

Neural Transmission of Information

Marcello M. DiStasio
SUNY Downstate Medical Center and
Polytechnic Institute of NYU

February 12, 2010

1 Introduction

Many definitions of ‘information’ in living tissue are in colloquial use, but in order to develop a framework for investigation into the principles underlying information processing by nervous systems we require a consistent definition of information as a quantity on an absolute scale, independent of content. One of the most general definitions of information (but not the only one) is ‘a measure of certainty at one time or place of the state of something at another time or place’. This is pretty general, and could in principle apply to any states that we can characterize with just two things: a variable, and a way of quantifying our certainty about that variable’s value. The application of a measure of information based on this definition requires some assumptions about the system in question, most importantly that the variables at one locus in the system *represent* variables at another locus. Transmission of information is only meaningful if there is something on the other end to receive it (with hopeful allusions towards SETI, there may be a signal careening across the universe to our electromagnetic receivers, but until we recognize it as meaningful it is just energy like any other stream – there is no information received when it dissipates on heating an intervening planet, for example). Thus, in the case of the brain, for our measure to have any chance of showing real utility, we must cast neurons as traffickers in information. An appeal to experimental neurobiology supports this assumption.

The behavioral output of organisms is quite clearly linked to sensory input. Bright lights cause pupillary constriction; rabbits can be conditioned to blink in response to a tone; if you call my name, I will likely turn around. The literature abounds with more specific and quantifiable examples of neurons functioning to coordinate responses to conditions internal and external to organisms. A correspondence has been demonstrated between neural activity and the states of many circumstances, including density and energy of photons incident on the retina [15, 16], current and past states of stretch in a muscle [11], impending motor activity [2], internal tissue damage [7], states of hunger or satiety [10], and innumerable other such conditions that could be characterized as ‘state variables’. A huge amount of metabolic energy is expended by most organisms on maintaining this neural correspondence. Because of the prevalence of these correlations, and the clear understanding of the functioning of

such mechanisms throughout the animal kingdom, the carrying of information by neurons is clearly not just an epiphenomenon, a side effect of some physiological process that runs independently of the correlation between action potential firing and external variable states. Instead, and since these states have such impact on survival, it makes sense to hypothesize that the nervous system is actively *representing* them in proxy so that they can be evaluated, processed, and used to generate outputs that increase the specificity (and thus likelihood of success) of responses.

Thus, a key area of investigation is how these representations can be inferred from discrete events (e.g. action potential firing times) or continuous quantities (e.g. transmembrane potential) measured from single neurons or ensembles of neurons. Because so little is known about this neural code, it makes sense to begin a general search by finding some limits on what the forms of the code could be. We want to make as few assumptions as possible about any particular input/output function for a neuron or group, since in experimental conditions we often have very incomplete knowledge about what the true range of inputs and outputs are. Quantitative measures of information that are based on statistics, rather than knowledge of a particular encoding/decoding algorithm, can be applied to certain quantities recorded from neurons to establish such limits. The measures discussed here, Fisher Information, Mutual Information, and Normalized Transfer Entropy, have been applied with success in the analysis of various experiments aimed at quantifying general properties of neural systems that dictate how well they transmit information from one place to another.

2 Fisher Information

Fisher information is a measure of the amount of information carried by observations of a random variable \mathbf{Z} about a parameter Θ that determines the likelihood function $L(\Theta) = f(\mathbf{Z}; \Theta)$. In statistics, the concept of likelihood allows estimation of unknown parameters of a distribution given known outcomes drawn from the distribution [19]. Fisher Information, $\mathcal{I}(\Theta)$, is defined as the second moment (variance) of the score. In statistics, the ‘score’ is the derivative of the log of the likelihood function. So roughly, the Fisher information $\mathcal{I}(\mathbf{Z})$ about Θ tells how sharply distributed the possible values of Θ are (given our observations

of \mathbf{Z}), and thus how certain we can be of its value. Formally, following the derivation in [Cover and Thomas Sec. 11.10](#) [6], Fisher information for an individual sample is given by

$$\mathcal{I}(\Theta) = E_{\Theta} \left[\frac{\partial}{\partial \theta} \ln f(\mathbf{Z}, \Theta) \right]^2 \quad (1)$$

