Are There Dimensions in the Brain?

Christian Balkenius

Lund University Cognitive Science
Kungshuset, Lundagård
S-222 22 LUND, Sweden

Abstract

If individual cells do not code dimensions, is it still possible to find representations of dimensions in the brain? It is suggested that perceptual quality dimensions are represented in the brain as a distributed activity pattern at multiple scales. It is shown that categories learned using such representations fulfil a soft convexity criterion when no counterexamples to convexity are presented. A multi-scale representation of properties also allows the same population of cells to support multiple representational spaces that can be selected by a priming mechanism.

1. Introduction

Gärdenfors (2000) proposes that objects can be represented as values on a number of quality dimensions such as length and width, weight and color. Apart from everyday wisdom that tells us that such quality dimensions are useful for describing objects, what evidence is there that the brain uses anything like dimensions to represent objects?

One common suggestion is that the topographic organization of the cells in cerebral cortex could be used to represent dimensions (Gärdenfors 1997, Gallistel 1990). Like a graph on a sheet of paper, the two dimensions of cortical tissue would correspond to two quality dimensions. This idea is supported by recordings from cortical areas where the location of a cell appears to covary with some sensory property such as the orientation of an edge or the location of a stimulus (Stein and Meredith, 1993).

There are however several reasons why this explanation for dimensions cannot be correct. First, the cortical surface is two-dimensional and can thus only represent two-dimensional spaces directly. Higher dimensional spaces could conceivably by coded by several cortical areas, but as the dimensionality of the representation increases, the topological relation between a stimulus and its representation is lost.

Second, a more serious objection is that the physical location of a representation does not in any way contribute to its meaning. A light-switch and a lamp may be located in the same room, but this does not make one control the other. Their causal connection is controlled by the wire between them and not by their spatial proximity. In the same way, the location of a cell cannot serve as a representation in the brain. If the topographic organization of cells in cortex has any significance, it is that it makes the routing of information easier, just as it is convenient to have a switch close to the light that it controls.

Third, it has been suggested that the topological organizations in for example superior colliculus represents a vector space (Gallistel, 1990). But the representations there are not vectors. In fact, none of the operations usually associated with such spaces are relevant. For example, the representations can not be added, subtracted or scaled. The likeness to a vector space is at best superficial.

Another suggestion is that individual cells code dimensions directly. This is certainly true if we accept dimensions such as the pressure applied to the tip of my left hand index finger or the frequency of light reaching a specific point on the retina. But is it possible to find more generally useful dimensions like pressure or color at a single cell? When the brain is probed in search for object representations, what is found is not dimensions but cells that react to very specific and often context dependent properties of sensory stimulation (Tanaka, 1993). There is seldom a simple relation between the response of a cell and a single property of an external object and cells coding something like quality dimensions appear to be lacking. Possible exceptions are the cells in inferior temporal cortex that react to faces (Desimone et al., 1984). But even in this case, there is little relation between the responses of the cells and any comprehensible dimension. Though there are cells that react to faces, none react to noses, and certainly not to the length of a nose or anything else that could constitute a quality dimension for the description of a face.

A more accurate description of the activity of cortical neurons is as tuned detectors (Martin, 1991, Balkenius, 1996). A cell reacts the most when a certain stimulus is presented and its activity decreases as the stimulus is gradually changed away from the optimal stimulus or prototype. The decay of the activity of the cell describes its tuning curve or generalization surface. A characteristic property of the tuning curves of neurons is that there often exist cells tuned to the same stimulus but with different tuning curves. A possible role for these cells may be to support different levels of generalization from a prototype. However, it is clear that dimensions can not reside at the level of the single cell. If we want to find dimensions in the brain we need to look elsewhere.

In Gärdenfors (1994), it was suggested that inferential processes can be seen at three different levels: a subsymbolic, a conceptual, and a symbolic. Since the brain works at a subsymbolic level, it may not be possible to directly observe conceptual representations such as quality dimensions or symbolic representations such as words. The problem of dimensions in the brain now becomes a question of how to interpret neural activity in the brain at a conceptual level as quality dimensions. The next sections describe a possible way in which this could be done and attempt to characterize the operations working on these dimensional representations. It is shown that a form of soft convexity criterion is fulfilled by the categories created using these dimensional representations.

