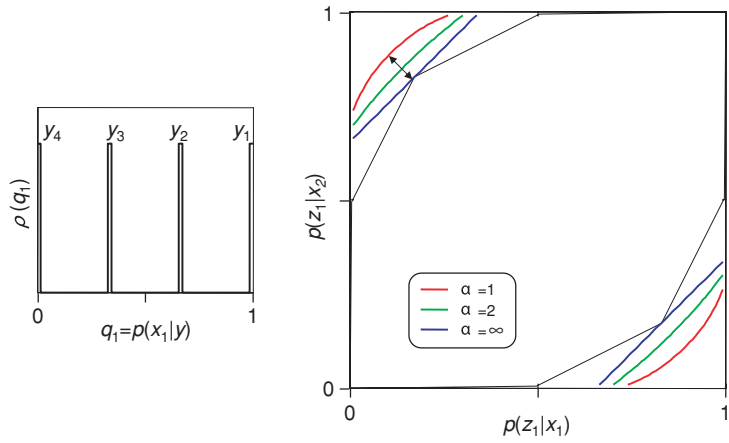

Figure 4: The range of uncertainties affects the choice of indices. The neural code (see inset) has four words y_1, y_2, y_3 and y_4, two of which have high certainty (y_1 and y_4) and two of which have low certainty (y_2 and y_3). The polygonal region in the main graph shows the range of behaviors that can be supported by the code and the bounds provided by several different indices of transmitted information. The Bayes bound ($\alpha = \infty$) is tighter than the Shannon bound ($\alpha = 1$). For further details, see the text.

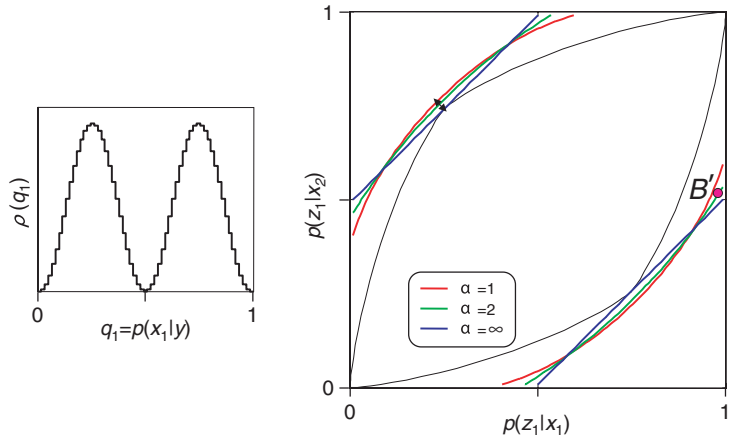

Figure 5: Complementary strengths of the Bayes and Shannon indices. The neural code words cover the entire range of a posteriori probabilities but are bimodally distributed (see inset). The lens-shaped region in the main graph shows the range of behaviors that can be supported by the code and the bounds provided by several different indices of transmitted information. The Bayes bound ($\alpha = \infty$) is tighter than the Shannon bound ($\alpha = 1$) near its the point of tangency to the range of supportable behaviors, but the Shannon bound is tighter than the Bayes bound away from this point. For further details, see the text.

We next consider the shapes of the bounds. The Bayes bound will always be a straight line, since it is determined by linear functions of the probabilities $p(z_1 \mid x_1)$ and $p(z_1 \mid x_2)$. In contrast, the Shannon bound will always be curved, since it is determined by a nonlinear function of these probabilities. Consequently, the Shannon bound will curve inward toward the region of supportable behaviors, while the Bayes bound departs from it at a tangent. Because the Shannon bound curves inward toward the region of supportable behaviors, it can be stronger than the Bayes bound away from the point at which the subject chooses a decision rule that maximizes the fraction of correct responses. That is, the Shannon bound is less sensitive to the decision criterion.

In sum, the bottom line can be stated simply: the Shannon index is weakened by fact that information is lost when a code is converted to a behavior, while the Bayes index is weakened if the decision rule is not known. However, the Bayes index can be readily pushed further, since its bound always comes into contact with the region of supportable behaviors. For further discussion, see appendix A.

2.7 Multiple Stimuli, Multiple Responses. The above examples considered scenarios with two stimuli and two behavioral responses. Here we show the implications of the analysis for scenarios in which there are more than two stimuli or more than two behavioral responses. As in the above analysis, the first step is to construct the set of supportable behaviors. When there are L stimuli and N behaviors, this is a space of $(L-1)N$ dimensions. This is because each of the N behaviors corresponds to an a posteriori distribution of the L stimuli, for which there are $L-1$ degrees of freedom (dimensions). Above, $L = N = 2$, so the supportable behaviors constituted a region in the plane. But here, the space of supportable behaviors has a higher dimension.

As in the simpler $L = N = 2$ scenario, the set of supportable behaviors must be a convex region within this space. Its boundary will have a more complex shape, parameterized by partitions of the space of a posteriori probabilities (this characterization is demonstrated in appendix D).

The Shannon index provides a bound that is curved inward and thus has the potential to follow the boundary of the complex shape closely. However, it is typically offset from the supportable behaviors because of conversion of the code into one of N possible behaviors.

In contrast, the Bayes index provides a bound that is a hyperplane tangent to the set of supportable behaviors. It is thus optimal at this point of tangency, but it is increasingly suboptimal away from that point.

Thus, as in the two-stimulus, two-response case, the Shannon index is weakened by the loss of information about certainty when the code is reduced to a behavior, while the Bayes index is weakened if the decision rule is not known exactly.

It is therefore straightforward to construct examples in which the Bayes index fails, or the Shannon index fails, or both. For larger values of L (the

number of stimuli) and N (the number of behavioral responses), there are separate factors that favor each of the two kinds of indices, so their complementary nature persists. As L increases, a factor that favors the Bayes index is that the dimensionality of the set of a posteriori probabilities associated with each code word is $L - 1$. Thus, there is an increasing information loss resulting from discretization of this high-dimensional characterization of uncertainty to one of N discrete responses. This loss due to discretization puts the Shannon index at an increasing disadvantage. But other factors favor Shannon-type indices. As N increases, the effect of response discretization decreases because more responses are available. Moreover, as both L and N increase, the variety of decision rules increases (see appendix D). Thus, sensitivity to the decision rule (which weakens the Bayes index) is an increasingly important factor. Related to this, as L or N increases, the possibility of near misses increases—that is, the possibility that a subject responds with an answer that is close to the correct answer but wrong. The Bayes index gives no credit for these close answers, only for correct answers. The Shannon index (as well as the intermediate indices) gives credit for answers that are systematic but wrong. When wrong answers are systematic, Shannon-like indices can provide a stronger test of a coding hypothesis than Bayes-like indices.

2.8 Other Indices of Concentration and Transmitted Information. Above, we have focused on the Shannon index, the Bayes index, and a natural continuum of indices for which they represent the extremes. However, this continuum does not exhaust the realm of useful indices. Moreover, these other indices have specific behavioral interpretations and are particularly useful when the subject's decision rule is known (see appendix A).

Any index of concentration can be modified by applying a nonnegative set of weights $\vec{w} = (w_1, \ldots, w_L)$ to the probabilities. That is, if $f(P) = f(p_1, \ldots, p_L)$ is an index of concentration, then so is

$$_{\vec{w}}f(P) \equiv f(w_1 p_1, \ldots, w_L p_L). \tag{2.11}$$

In the Bayes limit ($_{\vec{w}}f_{\max}$), the index of transmitted information corresponding to equation 2.11 measures the improvement in performance of a decoder that associates unequal values with each stimulus.

A second kind of generalization is useful for multi-alternative behavioral paradigms. When there are three or more possible behaviors ($N \geq 3$ in the above), a subject's pattern of errors may indicate a greater level of knowledge about the stimulus than merely the fraction of correct responses (Thomson & Kristan, 2005). This kind of systematic behavior captured by the index of concentration is $f^{(k)}_{\max}$, the sum of the k largest values of p_i:

$$f^{(k)}_{\max}(P) = \max_{i_1,\ldots,i_k} \left(\sum_k p_{i_k} \right) \quad (\text{all } i_k \text{ distinct}). \tag{2.12}$$

In particular, the index of transmitted information corresponding to equation 2.12 is the improvement in the performance of a Bayesian decoder that is allowed k attempts at a correct answer.

