Owen Marschall
DOB: 11/11/92
Neuroscience

Comparing measures of communication for the purpose of defining network optimality

Owen Marschall, Amherst College; Krešimir Josić, University of Houston

Introduction

Optimization problems in neural network models frequently appear in the theoretical neuroscience literature for a variety of purposes, such as putting upper bounds on network performance[1], formulating general principles of network organization[3,6], and even predicting how encoding strategies vary with changing stimulus statistics[1]. Each of these approaches raises several questions, but there is one common question among all of them: what is our definition of network performance? In other words, what exactly are we trying to optimize? We tackled this question directly and compared the **mutual information** and **minimum mean-square stimulus-estimation error**, two measures of network performance that are often used in neuroscience. Although Thomson and Kristan found in 2004 that knowing the mutual information of a distribution does not put an upper bound on its minimum mean-square error[4], we believe that there is a stronger relationship than this result suggests. We compared the two measures analytically in the special case of jointly Gaussian distributions and numerically in the Ising Model and Gaussian Channel. Since these measures are often used in optimization problems, it seemed most natural to compare them by investigating their extrema, or more generally, how small perturbations in the stimulus-response joint distribution affect both measures. Intuitively, we'd like to think that more mutual information would lead to a smaller MMSE; finding when this is true and when it is not will lead to a better understanding of these two measures.

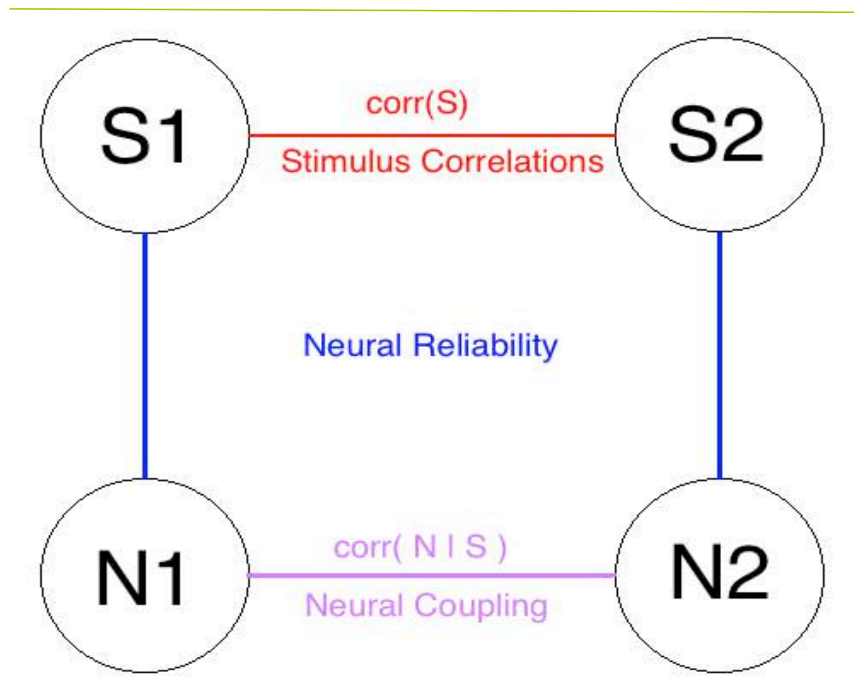


Figure 1: Diagram of a simple model of a two-neuron response to two stimuli. The model uses three parameters: stimulus correlation, neural reliability, and neural coupling.

A Simple Optimization Problem

Consider a two-neuron system as depicted in Figure 1. We take the input statistics and neural reliability to be fixed and the neural coupling **J** to be controlled by the network. The way we define "control" over the neural coupling parameter is by allowing it to vary in the following optimization problems:

$\arg\min_J MMSE$ and $\arg\max_J I$

We find these optimal neural couplings as a function of the parameters the network cannot control, namely the neural reliability and the input statistics. This setup is an important topic of research itself,[1] and it gives us a new approach to our fundamental question of "How are the MI and MMSE related?" by asking how the extrema of these functionals are related in models relevant to neuroscience.

Results

Motivating Example: Tkačik et al. (2010)

Tkačik et al. found that the mutual-information-maximizing neural coupling J_{12}^* for the two-neuron Ising Model varies smoothly with input correlation α and neural reliability β in a way that can be easily interpreted (see captions). We reproduce their results below for binary (left) and Gaussian (right) stimulus ensembles. We extended the analysis by asking two questions:

1. How does minimizing the MMSE in estimating **h** change the $J_{12}^*(\alpha,\beta)$ landscape?
2. Can we find similar results using a different model?

