Neural implementation of psychological spaces.

Włodzisław Duch
Department of Computer Methods, Nicholas Copernicus University, Grudziądzka 5, 87-100 Toruń, Poland; Email: duch@phys.uni.torun.pl

Abstract— **Psychological spaces give natural framework for construction of mental representations. Neural model of psychological spaces provides a link between neuroscience and psychology. Categorization performed in high-dimensional spaces by dynamical associative memory models is approximated with low-dimensional feedforward neural models calculating probability density functions in psychological spaces. Applications to the human categorization experiments are discussed.**

I. Introduction.

Although great progress has been made in recent years in understanding how the brain generates behavior reconciliation of language used in psychology and language used in neuroscience still remains one of the most important problems. Roger Shepard in a paper "Toward a universal law of generalization for psychological science" [1] wrote: "What is required is not more data or more refined data but a different conception of the problem", pointing out that psychological laws should be formulated in appropriate psychological spaces (P-spaces) [2]. Unified theory of mind in cognitive science that Allen Newell hoped for is still missing [3]. Clearly a set of new concepts, a mapping between neurophysiological and psychological events, is needed. How should the higher order cognitive processes, such as categorization, be reduced, at least in principle, to neurodynamics? How are the mental representations in the long-term memory formed? In this paper a model offering plausible solutions to these questions is described.

Categorization, or creation of concepts, is one of the most important cognitive processes. It is also one of the most difficult processes to understand if one tries to see it from the point of view of both psychology and neuroscience. Current research on category learning and concept formation frequently ignores constraints coming from neural plausibility of postulated mechanisms. Connectionist models are at best loosely inspired by the idea that neural processes are at the basis of cognition. An explanation given by a formal theory, even if it fits psychological data, may allow for predictions, but it may not give us more understanding of human cognition than a few-parameter fits allowing for prediction of sun eclipses gave the ancient astronomers.

Psychologists frequently use a language of psychological or feature spaces to describe results of categorization experiments. Shepard showed [1] the existence of universal scaling laws in psychological spaces. Therefore it should be very interesting to construct models of mental events taking place in P-spaces and to show how such models could be realized by neural dynamics. One of the mysteries in brain research is how are the mental representations acquired? Learning at the beginning involves many groups of neurons but after proficiency is gained brain's activity becomes localized. One solution to these problems is offered below.

II. Mind and neurodynamics.

There is growing theoretical and experimental evidence that the original idea of local reverberations in groups of cortical neurons coding the internal representations of categories, put forth by the psychologist Donald Hebb already in 1949, is correct [4]. Local circuits seem to be involved in perception and in memory processes. Analysis of integration of information from the visual receptive fields in terms of modules composed of dense local cortical circuitry [5] allows for explanation of a broad range of experimental data on orientation, direction selectivity and supersaturation. It would be most surprising if the brain mechanisms operating at the perceptual level were not used at higher levels of information processing. Neocortex has highly modular organization, with neurons arranged in six layers and grouped in macrocolumns that in turn contain microcolumns of about 110 neuron each. Successful models of memory, such as the tracelink model of Murre [6], make good use of this modular structure, postulating that each episodic memory is coded in a number of memory traces that are simultaneously activated and their activity dominates the global dynamics of the brain, reinstating similar neural state as was created during the actual episode.

How is then mind related to neurodynamics? In physics macroscopic properties results from microinteractions, in psychology behavior should also result from neurodynamics. In practice direct attempts at connecting neural dynamics with higher cognition seem to be hopelessly difficult. Macroscopic physics is possible because space-time, either Euclidean in classical physics, or described by differential geometry in relativistic physics, is a good arena for physical events. It seems fruitful to use P-spaces as an arena for mental events. A sketch of such theory was given recently [7].