3 Discussion

Information theory has been enormously useful for characterizing spike trains and proposing neural codes (reviewed in Dayan & Abbott, 2001; Rieke et al., 1997). Any feature of neural responses that depends systematically on the stimulus is a carrier of information and, in principle, a candidate neural code. Given that the ultimate goal is to reduce the space of candidate codes, that is, to close in on which codes the animal could actually be using, a logical next step is to consider whether any of these candidates can be eliminated (Nirenberg et al., 2006). Shannon's information works for this, but it is a loose bound, and it turns out that there are many related quantities that can be tighter bounds—that is, they can rule out more codes than can be eliminated by Shannon information. This is the focus of our article: the existence of these measures, their relationships to each other, and their properties for eliminating codes.

It is worth mentioning that using information to eliminate codes is distinct from using it to identify and characterize candidate codes. As a result, the properties that the indices must have are different. This is why the indices we discuss lack some of the properties of Shannon information associated with characterizing codes (Cover & Thomas, 1991; Rieke et al., 1997). However, these indices all retain one key property: the DPI. Consequently, they provide equally valid tests for the elimination of a code as provided by Shannon information, and in some cases, stronger tests.

3.1 Complementary Strengths and Weaknesses. We have shown that the indices we describe have complementary strengths and weaknesses: Shannon-like indices fail to exclude nonviable codes in scenarios in which the neural code words differ substantially in certainty, because these differences are suppressed (i.e., information is lost) when the code is reduced to a behavior (see Figures 3 and 4). Bayes-like indices yield stronger tests than Shannon-like indices in these scenarios, but the advantage may depend on knowing the decision rule precisely (see Figure 5). In addition, Bayes-like indices cannot take into account systematic error patterns. Systematic error patterns are particularly important in behavioral paradigms with multiple stimuli and multiple behaviors (Thomson & Kristan, 2005), such as the near misses that are likely to occur with reaching and eye movement tasks, or letter identification.

3.2 Indices Differ in Statistical Properties. The focus of this article is on the idealized case—that is, on the performance of an index when there are sufficient data so that its value can be determined exactly. In this limit, using

the index to test a code based on the DPI is rigorous, and the properties of the index depend in a simple way on the distribution of certainty of the responses, the number of behaviors, and the range of likely decision rules. In application to laboratory data, this limit may not be reached, and the statistical properties (i.e., the bias and variance of estimates of these indices) must also be taken into account. We do not analyze this in detail here because of the variety of approaches available to estimate information-theoretic quantities (Kennel, Shlens, Abarbanel, & Chichilnisky, 2005; Nemenman, Bialek, & de Ruyter van Steveninck, 2004; Nirenberg, Carcieri, Jacobs, & Latham, 2001; Paninski, 2004; Shlens, Kennel, Abarbanel, & Chichilnisky, 2007) and the many kinds of behavioral scenarios in which they might be applied. However, we do point out (see Figure 1) that these indices differ systematically in their bias and variance properties when naive (i.e., plug-in) estimates are used. Estimates of Bayes-like indices tend to have greater bias and variance for close-to-threshold responses than Shannon-like indices (see Figure 1), but smaller bias and variance away from threshold.

3.3 Eliminating Codes Versus Inferring Interaction Networks. A comparison of the problem studied here to the general problem of identifying interaction networks among several variables (Nemenman, 2004) provides further insight into the reason that non-Shannon quantities are useful for eliminating codes. For the problem of inferring interaction networks among genes, an approach based on Shannon mutual information outperforms Bayesian methods (Margolin et al., 2006). Yet for eliminating codes, Bayes-like indices can outperform Shannon-like indices.

There are two reasons for this difference: the goal of the analysis and the nature of the data. In the analysis of interaction networks, the goal is to identify the simplest, or most likely, relationship graph. Here, we have a different goal: in effect, we ask whether a particular relationship graph can be excluded. The second difference is that our problem lacks some of the symmetry of the interaction network problem: the three variables play distinguishable roles. The stimulus variable X is distinguished in that we know how it is correlated with each of the other two variables, but we do not necessarily know how the other two variables are correlated with each other. Moreover, the code variable Y and the behavior variable Z are distinguished from each other in that we are interested in whether the code can account for behavior, not vice versa. These asymmetries lead to the utility of measures $I_f(X, Y)$ and $I_f(X, Z)$ that are not symmetric in their arguments.

Appendix A: Taking into Account What the Subject Knows About the Task

In the main section of this article, we approached the problem of ruling out neural codes without making assumptions as to what the subject is thinking (e.g., what priors it is using for the task and what rules it is using for making

decisions). The DPI allows us to do this, since it places absolute limits on performance: any code that is ruled out by one of the indices I_f is truly ruled out, provided, of course, that the estimated value of the index I_f is accurate. However, it is intuitive that knowledge of the decision rule can allow us to take this approach a step further and allow even more codes to be ruled out. (This makes direct contact with ideal observer analysis; Geisler, 1989). Here, we show how this can be done. As in the main text, we focus on the two-stimulus, two-response scenario that we used in the examples of Figures 2 to 5. A hypothetical code Y has a limited range of behaviors that it can support: the diamond- or lens-shaped region in the main portion of each figure. Our task is to determine whether an observed behavior is inside this region. If it is not, our assumed code can be ruled out. That is, for a code to be viable, the observed behavior must be inside all of the tangents to this region.

We begin by showing that the tangents represent decision rules, and that each tangent also corresponds to a weighted Bayesian index (see equation 2.11). Thus, if the subject's decision rule is known, we can choose a specific index that is optimal for testing codes. To show this, we start by specifying the decision rule itself. Suppose that the subject makes a decision—to optimize the expected value of a response. So he assigns a value $v(x_j, z_k)$ to producing a behavior z_k in response to a stimulus x_j. The expected value is a linear combination of the quantities $v(x_j, z_k)$, each weighted by the probability $p(x_j, z_k)$ that the combination of the stimulus x_j and the behavior z_k occur together. Since $p(x_j, z_k) = p(z_k \mid x_j)p(x_j)$ and $p(x_j)$ is constant, the expected value is a linear function of the coordinates $p(z_1 \mid x_j)$. That is, the points that share the same expected value,

$$E = \sum_{j,k} v(x_j, z_k)p(x_j, z_k), \tag{A.1}$$

fall on a line in the $(p(z_1 \mid x_1),\ p(z_1 \mid x_2))$-plane, and the lines corresponding to different values of E are parallel. The line that maximizes E must be tangent to the region of supportable behaviors, since if it entered the interior of this region, then another line with a higher value of E could be positioned between this line and the boundary.

We next show that each such tangent line corresponds to a DPI test for an index ${}_{\vec{w}}I_{\max}$, where ${}_{\vec{w}}I_{\max}$ is the index of transmitted information that corresponds to a weighted Bayes-like index of concentration ${}_{\vec{w}}f_{\max}$ (see equation 2.11). To accomplish this, we first find a set of weights $\vec{w}$ for which the index ${}_{\vec{w}}I_{\max}$ is constant on line segments parallel to the desired tangent, and then we show that the criterion line ${}_{\vec{w}}I_{\max}(X, Z) = {}_{\vec{w}}I_{\max}(X, Y)$ contacts the region of behaviors that can be supported by the code.

To determine the weights $\vec{w}$ for an index that is constant on lines parallel to a given tangent, we use definition 2.4 to write out the index ${}_{\vec{w}}I_{\max}$

corresponding to $\vec{w} f_{\max}$:

$$\vec{w} I_{\max}(X, Z) = \sum_{k=1}^{2} p(z_k) \max(w_1 p(x_1 \mid z_k), w_2 p(x_2 \mid z_k)) - \max(w_1 p(x_1), w_2 p(x_2)). \quad \text{(A.2)}$$

With the usual rules for conditional probabilities and $p(z_2 \mid x_j) = 1 - p(z_1 \mid x_j)$, this can be rewritten as

$$\begin{aligned} \vec{w} I_{\max}(X, Z) = & \max(w_1 p(x_1) p(z_1 \mid x_1), w_2 p(x_2) p(z_1 \mid x_2)) \\ & + \max(w_1 p(x_1)(1 - p(z_1 \mid x_1)), w_2 p(x_2)(1 - p(z_1 \mid x_2))) \\ & - \max(w_1 p(x_1), w_2 p(x_2)). \end{aligned} \quad \text{(A.3)}$$

Equation A.3 is a piecewise linear function of the coordinates $p(z_1 \mid x_j)$. When $p(z_1 \mid x_1)$ is sufficiently large and $p(z_1 \mid x_2)$ is sufficiently small (or vice versa), it is constant along a line of slope $m = w_1 p(x_1)/w_2 p(x_2)$. Thus, to ensure that the locus for which $\vec{w} I_{\max}(X, Z) = C$ contains a line segment of the desired slope m, we choose the weights so that $w_1/w_2 = mp(x_2)/p(x_1)$. To see that this segment is tangent to the region of supportable behaviors (i.e., that $\vec{w} I_{\max}(X, Z) = \vec{w} I_{\max}(X, Y)$ can be achieved), we choose a decision rule that selects z_1 whenever $w_1 p(x_1 \mid y) \geq w_2 p(x_2 \mid y)$, and z_2 otherwise. This is the decision rule that an ideal observer would use to maximize the expected value (see equation A.1) with priors $p(x_1) = p(x_2)$ and $v(x_j, z_k) = \delta_{jk} w_j$.