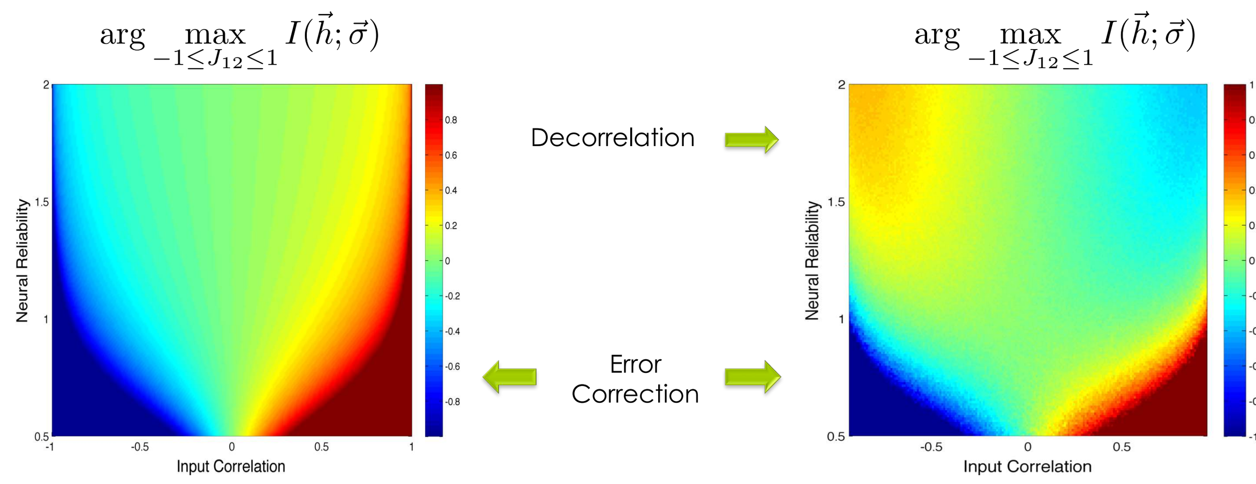


Figure 5: For a binary stimulus ensemble in which each h_i takes on a value from {+1,-1}, the optimal coupling J_{12}^* is plotted for different α and β. When there is a lot of noise, i.e. β is small, the optimal coupling takes on the same sign as the input correlation to correct for errors. This tendency gets weaker as neural reliability increases.

Figure 6: The same plot of $J_{12}^*(\alpha,\beta)$ but for a Gaussian stimulus ensemble. As in the binary ensemble, the optimal network performs error correction at low β, but at high β, J_{12}^* takes on the opposite sign of α, which can be interpreted as decorrelating the inputs.

Minimizing MMSE in Ising Model

With a continuous input, binary output model like the Ising Model, it seems like gaining mathematical information about a stimulus and being able to estimate it well are very similar tasks. Figure 7 has the exact same setup as Figure 6, except MMSE is minimized instead of MI being maximized. Qualitatively, the results are similar.

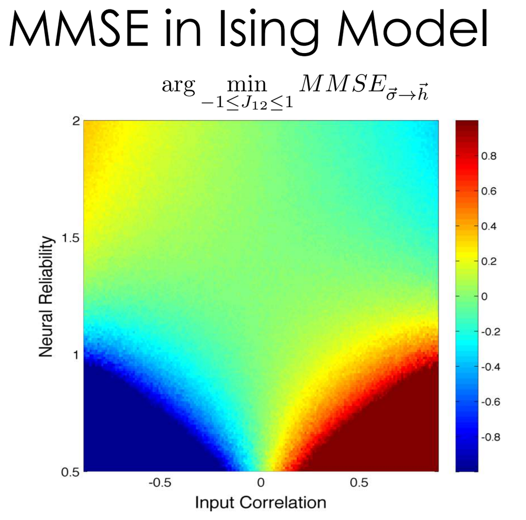


Figure 7: Same setup as Figure 6, but the optimal J_{12} is now the one that minimizes the MMSE in estimating **h**.

