Bits from brains for biologically inspired computing

Michael Wibral, Joseph T. Lizier, Viola Priesemann

(Presented by Sten Sootla)
Why information theory in neuroscience?

- Marr’s 3 levels of information processing: task, algorithmic and implementation levels.
- Shortcoming: results obtained at any of the levels does not constrain the possibilities at any other level.
- Missing relationships between Marr’s levels can be filled by information theory:
  1. implementation <-> task
  2. implementation <-> algorithmic
Motivating example: how much the response of a neuron varies across stimuli?
Your friend Bob can only speak 4 words.

He lives in Australia and you’d like to talk to him using as few bits as possible, because every bit costs money.

You design a **fixed-length** code:
Entropy

- However, you notice that Bob doesn’t use every word equally often:

- So you design a clever variable-length code that uses this information:
Entropy

- Since you now use a variable length code, the codewords can’t start with a common prefix. Otherwise it would be impossible to decipher the code.
- So by choosing a code, you make a sacrifice from the space of all possible codewords.

- Choosing the code “01”, you sacrifice $\frac{1}{4}$ of all possible codes:

<table>
<thead>
<tr>
<th></th>
<th>bit 1</th>
<th>bit 2</th>
<th>bit 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$\frac{1}{2^L} = \frac{1}{4}$
Entropy

- The **optimal** way to encode the information is to distribute our “budget” in proportion to how common an event is.
- The average message length using the optimal code is called **entropy**:
  \[
  H(R) = \sum_{r \in R} p(r) \log_2 \frac{1}{p(r)}.
  \]
- Another interpretation: the **amount of uncertainty** one has about a random variable.
Conditional entropy

- Neurons are noisy - their responses to repetitions of identical stimulus differ across trials.
- To quantify the noise, we use **conditional entropy**:

$$H(R|S) = \sum_{s \in S} p(s) \sum_{r \in R} p(r|s) \log_2 \frac{1}{p(r|s)}.$$  

- The noisier the neuron, the greater the $H(R|S)$. 
Mutual information

- How much of the information capacity in neural activity is robust to noise?
- To quantify it, we use **mutual information**:

\[
I(S:R) = \sum_{s,r} p(s,r) \log_2 \frac{p(s,r)}{p(s)p(r)} = H(R) - H(R|S).
\]

- It’s the **reduction in the uncertainty** of \( R \) due to the knowledge of \( S \).
Kullback-Leibler distance

- It measures the “distance” between 2 distributions:

\[ D_{KL}(p||q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)}. \]

- A measure of inefficiency of assuming that distribution is \( q \) when the true distribution is \( p \).
- Mutual information is just a **special case** of this: the KL-distance between the joint distribution \( p(x,y) \) and the product distribution \( p(x)p(y) \).
Local information theoretic quantities

- We have looked at **average** information content of random variables.
- We can also study **local** information theoretic quantities, which allow us to quantify information in a **single realization** of a random variable.
- Localized measures are trivially obtained by omitting the expectation over the whole distribution.
- For example:

\[ i(s : r) = \log_2 \frac{p(s, r)}{p(s)p(r)} \]
Analyzing neural codings

Crossing the bridge between the task level and implementation level.
Which neural responses carry information about which stimuli?

- Can be easily answered by computing $I(S:R)$.

- Example: we could extract features $F_i(R)$ from neural responses: time of the first spike and firing rate, and calculate $I(S:F_1(R))$ and $I(S:F_2(R))$. 
How much does an observer of specific neural response change its beliefs about the identity of stimulus from $p(s)$ to $p(s|r)$?

- Kullback-Leibler distance between $p(s)$ and $p(s|r)$.
- **Specific surprise:**
  
  \[ i_{sp}(S : r) = \sum_{s \in S} p(s|r) \log_2 \frac{p(s|r)}{p(s)} . \]

- $i_{sp}$ is a valid partition of mutual information into more specific, response dependent contributions, since:
  
  \[ I(S : R) = \sum_{r \in R} p(r) i_{sp}(S : r) . \]
Motivating example

1. How to quantify the *reduction in uncertainty* about the stimulus gained by a particular response $r$?
2. Which stimulus is reliably associated with the responses that are relatively unique for that stimulus?
How to quantify the reduction in uncertainty about the stimulus gained by a particular response \( r \)?

