CellPress

Questioning the role of sparse coding in the brain

Anton Spanne and Henrik Jörntell

Neural Basis of Sensorimotor Control, Department of Experimental Medical Science, Biomedical Center F10, Tornavägen 10, 221 84 Lund, Sweden

Coding principles are central to understanding the organization of brain circuitry. Sparse coding offers several advantages, but a near-consensus has developed that it only has beneficial properties, and these are partially unique to sparse coding. We find that these advantages come at the cost of several trade-offs, with the lower capacity for generalization being especially problematic, and the value of sparse coding as a measure and its experimental support are both questionable. Furthermore, silent synapses and inhibitory interneurons can permit learning speed and memory capacity that was previously ascribed to sparse coding only. Combining these properties without exaggerated sparse coding improves the capacity for generalization and facilitates learning of models of a complex and high-dimensional reality.

The concept of sparse coding is widely applied as an interpretational framework to understand the function of the neuronal circuitry in both the cerebellum and the neocortex. The main reason for the interest in this coding scheme is that sparse coding can have beneficial features for both memory capacity [1,2] and speed of learning [3], as illustrated in various simulations of brain circuitry function [4–9]. From a simulation point of view, it provides the additional advantage of reduced computational load, and is therefore popular in technical applications that for example involve artificial neural networks [10,11]. However, two things are not clear: (i) to what extent sparse coding could also lead to disadvantages for brain function, and (ii) whether the brain actually features circuitry mechanisms that enforce sparse coding.

A major difficulty for the elucidation of its possible role in the brain is that sparse coding is a relative term and a wide range of definitions are used by authors in the field. Indeed, in the widest possible theoretical sense, there is sparse activity in a neuronal population whenever the average activation ratio remains below 50% for binary neurons or below 100% for thresholded neurons that are continuously rate-coded when active. The wide definition allows almost any form of neural representations to be classified as sparse coded in a superficial analysis. Hence, highly disparate coding schemes can be labeled 'sparse coded' even though they have nothing in common other than low levels of neuronal firing under some circumstances. The situation is further complicated because the concepts of lifetime sparseness of a single neuron and population sparseness are used interchangeably, even though the first does not imply the other [12].

The interpretation of sparse coding is often applied uncritically, and a near-consensus appears to have developed that sparse coding can only bring advantages, with some of those advantages being unique to sparse coding. In the present paper we take a deeper look into the underlying constraints and trade-offs associated with sparse coding, as well as reviewing its current state of experimental support. The most fundamental of these trade-offs is the capacity for generalization [2], which in extreme sparse coding is very low and tend to lead to overtraining – in other words, the learning of meaningless contingencies [13]. Because we also find that the experimental support for sparse coding is questionable in both the cerebellum and the neocortex, we explore the consequences of the assumption that brain circuitry in reality implements alternative principles of coding, a scenario in which there are situations where sparse coding-like properties may arise as an epiphenomenon [14].

The meaning of the concept of sparse coding and implicit trade-offs

In the reasoning below we will refer to the concept of 'context', which results in a pattern of activation in a layer of input neurons. As in many theoretical studies exploring the properties of sparse codes [1–6], the context should elicit a specific response in an output neuron that samples the input neurons. From the point of view of an output neuron, the context is defined by all the sensory and other input signals that converge on the output neuron. From the point of view of the brain, a context equals a state that takes into account the configuration of the entire body, specifically the entire population of peripheral sensors and the entire population of neurons in the brain. This latter arises because every neuron of the brain has a potential contribution to the output of the brain (in terms of muscle activation), and the state thus comprises all sensor signals and all motor signals of the brain. Therefore, two different contexts, as defined here, can never coexist in time. An individual output neuron is typically part of a larger group of output neurons that belong to the same neuronal layer, and the individual neuron will therefore contribute by solving only a small part of the input–output puzzle that

Corresponding author: Jörntell, H. (henrik.jorntell@med.lu.se).

