Supporting Information

December 1, 2007

1 Minimum information in k^{th} order statistics

In the analysis of neural codes, we shall be specifically interested in the case where the known expected values are k^{th} order marginals of the true distribution $p(r|s)$. As we show below, such marginals are expected values of some function $\vec{\phi}(r)$, so that we can use the MinMI formulation. To see this, consider the case where the response is made up of N neurons, $R = R_1, \ldots, R_N$. Define the following set of functions $\vec{\phi}(r)$

$$\vec{\phi}_{i_1,\ldots,i_k,q_1,\ldots,q_k}(r_1,\ldots,r_N) \equiv \prod_{j=1}^{k} \delta(r_{i_j} - q_j) \ , \tag{1}$$

where $\delta(x)$ is Kronecker's delta function, $i_1, \ldots, i_k$ vary over all subsets of $\{1, \ldots, N\}$ of size k, and each of the q_i variables assumes all possible values for R_i. The expected value of this $\vec{\phi}(r)$ function can be seen to be the set of conditional $k^{(th)}$ order marginals,

$$\langle \vec{\phi}_{i_1,\ldots,i_k,q_1,\ldots,q_k}(r_1,\ldots,r_N) \rangle_{\hat{p}(r|s)} = \hat{p}(R_{i_1} = q_1, \ldots, R_{i_k} = q_k|s) \ . \tag{2}$$

The minimum information obtained when using the set of functions defined above will be denoted by $I^{(k)}$. The information-minimizing distribution in this case can be written in the following compact form

$$\hat{p}_{MI}(r|s) = \hat{p}_{MI}(r) e^{\gamma(s) + \sum_{i_1,\ldots,i_k} \psi(s, r_{i_1},\ldots,r_{i_k})} \ , \tag{3}$$

where, as before, the indices $i_1, \ldots, i_k$ correspond to all possible size k subsets of $\{1, \ldots, N\}$, and $\gamma(s)$ is a normalization factor. The notation $\psi(s, r_{i_1}, \ldots, r_{i_k})$ here is defined as

$$\psi(s, r_{i_1}, \ldots, r_{i_k}) = \psi_{i_1,\ldots,i_k,r_{i_1},\ldots,r_{i_k}}(s) \ . \tag{4}$$

In what follows, we shall specifically demonstrate how $I^{(1)}$ and $I^{(2)}$ may be used to study aspects of neural coding. The distribution in the $I^{(1)}$ case has the following form

$$\hat{p}_{MI}(r|s) = \hat{p}_{MI}(r) e^{\gamma(s) + \sum_{i=1}^{N} \psi(s, r_i)} \ , \tag{5}$$

where $\gamma(s)$ is a normalization factor. In the $I^{(2)}$ case, the distribution is given by

$$\hat{p}_{MI}(r|s) = \hat{p}_{MI}(r)e^{\gamma(s)+\sum_{1\leq i<j\leq N}\psi(s,r_i,r_j)} \ , \tag{6}$$

where the sum if over all distinct pairs of neurons. The above distributions bear some similarity to maximum entropy distribution, with some important differences, as explained in the next section.

2 Relation to Maximum Entropy Modeling

The MinMI method takes as input a set of expectation values and outputs the minimum possible information $I(S;R)$ subject to these expected values, and the distribution that achieves this minimum. A different approach to handling expected values is to seek the distribution that maximizes the entropy $H(S,R)$. The motivation for seeking this distribution is that it maximizes disorder in the population, given the expected values, and thus makes no additional modeling assumptions. This approach was first advocated by Jaynes [6], and has recently been introduced to neural code analysis in [7] and [12]. To illustrate this approach, let us focus on the case where the given expected values are responses of single neurons. Simple calculation then shows that the entropy maximizing distribution is the conditionally independent distribution

$$\hat{p}_{ME}(r|s) = \prod_{i=1}^{N} p(r_i|s) \ , \tag{7}$$

where $p(r_i|s)$ are the given first order statistics. This distribution is similar to the MinMI distribution in Equation 5 above in that it contains a product of independent factors, one for each neuron. However, the MaxEnt distribution does not contain the factor $\hat{p}_{MI}(r)$ and is thus mathematically very different from the MinMI one. One example of this difference is that the distribution $\hat{p}_{MI}(r)$ may contain zeros (see examples in the main text), whereas the MaxEnt distribution is never zero. As stated in the main text, this difference will result in MaxEnt saturating the information value as $N \to \infty$, whereas MinMI will not have this property.