where $E[\bullet]$ represents the expected value. It has the useful property that it is additive over independent identically distributed samples $\{z_1, \dots, z_n\} \in \mathbf{Z}$. This means that if we draw samples $\{z_1, \dots, z_n\}$, the information obtained about the parameter Θ taking all of them together, $\mathcal{I}_n(\Theta)$ is equal to $n \times \mathcal{I}(\Theta)$. This meaning of information makes sense, since with each sample drawn, we learn something about the distribution, which improves our ability to predict the subsequent samples. We are thus gaining knowledge about the generating process. If a receiver of a group of spike trains (i.e. a downstream neuron) is trying to obtain knowledge of the stimulus that provoked the spikes, finding out about the current value of a parameter that governs the distribution of inter-spike intervals (ISIs) may be way of finding out a the value of a parameter of the stimulus. Assume, for example, that a primary sensory neuron fires spikes with normally distributed ISIs with a constant variance whose mean μ varies in response to stimulus intensity. Then the Fisher information for each spike would give an indication of how many spikes it would take for downstream neurons to make an estimate of μ (and therefore the stimulus intensity) with some chosen level of accuracy [26]. The application of this method only applies when the distribution in question can be parametrized. So if the form of the distribution of the neural variable in question is completely unknown (or subject to change), Fisher Information and other likelihood-based methods (e.g. maximum likelihood) will not be appropriate.

3 Self-Information and Entropy

Since we have defined information as ‘certainty’ about a variable, it makes sense to measure it as a reduction in uncertainty. In 1948, Claude Shannon developed a mathematical definition for uncertainty (in the context of communication of variable values), and, at the suggestion of John von Neumann, termed it entropy. Entropy is defined as the expected value of the self-information a random variable contains about itself [25].

Self-information can be understood this way: Given the proclamation that a single outcome x_i has occurred out of a collection of possible outcomes represented by the random variable $\mathbf{X} = \{x_1, \dots, x_n\}$, and the probability of this occurrence is $p(x_i)$, then the the self-information $SI(x_i)$ is

$$SI(x_i) = \log \left(\frac{1}{p(x_i)} \right) = -\log(p(x_i)) \quad (2)$$

This definition has a couple attractive properties that make it useful as a basis for our information measure. One is that it depends only on the probability of the occurrence of the

outcome $p(x_i)$, which makes intuitive sense – it is more ‘surprising’ to find out a rare event has occurred rather than a common one, and if we have prior knowledge of their relative frequencies, we should need nothing else to quantify just how surprising it is. And, like Fisher Information, this measure is additive over independent events. In other words, if an event C is composed of two mutually independent events A and B then the amount of information contained in the knowledge that C occurred is equal to the sum of the information contained in the knowledge that A and B each occurred:

$$\begin{aligned} SI(C) &= -\log(p(C)) = -\log(p(A) \cdot p(B)) \\ &= -\log(p(A)) - \log(p(B)) \\ &= SI(A) + SI(B) \end{aligned}$$

Intuitive justification for this property comes from the fact that whether we have observed two outcomes separately, or a single other outcome that tells us that both of those have occurred, we have gained the same amount of information about what transpired. The use of the log function transforms the multiplication of independent probabilities into an addition of information.

The entropy of a random variable \mathbf{X} is just the expected value of $SI(\mathbf{X})$:

$$H(\mathbf{X}) = E(SI(\mathbf{X}))$$

where expected value is defined for a discrete random variable \mathbf{R} with probability mass function $P_{\mathbf{R}}(r)$ as

$$E(\mathbf{R}) = \sum_i r_i P_{\mathbf{R}}(r_i)$$

Thus, for a random variable \mathbf{X} with possible values $\{x_1, \dots, x_n\}$ and probability mass function $P_{\mathbf{X}}(x)$ we get a final expression for entropy $H(\mathbf{X})$:

$$H(\mathbf{X}) = \sum_i P_{\mathbf{X}}(x_i) \log_b P_{\mathbf{X}}(x_i) \quad (3)$$

