Robert Sneddon, Ph.D.
California Institute of Technology

Abstract:

Neural Representations are Invariant

A representation or "symbol" denotes a fundamental piece of information in a computation. In an actual computer, this information has a physical representation. For instance, in digital computers, the physical representations are the 5 volts and 0 volt outputs of a transistor – these voltages are the physical manifestations of 1s and 0s. Similarly, in neural computation, these symbols have a biophysical form. Specifically, they are manifested as spiking voltages.

This leads to the question, "exactly what sorts of spiking patterns are the manifestations of computational symbols?" That question is answered by a basic theorem which shows that neural computation on these representations must be invariant up to a constant. That is, the computation on the biophysical form of any element of the "alphabet of computational representations/symbols" must satisfy a basic functional equation. This equation is similar to an eigenfunction equation but is generalized to a non-linear system.

This theory is empirically tested on the thalamic relay neuron in the following manner. Assume that the thalamic relay neuron is performing computations. Then its outgoing spike trains must be the physical manifestations of computational symbols/representations. Therefore, the specific form of spike train outputs – outputs which represent computational symbols, can be deduced by solving the previously mentioned equation.

These predicted outgoing spike trains are a subset of all biophysically possible outputs. Thus, if the thalamic relay neuron is computing, only this set of spike train outputs will be observed in *vivo* and probably in *vitro*. The other spiking patterns, though biophysically possible, will never be observed.

Empirically, all predicted spiking patterns were observed in data and all patterns which were predicted to *not* occur, were not observed. This suggests that this theory of neural computation is an accurate one.