A reasonable hypothesis relating psychological concepts to brain activity seems to be the following: the activity of microcolumns shows quasidiscrete attractor dynamics. Several stable patterns of excitations may form, each coding a specific concept. Via axon collaterals of pyramidal cells, extending at distances of several millimeters, each microcolumn excites other microcolumns coding related concepts. These excitations should depend on the particular form of local dynamics. From the mathematical point of view the structure of local excitations is determined by attractors in the dynamics of neural cell assemblies. A collection of mode-locking spiking neurons provides a good model of such networks. Simple models of competitive networks with spiking neurons have been created to explain such psychological processes as access to consciousness (cf. [8]). Realistic simulations of the attractor dynamics of microcolumns, giving results comparable with experiment, should be possible, although they have not been done yet. In any case, possible attractor states of neurodynamics should be identified, basins of attractor outlined and transition probabilities between different attractors found. In the olfactory system it was experimentally found [9] that the dynamics is chaotic and reaches a cyclic attractor only when a proper external input is given as a cue. The same may be expected for the dynamics of a microcolumn. Specific external input provides a proper combination of features activating a microcolumn that partially codes a category. From the neurodynamical point of view external inputs push the system into a basin of one of the attractors.

A good approach connecting neurodynamics with mental events in higher cognition tasks should start from analysis of neural dynamics, find invariants (attractors) of this dynamics and represent the basins of attractors in P-space. Behavioral data may also be used to set a topography of psychological space. In the first step neural responses should be mapped to stimulus spaces. This may be done by population analysis or Bayesian analysis of multielectrode responses [10]. Conditional probabilities of responses $P(r_i|s), i = 1..N$ are computed from multi-electrode measurements. The posterior probability $P(s|r) = P$(stimulus|given response) is computed from the Bayes law:

$$P(s|r) = P\left(s|r_1, r_2..r_N\right) = \frac{P(s)\prod_{i=1}^{N} P(r_i|s)}{\sum_{s'} P(s')\prod_{i=1}^{N} P(r_i|s')}$$

Representing $P(s|r)$ probabilities in psychological spaces based on the feature of stimuli a number of "objects" representing recognized categories are created. Psychological research on categorization may provide additional behavioral data and both types of data may be used in one model.

It would be ideal to construct models of neurodynamics based on experimental data, describing how groups of neurons learn to categorize, and then to reduce these models to simplified, canonical dynamics (i.e. the simplest dynamics equivalent to the original neurodynamics) in the low-dimensional psychological space. So far there are no good neurodynamical spiking neuron models of the category learning process, but it is possible to create a simple attractor network models based on the Hopfield networks and use these models to understand some aspects of category learning in monkeys (cf. [4]). The internal state of these models is described by the activity of a large number of neurons. Since the input information $O(X)$ is uniquely determined by a point X in the psychological space it is possible to investigate the category that the attractor model $A(O(X))$ will assign to each point in the psychological space. Thus an image of the basins of attractor dynamics in the psychological space may be formed. Attractors do not have to be point-like, as long as a procedure to assign categories, or probability of different categories, to a specific behavior of the attractor network is defined. To characterize the attractor dynamics in greater details probabilities $p_i(X)$ may be defined on the P-space. In a K-category problem there are $K-1$ independent probabilities. Other functions that one may define on P-space may measure the time the dynamical system needs to reach the asymptotic categorization probability value. Functions on P-spaces may be modeled using conventional feedforward neural networks.

More detailed models of this kind, which I have called previously [7] "Platonic models" (Plato thought that mind events are a shadow of ideal reality, here probability maxima representing categories of input objects are shadows of neurodynamics), should also preserve similarities between categories learned by neural systems. Similarity of categories represented in feature spaces by peaks of high probability clusters should be proportional to some measure of distance between them. In neural dynamics this is determined by transition probability between different attractor states, determining how "easy" it is to go from one category to the other. However, there is no reason why such transition probabilities should be symmetric. As a consequence distance $d(A, B)$ between two objects A and B in the feature space should be different than distance $d(B, A)$. Euclidean geometry cannot be used in such case. A natural generalization of distance is given by the action integral in the Finsler spaces [11]:

$$s(A, B) = \min \int_A^B L(X(t), dX(t)/dt)dt$$

where $L(\cdot)$ is a Lagrangian function. Attractor basins correspond to regions of high values of probability densities in P-spaces. The dynamics in P-spaces is represented by the movement of a point called the state S, going from one category to another, following the underlying neurodynamics. Dynamics should slow down or stabilize around probability peaks, corresponding to the time that the mind spends in each category coded by an attractor state of neurodynamics. Only a small part of the overall neurodynamics of the brain is modeled, the rest acting as a source of noise. Point, cyclic and strange attractors may be interpreted as recognition of categories. Point attractors correspond to infinite times spend on one category. The distance between such categories should in this case grow infinitely – if interactions with other parts of the brain are neglected point attractors behave like "black holes", trapping the mind state forever.