We have basically just summed over the probabilities of all possible outcomes (with each probability scaled by it’s own logarithm). We can choose any base b we want for the logarithm, but it is common to choose 2 giving entropy in units of ‘bits’, or e giving units of ‘nats’. Bits are the most commonly used unit. This measure of entropy is extremely useful, since for any condition (neural or otherwise) that we can cast as a random variable and assign a probability mass function (often the most difficult part), we immediately have access to the uncertainty about a report of its state. In a neural system, the entropy computed using spike trains depends on how we define the random variable. It is common to take \mathbf{X} as the number of spikes observed during a single time bin (whose distribution we estimate empirically based on observation). Entropy then gives us a measure of the number of possible signals that could in theory be distinguished from one another using this temporal resolution. This places a theoretical upper bound on the number of possible states the spike train could represent if it is using an *ideal code*. The following measures build upon this knowledge to quantify changes in uncertainty under particular conditions.

4 Mutual Information

One useful question we can use this measure of uncertainty to ask is the one suggested by our general definition of information above (‘a measure of certainty at one time or place of the state of something at another time or place’). To do this, we take for one random variable \mathbf{X} the quantity being reported on and for another random variable \mathbf{Y} the quantity representing the report. For example, in a primary sensory neuron, the intensity of the neuron’s preferred stimulus during a given interval could be represented by \mathbf{X} and the number of action potentials fired during this interval could be \mathbf{Y} . These are arbitrary selections; any physical quantity could be used in principle, but, as mentioned above, in experiments the probability distribution (mass functions) for these variables are often prohibitively difficult to control, measure, or even estimate. Putting aside these concerns for now and assuming we know the distributions, we can define mutual information as the amount of information shared between two variables. In other words it is the amount by which our uncertainty about one variable is reduced by knowing the value of the other. Thus, mutual information is given by

$$I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}) \quad (4)$$

which can be expressed more directly by its equivalent in terms of probability mass functions:

$$I(\mathbf{X}; \mathbf{Y}) = \sum_{y \in \mathbf{Y}} \sum_{x \in \mathbf{X}} P_{\mathbf{X}\mathbf{Y}}(x, y) \log_2 \left(\frac{P_{\mathbf{X}\mathbf{Y}}(x, y)}{P_{\mathbf{X}}(x)P_{\mathbf{Y}}(y)} \right) \quad (5)$$

Note that this equation implies that the measure is completely symmetrical: $I(\mathbf{X}; \mathbf{Y}) = I(\mathbf{Y}; \mathbf{X})$. Again this is intuitive, since mutual information is a measure of *shared* information, which makes no statement about causality.

4.1 Application to Neural Systems

Knowing the entropy of spiking pattern observed in a neuron places a limit on the maximum amount of information that the neuron could be transmitting; the entropy represents the number of bits required to specify all possible responses under all stimulus conditions. If we call the response \mathbf{R} and the stimulus \mathbf{S} , then the conditional entropy $H(\mathbf{R}|\mathbf{S})$ can be thought of as ‘neuronal noise’ - the amount of uncertainty remaining in the neural response when the stimulus is known. Thus, comparison of $H(\mathbf{R}|\mathbf{S})$ to $H(\mathbf{R})$ gives a measure of how much uncertainty about the neuron’s activity is reduced upon presentation of the stimulus. This is a good measure of the average efficiency of neural encoding. Measurement of neural response \mathbf{R} in experiments is highly dependent on the stimuli applied, since this directly affects which responses occur. In order to fully characterize the distribution of possible responses, stimuli with maximum entropy can be applied, in order to evoke as many different responses as possible and give the closest estimate of true response entropy. Calculation of the opposite quantity $H(\mathbf{S}|\mathbf{R})$ is meaningful as well. It represents ‘stimulus equivocation’, or the amount of uncertainty remaining about the

value of the stimulus when the neural response is known. If $H(\mathbf{S}|\mathbf{R})$ is comparably large with $H(\mathbf{S})$, then the response is doing a poor job of discriminating any given stimulus from the other possible stimuli. By its symmetry, $I(\mathbf{R}; \mathbf{S})$ measures both neural noise and stimulus equivocation, giving a good picture of how well responses discriminate among all possible stimuli [1].