2. The Representation of Dimensions

In this section a new view of the representations of dimensions in the brain will be presented. It will be assumed that there exist cells that react to each sensory stimulus with different tuning curves (Balkenius, 1996). A quality dimension is suggested to consist of a collection of detectors covering all possible values of a quality dimension at multiple scales.

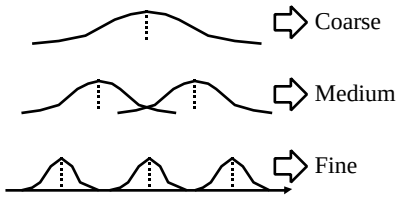

Figure 1. The tuning curves for a population of cells.

Figure 1 illustrates a set of tuned detectors with different tuning curves. When a stimulus is received, all cells with a tuning curve overlapping the stimulus value will become activated proportional to the level of the tuning curve at that value. Since different cells have differently shaped tuning curves, the stimulus value will be represented at different scales (for example, coarse, medium and fine).

Let $G_i(x)$ be a collection of tuned detectors with normalized Gaussian tuning curves centered at c_i with variance s_i . The argument x is a point at a certain quality dimension corresponding to a certain property (Gärdenfors, 1991). The neural representation vector R for the property p is given by,

$$R_i(p) = G_i(p\text{-}c_i) \qquad \text{(Eq. 1)}$$

This idea is consistent with the view that the nervous system uses representation by place, that is, activity at a certain location in the nervous system corresponds to some specific stimulus property (Martin 1991). For example, a specific set of neurons may react each time a tone of a certain pitch is heard.

It follows that two sensory representations are similar to the extent that their neural realizations physically overlap (Balkenius, 1994). The natural similarity measures for such representations resemble similarity based on features (Tversky, 1977), except that the features (i. e. the tuned detectors) are very context dependent. This view also agrees with the classical view of generalization presented by Pavlov (1927).

Representations of this kind can be called semi-local since they share properties with both local and distributed coding schemes (Thorpe, 1995). They are distributed in the sense that many nodes are

necessary to represent the stimulus, but local in the sense that the activity of each node can be given meaning independently.

The relation described by Eq. 1 shows the connection between a quality dimension and its neural representations. It does not imply that neurons use that equation to set its level of activity and it can not, in general, be used to recover a quality dimension from its neural representations. Mathematically speaking, a quality dimension is a subspace of the total representational space of the brain, given by the activity of all neurons, and to make the quality dimension recoverable, the set of tuned detectors must constitute a basis for that subspace.

A more beautiful, though possibly less accurate, model can be constructed by replacing the Gaussians above with an orthogonal wavelet basis (Strang and Nguyen, 1996). In that case, the distributed representation of the dimension is given by a wavelet transform of a Dirac function $D_p(x)$, with unit value at x=p and 0 elsewhere, and the recovery of the original value is straight forward using the inverse transform.

3. Attending to a Dimension

A single property, such as color, can be represented in many different spaces. For example in computer graphics, there is often reason to work with many types of color spaces such as RGB, Lab, CMYK, and HLS to mention only a few. In these different spaces, different dimensions are used although the color is exactly the same. There is no reason to assume that one of these spaces is more accurate than any other though the ways they represent colors have different possibilities and limitations. That we can easily shift between these spaces illustrates that humans are not fixed to one representational space for a single property but can shift between them when needed.

The multi-scale representation of a dimension suggested above may explain how the brain is able to represent dimensions in a distributed manner but does not describe how we can use different dimensions at a symbolic level. A possible link between these levels can be seen in the attentional priming of a neural population (Fuster, 1997, Johnston and Dark, 1986). If the population in question makes up a dimensional space, we have a mechanism that can direct attention to a dimension (figure 2). This prime, in itself, could act as a symbolic representation of the dimension that could prime other symbols for values of that dimension. The suggested mechanism is similar to the k-lines proposed by Minsky (1987). It is also possible to prime only a part of a dimension corresponding to a specific property (Gärdenfors, 1991).