0166-2236/

 http://dx.doi.org/10.1016/j.tins.2015.05.005

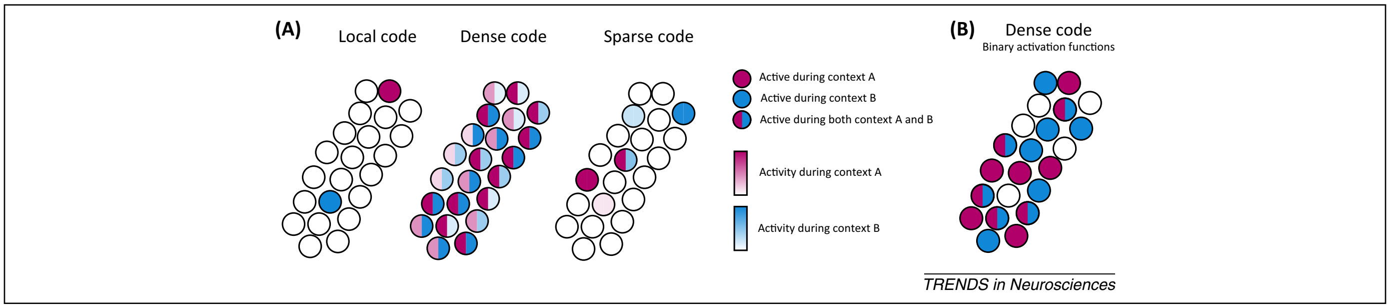

Figure 1. Sparse code is a compromise between local code and dense code. **(A)** Comparison of coding schemes that differ in their ratio of active neurons: in other words, in their sparseness. The activities within the population during two hypothetical contexts (context A and context B) are shown as examples of how different contexts are represented within the population. Note that by our definition only a single context would be active at any time because a context represents the global brain state (i.e., all the neurons). In local code, a context is represented by the activity of a single neuron, or a small subset of neurons, and different contexts are represented by different neurons. Notably, the activities of the neurons are not independent because if a neuron is responding to context A, it will not respond to any other context. In dense code, all neurons are active and their combined activity is used to encode each context. Any state in between the two extreme cases of local and dense code can in principle be labeled sparse code. The reduction of average activation leads to a reduction in the overlap or interference between the activation during different contexts. **(B)** In the special case of binary activation functions, maximal representational capacity is obtained if 50% of the neurons are active during each context. For this reason an average activation of 50% is usually considered dense code in the binary case.

the local output layer needs to solve (the brain consists of many such local output layers). In each context, or state, there is only one optimal output for each output neuron. Failure to find this optimal output will degrade the performance of the system, but may not mean that the performance breaks down completely.

An extreme case of sparseness is the so-called local code, where only one or a small set of the neurons in an input layer are active during a given context. Each neuron participates in the representation of a single context only (Figure 1A), and this eliminates any interference between representations of different contexts. Neurons with this type of specificity are commonly referred to as 'grandmother neurons' [15]. Because each context is encoded by separate neurons, an output neuron that is innervated by the layer of local coding neurons can easily learn to decode the signal. In principle, the network could already learn to respond correctly to a context after a single trial, by adjusting the weights of the active synapses only, because there is no overlap or interference of the activity between different contexts. Therefore, learning in local code can be extremely fast, at least in cases where there is also a system to supervise the learning – in other words, to provide information to the output neuron on how to respond in a given context. However, local code comes with major drawbacks, in particular a low representational capacity – the layer of input neurons can maximally encode one context per neuron [2].

The opposite of local code in terms of sparseness is a dense code where each context is represented by the combined activity of all (100%) of neurons (Figure 1A, middle). In general, a network with a dense code can encode M^N contexts, where M is the number of distinct states of the neurons, and N is the number of neurons. Hence, the number of possible encoded contexts using dense code will quickly reach astronomical values as the number of neurons increases. However, a dense code also comes with limiting drawbacks. Interference between contexts, a natural consequence because all representations can overlap, increases the possible complexity of the relationship the output neuron needs to learn, which therefore leads to a decrease in learning speed. The relationship between the speed of learning and degree of sparseness has been shown using models of the cerebellar circuitry and theoretical reasoning based on gradient descent [3] (but see the caveat discussed in 'Model complexity', below). For completeness, the special case of binary activation functions should also be considered because maximal representational capacity in this case is obtained when there is an equal probability that each neuron will be either on or off under each context (Figure 1B). Consequently, an average activity of 50% within a population of binary neurons is considered as dense.

Sparse coding can be described as a trade-off between the benefits and drawbacks of the dense and local codes [16] – for example, the speed of learning of the local code is traded against the high representational capacity allowed under dense code. Nevertheless, there are also other trade-offs between the two coding schemes. As detailed below, local code offers better memory capacity than dense code in some settings, but provides poorer fault-tolerance and no generalization. Note that because local code and dense code are opposite extremes, in principle any case in between these could be labeled sparse coding (i.e., below 50% active binary neurons or below 100% active rectified continuous neurons) (Figure 1A).

Because of potential redundancy, fault-tolerance (i.e., the capacity to handle neuronal noise, or loss of a subset of the neurons) improves as code density increases, whereas local code is very sensitive to any error or noise. As local code transforms into sparse code, fault-tolerance improves owing to increased redundancy of the input signal.

Whereas Figure 1 considers different properties arising from the code used within the input layer, we next consider the decoding properties of the output neuron (Figure 2). Assuming that information is carried only by excitatory neurons, while the inhibitory interneurons only provide blanket inhibition, the memory capacity (i.e., the total number of contexts that can be stored in the network) increases with the sparseness in the input layer [1,17]. In this scenario, dense coding is associated with low memory capacity because of the large interference between contexts. In sparse coding, the constraint of having only excitatory synapses carrying information has the consequence that the output neuron, in a learnt stage with optimal memory capacity, will have a large