Another useful way to analyze neural activity is to model noisy neural processes in response to a noisy input as a ‘Gaussian channel’ [9]. Imagine a neuron responds to some external signal s with a response y . If we hypothesize that the neuron’s response is proportional to the signal but with some added noise η referred to the input, we can model the process as

$$y = g(s + \eta)$$

where g is a gain factor, and the noise is drawn from a Gaussian distribution ($\eta \in \mathbf{N} \sim N(\mu, \sigma)$). This means that the output is completely determined by the signal and the noise. Also note that we make no assumption about the distribution of the signal.

In the case of a Gaussian channel, the upper bound on mutual information can be calculated directly from the signal-to-noise ratio (SNR). This very useful property is the result of a long derivation that can be found in section 3.1.3 of [Spikes], and relies on the fact that it can be proven that a Gaussian distribution has the greatest entropy of any distribution. It states that

$$I(\mathbf{S}; \mathbf{Y}) \leq \frac{1}{2} \log_2 [1 + SNR] \quad (6)$$

The signal-to-noise ratio for a Gaussian channel is defined as $SNR = \frac{\langle s^2 \rangle}{\langle \eta^2 \rangle}$, where $\langle \bullet \rangle$ denotes taking a mean. This comes from the definition in signal theory of SNR as signal power divided by noise power. The inequality above becomes an equality (which is the maximum information) if the signal is Gaussian. From here we can also derive an expression for the *rate* of information transmission in terms of the power spectra of the signal and noise, $S(\omega)$ and $N(\omega)$:

$$R = \int_0^\infty \log_2 \left[1 + \frac{S(\omega)}{N(\omega)} \right] d\omega \quad (7)$$

which has units of bits per second.

One useful question to ask is ‘What is the maximum amount of information a neuron can transmit, in the presence of noise?’. The expression above tells us that we can find out if we are able to measure the signal-to-noise ratio for the neuron’s output. de Ruyter van Steveninck and Laughlin [22] applied this method to the analysis of the transmission of information through photoreceptor chemical synapses onto large monopolar cells in the blowfly. This is a graded synapse, meaning the post-synaptic depolarization due to neurotransmitter release varies continuously (in analog fashion). They determined the noise in the synapse by repeatedly presenting the same stimulus (a sequence of light flashes resulting in a known contrast profile) and recording membrane voltage from the photoreceptor and from the post-synaptic (monopolar) cell. The mean of the voltage

response $\langle v(t) \rangle$ was taken as the signal power, while the average of the power spectra of the demeaned responses was taken as the noise power. Performing the above calculation (7) yielded a value of 1650 bits per second for the rate of transmission across this ‘analog’ synapse, which is five times greater than the rates estimated for spiking neurons. They also were able to estimate the information rate at each synaptic active zone to be approximately 50 bits per second, assuming that each zone operates independently of the others. This is a significantly higher rate of transmission than that estimated for synapses of spiking neurons (in the neighborhood of 50-100 bits per second for the whole synapse, which still translates to > 1 bit per spike in many cases).

5 Transfer Entropy

Transfer entropy was devised to provide a measure of information that differentiates between information that is actually exchanged and information that is shared due to a common history or common inputs [24]. By using Shannon entropy calculations on variables that are conditioned on their appropriate transition probabilities, transfer entropy gives a measure of only the information in one responding variable that is due to the value of another driving variable. Thus, it is an asymmetrical measure: $TE_{\mathbf{X} \rightarrow \mathbf{Y}} \neq TE_{\mathbf{Y} \rightarrow \mathbf{X}}$. This is highly useful in complex neural systems, where causality cannot often be determined directly or assumed. Transfer entropy can be applied to a system that can be modeled by a Markov process of arbitrary order k . This means that the probability of finding a given state of the system at any time n is independent of the states of the system at time $n-k$ (or further back in the past). The system must also be stationary, meaning that the statistical distributions of the random variables involved do not change over time. These are very rough approximations to the nervous system, since we know that biological processes may be influenced by events occurring further back in the past than we can account for, and the probabilities of cells firing may change over time, in response to neuromodulatory effects, for example. Nonetheless, if we take care to minimize such effects, modeling spike trains as stationary Markov processes is a useful approximation. According to this model, the formulation of transfer entropy for two time series of samples $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$ drawn from probability mass functions \mathbf{X} and \mathbf{Y} , respectively, is