In sum, then, what we have shown are three related facts: (1) the tangents represent decision rules; (2) when we know a decision rule, we know what the ideal observer would do, given that decision rule; and (3) the tangents also correspond to weighted Bayesian indices. Thus, if we known the subject's decision rule, we can choose the index corresponding to ideal observer's behavior.

The significance of this is that when we know the decision rule, we do not have to deal with the problem of using many indices and face the potential problems of multiple comparisons—instead we can cut to the chase and choose the index that compares behavior with the ideal observer limit. Note that in the main text, we avoid the multiple-comparison problem by other means—choosing a test that leads to a curved bound; here we are describing how knowledge of the decision rule also provides a way to avoid the problem.

Finally, when there are multiple stimuli and multiple behavioral responses, the correspondence between ideal-observer analysis and tests based on indices becomes more complex. There are two reasons for this. First, there are extreme behaviors (i.e., points on the boundary of the set of supportable behaviors) that do not correspond to optimal decision rules for

any set of values. This is shown in appendix D. Second, the set of optimal decision rules is larger. As a consequence, one must look beyond the weighted Bayesian indices to find the equivalent test based on the DPI. For example, a decision rule that maximizes the expected value of a second guess requires an index derived from equation 2.12 to provide the equivalent DPI test.

Appendix B: The Direct Approach: The Problem of Multiple Comparisons

The reader may wonder why we do not take a more direct approach to the problem of ruling out codes. By "direct," we mean the following: having determined the range of behaviors that can be supported by a code (as in Figures 2–5), why not simply ask whether the observed behavior lies within this convex set?

While at first glance the direct approach might appear the most straightforward, it in fact has a substantial disadvantage: it leads to a problem of multiple comparisons. To be sure that a behavior is inside the convex set, one must test whether it is on the correct side of each tangent to the set (as discussed in appendix A). Each tangent therefore corresponds to a separate statistical test that must be satisfied. That is, determining "directly" whether a behavior can be supported by a code is, implicitly, a multiple-comparison problem. This problem is exacerbated when there are more than two behaviors or more than two stimuli, since the region of supportable behaviors is high-dimensional (see appendix D).

The multiple-comparison problem is particularly difficult because the individual comparisons are highly interdependent but nevertheless distinct. That is, most, but not all, of the codes that are excluded by one test are also excluded by another. In the main text, we circumvent the multiple-comparison problem by choosing a single index. When this strategy is taken, Shannon-like indices, which do not correspond to any decision rule, can be more effective than Bayes-like indices (e.g., Figures 2 and 5), since their bounds curve inward. In appendix A, we describe another way to solve the multiple-comparison problem: if the subject's decision rule is known, then a single Bayes-like index becomes more effective: in fact, it becomes nearly ideal.

One might also hope to avoid the multiple-comparisons problem by formulating a single "compound" hypothesis to test whether the behavior lies within the convex set supported by a putative code. That is, if we had a priori knowledge of the shape of the convex set, we could formulate a single test statistic that would accurately indicate whether the behavior was inside the convex set. The problem with this approach is that in typical experimental situations, the shape of the convex set is not known in advance. Rather, as illustrated in Figures 2 to 5, its shape is determined by the fraction of code words y with each ratio of a posteriori probabilities $p(x_1 \mid y)/p(x_2 \mid y)$. Thus, a multiple-comparisons problem has been avoided, but it has been

replaced by an equivalent problem: the estimation of the number of code words with each a posteriori probability ratio.

Appendix C: Properties of Naive Estimators of Indices of Concentration and Transmitted Information

In this appendix, we (1) show that naive estimators of indices of concentration are upwardly biased, (2) show that the bias decreases monotonically as sample size increases, (3) develop asymptotic expressions for their bias and variance, and (4) discuss how these results extend to estimators of transmitted information. The analysis of the bias of naive estimates of indices of concentration hinges on the convexity property, equation 2.5.

C.1 Upward Bias of Naive Estimates of Indices of Concentration. The naive estimate is formed in the following way. Suppose that N observations $\vec{x} = (x_1, \ldots, x_N)$ are drawn from a discrete distribution P on L symbols Each of the samples of $\vec{x}$ is one of the discrete symbols $1, \ldots, L$. From the set of observations $\vec{x}$, we construct an empirical distribution $P_{\vec{x}}$, in which the probabilities match the observed frequencies in $\vec{x}$. That is, in $P_{\vec{x}}$, the probability assigned to the kth symbol ($1 \leq k \leq L$) is $c_k(\vec{x})/N$, where $c_k(\vec{x})$ is the count of occurrences of the symbol k in $\vec{x}$. By definition, the naive estimate of $f(P)$ is $f(P_{\vec{x}})$.

The expected value of the naive estimate from data sets of size N is

$$E_N(f) = \sum_{\vec{x}=(x_1,\ldots,x_N)} P(\vec{x}) f(P_{\vec{x}}), \tag{C.1}$$

where $P(\vec{x})$ denotes the probability of the set of observations $\vec{x}$ in the true distribution P:

$$P(\vec{x}) = \prod_{n=1}^{N} p_{x_n}. \tag{C.2}$$

The true distribution P is a mixture of the empirical distributions $P_{\vec{x}}$, each weighted by their probabilities $P(\vec{x})$:

$$\sum_{\vec{x}=(x_1,\ldots,x_N)} P(\vec{x}) P_{\vec{x}} = P. \tag{C.3}$$

Therefore, with weights λ chosen to be the probabilities $P(\vec{x})$, the convexity property, equation 2.5, implies that $E_N(f) \geq f(P)$. That is, that the expected value of the naive estimator is greater than the true value of the index of concentration.

C.2 Expected Values of Estimators Decrease Monotonically with Sample Size. Having shown that naive estimates are upwardly biased, we now prove the stronger statement that as the number of observations N increases, the expected value of the naive estimate descends monotonically to its final value. Specifically, we show that $E_N(f)$, equation C.1 is a nonincreasing function of N.

The argument hinges on the observation that an empirical probability distribution $P_{\vec{x}}$ derived from N observations $\vec{x}$ is a mixture of probability distributions derived from $N-1$ observations, that is, probability distributions with one observation missing. We use $\vec{x}(n)$ to denote the sequence of observations of $\vec{x}$ with the nth observation missing, namely, $\vec{x}(n) = (x_1, \ldots, x_{n-1}, x_{n+1}, \ldots, x_N)$. Then,

$$P_{\vec{x}} = \frac{1}{N} \sum_{n=1}^{N} P_{\vec{x}(n)}, \tag{C.4}$$

where $P_{\vec{x}}$ is the empirical probability distribution formed from the full data set, and $P_{\vec{x}(n)}$ is the empirical probability distribution formed from the data set with the nth observation missing.

Equation C.4 follows from a simple counting argument. The probability assigned to the kth symbol in $P_{\vec{x}}$ is $P_{\vec{x}}(k) = c_k(\vec{x})/N$, where $c_k(\vec{x})$ is the count of occurrences of the symbol k in $\vec{x}$. The probability assigned to the kth symbol in $P_{\vec{x}(n)}$ is $P_{\vec{x}(n)}(k) = c_k(\vec{x}(n))/(N-1)$. After multiplication by $N(N-1)$, equation C.4 is equivalent to

$$(N-1)c_k(\vec{x}) = \sum_{n=1}^{N} c_k(\vec{x}(n)). \tag{C.5}$$

Equation C.5 holds because each occurrence of k in the full data set $\vec{x}$ (say, $x_r = k$) corresponds to an occurrence in all of the missing-observation data sets $\vec{x}(n)$ except $\vec{x}(r)$. So each contribution to $c_k(\vec{x})$ is counted $N-1$ times.