Similarly, for second order statistics the MaxEnt distribution is

$$\hat{p}_{ME}(r|s) = e^{\gamma(s)+\sum_{1\leq i<j\leq N}\psi(s,r_i,r_j)} \ , \tag{8}$$

where $\gamma(s)$ is a normalization constants, and $\psi(s,r_i,r_j)$ are chosen to fit the given expected values. This distribution is again similar to the MinMI one in Equation 6, but without the $\hat{p}_{MI}(r)$ factor.

3 Algorithm Convergence Proof

We now prove the convergence of the information minimizing algorithm presented in the main text. Since $\hat{p}_{t+1}(r|s)$ is the *I-projection* of $\hat{p}_t(r)$ on $\mathcal{F}(\vec{\phi}(r),\vec{a}(s))$

and $\hat{p}_t(r|s)$ is also in $\mathcal{F}(\vec{\phi}(r), \vec{a}(s))$ by the definition of the previous iteration, we have by the Pythagorean equality for *I-projections* on expectation constraints [4] that

$$D_{KL}[\hat{p}_t(r|s)|\hat{p}_t(r)] = D_{KL}[\hat{p}_t(r|s)|\hat{p}_{t+1}(r|s)] + D_{KL}[\hat{p}_{t+1}(r|s)|\hat{p}_t(r)] \ .$$

Averaging the above over $p(s)$ and rearranging

$$I_{\hat{p}_t}(S;R) = \langle D_{KL}[\hat{p}_t(r|s)|\hat{p}_{t+1}(r|s)]\rangle_{p(s)} + I_{\hat{p}_{t+1}}(S;R) + D_{KL}[\hat{p}_{t+1}(r)|\hat{p}_t(r)] \ .$$

The right hand side has $I_{\hat{p}_{t+1}}(S;R)$ plus some positive quantity. We can thus conclude that

$$I_{\hat{p}_{t+1}}(S;R) \leq I_{\hat{p}_t}(S;R) \tag{9}$$

and therefore the algorithm reduces the mutual information in each iteration.

To see that the algorithm indeed converges to the minimum, note that since $I_{\hat{p}_t}(S;R)$ is a monotone, lower bounded, decreasing series, its difference series converges to zero as t goes to infinity

$$\langle D_{KL}[\hat{p}_t(r|s)|\hat{p}_{t+1}(r|s)]\rangle_{p(s)} + D_{KL}[\hat{p}_{t+1}(r)|\hat{p}_t(r)] \to 0$$

The above expression is zero if and only if $\hat{p}_t(r|s) = \hat{p}_{t+1}(r|s)$ and $\hat{p}_t(r) = \hat{p}_{t+1}(r)$ This implies that at the limit $\hat{p}_\infty(r)$ and $\hat{p}_\infty(r|s)$ are invariant to the MinMI iterations. Thus $\hat{p}_\infty(r|s)$ is the *I-projection* of $\hat{p}_\infty(r)$ and therefore these two distributions satisfy Equation 4 in the main text, which characterizes the information minimizing distribution that we sought to calculate.

4 Calculating the CI information $I_{CI}(S;R)$

The conditionally independent distribution is given by $p_{CI}(r|s) \equiv \prod_{i=1}^{n} p(r_i|s)$, and the stimulus prior $p(s)$ is assumed to be known. The mutual information under this distribution can be expressed as

$$I_{CI}(S;R) = H_{CI}(R) - H_{CI}(R|S) \tag{10}$$

The term $H_{CI}(R|S)$ can be easily calculated since it decomposes into $H_{CI}(R|S) = \sum_i H_{CI}(R_i|S)$. However $H_{CI}(R)$ is more complicated since it requires the evaluation of the distribution

$$p(r) = \sum_s \prod_{i=1}^{n} p(r_i|s)p(s) \ . \tag{11}$$

When n_r is large, one cannot simply calculate the sum $H(R) = -\sum p(r)\log p(r)$. To approximate this sum, we use a sampling approach [1]. We draw a sample of T points from the distribution $p(r)$, and denote this sample by $r^{(1)}, \ldots, r^{(T)}$ (each $r^{(i)}$ consists of an assignment to the variable $r_1, \ldots, r_n$), and calculate the average: $-\frac{1}{T}\sum_i \log p(r_i)$. This sum will approach $H(R)$ as the sample size grows. In practice, we keep increasing the sample until the sum converges.