$$TE_{\mathbf{X} \rightarrow \mathbf{Y}} = \sum_{i=1}^{n-1} P(y_{i+1}, y_i^{(k)}, x_i^{(l)}) \log \frac{P(y_{i+1} | y_i^{(k)}, x_i^{(l)})}{P(y_{i+1} | y_i^{(k)})} \quad (8)$$

where $r_i^{(m)} = \{r_i, \dots, r_{i-m+1}\}$. This means that the selection of k and l in the formula allows one to choose how much of each variable’s past to incorporate into the calculation independently from the other. The probability conditional on the past in the denominator, $P(y_{i+1} | y_i^{(k)})$, means that any predictive value about the value of \mathbf{Y} that comes from its own past (which encompasses common inputs to both) is factored out. All that remains is information transmitted

from \mathbf{X} to \mathbf{Y} , though this may include any intermediary processes.

5.1 Numerical implementation for spike data

For finite discrete time signals (e.g. binned spike data, such that the probability of occurrence of a spike in any given bin is $\ll 1$), an estimation that converges on the transfer entropy is

$$TE_{\mathbf{X} \rightarrow \mathbf{Y}} = \sum_u \sum_v \sum_w P(Y^F = u, Y^P = v, X^P = w) \times \log \left[\frac{P(Y^F = u, Y^P = v, X^P = w) P(Y^P = v)}{P(Y^P = v, X^P = w) P(Y^F = u, Y^P = v)} \right] \quad (9)$$

for $u, v \in \{y_1, y_2, \dots, y_{max}\}$ and $w \in \{x_1, x_2, \dots, x_{max}\}$ where y_{max} and x_{max} are the maximum observed values in the signals \mathbf{Y} and \mathbf{X} , respectively [4]. Basically u, v , and w range up to infinity, but the probabilities at values greater than the maximum values observed are zero and are thus unnecessary to compute. The variables Y^F , Y^P , and X^P represent the states of the random variables at times in the future and past. Thus, if \mathbf{Y} represents a spike train, that at any given time t , Y^P is the number of spikes that occurred in train \mathbf{Y} during the past time interval $[t - \tau_{past}, t]$ and Y^F is the number of spikes that occur in the upcoming interval $[t, t + \tau_{future}]$. Likewise for X^P . Note that we can choose our past and future intervals independently, allowing us to set how much ‘memory’ and ‘foresight’ the calculation employs. By choosing the appropriate τ values information transfer on arbitrarily large or small time scales can be investigated, as long as the statistics can be assumed to be stationary on the chosen scale.

5.1.1 Normalization

The transfer entropy measure above has a tendency to drift upwards as a function of the window size $[\tau_{future}, \tau_{past}]$ used to calculate it in practice. As such, it is necessary to remove this bias from it by recomputing the measure with a shuffled \mathbf{X} in order to disconnect \mathbf{X} and \mathbf{Y} without changing the inter-spike distribution of either [4]. To maintain the interpretation as information in \mathbf{Y} not explained by its own past but explained by the past of \mathbf{X} , the result should be normalized by the conditional entropy in \mathbf{Y} of its future on its past: $H(Y^F | Y^P)$. Thus the normalized transfer entropy is:

$$NTE_{\mathbf{X} \rightarrow \mathbf{Y}} = \frac{TE_{\mathbf{X} \rightarrow \mathbf{Y}} - TE_{\mathbf{X} \rightarrow \mathbf{Y}}^{shuffled}}{H(Y^F | Y^P)} \quad (10)$$

This defines a computational procedure that can be implemented on the computer to determine the information transfer in one direction or the other between any given pair of spike trains. Note NTE will, by definition, fall within the interval $[0, 1]$, allowing comparisons between multiple pairs. It is important to realize, however that the distribution of

NTE values is difficult to estimate, meaning that one should not use means of NTEs (or z-scores) for comparisons. Instead non-parametric methods could be employed.