Combining the convexity property equation 2.5 (with $\lambda_n = 1/N$) and equation C.4 yields

$$\sum_{\vec{x}=(x_1,\ldots,x_N)} \sum_{n=1}^{N} \frac{1}{N} P(\vec{x}) f(P_{\vec{x}(n)}) \geq \sum_{\vec{x}=(x_1,\ldots,x_N)} P(\vec{x}) f(P_{\vec{x}}). \tag{C.6}$$

The right-hand side of equation C.6 is $E_N(f)$. We will show that the left-hand side is $E_{N-1}(f)$. To do this, we simplify the left-hand side by (1) interchanging the order of summation, (2) breaking the sum over $\vec{x}$ into a component that depends on only the $N-1$ retained observations $\vec{x}(n)$ and

an inner sum that depends on only the value of missing observation $k = x_n$, (3) noting that if the omitted nth value x_n is k, then $P(\vec{x}) = p_k P(\vec{x}(n))$, and (4) noting that $\sum_{k=1}^{L} p_k = 1$. That is,

$$\sum_{\vec{x}=(x_1,\ldots,x_N)} \sum_{n=1}^{N} \frac{1}{N} P(\vec{x}) f(P_{\vec{x}(n)}) = \sum_{n=1}^{N} \sum_{\vec{x}(n)} \sum_{k=1}^{L} \frac{1}{N} P(\vec{x}(n)) f(P_{\vec{x}(n)}) p_k$$

$$= \sum_{n=1}^{N} \sum_{\vec{x}(n)} \frac{1}{N} P(\vec{x}(n)) f(P_{\vec{x}(n)}). \tag{C.7}$$

The final expression of equation C.7 is a sum over N replicas of the same quantity, since, across all full data sets $\vec{x}$, the collection of omitted-sample data sets will be independent of which sequential sample n is omitted. Thus,

$$\sum_{\vec{x}=(x_1,\ldots,x_N)} \sum_{n=1}^{N} \frac{1}{N} P(\vec{x}) f(P_{\vec{x}(n)}) = \sum_{\vec{y}=(y_1,\ldots,y_{N-1})} P(\vec{y}) f(P_{\vec{y}}) = E_{N-1}(f). \tag{C.8}$$

Finally, combining equations C.6 and C.8 yields

$$E_{N-1}(f) \geq E_N(f). \tag{C.9}$$

Moreover, the above argument shows that inequality C.9 is strict (i.e., that the sequence $E_N(f)$ is strictly decreasing) whenever there is at least some pair of distributions $P_{\vec{x}}$ and $P_{\vec{y}}$ for which the inequality of equation C.6 is strict. This is typical of any nontrivial index of concentration.

For the indices of concentration considered in the main text, it is straightforward to show that the expected value of the naive estimate $E_N(f)$ converges to the true value $f(P)$.

C.3 Asymptotic Analysis. We can gain insight into the qualitative behavior of the bias of naive estimates by expanding $f(Q) = f(q_1, \ldots, q_L)$ as a Taylor series for $Q = P_{\vec{x}}$ near P:[7]

$$f(Q) = f(P) + \sum_{k=1}^{L} \frac{\partial f}{\partial p_k}(q_k - p_k)$$

$$+ \frac{1}{2} \sum_{k=1}^{L} \sum_{m=1}^{L} \frac{\partial^2 f}{\partial p_k \partial p_m}(q_k - p_k)(q_m - p_m) + \cdots. \tag{C.10}$$

[7] For the Taylor expansion, we consider the arguments $q_1, \ldots, q_L$ of f to be independent, not constrained to sum to unity. This constraint is expressed by the fact that the covariance matrix, equation C.11, is singular.

Here, q_k is the naive estimate of p_k derived from the observations $\vec{x} = (x_1, \ldots, x_N)$. That is, $q_k = c_k(\vec{x})/N$, where $c_k(\vec{x})$ is the number of occurrences of the symbol k in $\vec{x}$. The first-derivative term does not contribute to the bias, since the expected value of q_k is p_k. Bias arises from the second term, since the covariance of two estimates q_k and q_m is nonzero. In particular, from just a single trial ($N = 1$), the covariances of the counts $c_k(\vec{x})$ are

$$C = \begin{pmatrix} p_1(1-p_1) & -p_1 p_2 & \ldots & -p_1 p_L \\ -p_2 p_1 & p_2(1-p_2) & \ldots & -p_2 p_L \\ \vdots & \vdots & \ddots & \vdots \\ -p_L p_1 & -p_L p_2 & \ldots & p_L(1-p_L) \end{pmatrix}. \tag{C.11}$$

Since successive observations are independent, the covariance matrix of the counts on N trials is CN, and the covariance matrix of the probability estimates $q_k = c_k(\vec{x})/N$ is $CN/N^2 = C/N$. It now follows from equation C.10 that the bias of $E_N(f) = \langle f(Q) \rangle$ may be estimated by

$$\begin{aligned} \langle f(Q) \rangle - f(P) &\approx \frac{1}{2} \sum_{k=1}^{L} \sum_{m=1}^{L} \frac{\partial^2 f}{\partial p_k \partial p_m} \langle (q_k - p_k)(q_m - p_m) \rangle \\ &= \frac{1}{2N} \sum_{k=1}^{L} \sum_{m=1}^{L} \frac{\partial^2 f}{\partial p_k \partial p_m} C_{km}. \end{aligned} \tag{C.12}$$

For $f_{\max}$,

$$\frac{\partial^2 f_{\max}}{\partial p_k \partial p_m} = 0 \quad \text{if } p_k \neq p_m, \tag{C.13}$$

so the asymptotic bias, equation C.12, is zero when all of the p_k's are distinct.

In the Shannon limit, $f_{Shannon}(P) = \lim_{\alpha \to 1} \frac{f_\alpha(P)-1}{\alpha-1} = -H_1(P) = \sum_{i=1}^{L} p_i \log p_i$ (see equation 2.2),

$$\frac{\partial^2 f_{Shannon}}{\partial p_k \partial p_m} = 0 \quad \text{if } k \neq m \quad \text{and} \quad \frac{\partial^2 f_{Shannon}}{\partial p_k^2} = \frac{1}{p_k}. \tag{C.14}$$

This recovers from equation C.12 the well-known result (Carlton, 1969; Miller, 1955; Treves & Panzeri, 1995; Victor, 2000) that the bias is asymptotically independent of the probabilities p_k:

$$\langle f(Q) \rangle - f(P) \approx \frac{L-1}{2N}. \tag{C.15}$$

The customary factor of $\log 2$ is missing from the denominator since the quantities are calculated with natural logs, not bits.

The Taylor expansion, equation C.10, also provides an asymptotic estimate for the variance of the estimate $E_N(f)$:

$$\langle (f(Q) - \langle f(Q) \rangle)^2 \rangle \approx \frac{1}{N} \sum_{k=1}^{L} \sum_{m=1}^{L} \frac{\partial f}{\partial p_k} \frac{\partial f}{\partial p_m} C_{km}. \tag{C.16}$$

Figure 1 shows the behavior of the coefficient of $1/N$ in the bias (see equation C.12) and variance (see equation C.16) of estimators of indices f_α (see equation 2.3).

C.4 Estimators of Generalized Transmitted Information. The analysis in sections C.1 and C.2 extends to the first term in the definition (see equation 2.4) of the transmitted information I_f, namely, the sum over the conditional probabilities:

$$\sum_{j=1}^{N} p_{Y;\,j} f(P_{X|Y=j}). \tag{C.17}$$

As above, we relate estimates from a data set containing N observations to estimates from N data sets containing $N-1$ observations. To do this, we drop the rth observation from the larger data set and compare the estimates of expression C.17 to the estimates obtained from the resulting smaller data sets. Say that for the rth observation, $\vec{y}_r = k$. Dropping this observation does not change any of the estimates involving the probabilities conditioned by the other values $j \neq k$ in Y. That is, estimates of $f(P_{X|Y=j})$ are unchanged, for $j \neq k$. For $f(P_{X|Y=k})$, the above arguments (applied to the subset within the N observations that have $Y = k$) imply that the bias of this term is positive and decreases when the rth observation with $\vec{y}_r = k$ is included. Thus, estimators of expression C.17 have a positive bias that decreases monotonically with sample size.

However, $I_f(X, Y)$ is the difference between expression C.17 and the concentration $f(X)$. If X is known exactly, then so is $f(X)$, and the statistical properties of estimators of $I_f(X, Y)$ are determined by the properties of estimators of expression C.17 as described in the above paragraph. But if X must be determined from the sample, the statistics of estimators of $f(X)$ also have to be considered. In this case, the bias of naive estimates $I_f(X, Y)$ are not guaranteed to be monotonic decreasing.

Appendix D: The Range of Behaviors That Can Be Supported by a Neural Code

In this appendix, we determine the range of stimulus-behavior relationships that can be supported by a given neural code. We show that the

stimulus-behavior relationships that can be supported by a code form a convex set (as illustrated in Figures 2–5), and we characterize the boundary of this set in terms of the decision rules that generate these behaviors—the "extreme" decision rules. Surprisingly, although many extreme decision rules can be described in terms of optimizing the "value" of a behavior, not all extreme decision rules can be described in this fashion when the number of stimuli and behaviors is sufficiently large.