The model of forming mental representations proposed here assumes that categorization is initially done by the brain using many collaborating microcolumns in the associative areas of the cortex or in the hipocampus. This process should be described by an attractor network. Categorization, or interpretation of the states of this network, is done by distal projections to cortical motor areas. Longer learning leads to development of a specialized feedforward neural network that matches higher-level complex features with categorization decisions of the attractor network. Mental representations are defined and interpreted in the low-dimensional P-spaces, not in the high-dimensional patterns of activity of the attractor network dynamics. Higher-level complex features are created by combination of lower-level features (mechanisms of attention should play a role here). Alternatively they may be formed by a neural realization of multidimensional scaling procedure [12]. Preservation of similarities is the only requirement for the dimensionality reduction. Mental representations - objects in P-spaces - are formed slowly transferring the knowledge about categorization from the attractor networks to simpler feedforward networks.

Geometrical characterization of P-spaces and of the landscapes created by the probability density functions defined on these spaces and obtained as an approximation to neurodynamics lead to an alternative model of mind. Such P-spaces offer an arena for mind events that is acceptable to psychology and understandable from neurobiological point of view.

III. Encoding categories in feature spaces.

The model presented in the previous section may be applied to categorization in psychology. Although the examplar theory of categorization is usually presented as an alternative to the prototype theory [13] neurodynamics lies at the basis of both theories. Is it possible to distinguish between categorization based on prototypes and exemplars? In the first case basins of attractors should be large and the corresponding objects in P-spaces should be large and fuzzy. A prototype is not simply a point with average features for a given set of examples, but a complex fuzzy object in the P-space. If categorization is based on exemplars there are point-like attractors corresponding to these exemplars and the P-space objects are also point-like. Intermediate cases are also possible, going from set of points representing exemplars, to a fuzzy object containing all the exemplars. Although representation is different both theories may give similar behavioral results if processes acting on these representations are different. If the neural dynamics is noisy exemplars become so fuzzy that a prototype is formed. Neural dynamic models physical processes at the level of brain matter while dynamic in the P-spaces models a series of successive categorizations, or mental events, providing precise language useful from psychological perspective.

A classic category learning task experiment has been performed by Shepard *et.al.* [14]. Subjects were tested on six types of classification problems of increasing complexity. 8 distinct objects had two kinds of shape, two colors and two sizes. In type I problems only a single feature was relevant, for example category A included all squared-shaped objects and category B all triangle shaped objects. In type II problems two features were relevant for categorization, for example shape and color, but not size of the objects. The logic behind category assignment could be AND, OR, XOR. Type II-VI problems involve all three features with various logic behind the assignment.

Since the details of neurodynamics are not important to understand such categorization experiments, it should be sufficient to investigate canonical form of simplified neurodynamics. One may claim that any neural dynamics responsible for categorization in problems with two relevant features is in principle reducible to one of the simplified dynamical systems defined in the 3-dimensional psychological spaces (two features plus the third dimension labeling categories). Parameters defining such simplified dynamics should allow to reproduce observed behavior. Prototype dynamics for all logical functions used in categorization experiments has been found. For example, Type II problems are solved by the following prototype dynamical system:

$$\begin{aligned}
V(x,y,z) &= 3xyz + \frac{1}{4}\left(x^2+y^2+z^2\right)^2 \\
\dot{x} = -\frac{\partial V}{\partial x} &= -3yz - \left(x^2+y^2+z^2\right)x \\
\dot{y} = -\frac{\partial V}{\partial y} &= -3xz - \left(x^2+y^2+z^2\right)y \\
\dot{z} = -\frac{\partial V}{\partial z} &= -3xy - \left(x^2+y^2+z^2\right)z
\end{aligned}$$