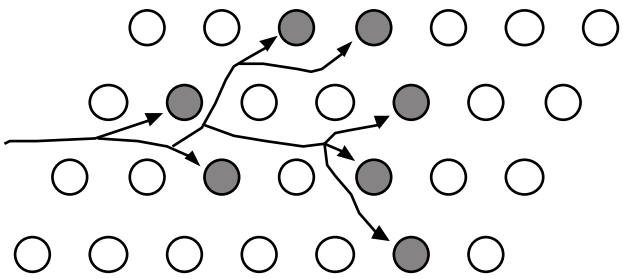

Figure 2. Priming of a dimension. Cells are selected that code for properties along a single dimension.

4. Categorization

Dimensions on their own are useless unless it can be shown that they can be used to support categorization. A classification function is defined as a function C(x) that assigns a value 0 or 1 to each example vector x. Vectors with C=1 are members of the category while C=0 indicates an example outside the category. A soft categorization is given by a function S(x) where the range is values in [0, 1]. A higher value indicates a higher degree of membership in the category. A strict classification function can

be derived from a soft by selecting a threshold where the soft classification gives strict category membership.

A model of categorization in accord with the above view of dimensions is that given by the radial basis function (RBF) networks (Poggio and Girosi, 1989). In a RBF network, a set of Gaussian classifiers is connected to a single node which weigh the contribution of each Gaussian classifier into a linear sum. Given sufficient number of classifiers with appropriate distribution of centers and variances, this sum can approximate an arbitrary function (Poggio and Girosi, 1989, Girosi et al., 1993).

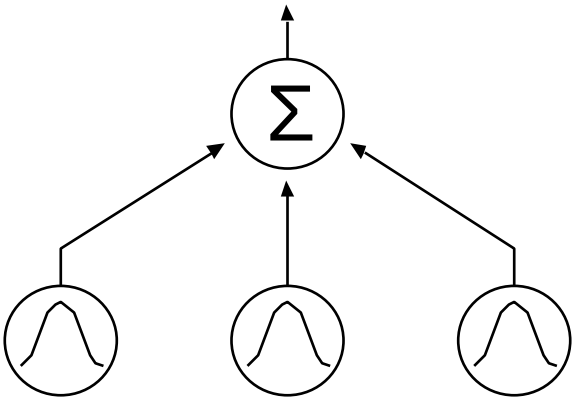

Figure 3. A RBF Network.

It was suggested by Poggio and Girosi (1989) that Gaussians with different variances could be mixed in an approximation task, but the theory was not developed further. To model learning in the network a simple delta-rule can be used (Widrow and Hoff, 1960/1988). The quality of the approximation does however depend on the number of Gaussians used and their location and variances.

In Balkenius (1996) it was shown that categorization with multiple scales increases the average learning rate from O(n) to O(log n) in a spatial sequential decision problem where n is the number of states. Preliminary computer simulations suggest that this is also the case for categorization and general function approximation.

The relation between the Gaussian classifiers and the output of the summing node is suggested to model the relation between two cortical areas. The distributed pattern in one area is used to generate the activity of one node in the other area. It follows from the general approximation property of RBF networks that an architecture of this kind can learn an arbitrary classification of patterns in the first area.

5. Soft Convexity

The general approximation property of RBF networks assures that an arbitrary classification can be learned but it does not tell us what happens when the classification is underspecified. In Gärdenfors (2000), it was proposed that natural categories are convex, that is, if two examples of have a certain property p, then, and example in between those two, will also have that property. Assuming the examples resides in a linear space, we can define convexity of a category as, $C(x) = 1$ and $C(y)=1$ implies that $C(ax+(1-a)y) =1$, where a lies between 0 and 1. In a learning system, this would imply that if two examples x and y are learned as instances of a specific category, this classification should be generalized to instances between them. Let us now relax the condition of convexity somewhat and define soft convexity.