Analytical Results

Directly comparing the MI and MMSE functionals for the special case of Gaussian joint distributions, I found

$$I = \frac{1}{2}\log\left(\frac{1}{m}\right) \qquad m \equiv \frac{MMSE_{y\to x}}{var(x)}$$

where m is the *normalized MMSE*; that is, the MMSE divided by the marginal variance of the variable you want to estimate. This result holds as long as X is univariate, even if Y is multivariate, and it gels with our intuition, because the MI is a monotonically decreasing function of the normalized MMSE: so more MI (good) means less MMSE (good), and less MI (bad) means more MMSE (bad). I also have a general relationship for the case where the stimulus X is multivariate:

$$I = \frac{1}{2}\log\left(\frac{|\Sigma_{xx}|}{|\Sigma_{xx}-\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}|}\right) \qquad m = \frac{tr\{\Sigma_{xx}-\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\}}{tr\{\Sigma_{xx}\}}$$

$$\Sigma_{xx} \equiv E\{\vec{X}\vec{X}^T\} \quad \Sigma_{yy} \equiv E\{\vec{Y}\vec{Y}^T\} \quad \Sigma_{yx}^T = \Sigma_{xy} \equiv E\{\vec{X}\vec{Y}^T\}$$

Gaussian Channel

If we take the tuning curves to be the identity mapping and

$$\Sigma_s = \begin{pmatrix}1 & \alpha\\ \alpha & 1\end{pmatrix}, \quad \Sigma_N = \begin{pmatrix}1 & K\\ K & 1\end{pmatrix}$$

then we get the following results for optimizing the MI and MMSE with respect to K, the noise correlation.

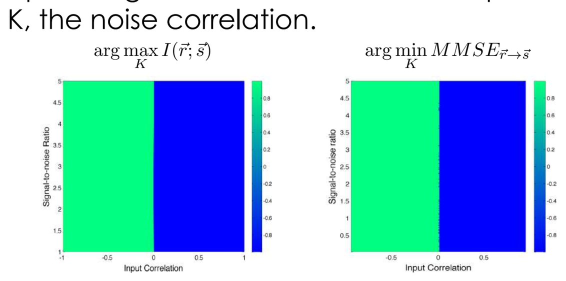


Figure 8: Side-by-side comparison of the two optimization problems we can make out of the Gaussian Channel without tuning curves using MI and MMSE. While both turn out to be trivial, it is still worth noting that *they are trivial in the same way*.

With Tuning Curves

Suppose the Gaussian-distributed stimulus designates the mean neural response by some nonlinear "tuning curve":

$$\vec{r} = \sqrt{\gamma}\begin{bmatrix}\cos(s_1)\\ \cos(s_2-\Delta)\end{bmatrix} + \vec{N}$$

Here cosines are used, which is common. If we use the tuning curve separation Δ as our definition of neural coupling, we get the following results, which are by themselves are difficult to interpret but are an important counterexample to the trend I've shown up until now:

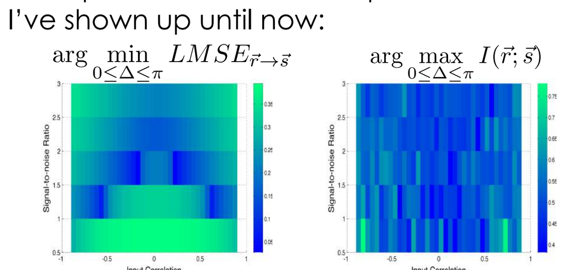


Figure 9: Side-by-side comparison of the optimal Δ (normalized by π) for the MI and mean square error associated with the optimal *linear* estimator. Even qualitatively they have little resemblance.

Conclusions/Future Research Directions

We conclude that, despite Thomson and Kristan's findings that the MI puts no bound on the MMSE for continuous-valued stimuli,[4] there is still a strong relationship between the two worth exploring, especially if the MMSE is normalized. There are a lot of directions we could go in next, but I think it would be particularly interesting to apply Thomson and Kristan's method to the MI and normalized MMSE, to see if perhaps the MI puts an upper bound on the normalized MMSE.