This system has 5 attractors (0,0,0), (-1,-1,-1), (1,1,-1); (-1,1,1), (1,-1,1); the first attractor is of the saddle point type and defines a separatrix for the basins of the other four. Such dynamical system may be realized by different neural networks. In this example, as well as in the remaining five types of classification problems of Shepard *et.al.* [14], it is easy to follow the path from neural dynamics to the behavior of experimental subjects during classification task. Starting from examples of patterns serving as point attractors it is always possible to construct a formal dynamics and realize it in the form of a set of frequency locking nonlinear oscillators [15].

Although polynomial form of canonical dynamical system is for the XOR case very simple and has only one saddle point for other useful functions it is more complex. Modeling point attractors using functions $G(X_i, s_i)$ localized around the K attractors, leads to the following equations:

$$\begin{aligned}
V(X) &= \sum_{i=1}^{K} W_i G(X_i, s_i) \\
\dot{X}_i &= -\frac{\partial V}{\partial X_i}
\end{aligned}$$

This form allows us to model the potential by changing the positions and fuzziness (controlled by s_i parameters) of the attractors and their relative weights W_i. Functions $G(X_i, s_i)$ may either be Gaussian or, if neural plausibility is required, a sum of combination of pairs of sigmoidal functions $\sum_i(\sigma(X_i+s_i)-\sigma(X_i-s_i))$ filtered through another sigmoid. Using this form of the potential one may create basins of attractors with desired properties and set up the parameters of these functions to account for experimental data.

People learn relative frequencies (base rates) of categories and use this knowledge for classification. The inverse base rate effect [16] shows that in some cases predictions contrary to the base rates are made. This effect may be explained using specific shapes of the basins of attractors of neural dynamics, but it is much easier to understand it representing these attractors and the decision boundaries between them in the P-space. Thus the same event may be seen from psychological and from the neurodynamical point of view. Learning of the base rates changes synaptic connections in the neural models, creating larger and deeper basins of attractors - in the P-space objects (high probability values) corresponding to these attractors are large. Inverse base rate effects result from deeper, localized attractors around rare categories, or smaller, localized objects in P-spaces.

Processes acting on representations in feature spaces define certain physics of mental events, with forces reflecting the underlying neural dynamics. The state of the system, called "the mind state" [7], is a point moving in the P-space. Base rate effects influence the size of the basins of attractors, represented by the size of objects in the P-space. They also influence transition probabilities: specifying value of a feature that frequently appears in combination with other features gives momentum to the mind state in the direction parallel to the axis of this feature, initiating a search for a category completing the missing features (for application of the searches in feature spaces see [17]).

IV. Summary and related work.

The need for psychological space treated as an arena for psychological events is evident in recent psychological literature. The static picture of P-spaces with probability density functions defined on them is useful not only for categorization but also object recognition [18]. In psycholinguistics problems such as the word sense disambiguation and learning the semantic contents of new words by children are solved placing categories (words) in P-spaces. Landauer and Dumais [19] analyzed a dictionary of 60.000 words and using the Latent Semantic Analysis model estimated effective dimension of the P-space that is needed to preserve similarity relations to be about 300. Linguists use also the idea of non-classical feature spaces, calling them "mental spaces" (cf. [20]).

Static version of the Platonic model should be sufficient for description of a short-term response properties of the brain, "intuitive" behavior or memory-based responses, but understanding behavior in the time frame longer than a few seconds must include dynamical aspects. The dynamic Platonic model, introduced in [7], goes in the same direction as Elman's "language as a dynamical system" idea in psycholinguistics and "mind as motion" ideas in cognitive psychology (cf. [21]). Stream of thoughts forming a sentence create a trajectory of visited (or "activated") objects in psychological space. General properties of such trajectories reflect grammar of the language. Further simplification of the Platonic model lead to the Bayesian networks and Hidden Markov Models in which the dynamics is completely neglected and only the probability of transitions between states/objects remains. Reasoning in symbolic models of mind, such as SOAR, is based on problem spaces, which are metric spaces rather than vector spaces. It is not yet clear what are the precise restrictions of modeling psychological spaces using vector space structure.