Definition (Soft Convexity): A soft categorization function $S(x)$ shows soft convexity if there exists a threshold $T>0$ such that $S(x) > T$ and $S(y) > T$ implies that $S(ax+(1-a)y) > T$, where a lies between 0 and 1.

Figure 4 illustrates three types of generalization after learning two examples from the same category. The examples are called A and B and the graphs show the level of category membership for different examples along a single dimension. In (a), the two examples have been learned individually and have their own

generalization gradients. The examples have a membership of 1 and the certainty decreases as the example is moved away from A or B. In (b), convexity is used to determine category membership. Both A and B are members of the category and this property is generalized to all examples between them. In higher dimensions this region corresponds to the convex hull of the examples. Finally, in (c) the generalization shows soft convexity. The memberships of all examples between A and B have increased after training with A and B but there are still generalization gradients centered at the trained examples. In the next section, a computer simulation of simple category learning is reported which fulfills soft convexity.

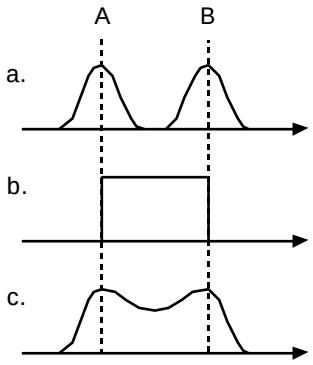

Figure 4. Three types of generalization gradients (a) Local generalization. (b) Convexity. (c) Soft convexity.

6. Simulation

Category learning in a single quality dimension has been computer simulated with a multi-scale RBF-network. The network used 70 different cells distributed at 5 scales. The two positive examples were set to 0.3 and 0.7. The generalization function is shown in figure 5. The multi-scale network automatically generalizes to instances between the learned examples and the learned category shows soft convexity as defined above.

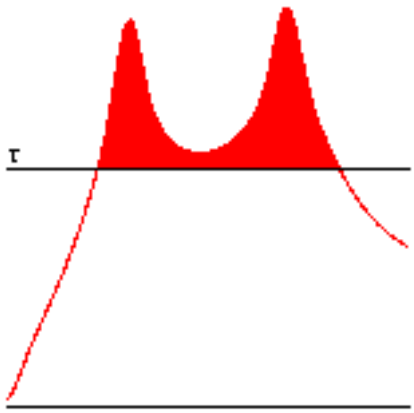

Figure 5. Simulation with category learning with a multi-scale RBF network. T is a threshold that makes the learned category convex.

The soft convexity is not an absolute consequence of the multi-scale representation however. If an explicit counter example is learned the expectation of convexity will be broken. Figure 6 shows such a simulation where a counter example at 0.5 was also presented to the network. In this case, the categorization is no longer softly convex.

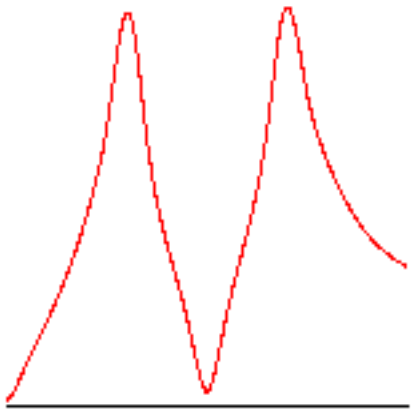

Figure 6. The expectation of convexity is broken when an explicit counter example is presented.

7. Conclusion

If the above account for the representation of dimensions are correct then it would appear that dimensions do exist in the brain but must be approached at a conceptual rather than at a subsymbolic level. Quality dimensions consist of collections of tuned detectors and a property or a value on a dimension corresponds to a distributed activity value at those detectors.

To further support the notion of dimensions as collections of detector readings, it was shown how a simple categorization mechanism could act on those representations. The linear learning method automatically learns concepts that are softly convex in the sense that learning is generalized to examples between the training examples. However, to bear out the notion of dimensions presented here, it will be necessary to explain how other operations on quality dimensions can operate on the suggested representation. An interesting extension of this work would be to merge the model of non-monotonic inferences in neural networks described by Balkenius and Gärdenfors (1991) with the view of conceptual representations presented above.