D.1 Extreme Decision Rules Correspond to Convex Polyhedral Partitions. We consider scenarios in which the stimulus set X has L discrete elements, and the behavioral response set Z has N discrete elements. The neural code Y can be either discrete or continuous. Our goal is to determine the range of stimulus-behavior relationships $P_{Z|X}$ that can be supported by a given X, Y, and $P_{X|Y=y}$.

A stimulus-behavior relationship is the net result of an encoding process that transforms the stimuli X into the neural code Y and a decision rule that generates behaviors Z from the code words of Y. In general, the decision rule may be probabilistic, that is, it is specified by the conditional probability distributions $P_{Z|Y}$. The decision rule $P_{Z|Y}$ and the encoding process $P_{Y|X}$ together determine the stimulus-behavior relationship $P_{Z|X}$:

$$P_{Z|X=x;z} = \sum_{y} P_{Z|Y=y;z} P_{Y|X=x;y}. \tag{D.1}$$

Equation D.1 states that the stimulus-behavior relationship $P_{Z|X}$ is a linear transformation of the decision rule $P_{Z|Y}$.

Note that even though individual decision rules may be highly nonlinear, decision rules can be considered to combine in a linear fashion—by mixture. That is, a mixture of two decision rules is a decision rule in which the subject uses one decision rule on some fraction of the trials and another decision rule on the rest of the trials. In this sense, decision rules form a convex set.

Because decision rules form a convex set, the linearity of equation D.1 implies that the set of supportable stimulus-behavior relationships is also convex. We therefore focus on determining the boundaries of this set. These are the "extreme" stimulus-behavior relationships—those for which there is no (nontrivial) decomposition as a mixture:

$$P_{Z|X} = \sum_{n=1}^{N} \lambda_n P_{Z_n|X}. \tag{D.2}$$

The linear relationship D.1 maps a mixture of decision rules into a mixture of behaviors. Therefore, extreme stimulus-behavior relationships must have extreme decision rules—decision rules that are not mixtures of other rules.

Extreme decision rules must be deterministic. This is because nondeterministic decision rules are mixtures of deterministic ones. To see this, suppose that some code word y_0 can lead to several behaviors $z_1, \ldots, z_m$, each with nonzero probability $p_{Z|Y=y_0;z_n}$. These nonzero probabilities can be viewed as weights $\lambda_n = p_{Z|Y=y_0;z_n}$, which express $P_{Z|Y}$ as a convex mixture of rules $P_{Z_n|Y}$ that are deterministic for $y = y_0$ and match $P_{Z|Y}$ for $y \neq y_0$. Conversely, a decision rule that is deterministic for code word y cannot be a nontrivial mixture (since the mixing process implies that a single neural symbol $y \in Y$ can map to more than one behavior in Z).

To sum up, all supportable behaviors are mixtures of extreme behaviors, and extreme behaviors correspond to extreme decision rules, which are necessarily deterministic. We now show that it suffices to consider only a small subset of deterministic rules.

To determine this subset, we introduce a parameterization of decision rules. A decision rule (the probability of choosing a behavior z, given the code word y) can be thought of as acting on the a posteriori distribution $P_{X|Y=y}$ rather than on the identity y of the code word. A decision rule is therefore characterized by the (possibly stochastic) mapping from each a posteriori distribution $\vec{q}$ to the behaviors in Z. We denote this mapping by $r(\vec{q}, z)$, where $\vec{q}$ ranges over all a posteriori distributions $P_{X|Y=y}$, and z ranges over all behaviors Z. That is, given a code word y with a posteriori distribution $\vec{q} = P_{X|Y=y}$, $r(\vec{q}, z)$ is the probability that the behavioral outcome is z. When there are L input symbols, an a posteriori distribution $P_{X|Y=y}$ is a list of L probabilities, that is, a vector $\vec{q} = (q_1, \ldots, q_L)$ whose components are nonnegative and sum to 1. A decision rule is completely described by mappings from such vectors to behaviors in Z, namely, $r(\vec{q}, z)$.

We next rewrite equation D.1, the stimulus-behavior relationship, in terms of $r(\vec{q}, z)$. To do this, we introduce $\rho_{Y|X=x_j}(\vec{q})$ to represent the probability that a trial with stimulus x_j will produce any code word in Y for which $P_{X|Y=y} = \vec{q}$. Since all of these symbols lead to behaviors as determined by $r(\vec{q}, z)$, we may rewrite equation D.1 as

$$P_{Z|X=x_j;z} = \int_{\vec{q} \in J} \rho_{Y|X=x_j}(\vec{q}) r(\vec{q}, z)\, d\vec{q}, \tag{D.3}$$

where J is the space of a posteriori probabilities, $\vec{q} = (q_1, \ldots, q_L)$ When the set of code words Y is discrete, then so is the density $\rho_{Y|X=x_j}(\vec{q})$, and the integral, equation D.3, becomes a sum.

Equation D.3 can be put into a form that avoids the stimulus conditioning in the density $\rho_{Y|X=x_j}(\vec{q})$. To begin, let $\rho_Y(\vec{q})$ be the probability that any trial will produce a code word y for which $P_{X|Y=y} = \vec{q}$. (In the text examples, $\rho_Y(\vec{q}) = \rho(q_1)$, the quantity plotted in the insets of Figures 2–5.) We now relate $\rho_Y(\vec{q})$ to $\rho_{Y|X=x_j}(\vec{q})$. The joint probability of a stimulus x_j and a

code word $y \in Y$ with $P_{X|Y=y} = \vec{q}$ is $\rho_Y(\vec{q}) \cdot q_j$, since this expression is the probability that such a code word y was present ($\rho_Y(\vec{q})$), times the conditional probability of x_j given this code word (q_j). But this joint probability can also be calculated from the product of the conditional density $\rho_{Y|X=x_j}(\vec{q})$ and the a priori probability of x_j (p_j). Thus,

$$p_j \rho_{Y|X=x_j}(\vec{q}) = q_j \rho_Y(\vec{q}). \tag{D.4}$$

Substitution of equation D.4 into equation D.3 yields

$$P_{Z|X=x_j;z} = \frac{1}{p_j} \int_{\vec{q} \in J} q_j \rho_Y(\vec{q}) r(\vec{q}, z)\, d\vec{q}. \tag{D.5}$$

The above equations are linear in $r(\vec{q}, z)$, so linear combinations of r correspond to linear combinations of the stimulus-behavior relationship. We are now set up to characterize the extreme stimulus-behavior relationships in terms of r.

As we have seen above, for extreme stimulus-behavior relationships, the decision rule is deterministic. For a deterministic rule, $r(\vec{q}, z)$ is concentrated on a single value in Z for each $\vec{q} \in J$. We use $J(z)$ to denote the "indicator region" for z, namely, the region of J for which $r(\vec{q}, z) = 1$. Equation D.5 can be then be rewritten as

$$P_{Z|X=x_j;z} = \frac{1}{p_j} \int_{\vec{q} \in J(z)} q_j \rho_Y(\vec{q})\, d\vec{q}. \tag{D.6}$$

Appendix E shows, based on this representation, that any extreme stimulus-behavior relationship is the result of a rule in which each $J(z)$ is convex. That is, extreme stimulus-behavior relationships correspond to partitions of the space J into convex indicator regions $J(z)$. In the main portions of Figures 2 to 5, the region of behaviors that can be supported by each code is derived from equation D.6, with indicator regions $J(z_1)$ and $J(z_2)$ that consist of partitions of the interval $[0, 1]$ into disjoint convex sets, namely, $[0, q)$ and $[q, 1]$.

