Four Top Reasons Mutual Information Does Not Quantify Neural Information Processing

Don H. Johnson
Rice University
Houston, Texas

Oral or Poster Presentation

1 Introduction

The mutual information between the stimulus S and the response R, whether the response be that of a single neuron or of a population, has the form

$$\mathcal{I}(S;R) = \iint p(R \mid S)p(S) \log \frac{p(R \mid S)}{\int p(R \mid \sigma)p(\sigma)\,d\sigma}\,dS dR\,, \tag{1}$$

where $p(R \mid S)$ is the probability distribution of the response for a given stimulus condition and $p(S)$ is the probability distribution of the stimulus. It is important to note that this probability distribution is defined over the entire stimulus space. For example, if black-and-white images serve as the stimulus, $p(S)$ is defined over *all* positive-valued signals having a compact two-dimensional domain. To measure mutual information, the experimenter defines a stimulus set $\{S_1, \ldots, S_N\}$ and, from the measured response, estimates $p(R \mid S_n)$, the probability distribution of the response under each stimulus condition. The mutual information is estimated as [1]

$$\widehat{\mathcal{I}}(S;R) = \sum_n \int p(R \mid S_n)\,\Pr(S_n) \log \frac{p(R \mid S_n)}{\sum_l p(R \mid S_l)\,\Pr(S_l)}\,dR \tag{2}$$

where $\Pr(S_n)$ is the probability of the n^{th} stimulus occurring.

In information theory, the mutual information is seldom used save for finding the capacity $\mathcal{C}$, defined to be the maximum value of the mutual information over all stimulus probability distributions.

$$\mathcal{C} = \max_{p(S)} \mathcal{I}(S;R)$$

What other uses mutual information might have can be questioned. Because of the way mutual information is calculated and used, several important issues *not* involving empirical concerns arise.

Mutual information depends on the stimulus probabilities. As can be seen from its definition (1) or its estimate (2), mutual information depends on the stimulus probabilities. Mutual information measures how different, in a statistical sense, the stimulus and response are. It equals zero when the response is statistically independent of the stimulus and equals either $\mathcal{H}(S)$ (2) or ∞ (1) when the response directly reflects the stimulus.[1] *Mutual information does not depend solely the stimulus-response relationship.* In computing (2), picking *any* particular set of stimulus probabilities $\Pr(S_n)$ is difficult (impossible) to justify. The usual choice is to make them equal or to correspond to the presentation probabilities. However, what does this arbitrary choice have to do with the capability of the system under study to represent the stimulus? As the stimulus probabilities vary, mutual information can vary between zero and the theoretical upper limit (capacity). By finding the capacity, we obtain a measure that does characterize the stimulus-response relationship. This approach would seem to provide a measure of information processing capability.

[1] Note the difference here between measured and theoretical mutual information calculations. This discrepancy shows that in the simple case of the response being a unique deterministic function of the stimulus, the true and estimated values of mutual information *cannot* agree.

Finding capacity depends on the stimulus set. Stein [4] studied the capacity of several rate-coding models. In these models, stimulus S_n produced response rate R_n, which controlled the count distribution according to a model (e.g., Poisson). Not only did he maximize mutual information with respect to $\Pr(S_n)$, he also found the response rates within a specified rate range that maximized mutual information. For example, for the average spike count constrained to lie between 1 and 10, the maximal mutual information occurred when $\{R_n, \Pr(S_n)\}$ were $(1, 0.42)$, $(4.4, 0.18)$ and $(10, 0.40)$. What his paper does *not* say is that when more than three stimuli are used to produce responses in the same average-count range, mutual information is maximized by the same set of three response rates and stimulus probabilities with the "extra" stimuli assigned zero probability. Thus, it is theoretically possible that in an attempt to maximize mutual information, *the result may be that the experimenter should not have presented some of his/her stimuli*. This effect makes little sense, but is a property of estimated mutual information (2) and the capacity calculation.