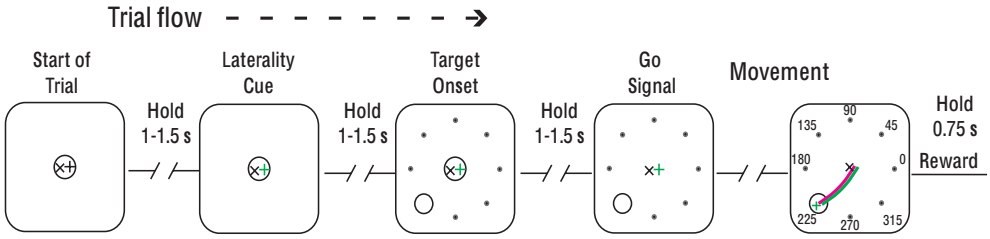

Figure 1: The center out movement task performed by behaving monkeys. The monkey is initially presented with a laterality cue, followed by the target location, and a go signal. These three signals are separated by hold periods of $1 - 1.5$ seconds.

5 Experimental Procedures and Data Analysis

In this section we provide details about the data used in the *temporal coding* section in the main text.

5.1 Experimental Procedures

The data was obtained from experiments studying motor control in monkeys. Monkeys were trained to perform unimanual movements by operating two X-Y manipulanda. The movements were standard center-out reaching tasks with eight movement directions [5], and a delayed GO signal [1]. For comprehensive experimental details, see [11].

In each trial, the monkey was first presented with a signal indicating which hand will perform the movement (*Laterality Cue*). After a hold period ($1000 - 1500$ ms), a signal (*Target Onset*) was given indicating the direction to move to (one of eight). This was followed by an additional hold period ($1000 - 1500$ ms). During the hold periods the monkey was instructed to hold the cursor in a circle located at the center of the screen. After the second hold period, this circle disappeared (*Go Signal*), and the monkey was instructed to move to the specified target. The trial flow is illustrated in Figure 1.

During task performance, single-unit activity was recorded extra-cellularly from the primary and supplementary motor areas of the cortex. The data presented here is taken from primary motor cortex recordings. The number of successful repetitions per movement direction varied in the range $5 - 20$, resulting in $40 - 160$ repetitions for each laterality cue.

5.2 Quantization and Bias Correction

Spike trains were represented using the total number of spikes (i.e. spike count) in windows of different sizes (depending on the application), with a maximum

[1] The complete behavioral paradigm included other components such as learning visuomotor transformations, but these are not relevant for our analysis.

size of 600 ms. However, due to the low number of repetitions, we could not estimate $p(r_i|y)$ reliably when r_i was the raw spike count. To clarify why, consider for example a case where the spike counts vary in the range $0-20$ spikes, and the number of repetitions per stimulus is 40. Then each spike count will be repeated two times on average, precluding reliable estimate of its appearance probability.

To overcome this difficulty, we choose to quantize the spike counts into a small number of bins. There are several possible binning strategies, for example uniform (i.e. $1, 2, 3, 4, \ldots, 19, 20$) , equal number of samples per bin, log scale etc. We chose a greedy scheme, where one searches for the binning which maximizes the mutual information. Because the number of possible binning schemes is exponential in the maximum spike count, we used a suboptimal search algorithm which unified neighboring bins that increased information. In calculating information we applied the Panzeri-Treves bias correction term ([10]; see also [9] for a comprehensive review of bias estimation).[2] This partly compensated for the increase of information with the number of bins. A similar quantization scheme was recently used in [8].

The above quantization algorithm was used to determine the bin sizes to be used in calculating the observed marginals, and the associated minimum information values. As in any MI estimation from finite samples, the MI statistic has a positive bias.[3]. To overcome this bias, we used a bootstrap scheme [13]: the entire information calculation (including quantization) was repeated N times, where each repetition was on a new sample which was generated by drawing trials uniformly with repetitions from the original set of trials. This resulted in N bootstrapped information values $I(1), \ldots, I(N)$. An estimate of the bias was then given by $I(0) - \frac{1}{N}\sum_{k=1}^{N} I(k)$ where $I(0)$ is the information calculated from the original sample. This resulted in an effective removal of the bias term, as verified in simulation experiments.

6 Relation to the Data Processing Inequality

The current section illustrates the relation between MinMI and the Data Processing Inequality [3], and shows that the latter in fact constitutes a specific example of the MinMI principle. The Data Processing Inequality states that for any function $f(R)$ of the response variable R, the information $I(S; f(R))$ is a lower bound on $I(S; R)$. Information estimation based on this fact proceeds by choosing some function $f(R)$ (e.g., the total spike count in a spike train) and using $I(S; f(R))$ as a lower bound on the true $I(S; R)$ (e.g., see [2]). In order for the method to be practical, the function $f(R)$ must take a

[2] Although resampling could also be used to remove bias here, we preferred the Panzeri-Treves method since it is faster, and since for the quantization we were not interested in the exact mutual information value.