The partitioning of the domain J into convex regions $J(z_j)$ has an intuitive interpretation. Imagine that on a single trial, the decision process is uncertain as to whether a trial resulted in neural symbol y_a or y_c (with a posteriori probabilities $\vec{q}_a = P_{X|Y=y_a}$ and $\vec{q}_c = P_{X|Y=y_c}$), but that both neural symbols are in the same indicator region $J(z_j)$. Even though the subject is uncertain about the a posteriori probabilities, the subject is nevertheless sure that they lie somewhere on the line between $\vec{q}_a$ and $\vec{q}_c$. The convexity property states that under these circumstances, the decision rule yields the output symbol z_j. That is, if the subject is uncertain as to

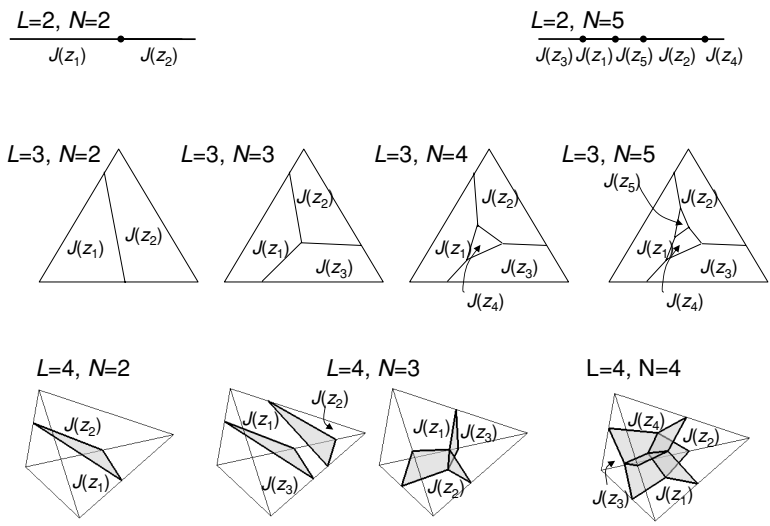

Figure 6: Geometric characterization of extreme stimulus-behavior relationships in scenarios with L stimuli and N behavioral responses. The extreme stimulus-behavior relationships are parameterized by the partitions of the $L-1$-dimensional space of a posteriori probabilities into N convex subsets.

which of two neural symbols was present but either would have resulted in the same behavior z_j, then that behavioral response will always be produced.

Since any two adjacent indicator regions $J(z_j)$ and $J(z_k)$ within this partition are both convex, their mutual border must be flat (e.g., a straight line segment or a region of a hyperplane). Consequently, for any extreme decision rule, the domain J is partitioned into convex polyhedrons $J(z_j)$, one for each output symbol z_j. This is a much stronger condition than merely requiring that r is deterministic, since deterministic rules can have arbitrary shapes for the indicator regions $J(z)$.

In sum, the boundary of the stimulus-behavior relationships $P_{Z|X}$ that can be supported by a neural code Y is determined by the "extreme" decision rules, which are in turn parameterized by the partitions of the space J of a posteriori probabilities into convex polyhedral indicator regions $J(z_j)$. Examples of partitions of J with L stimuli in X and N behaviors in Z are shown in Figure 6.

D.2 Optimal Decision Rules. The above framework readily provides for a characterization of "optimal" decision rules: decision rules that maximize the expected value of a behavior. But as the examples below will show,

extreme decision rules need not be "optimal" in this sense. This somewhat surprising result underscores the point that the range of systematic strategies that a subject might choose is very large.

Optimal decision rules are characterized by convex polyhedral indicator regions. We consider an optimal decision rule for a set of values $v(x_j, z_k)$ associated with producing a behavior z_k when the stimulus is x_j. Given a neural response y with a posteriori probabilities $\vec{q}$, the expected value $V_k(\vec{q})$ associated with a behavior z_k is

$$V_k(\vec{q}) = \sum_{j=1}^{L} v(x_j, z_k) q_j. \tag{D.7}$$

Maximizing the expected value of the behavior corresponds to choosing the behavior z_k that maximizes $V_k(\vec{q})$.

Each $V_k(\vec{q})$ is a linear function on the domain J of $\vec{q}$. At a typical point $\vec{q}$ in J, one of the $V_k(\vec{q})$ will be larger than all other $V_m(\vec{q})$. For each k, the set of points for which $V_k(\vec{q})$ is larger than all other $V_m(\vec{q})$ constitutes the interior of $J(z_k)$, the indicator region of J on which the response symbol z_k is chosen. The boundary between two indicator regions is the points of J at which two (or more) of the $V_k(\vec{q})$ are identical. Since $V_k(\vec{q}) = V_m(\vec{q})$ is a linear relationship among the $\vec{q}$'s, these boundaries will be flat.

D.3 Examples. We now consider how this analysis applies to scenarios in which there are a specific number of stimuli L in X and a specific number of behaviors N in Z. When there are only two stimuli, or only two behaviors, the optimal decision rules and the extreme decision rules are identical. But in general, there are extreme decision rules that do not correspond to any optimal decision rule.

The set J of all possible a posteriori probabilities $\vec{q}$ consists of all L-element vectors $\vec{q} = (q_1, \ldots, q_L)$ with nonnegative components that sum to 1. For example, with $L = 2$ symbols, J is a line segment. With $L = 3$ stimuli, J is the set of three-element vectors $\vec{q} = (q_1, q_2, q_3)$ in the triangle with vertices at $(0, 0, 1)$, $(0, 1, 0)$, and $(1, 0, 0)$. To characterize the extreme stimulus-behavior relationships, we need to determine the ways that we can partition this space into N convex subsets.

D.3.1 Two Stimuli, Two or More Responses. For two stimuli ($L = 2$), J is the set of two-element vectors $\vec{q} = (q_1, q_2)$ on the line segment from $(0, 1)$ to $(1, 0)$, Convex subsets of J consist of shorter segments that it contains. Therefore, extreme decision rules with N behavioral responses correspond to partitions of the unit line segment into N smaller segments (see Figure 6, top row). Any such partition corresponds to a choice of $N - 1$ cutpoints, followed by an assignment of the resulting N segments to the N behavioral

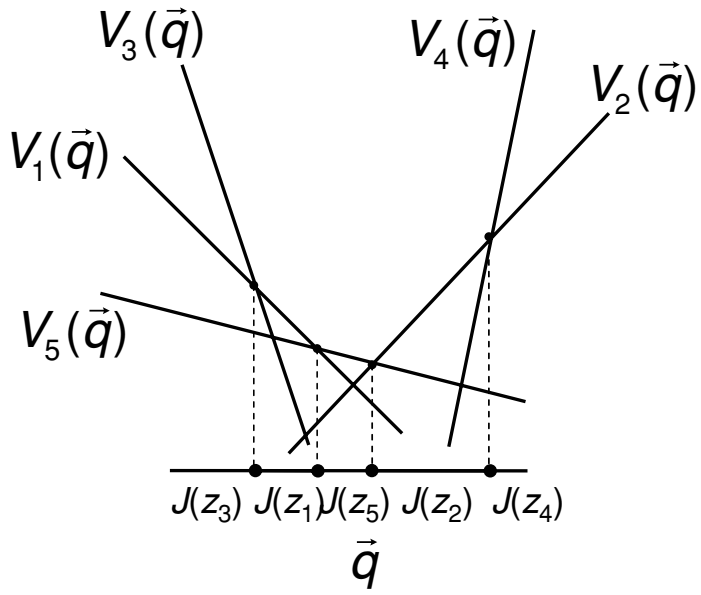

Figure 7: Correspondence between extreme rules and optimal rules for scenarios with $L = 2$ stimuli. The case of $N = 5$ behavioral responses is shown. The expected value associated with each behavior $z_1, \ldots, z_5$ is a linear function V of the a posteriori probabilities $\vec{q}$. Each of these linear functions is maximal over a different interval of the range of a posteriori probabilities.

responses. Note that the symbols can be assigned in any order (as illustrated in the example for $L = 2, N = 5$). For any such partition, it is possible to find a corresponding set of stimulus-behavior values $v(x_j, z_k)$—a set of values $v(x_j, z_k)$ for which the corresponding expected values $V_k(\vec{q})$ (see equation D.7) intersect at these cutpoints, as illustrated in Figure 7.

Note that for all values of N, many choices of the values $v(x_j, z_k)$ will lead to the same cutpoints, since the cutpoints are determined solely by the values of $\vec{q}$ at which the $V_k(\vec{q})$ intersect. The slopes of the successive $V_k(\vec{q})$ must be monotonically increasing, but they are otherwise free to vary.