5.1.2 Application to model spike trains

To test the NTE measure, two series of spike arrival times **Signal 1** and **Signal 2'** were generated using Poisson arrival times as a model for the firing statistics. These times are initially uncorrelated. Next, a set fraction of spike times in **Signal 2'** were randomly selected and removed. They were then replaced with the same fraction of instances of spike times taken from **Signal 1** but shifted forward in time (all by the same fixed amount). The result is **Signal 2**. This deliberately introduces a degree of causality of **Signal 1** upon **Signal 2**. The NTE algorithm was applied to calculate $NTE_{\text{Signal 1} \rightarrow \text{Signal 2}}$. This whole procedure was repeated for series of fractions of spikes replaced (i.e. introducing more and more causality) and the results are shown here:

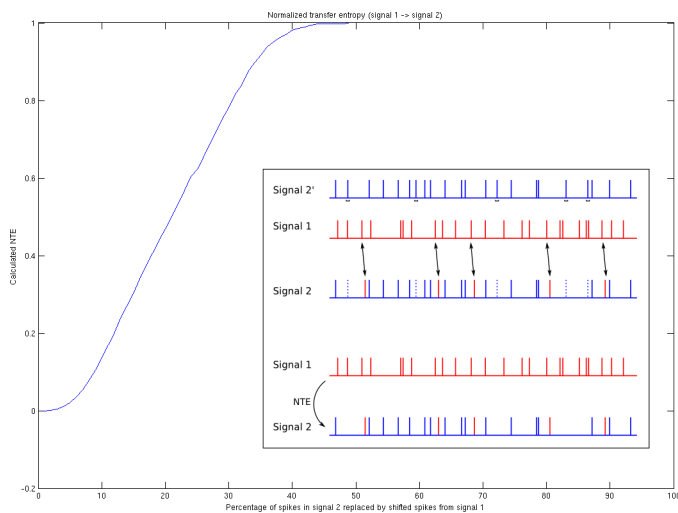


Figure 1: **Testing NTE on model spike trains.** The inset is a cartoon showing the process of generating the model data (see text).

Notice that NTE increases monotonically as the fraction of ‘directly influenced’ spikes increases, up to its maximum value when the fraction replaced is 50%. The particular form of this curve is specific to this model and does not apply to all information sources, but the monotonic increase with increasing causality is inherent, as is the scale. This example simply serves to validate our intuition about the behavior of the measure on discrete time series data.

6 Information in Ensembles of Neurons

Decoding neural networks amounts to identifying channels along which groups of neurons communicate, and figuring out what information is being transmitted via these channels. If our hypothesis is that the channels for communica-

tion are in the form of spike times, we may wish to record a large group of these simultaneously; from an array of electrodes placed in the cortex, for example. Faced with a large ensemble of neural spike trains recorded from such areas in the brain where inputs and outputs are not well defined, it becomes extremely useful to be able to guess the direction of information flow based purely on observation of spike times. Pairwise comparison of spike trains against each other can give a map of information transfer, showing which neurons’ spike trains are most influential on any given cell, and which cells in the group its spikes most influence.

6.1 Application to experimental data

Calculations of NTE were applied to simultaneous spike trains recorded extracellularly from the thalamus and cortex of awake behaving macaques. The time resolution of the sampling was 1 ms. Spikes were sorted offline using principle component analysis. The NTE measure was applied to every pairwise combination of spike trains in order to generate an information transfer matrix that shows the direction of information flow, taking each channel as a potential source or receiver. For these calculations, parameters $\tau_{past} = \tau_{future} = 10\text{ms}$ were used.

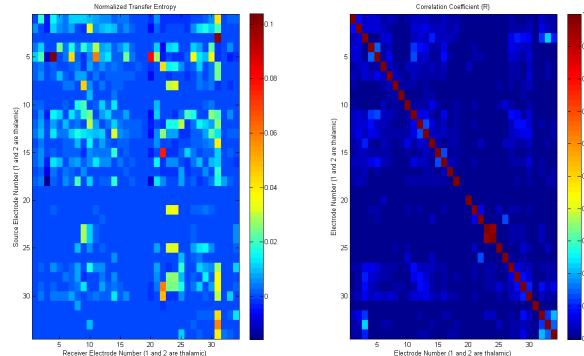


Figure 2: **Comparing NTE with cross-correlation on spike data from thalamus and cortex of a monkey.** On the left side, the sources are arranged on the ordinate and the receivers on the abscissa. The color value for each pair represents NTE from the source to the receiver, as a percentage of the maximum NTE observed. The first two channels in each are thalamic, while the rest are cortical. On the right the cross correlation matrix (R values) for the same units is shown for comparison.