[3] To see why, consider a distribution with zero MI Since information is non-negative, any estimation from finite samples will generate a non-negative value, and therefore the mean estimate will be strictly positive, yielding a biased estimate.

relatively small number of values so that $p(s, f(r))$ can be reliably estimated. We next show that the distribution $p(s, f(r))$ can be interpreted as a set of expected values, and that the minimum information subject to these expected values is exactly $I(S; f(R))$. Estimating $p(s, f(r))$ is equivalent to estimating the expected values of the functions $\phi_k(r) = \delta(f(r) - k)$ (where $r = 1, \ldots, n_r$, $k = 1, \ldots, n_f$, and n_f is the number of different values $f(r)$ may take), since $\langle \phi_k(r) \rangle_{p(r|s)} = p(f(r) = k|s)$. The following proposition states that the minimum information subject to constraints on the expected values of $\phi_k(r)$ is in fact $I(S; f(R))$.

Proposition 1 *$I_{min}\left[\vec{\phi}(r), p(f(r)|s)\right] = I_p(S; f(R))$, where $\vec{\phi}(r)$ is the set of functions defined above, and $p(f(r)|s)$ are their expected values.*

Proof: *Denote by n_k the number of values of r such that $f(r) = k$. Now consider the conditional distribution $q(r|s) = \frac{p(f(r)|s)}{n_{f(r)}}$, and the joint distribution $q(s, r) = p(s)q(r|s)$. Note that $q(r) = \sum_y p(s)\frac{p(f(r)|s)}{n_{f(r)}} = \frac{p(f(r))}{n_{f(r)}}$*

It is easy to see that $q(s, r) \in \mathcal{P}\left(\vec{\phi}(r), p(f(r)|s)\right)$. In other words, $q(s, r)$ has the desired expected values of $\vec{\phi}(r)$, and is thus in the set we are minimizing over. Next, observe that the information $I_q(S; R)$ is equal to $I_p(S; f(R))$ since

$$\begin{aligned} I_q(S; R) &= \sum_s p(s) \sum_r q(r|s) \log \frac{q(r|s)}{q(r)} \\ &= \sum_s p(s) \sum_k n_k \frac{p(f(r) = k|s)}{n_k} \log \frac{p(f(r) = k|s)}{p(f(r) = k)} = I_p(S; f(R)) \end{aligned}$$

Finally, the Data Processing Inequality guarantees that $I_p(S; f(R)) \leq I_w(S; R)$ for any distribution $w(s, r) \in \mathcal{P}\left(\vec{\phi}(r), p(f(r)|s)\right)$ (since any such w will reduce to the distribution $p(s, f(r))$ over the reduced alphabet $f(r)$), and thus $I_p(S; f(R)) \leq I_{min}$. Since we have found a distribution $q(s, r)$ in $\mathcal{P}\left(\vec{\phi}(r), p(f(r)|s)\right)$ where this is an equality, it must be that $I_{min} = I_p(S; f(R))$.

Thus, calculating a lower bound on the information via the Data Processing Inequality is equivalent to the MinMI bound obtained from the statistics of the quantized response.

References

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

[2] A. Borst and F.E. Theunissen. Information theory and neural coding. *Nature Neuroscience*, 2(11):947–957, 1999.

[3] T.M. Cover and J.A Thomas. *Elements of information theory*. Wiley, 1991.

[4] I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, 1975.

[5] A.P. Georgopoulos, R.E. Kettner, and A.B. Schwartz. Primate motor cortex and free arm movements to visual targets in three dimensional space. i. relations between single cell discharge and direction of movement. *J. Neuroscience*, 8:2913–2927, 1988.

[6] E.T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.

[7] L. Martignon, G. Deco, K.B. Laskey, M.E. Diamond, W. Freiwald, and E. Vaadia. Neural coding: Higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Computation.*, 12(11):2621–2653, 2000.

[8] I. Nelken, G. Chechik, T.D. Mrsic-Flogel, A.J. King, and J.W.H. Schupp. Encoding stimulus information by spike numbers and mean response time in primary auditory cortex. *J. Comp. Neurosci.*, 19:199–221, 2005.

[9] L. Paninski. Estimation of entropy and mutual information. *Neural Computation.*, 15:1191–1254, 2003.

[10] S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7:87–107, 1996.

[11] R. Paz, T. Boraud, C. Natan, H. Bergman, and E. Vaadia. Preparatory activity in motor cortex reflects learning of local visuomotor skills. *Nature Neuroscience*, 6:882–890, 2003.

[12] E. Schneidman, M.J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440:1007–1012, 2006.

[13] E. Stark and M. Abeles. Applying resampling methods to neurophysiological data. *J. Neuroscience Methods*, 145(1):133–144, 2005.