D.3.2 Two Responses, Two or More Stimuli. For L stimuli, J is a simplex of dimension $L - 1$ with vertices at $(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$. Extreme decision rules with $N = 2$ behavioral responses correspond to partitions of J into two convex sets—to all the hyperplanes that intersect J (first column of Figure 6). As in the above case of two stimuli, it is possible to find many sets of stimulus-behavior values $v(x_j, z_k)$ that lead to this partition. To do this, it suffices to choose $v(x_j, z_k)$so that the equation of the separating hyperplane

between $J(z_1)$ and $J(z_2)$ is given by

$$\sum_{j=1}^{L} \big(v(x_j, z_1) - v(x_j, z_2)\big) q_j = 0. \tag{D.8}$$

D.3.3 More Than Two Stimuli and More Than Two Behavioral Responses. This situation is noteworthy because it is no longer the case that partitions of J into N convex sets necessarily correspond to an optimal decision rule for some set of values $v(x_j, z_k)$. To demonstrate this, we show that there are more degrees of freedom required to specify the partition than can be supported by the number of degrees of freedom provided by choices of the values $v(x_j, z_k)$. Let us assume that any partition can be generated by optimizing the expected value $V_k(\vec{q})$ derived from some set of values. Consider the partitions for ($L = 3$, $N = 3$) and ($L = 3$, $N = 4$) in Figure 6. For the ($L = 3$, $N = 3$) partition, the boundaries between each pair of indicator regions correspond to the condition that two expected values are equal: $V_1(\vec{q}) = V_2(\vec{q})$, $V_1(\vec{q}) = V_3(\vec{q})$, and $V_2(\vec{q}) = V_3(\vec{q})$. The point at which all three indicator regions meet is determined by $V_1(\vec{q}) = V_2(\vec{q}) = V_3(\vec{q})$. The configuration shown for ($L = 3$, $N = 4$) is derived from the configuration from ($L = 3$, $N = 3$) in the following manner. First, the point at which the three indicator regions meet is replaced by a small triangle, $J(z_4)$. Second, the line segments that form the boundaries of the first three indicator regions are shifted by arbitrary small amounts, since (because of the presence of $J(z_4)$) they are no longer constrained to meet at a point. Thus, the ($L = 3$, $N = 4$) configuration has four more degrees of freedom than the ($L = 3$, $N = 3$) configuration: for ($L = 3$, $N = 4$), the coordinates of the corners of $J(z_4)$ must be specified (6 degrees of freedom), while for the ($L = 3$, $N = 3$) configuration, the coordinates of vertex common to $J(z_1)$, $J(z_2)$, and $J(z_3)$ must be specified. This process can be continued indefinitely, adding an indicator region by replacing a vertex with a triangle and allowing the positions of its corners to vary slightly. Each new indicator region thus adds 4 degrees of freedom. However, adding a new behavior allows only three additional values: $v(x_1, z_N)$, $v(x_2, z_N)$, and $v(x_3, z_N)$. Consequently, for a sufficiently large number of behaviors N, at least some configuration cannot be realized as any optimal decision rule. For $L = 3$, this point is reached at $N = 4$, since 9 degrees of freedom are required to specify a polygonal partition (2 degrees of freedom for each of the three internal intersections, 1 degree of freedom for each of the points on the edges), while at most $LN - L - 1 = 8$ degrees of freedom are available from the choices of values.[8]

[8]The array of values $v(x_j, z_k)$ has NL parameters, but some sets of values result in the same decision rules. This results in a loss of $L + 1$ degrees of freedom for the decision rules: the hyperplanes determined by equation D.8 are unchanged if an arbitrary constant

The existence of extreme decision rules that are not optimal decision rules also occurs for $L > 3$ stimuli. Once a sufficient number of indicator regions are present, addition of a new indicator region (a new behavior) replaces a point common to L indicator regions with a small new simplex. The L vertices of the new simplex can be independently specified in the $(L-1)$-dimensional domain J, gaining $L(L-1)$ degrees of freedom. But since the new simplex covered up the common point, the $L-1$ degrees of freedom required to specify the coordinates of the common point are lost. Thus, there is a net gain of $L(L-1)-(L-1) = (L-1)^2$ degrees of freedom. Since adding a new behavior allows only L additional values $v(x_j, z_{N+1})$, the number of degrees of freedom required to specify a convex partition will eventually exceed the number of degrees of freedom available to specify an optimum decision rule via a set of values.

Appendix E: Convexity of Indicator Regions for Extreme Decision Rules

Here we prove, as required for appendix D, that any extreme stimulus-behavior relationship can be represented by a decision rule in which each $J(z)$ is convex. We argue by contradiction. Assume we have an extreme stimulus-behavior relationship for which the sets $J(z)$ that characterize the deterministic rule are not convex. We show how this implies that the behavior is a nontrivial mixture. In particular, assume that we have a rule described by a deterministic $r(\vec{q}, z)$ with a nonconvex indicator region $J(z_1)$ (see Figure 8). That is, we assume that there are points $\vec{q}_a$ and $\vec{q}_c$ in $J(z_1)$ and an intermediate point $\vec{q}_b$ in a distinct region $J(z_2)$. We will construct a mixture of a trio of deterministic rules $r^{[m]}(\vec{q}, z)$ $(m = 1, 2, 3)$ that yields the same behavior as $r(\vec{q}, z)$ via equation D.5. The strategy is to shift some of the mass in $J(z_2)$ at $\vec{q}_b$ to one of the flanking points $\vec{q}_a$ or $\vec{q}_c$ (see Figure 8).

Since the behaviors corresponding to the rules $r^{[m]}(\vec{q}, z)$ (via equation D.5) are distinct, we have exhibited the assumed extreme behavior $r(\vec{q}, z)$ as a mixture—and therefore have derived a contradiction from the assumed existence of an intermediate point $\vec{q}_b$ in a distinct region $J(z_2)$.

In order for this shift to leave the stimulus-behavior linkage unchanged, $r^{[m]}(\vec{q}, z)$ and their mixture weights w_m (with $\sum_{m=1}^{3} w_m = 1$) need to satisfy

$$\frac{1}{p_j} \sum_{h=a,b,c} q_{h,j}\, \rho_Y(\vec{q}_h) r(\vec{q}_h, z) = \frac{1}{p_j} \sum_{m=1}^{3} w_m \left(\sum_{h=a,b,c} q_{h,j}\, \rho_Y(\vec{q}_h) r^{[m]}(\vec{q}_h, z) \right), \tag{E.1}$$

c_j is added to each $v(x_j, z_k)$ or if all of the $v(x_j, z_k)$ are multiplied by a positive constant. Thus, at most $NL - L - 1$ degrees of freedom are available for the number of optimal decision rules.

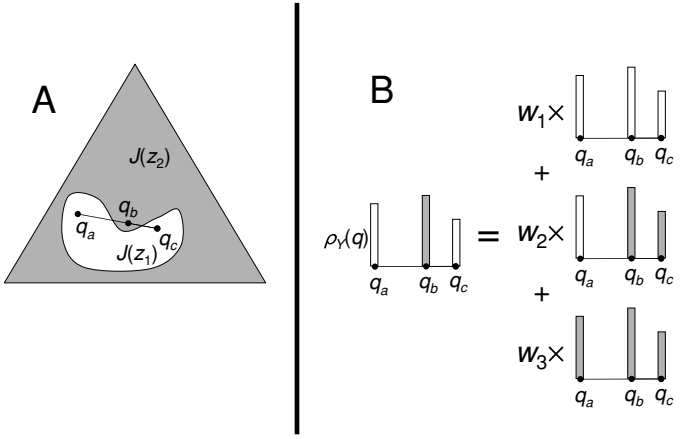

Figure 8: Diagram illustrating shifting of mass for proof of characterization of extreme decision rules (see appendix E). (A) A decision rule r with a nonconvex indicator region $J(z_1)$. (B) At the code words with a posteriori probabilities $\vec{q}_a$, $\vec{q}_b$, and $\vec{q}_c$, r can be represented as a weighted sum of three decision rules $r^{[1]}$, $r^{[2]}$, and $r^{[3]}$. The heights of the bars indicate the density $\rho(\vec{q})$, and the shading indicates assignment to $J(z_1)$ (open) or $J(z_2)$ (filled).

where $r(\vec{q}_a, z_1) = 1$ and $r(\vec{q}_c, z_1) = 1$ (since $\vec{q}_a$ and $\vec{q}_c$ are in $J(z_1)$), and $r(\vec{q}_b, z_2) = 1$ (since $\vec{q}_b$ is in $J(z_2)$), and all other $r(\vec{q}_h, z) = 0$. This follows from the discrete version of equation D.6.