Estimating capacity cannot yield the true capacity. The estimate of (2) cannot yield the true value expressed by (1). Some have argued that by presenting a complete set of stimuli[2] and measuring mutual information, the stimulus-response relation is fully characterized. Consequently, if we optimized (2) to remove stimulus probability effects, we have a capacity estimate $\widehat{\mathcal{C}}$. *Unfortunately, optimizing over stimulus probabilities* $\Pr(S_n)$ *is not equivalent to optimizing over all stimulus probability distributions* $p(S)$. In our example taken from Stein, as we try to add more stimuli, the optimal solution is to have only three, which will never approximate a continuous distribution. If, instead of optimizing with respect to the rates as well as to the stimulus probabilities, equally spaced rates are assumed and the maximizing probabilities, the capacity increases to maximum then decreases as the number of stimuli increases. Stein [4] argues that the capacity-achieving probability distribution is proportional to $1/\sqrt{R}$. It is impossible to find this result from the experimental approach implied by (2).

Inferring information processing capability from capacity is difficult. Even if the true capacity could be computed, what that result means is unclear. *If* the stimulus was somehow being used as a digital communication medium (i.e., used to communicate sequences of bits) and the responses were used to determine the bit sequence, the capacity $\mathcal{C}$ would define how many bits per stimulus presentation could be reliably communicated (Shannon's Noisy Channel Coding Theorem). However, most stimuli represent the external environment, which is not a bit sequence. The Noisy Channel Coding Theorem also applies to analog (continuous-valued) communication, but in a much more complicated way [3]. His framework is that the stimulus is encoded in a rate λ, this rate and the count statistics determine the response R, and this quantity decoded as S'.[3] Shannon defines what is now known as a *distortion measure* ν according to

$$\nu = \iint \rho(S, S') p(S, S')\, dS dS' ,$$

where $\rho(S, S')$ is the *distortion function* that measures how similar the original and decoded signals are. For example, $\rho(S, S')$ could be $(S - S')^2$, in which case ν is the mean-squared distortion. $p(S, S')$ is the joint distribution between the stimulus and the decoded stimulus. Finding this quantity requires a specification of the stimulus probability distribution $p(S)$ and sending stimuli according to that probability law through the sensory system to determine the statistical characteristics $p(S' \mid S)$ of the decoded output. Shannon then defines $\mathcal{R}$ to be the rate at which information can be reproduced to a given distortion ν_0 as

$$\mathcal{R} = \min_{p(S,S')} \mathcal{I}(\lambda; R) \text{ with } \nu \leq \nu_0 .$$

[2] A *complete* set means that all stimuli can be represented as a weighted linear combination of elements in the set.

[3] For simplicity, I use the rate code as an example. The same result applies to more general point-process models of the response.

This constrained minimization with respect to the joint probability function $p(S, S')$ can be quite difficult to compute. The Noisy Channel Coding Theorem now becomes that so long as $\mathcal{R} < \mathcal{C}$, where capacity is computed with respect to the pair (λ, R), the input can be encoded in such a way that the required distortion criterion is met. The more stringent the criterion (a smaller ν_0), the larger $\mathcal{R}$ becomes until one cannot send information through the channel and meet the distortion criterion. Said another way, finding the capacity of a stimulus-response relationship with respect to analog stimuli yields a limit on how effectively a source can be encoded according any distortion criterion. Note that in contrast to the digital version of the Noisy Channel Coding Theorem, error-free communication is not possible in this case. We could assess information processing by translating the measured capacity into a value for the distortion measure. Note that we chose the distortion measure; what distortion measures sensory systems employ are difficult to characterize and using the "right" one will almost certainly lead to analytic difficulties in finding ν. Thus, it would seem that exploiting capacity is difficult if not impossible.

2 Conclusions

Because of these difficulties, communication engineers do not use mutual information. It is best used as a vehicle for defining capacity, but the utility of using capacity to characterize analog sensory systems is fraught with difficulties. Quantities other than mutual information should be used, one possibility being the approach described in [2].

References

[1] A. Borst and F.E. Theunissen. Information theory and neural coding. *Nature Neuroscience*, 2:947–957, 1999.

[2] D.H. Johnson, C.M. Gruner, K. Baggerly, and C. Seshagiri. Information-theoretic analysis of neural coding. *J. Comp. Neuroscience*, 10:47–69, 2001.

[3] A.N. Kolmogorov. On the Shannon theory of information transmission in the case of continuous signals. *IRE Trans. Info. Th.*, 3:102–108, 1956.

[4] R. B. Stein. The information capacity of nerve cells using a frequency code. *Biophysical J.*, 7:67–82, 1967.