● “In contrast to the previous question, here we ask whether the response increases or reduces uncertainty about the stimulus.”

● **Response-specific information:**

\[
i_r(S : r) = H(S) - H(S| r).
\]

● \( i_r \) is a valid partition of mutual-information:

\[
I(S : R) = \sum_{r \in R} p(r)i_r(S : r).
\]
Which stimulus leads to responses that are informative about the stimulus itself?

- Response is informative if it has a large $i_r(S:r)$.
- We then ask how informative the responses for a given stimulus are on average over all responses that the stimulus elicits with probabilities $p(r|s)$.
- We obtain the stimulus specific information:

\[
i_{SSI}(s:R) = \sum_{r \in R} p(r|s)i_r(S:r).
\]

- $i_{SSI}$ is a valid partition mutual information:

\[
l(S:R) = \sum_{s \in S} p(s)i_{SSI}(s:R).
\]
Specific example

- Two competing interpretations:
  1. The stimuli that evoke the **highest firing rates** are most important to the neuron.
  2. Nearby stimuli are most easily discriminated in **high-slope regions** of the tuning curve.

- Both interpretations are correct, depending on the amount of neuronal variability (noise).

(Tuning curves, Neuronal Variability and Sensory Coding. 2006. Daniel A. Butts, Mark. S. Goldman.)
Relevance to BICS: “Encoding of an environment in a computing system may be modeled on that of a neural system that successfully lives in the same environment.”
Motivating example: How is information about a stimulus distributed in the brain between the neurons?
Intuition

- \( H(S) = 2 \)
- \( H(R_i) = 1 \)
- \( I(S;R_1,R_2,R_3) = 2 \)
- \( I(S;R_i) = 1 \) (!)
- **Redundancy** - neurons 1,2 show identical responses
- **Synergy** - how much information neurons 1,3 have about a single short bar?
- **Unique information** - consider neurons 1,3 and stimuli 1,3
Mathematical formulation

Let’s consider input variables $R_1$ and $R_2$ and an output variable $S$. Then we can decompose:

$$I(S : R_1) = SI(S : R_1; R_2) + UI(S : R_1 \setminus R_2),$$

$$I(S : R_2) = SI(S : R_2; R_1) + UI(S : R_2 \setminus R_1),$$

$$I(S : R_1, R_2) = UI(S : R_1 \setminus R_2) + UI(S : R_2 \setminus R_1) + SI(S : R_1; R_2) + CI(S : R_1; R_2).$$

Only one of the terms needs to be defined!
Proposed axioms for redundant information

- **Symmetry**: the redundant information that variables $R_1, R_2, \ldots, R_n$ have about $S$ is symmetric under permutations of $R_1, R_2, \ldots, R_n$.

- **Self-redundancy**: the redundant information that $R_1$ shares with itself about $S$ is the mutual information $I(S:R1)$.

- **Monotonicity**: the redundant information that variables $R_1, R_2, \ldots, R_n$ have about $S$ is smaller than or equal to the redundant information that variables $R_1, R_2, \ldots, R_{n-1}$ have about $S$. 
Mathematical formulation

- Additional assumption: **unique information** should depend only on the **marginal distributions** \( P(s,r_1) \) and \( P(s,r_2) \).
- So we can define:

\[
\tilde{U}I(S:R_1 \setminus R_2) = \min_{Q \in \Delta_p} I_Q(S:R_1|R_2),
\]

where \( \Delta_p = \{ Q \in \Delta : Q(S = s, R_1 = r_1) = P(S = s, R_1 = r_1) \) and \( Q(S = s, R_2 = r_2) = P(S = s, R_2 = r_2) \} \forall s \in S, r_1 \in R_1, r_2 \in R_2 \).
Mathematical formulation

- Redundant information:

\[
\tilde{I}(S : R_1; R_2) = \max_{Q \in \Delta_p} [I(S : R_1) - I_Q(S : R_1 | R_2)].
\]

- Shared information:

\[
\tilde{C}(S : R_1; R_2) = I(S : R_1, R_2) - \min_{Q \in \Delta_p} I_Q(S : R_1, R_2).
\]

- All these measures can be found by convex optimization, and they are **positive**.
Relevance to BICS: “Knowing how information is distributed over the agents can inform the designer of BICS about strategies to distribute the relevant information about a problem over the available agents.”