At all points z other than z_1 or z_2, equation E.1 is readily satisfied by choosing $r^{[m]}(\vec{q}, z) = 0$ unless $z = z_1$ or $z = z_2$. For $z = z_1$ or $z = z_2$, equation E.1 represents L pairs of equations, with one pair for each coordinate j in Q:

$$q_{a,j}\rho_Y(\vec{q}_a) + q_{c,j}\rho_Y(\vec{q}_c) = \sum_{m=1}^{3} w_m \sum_{h=a,b,c} q_{h,j}\rho_Y(\vec{q}_h) r^{[m]}(\vec{q}_h, z_1) \tag{E.2}$$

and

$$q_{b,j}\rho_Y(\vec{q}_b) = \sum_{m=1}^{3} w_m \sum_{h=a,b,c} q_{h,j}\rho_Y(\vec{q}_h) r^{[m]}(\vec{q}_h, z_2). \tag{E.3}$$

However, these L pairs of equations are highly degenerate because $\vec{q}_b$ is between $\vec{q}_a$ and $\vec{q}_c$ on a line. That is, we can write $q_{h,j} = q_{a,j} + \mu(q_{c,j} - q_{a,j})$, with $\mu = 0$ for $\vec{q}_a$, $\mu = 1$ for $\vec{q}_c$, and some $\mu = \lambda$ for $\vec{q}_b$, with $0 < \lambda < 1$. With these substitutions and matching the coefficients of μ^0 and μ^1, the L pairs

of equations, E.2 and E.3 (one for each j), reduce to four equations:

$$\rho_Y(\vec{q}_a) + \rho_Y(\vec{q}_c) = \sum_{m=1}^{3} w_m \sum_{h=a,b,c} \rho_Y(\vec{q}_h) r^{[m]}(\vec{q}_h, z_1), \tag{E.4}$$

$$\rho_Y(\vec{q}_c) = \lambda \rho_Y(\vec{q}_b) \sum_{m=1}^{3} w_m r^{[m]}(\vec{q}_b, z_1) + \rho_Y(\vec{q}_c) \sum_{m=1}^{3} w_m r^{[m]}(\vec{q}_c, z_1), \tag{E.5}$$

$$\rho_Y(\vec{q}_b) = \sum_{m=1}^{3} w_m \sum_{h=a,b,c} \rho_Y(\vec{q}_h) r^{[m]}(\vec{q}_h, z_2), \tag{E.6}$$

$$\lambda \rho_Y(\vec{q}_b) = \lambda \rho_Y(\vec{q}_b) \sum_{m=1}^{3} w_m r^{[m]}(\vec{q}_b, z_2) + \rho_Y(\vec{q}_c) \sum_{m=1}^{3} w_m r^{[m]}(\vec{q}_c, z_2). \tag{E.7}$$

Without loss of generality, we may assume that

$$\lambda \rho_Y(\vec{q}_a) \geq (1 - \lambda) \rho_Y(\vec{q}_c), \tag{E.8}$$

since if not, it suffices to reverse the roles of a and c.

To find weights in $[0, 1]$ that solve equations E.4 to E.7, we make the following assignments for the deterministic rules $r^{[m]}(\vec{q}, z)$: $r^{[1]}$ chooses response z_1 at all $\vec{q}_h$, $r^{[2]}$ chooses response z_1 at $\vec{q}_a$ but z_2 at $\vec{q}_b$ and $\vec{q}_c$, and $r^{[3]}$ chooses z_2 at all $\vec{q}_h$ (see the shading in Figure 8B). That is,

$$\begin{aligned} r^{[1]}(\vec{q}_a, z_1) &= r^{[1]}(\vec{q}_b, z_1) = r^{[1]}(\vec{q}_c, z_1) = 1 \\ r^{[2]}(\vec{q}_a, z_1) &= r^{[2]}(\vec{q}_b, z_2) = r^{[2]}(\vec{q}_c, z_2) = 1 \\ r^{[3]}(\vec{q}_a, z_2) &= r^{[3]}(\vec{q}_b, z_2) = r^{[3]}(\vec{q}_c, z_2) = 1 \end{aligned} \tag{E.9}$$

with all other values $r^{[m]}(\vec{q}_h, z) = 0$. These choices reduce equations E.4 to E.7 to

$$\rho_Y(\vec{q}_a) + \rho_Y(\vec{q}_c) = w_1 \left(\rho_Y(\vec{q}_a) + \rho_Y(\vec{q}_b) + \rho_Y(\vec{q}_c)\right) + w_2 \rho_Y(\vec{q}_a), \tag{E.10}$$

$$\rho_Y(\vec{q}_c) = \lambda w_1 \rho_Y(\vec{q}_b) + w_1 \rho_Y(\vec{q}_c), \tag{E.11}$$

$$\rho_Y(\vec{q}_b) = (w_2 + w_3) \left(\rho_Y(\vec{q}_b) + \rho_Y(\vec{q}_c)\right) + w_3 \rho_Y(\vec{q}_a), \tag{E.12}$$

$$\lambda \rho_Y(\vec{q}_b) = (w_2 + w_3) \left(\lambda \rho_Y(\vec{q}_b) + \rho_Y(\vec{q}_c)\right). \tag{E.13}$$

Straightforward algebra verifies that the choices

$$
\begin{aligned}
w_1 &= \frac{\rho_Y(\vec{q}_c)}{\lambda\rho_Y(\vec{q}_b) + \rho_Y(\vec{q}_c)} \\
w_2 &= \frac{\rho_Y(\vec{q}_b)}{\rho_Y(\vec{q}_a)} \frac{\lambda\rho_Y(\vec{q}_a) - (1-\lambda)\rho_Y(\vec{q}_c)}{\lambda\rho_Y(\vec{q}_b) + \rho_Y(\vec{q}_c)} \\
w_3 &= \frac{\rho_Y(\vec{q}_b)}{\rho_Y(\vec{q}_a)} \frac{(1-\lambda)\rho_Y(\vec{q}_c)}{\lambda\rho_Y(\vec{q}_b) + \rho_Y(\vec{q}_c)}
\end{aligned}
\tag{E.14}
$$

satisfy equations E.10 through E.13 and that $\sum_{m=1}^{3} w_m = 1$. Since (via equation E.8) all of the $w_m \geq 0$, it also follows that $w_m \leq 1$.

Acknowledgments

We thank Ifije Ohiorhenuan and Rebecca Jones for comments on the manuscript. S.N. is supported by EY12978. J.V. is supported by EY9314 and by MH68012 (to Dan Gardner).

References

Abramowitz, M., & Stegun, I. A. (1970). *Handbook of mathematical functions*. New York: National Bureau of Standards.

Berger, W. H., & Parker, F. L. (1970). Diversity of planktonic foraminifera in deep-sea sediments. *Science*, *168*(3937), 1345–1347.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., & Warland, D. (1991). Reading a neural code. *Science*, *252*(5014), 1854–1857.

Carlton, A. G. (1969). On the bias of information estimates. *Psychol. Bull.*, *71*, 108–109.

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

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

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

Kennel, M. B., Shlens, J., Abarbanel, H. D., & Chichilnisky, E. J. (2005). Estimating entropy rates with Bayesian confidence intervals. *Neural Comput.*, *17*(7), 1531–1576.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, *7*, S7.

McClurkin, J. W., Optican, L. M., Richmond, B. J., & Gawne, T. J. (1991). Concurrent processing and complexity of temporally encoded neuronal messages in visual perception. *Science*, *253*(5020), 675–677.

Miller, G. A. (1955). Note on the bias on information estimates. *Information Theory in Psychology: Problems and Methods, II-B*, 95–100.

Nemenman, I. (2004). *Information theory, multivariate dependence, and genetic network inference* (Tech. Rep. NSF-KITP-04-54). Santa Barbara: Kavli Institute for Theoretical Physics, University of California, Santa Barbara.

Nemenman, I., Bialek, W., & de Ruyter van Steveninck, R. (2004). Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, *69*(5 Pt. 2), 056111.

Nirenberg, S., Carcieri, S. M., Jacobs, A. L., & Latham, P. E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature*, *411*(6838), 698–701.

Nirenberg, S., Jacobs, A., Fridman, G., Latham, P., Douglas, R., Alam, N., et al. (2006). Ruling out and ruling in neural codes. *Journal of Vision*, *6*, 889a.

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

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 547–561). Berkeley: University of California Press.

Rényi, A. (1970). *Probability theory*. Amsterdam: Elsevier North-Holland.

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

Shlens, J., Kennel, M. B., Abarbanel, H. D., & Chichilnisky, E. J. (2007). Estimating information rates with confidence intervals in neural spike trains. *Neural Comput.*, *19*(7), 1683–1719.

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*(1), 197–200.

Thomson, E. E., & Kristan, W. B. (2005). Quantifying stimulus discriminability: A comparison of information theory and ideal observer analysis. *Neural Comput.*, *17*(4), 741–778.

Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Comput.*, *7*, 399–407.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.*, *52*, 479–487.

Victor, J. D. (2000). Asymptotic bias in information estimates and the exponential (Bell) polynomials. *Neural Comput.*, *12*(12), 2797–2804.

Victor, J. D., & Purpura, K. P. (1996). Nature and precision of temporal coding in visual cortex: A metric-space analysis. *J. Neurophysiol.*, *76*(2), 1310–1326.

Wehrl, A. (1978). General properties of entropy. *Reviews of Modern Physics*, *50*(2), 221–260.

Zhao, Y.-B., Fang, S.-C., & Li, D. (2005). *Constructing generalized mean functions using convex functions with regularity conditions*. Available online at http://www.optimization-online.org/DB_FILE/2005/06/1164.pdf.

Received October 29, 2007; accepted February 12, 2008.