Measures

Mutual Information

The mutual information is a symmetric functional that maps a joint probability distribution to the positive real numbers:

$$I(X,Y) = E\left\{\log\frac{P(X,Y)}{P(X)P(Y)}\right\}_{X,Y}$$

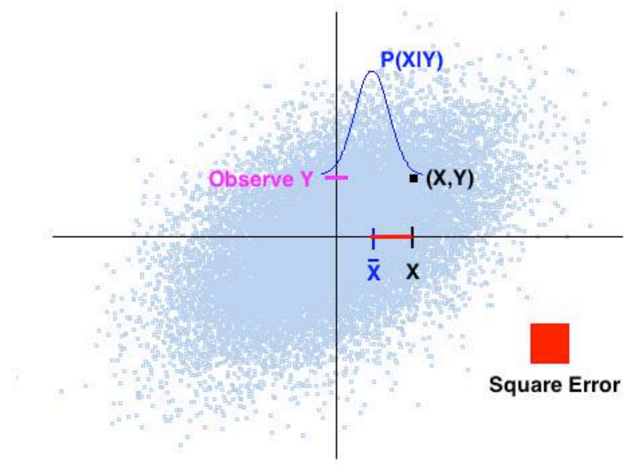
The mutual information gives the average reduction in "uncertainty" about one random variable when another random variable is known. "Uncertainty" means the average number of bits of information needed to report a list of outcomes of that random variable, if you're being as clever as possible in your coding scheme. One of the biggest questions in information-theoretic neuroscience is whether or not neural codes are optimal.

Minimum Mean Square Error

The minimum mean square error is, not surprisingly, the smallest possible mean square error you can achieve on average in estimating one random variable given another. It is simple to prove that the best possible guess you can make is the expectation of the posterior. Therefore the MMSE is written as

$$MMSE_{Y\to X} = E\left\{||\vec{X} - E\{\vec{X}|\vec{Y}\}||^2\right\}$$

Figure 4: Some point (X,Y) is picked from a joint distribution, and the true value of Y is revealed to the ideal observer, from which she must guess the value of X. If she wants to minimize the mean square error of her guesses averaged over all pairs (X,Y), she must pick the mean of the posterior.



References

1. Tkačik, G., Prentice, J. S., Balasubramanian, V., Schneidman, E. (2010) Optimal population coding by noisy spiking neurons. *Proc Natl Acad Sci USA* 14419-14424
2. Weber F., Eichner H., Cuntz H., Borst A. (2008) Eigenanalysis of a neural network for optic flow processing. *New Journal of Physics* **10** 015013
3. Sharpee T., Bialek W. (2007) Neural Decision Boundaries for Maximal Information Transmission. *PLoS ONE* 2(7): e646. doi:10.1371/ journal.pone.0000646
4. Thomson, E., Kristan, W. (2004) Quantifying Stimulus Discriminability: A Comparison of Information Theory and Ideal Observer Analysis. *Review*
5. Salinas A., Abbot L.F. (1994) Vector Reconstruction from Firing Rates. *Journal of Computational Neuroscience* 1, 89–107
6. Fitzgerald, J.D., Sharpee, T.O. (2009) Maximally informative pairwise interactions in network. *Physical Reivew* E **80,** 031914

Acknowledgements

I am deeply indebted to the Mathematical Biosciences Institute, and especially to Dennis Pearl, for their thoughtful organization and generous funding of this program; to the NSF for the grant covering my work; to my research advisor Krešimir Josićfor all the time he spent patiently working with me every day; and to a number of UH graduate students—in particular Manuel Lopez, Manisha Bhardwaj, and James Trousdale—who gave way too much of their time to help me with mathematical, computational, and transportation-related problems I encountered.

Models

Ising Model

The Ising Model, borrowed from statistical physics, gives each neuron σ_i a value of +1 for firing or -1 for not firing in a small time bin. For N neurons, there are 2^N possible configurations, and the probability of each is given by

$$P(\vec{\sigma}|\vec{h}) = \frac{e^{\beta(\sum_i h_i\sigma_i + \sum_{i,j} J_{ij}\sigma_i\sigma_j)}}{\sum_{\hat{\sigma}} e^{\beta(\sum_i h_i\hat{\sigma}_i + \sum_{i,j} J_{ij}\hat{\sigma}_i\hat{\sigma}_j)}}$$

where the h_i's, which represent the stimulus, give the tendency for each neuron to fire, and the J_{ij}'s, which represent the neural couplings, give each pair of neurons the tendency to fire together or not. These parameters (indirectly) designate the mean firing rates and pairwise correlations. The Ising Model is the maximum entropy probability distribution on binary variables for fixed values of these parameters, which is why it is used: mean firing rates and pairwise correlations are easy to measure empirically.[1] There is a third parameter β, prescribing the neural reliability, or the inverse temperature in physics.