Example: reliability vs capacity.
Analyzing distributed computation in neural systems

Crossing the bridge between the implementation level and algorithmic level.
Motivation

- If we probe the system beyond early sensory or motor areas, we have little knowledge of what is actually encoded by the neurons in deeper inside the system.
- The gap between the task- and implementation level may become too wide.
- We may use information theory to link the implementation and algorithmic level, by retrieving a “footprint” of the information processing.
Partitioning of information processing

- Information **transfer** - how much information is transferred from source process to target process.
- Information **storage** - how much information in a process is predictable from its past.
- Information **modification** - quantifies the combination of information from various source processes into a new form that is not trivially predictable from any subset of these source processes.
State space reconstruction

- An optimal prediction of future realizations of a process typically requires looking at many past realizations of random variables of this process (pendulum example).
- A vector of past realizations, such that they’re sufficient for prediction is a state of the system.
- Formally, we have to form the smallest collection of variables \( X_t = (X_t, X_{t_1}, \ldots, X_{t_i}, \ldots) \) with \( t_i < t \) that jointly make \( X_{t+1} \) conditionally independent of all \( X_{t_k} \) with \( t_k < \min(t_i) \):

\[
p(x_{t+1}|x_{t_k}, x_t) = p(x_{t+1}|x_t).
\]
1. Information transfer

- Information transfer from source process $X$ to a target process $Y$:

$$TE(X_{t-u} \rightarrow Y_t) = I(X_{t-u} : Y_t | Y_{t-1}) = H(Y_t | Y_{t-1}) - H(Y_t | Y_{t-1}, X_{t-u})$$

- Interaction delay parameter need not be chosen \textit{ad hoc}:

$$\delta = \arg\max_u [TE(X_{t-u} \rightarrow Y_t)]$$

- \textbf{Not} optimal for inference about \textit{causal} interactions.
Problems with transfer entropy

In reality, we have more than 2 interacting processes.

1. **Common driver effect:**
   \[ \delta_{Z \rightarrow X} < \delta_{Z \rightarrow Y} \implies TE(X_{t-u} \rightarrow Y_t) > 0 \]

2. **Cascade effect:** if information if transferred from X to Y and then from Y to Z, bivariate analysis will also indicate information transfer from X to Z.

3. Two sources can transmit information **synergistically**.
2. Information storage

- We’re concerned with **active** information storage (information stored in neural activity), not passive (synaptic weights):

\[ A_{X_t} = \lim_{k \to \infty} I(X_{t-1}^k : X_t), \]

where \( X_{t}^{k-} = \{X_t, X_{t-1}, ..., X_{t-k+1}\}. \)

- The limit can be replaced by finite \( k_{\text{max}}. \)
3. Information modification

- Information modification is an interaction between transmitted and/or stored information that results in modification of one of or the other.

- **Local separable information:**

\[ s_{X_t} = a_{X_t} + \sum_{z_{t-} \in V_{X_t \backslash X_{t-1}}} i(x_t : z_{t-} | x_{t-1}) , \]

with \( V_{X_t \backslash X_{t-1}} = \{Z_{t-}, ..., Z_{t-}, G\} \) indicating the set of \( G \) past state variables of all processes \( Z_{t-},i \) that transfer information into the target variable \( X_t \).
Relevance to BICS: “These measures of how information is processed allow us to narrow in (derive constraints) on the algorithms being implemented in the neural system.”
Conclusion

- Neural system processing can be quantitatively partitioned into information -storage, -transfer and -modification. These observations allow us to derive constraints on possible neural algorithms.
- The representation that these algorithms operate on can be guessed by analyzing the neural codes.
- Care must be taken when analyzing neural codes because separation of how neurons code uniquely, redundantly, and synergistically has not been solved completely.