A unified paradigm for cognitive science requires elucidation of the structure of psychological spaces, search for low dimensional representations of behavioral data and for connections with neural dynamics. Linking neural dynamics with psychological models based on feature spaces leads to a complementary description of brain processes and mental events. The laws governing these mental events result from approximations to neural dynamics, similarly as the laws of classical physics result from approximations to quantum mechanics. These modified feature space models are useful in analysis of psychological experiments, explaining data on judgments of similarity between objects and abstract concepts, as well as results of experiments on categorization. Perhaps at the end of this road a physics-like theory of events in mental spaces is possible?

References

[1] R. Shepard, *Toward a universal law of generalization for psychological science*, Science 237 (1987) 1318-1323

[2] J. Eliot, *Models of Psychological Space*, Springer 1987

[3] A. Newell, *Unified theories of cognition.* Harvard University Press, Cambridge, Massachusetts, 1990,

[4] D.J. Amit, *The Hebbian paradigm reintegrated: local reverberations as internal representations*. Brain and Behavioral Science 18 (1995) 617-657

[5] D. C. Somers, Emanuel V. Todorov, Athanassios G. Siapas, Mriganka Sur, *Vector-space integration of local and long-range information in visual cortex*. MIT AI memo 1556, November 1995.

[6] J. Murre, *A model of amnesia and consolidation of memory*. Hippocampus 6 (1996) 675-684

[7] W. Duch, *Platonic model of mind as an approximation to neurodynamics*, in: Brain-like computing and intelligent information systems, ed. S-i. Amari, N. Kasabov (Springer, Singapore 1997), chap. 20, pp. 491-512; *Computational physics of the mind*, Comp. Phys. Comm. 97 (1996) 136-153

[8] J.G. Taylor, F.N. Alavi, *Mathematical analysis of a competitive network for attention*. In: J.G. Taylor, ed. Mathematical Approaches to Neural Networks (Elsevier 1993), p.341-382

[9] W.J. Freeman, *Simulation of chaotic EEG patterns with a dynamic model of the olfactory system*. Biolog. Cybernetics 56 (1987) 139-150

[10] P. Földiak, *The 'Ideal homunculus': statistical inferences from neural population responses.* In: Eeckman F.H, Bower J.M (Eds.), Computation and neural systems (Kluver 1993), pp. 55-60

[11] P.L. Antonelli, R.S. Ingarden, M. Matsumoto, *The Theory of Sprays and Finsler Spaces with Applications in Physics and Biology* (Kluver 1993)

[12] M.D. Lee, *The Connectionist Construction of Psychological Spaces*, Connection Science 9 (1997) 323-352

[13] I. Roth, V. Bruce, *Perception and Representation*. Open University Press, 2nd. ed, 1995.

[14] R.N. Shepard, C.I. Hovland and H.M. Jenkins (1961) *Learning and memorization of classifications*. Psychological Monographs 517

[15] H. Haken, *Synergetic Computers and Cognition*. Springer 1991

[16] D.L. Medin, S.M. Edelson, *Problem structure and the use of base-rate information from experience*. Journ. of Exp. Psych: General 117 (1988) 68-85

[17] W. Duch, G.H.F. Diercksen, *Feature Space Mapping as a universal adaptive system*. Comp. Phys. Comm. 87 (1995) 341-371

[18] N. Intrator, S. Edelman, *Learning low dimensional representations of visual objects with extensive use of prior knowledge*, Network 8 (1997) 259-281

[19] T. Landauer, S. Dumais, *Latent Semantic Analysis Theory*, Psych. Rev. 104 (1997) 211-232

[20] G. Fauconnier, *Mental Spaces* (Cambridge U.P. 1994)

[21] R.F. Port, T. van Gelder, eds, *Mind as motion*. (MIT Press 1995)