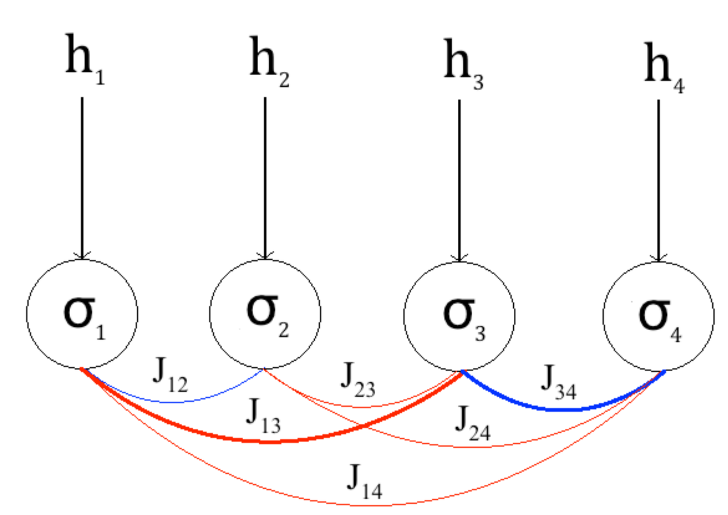


Figure 2: Diagram of a four-neuron Ising Model, with some arbitrary input **h** and coupling matrix **J** given by the color (blue is excitatory/positive, red inhibitory/negative) and thickness of the connecting curves:

$$J = \begin{pmatrix}0 & 1 & -2 & -1\\ 1 & 0 & -1 & -1\\ -2 & -1 & 0 & 2\\ -1 & -1 & 2 & 0\end{pmatrix}$$

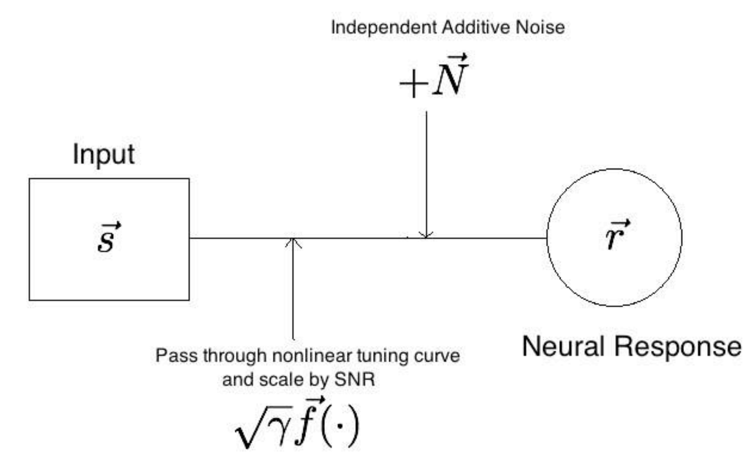


Figure 3: A diagram of the Gaussian Channel. The input **s**, which we take to have zero mean and some covariance structure, passes through some generally nonlinear tuning curve before being scaled by the square-root of the signal-to-noise ratio γ and subjected to Additive White Gaussian Noise, which also has zero mean and some covariance structure.

Gaussian Channel

The Gaussian Channel, borrowed from information theory, models a vector of neurons giving a continuous-valued response to some vector of input. The mean of the conditional response is given by the tuning curve scaled by the square root of the SNR, and the shape of the conditional response entirely determined by the covariance structure of the noise.

$$\vec{r} = \sqrt{\gamma}\vec{f}(\vec{s}) + \vec{N} \qquad \vec{s} \sim \mathcal{N}(\vec{0},\Sigma_s) \qquad \vec{N} \sim \mathcal{N}(\vec{0},\Sigma_N)$$

The neurons' output can be any real number, which might seem odd: don't neurons encode information in spikes and bursts? Most of the time, yes, but in the fly visual system, network activity is best modeled as the mostly passive diffusion of membrane potentials across various neurons' axonal and dendritic compartments.[2] In this model, we take the *noise correlation* to be our operational definition of neural coupling, because it controls the degree to which the neural responses correlate independent of the input correlations. Parameters defining the tuning curves can also be used as definitions of neural coupling. The input correlations are given by the input covariance matrix, and the neural reliability is given by the signal-to-noise ratio γ.