Notice that channels that show higher correlation also have a tendency to show an increase in transfer entropy in one direction or the other. The cross correlation gives us no information about *which* direction, however – its graph is perfectly symmetrical. NTE, on the other hand shows a number of channel pairs where information is transmitted preferentially or exclusively in one direction. Some sources show information transfer to a large number of receivers (they appear as bright horizontal lines), and some receivers show information transfer from a large number of sources

(bright vertical lines). The 'memory' parameters τ_{past} and τ_{future} were chosen arbitrarily for this calculation. It would be interesting to see how the information transmission patterns change as larger spans (moments) in the past and future are included in the calculation. Maybe some cells will be sensitive to many others for a very brief period in the past, but only sensitive to one or a few over longer periods. It would also be interesting to look at these maps under different experimental conditions. Maybe some cell pairs will show information transfer in one direction in under some circumstances, and in the opposite direction under others. Finally, it should be clear that this measure does not make any assumptions about potential intermediate cells or other processes, as long as the distributions from which the measured inter-spike intervals are drawn are stationary.

7 Conclusion

Research into the nature of the neural code focuses on three main questions [1]: What is being encoded? How is it being encoded? With what precision is it being encoded? The first two questions can only be answered by insightful experimental design and probing the activity of neurons directly using relevant stimuli. The third question, however, is amenable to analysis with the information theoretical techniques developed above, and has important implications for the design of experiments designed to answer the other two. Comparing the behavioral responses to a stimulus to the responses of a theoretical 'ideal' observer based on a model of the neural processes observed gives a useful measuring stick for whether the neural activity measured and modeled is really the basis for the behavior, or whether other important processes are being missed. Thus, if an investigator has developed a model for how processing of a sensory input works, a key component of its validation should be an information theory based investigation of the precision with which it handles inputs on the relevant time scales. If the results show an upper bound on information transfer that does not permit enough data to reach stages in the model where it is required, the model must be revised.

Fisher information is well-suited to problems where the general form of the distribution of inputs is known, but the parameters that dictate the specific activity are not. In sensory systems, such a parameter (the mean over a certain period of an input with known variance) may be exactly the piece of information required by downstream processing and perceptual systems. Thus, it makes sense to estimate the information transmitted this way when we have a good idea about what the information input is, and what form its probability distribution takes. Relaxing the requirement for knowledge of the form of the distribution brings us to entropy based measures like mutual information. We see that mutual information is well-suited to investigating the properties of communication channels and methods, as in the establishment of limits on the amount of information transmitted through a given system with known noise. Due to its symmetrical nature, however, mutual information gives no

indication of directionality or causality. Causality could be suggested (but not proven) by significant mutual information between one event and another that preceded it (since it is unlikely that any event is affected by the future, barring quantum mechanical effects). Still, it is highly useful to employ an inherently asymmetrical measure like transfer entropy in situations where we are asking about which neurons drive others and under what conditions (particularly in large networks). Transfer entropy, like mutual information, makes no assumptions about the form of the underlying probability distributions that generate observed activity other than that they are stationary (i.e. whatever the distributions are, they do not change over time).

These tools give us useful methods for examining the properties of candidate neural codes, and the representation of variables using them. However, even the constituents of such encoded representations are not yet unquestionably established. For any given variable to be represented, we must ask: are the spike times the most important variable? Average numbers of spikes occurring within certain spatiotemporal windows? Local field potentials? Spike times relative to the period of field potential oscillations? Intracellular calcium concentrations in certain microdomains? Information theory will help eliminate (and potentially validate) choices for the appropriate encoders, out of the many possible variables to consider.

All of the measures here are based on characterizing probability distributions for inputs and outputs, which is notoriously difficult in experiments on complex neural systems. The statistics of such systems are obstinately non-stationary (one cell may exhibit different firing modes, bursting vs. tonic, for example, in response to the effects of certain neuromodulators, and these modes may have markedly different statistics for inter-spike intervals). Thus, experiments must be precisely designed to ensure that analysis is done over intervals wherein stationarity can be roughly assumed.

Finally, we must be aware that none of the measures discussed here make any claims about the content of any given spike interval or other individual transmission. They are all properties of the communication process on the whole, determined statistically. The interpretation of their meaning and the validity of their application to problems in neuroscience is highly dependent on the appropriate selection of inputs, outputs, and parameters of the assumed distributions, if any. However, if care is taken, they can yield important conclusions about the nature of information transmission and representation in neural systems [5, 21, 18, 12, 13, 23].