It has finally been hinted that a lot of interesting mathematics is lurking around the corner connecting the theory of convexity, interpolation and approximation theory, multi-resolution analysis and wavelets.

References

Balkenius, C. (1994). Some properties of neural representations. In L. F. Niklasson and M. B. Bodén (eds.) *Connectionism in a broad perspective*, New York: Ellis Horwood, 79-88.

Balkenius, C. (1996). Generalization in instrumental learning. In Maes, P., Mataric, M., Meyer, J.-A., Pollack, J., Wilson, S. W. (Eds.) *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior* , Cambridge, MA: The MIT Press/Bradford Books.

Balkenius, C. and Gärdenfors, P. (1991). Non-monotonic inferences in neural networks. In Allen, J. A., Fikes, R., Sandewall, E. (Eds.) *Principles of knowledge representation and reasoning: Proceedings of the second international conference*, San Mateo, CA: Morgan Kaufman, 32-39.

Desimone, R., Albright, T. D., Gross, C. G. and Bruce, C. J. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 8, 2051–2068.

Fuster, J. M. (1997). *Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe*, Lippincott.

Gallistel, C. R. (1990). *The Organization of Learning*, Cambridge, MA: Bradford Books, MIT Press.

Gärdenfors, P. (1991). Frameworks for properties: Possible worlds vs. conceptual spaces, *Language, Knowledge and Intentionality* (Acta Philosophica Fennica, vol. 49) , ed. by L. Haaparanta, M. Kusch, and I. Niiniluoto, Helsinki, 383–407.

Gärdenfors, P. (1994). Three levels of inductive inference, in: *Logic, Methodology, and Philosophy of Science IX*, D. Prawitz, B. Skyrms and D. Westerståhl, eds., Elsevier Science, Amsterdam, 427-449.

Gärdenfors, P. (1997). Symbolic, conceptual and subconceptual representations, In *Human and machine perception: information fusion*, ed. by V. Cantoni, V. di Gesù, A. Setti and D. Tegolo, Plenum Press, New York, 255-270.

Gärdenfors, P. (2000). *Conceptual spaces*, Cambridge, MA: Bradford Books, MIT Press, to appear.

Girosi, F., Jones, M., and Poggio, T. (1993). Priors, stabilizers and basis functions: from regularization to radial, tensor and additive splines, *MIT AI Memo No 1430*.

Johnston, A. J. and Dark, V. J. (1986). Selective attention, *Ann. Rev. Psychol.*, 37, 43-75.

Martin, J. H. (1991). Coding and processing of sensory information. In Kandel, E. R., Schwartz, J. H., Jessel, T. M. (Eds.) *Principles of neural science*, New York: Elsevier, 329-340.

Minsky, M. (1987). *The society of mind*, London: Heineman.

Pavlov, I. P. (1927). *Conditioned reflexes*, Oxford: Oxford University Press.

Poggio, T. and Girosi, F. (1989). A theory of networks for approximation and learning, *MIT AI Memo No 1140*.

Stein, B. E. and Meredith, M.A. (1993). *The merging of the senses*, Cambridge, MA: MIT Press.

Strang, G. and Nguyen, T. (1996). *Wavelets and filter banks*, Cambrdige, MA: Wellesley-Cambridge Press.

Tanaka, K. (1993). Neuronal mechanisms of object recognition, *Science*, 262, 685-688.

Thorpe, S., (1995), Localized versus distributed representations. In M. A. Arbib (ed.) *The handbook of brain theory and neural networks*, Cambridge, MA: MIT Press.

Tversky, A. (1977). Features of similarity, *Psychological Review*, 84, 327-352.

Widrow, B. and Hoff, M. E. (1960/1988). Adaptive switching circuits. In Anderson, J. A., Rosenfeld, E. (Eds.) *Neurocomputing: foundations of research*, Cambridge, MA: Mit Press, 123-134.