References

- [1] F E Theunissen A Borst. Information theory and neural coding. *Nature Neuroscience*, 2(11):947–57, Nov 1999.
- [2] R Caminiti J T Massey A P Georgopoulos, J F Kalaska. On the relations between the direction of two-dimensional arm movements and cell discharge in

- primate motor cortex. *J. Neurosci.*, 2(11):1527–1537, 1982.
- [3] G Faucon B Gourevitch, R Le Bouquin-Jeannes. Linear and nonlinear causality between signals: methods, examples and neurophysiological applications. *Biol Cybern.*, (95):349–369, Sep 2006.
- [4] J J Eggermont B Gourevitch. Evaluating information transfer between auditory cortical neurons. *J Neurophysiol*, 97:2533–2543, Mar 2007.
- [5] A Borst. Noise, not stimulus entropy, determines neural information rate. *J Comput Neurosci*, 14(1):23–31, Jan 2003.
- [6] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley and Sons. New Jersey, 2nd edition, 2006.
- [7] R Dubner and G J Bennett. Spinal and trigeminal mechanisms of nociception. *Annu Rev Neurosci*, 6:381–418, Jan 1983.
- [8] Chris Eliasmith and Charles H. Anderson. *Neural Engineering. Computation, Representation, and Dynamics in Neurobiological Systems*. MIT Press, 2003.
- [9] Rob de Ruyter van Steveninck Fred Rieke, David Warland and William Bialek. *Spikes. Exploring the Neural Code*. MIT Press., 1997.
- [10] R L Townsend L M Patterson H Zheng H R Berthoud, G M Sutton. Brainstem mechanisms integrating gut-derived satiety signals and descending forebrain information in the control of meal size. *Physiol Behav*, 89(4):517–24, Nov 2006.
- [11] M Hulliger. The mammalian muscle spindle and its central control. *Reviews of Physiology, Biochemistry and Pharmacology*, 101:1–110, 1984. PMID: 6240757.
- [12] R R de Ruyter van Steveninck I Nemenman, W Bialek. 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, May 2004.
- [13] W Bialek R R de Ruyter van Steveninck I Nemenman, G D Lewen. Neural coding of natural stimuli: information at sub-millisecond resolution. *PLoS Computational Biology*, 4(3):1–12, Mar 2008.
- [14] A N Kolmogorov. Three approaches to quantitative definition of information. *Problemy Peredachi Informatsii*, 1(1):3–11, Oct 1965.
- [15] S W Kuffler. Discharge patterns and functional organization of mammalian retina. *J Neurophysiol*, 16(1):37–68, Jan 1953.
- [16] D Baylor L Lagnado. Signal flow in visual transduction. *Neuron*, 8(6):995–1002, Jun 1992.
- [17] M Meister M R DeWeese. How to measure the information gained from one symbol. *Comput. Neural Syst*, 10:325–34, Nov 1999.
- [18] K Aihara N Masuda. Ergodicity of spike trains: when does trial averaging make sense? *Neural Computation*, Jan 2003.
- [19] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 2002.
- [20] John R. Pierce. *An Introduction to Information Theory. Symbols, Signals and Noise*. Dover. New York, 2nd edition, 1980.
- [21] S P Strong R Koberle W Bialek R R de Ruyter van Steveninck, G D Lewen. Reproducibility and variability in neural spike trains. *Science*, 275(5307):1805–1808, Mar 1997.
- [22] S B Laughlin R.R de Ruyter van Steveninck. The rate of information transfer at graded-potential synapses. *Nature*, 379:642–645, Feb 1996.
- [23] W Bialek R Koberle S P Strong, R R de Ruyter van Steveninck. On the application of information theory to neural spike trains. *Pac Symp Biocomput*, pages 621–32, Jan 1998.
- [24] T Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461–464, 2000. PMID: 10991308.
- [25] C E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:243–379, Jul 1948.
- [26] S Amari T Toyozumi, K Aihara. Fisher information for spike-based population decoding. *Phys. Rev. Lett.*, 97(9):098102, Aug 2006.