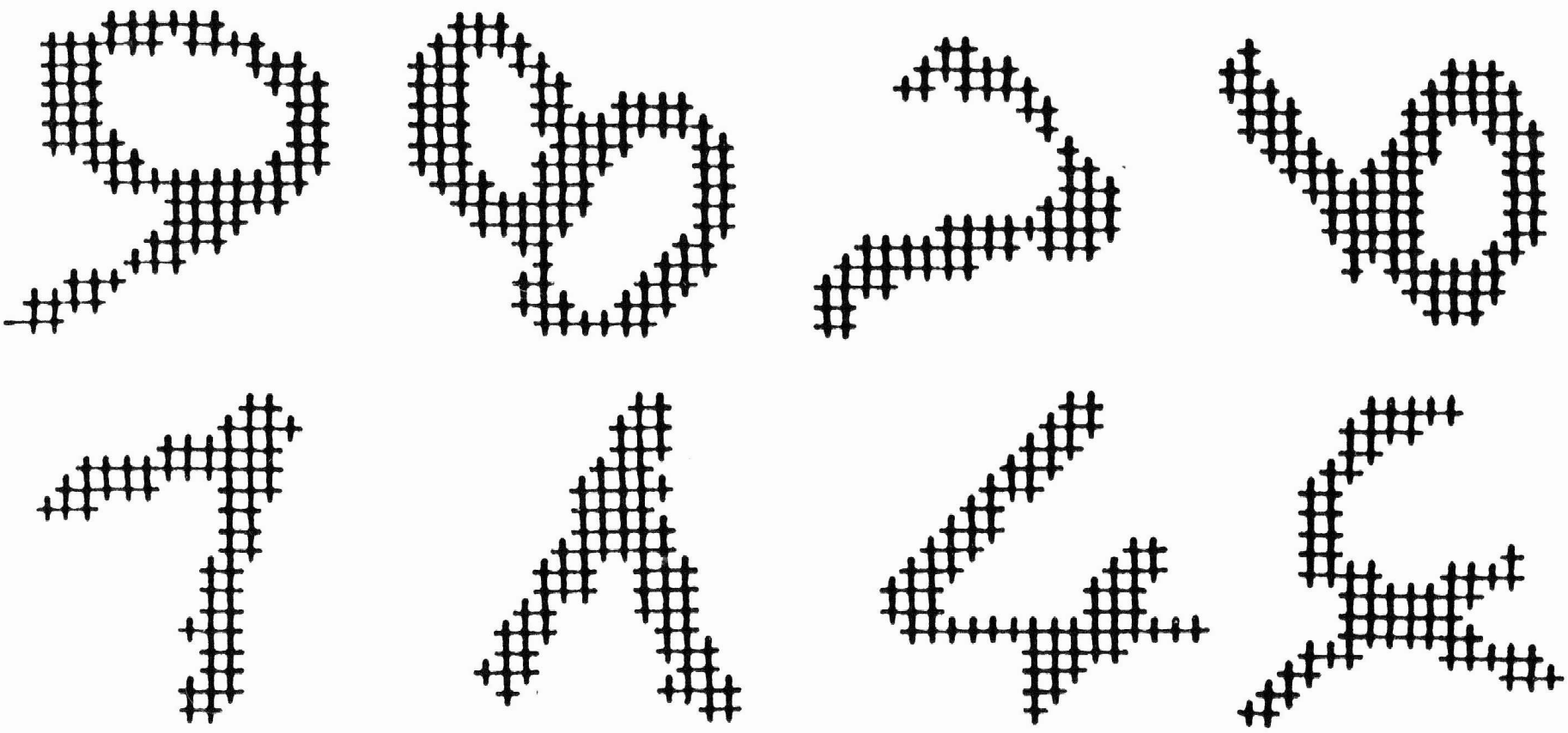
IDG

NEURAL NETWORK WORLD

International Journal on Neural and Mass-Parallel Computing and Information Systems

VOLUME 1 | 1991 | NUMBER 2



NEURAL NETWORK WORLD is published in 6 issues per annum by the IDG Company, Czechoslovakia,
120 00 Prague, Blanická 16, Czechoslovakia,
the member of the IDG Communications, USA.

Editor-in-Chief: Dr. Mirko Novák

Associate Editors: Prof. Dr. V. Hamata,
Dr. M. Jiřina,
Dr. D. Húsek

Institute of Computer and Information Science, Czechoslovak Academy of Sciences, 182 07 Prague, Pod vodárenskou věží 2, Czechoslovakia.

Phone: (00422) 82 16 39, (00422) 815 20 80, (00422) 815 31 00

Fax: (00422) 85 85 789,

E-Mail: CVS35@CSPGCS11. BITNET

General Manager of the IDG Co., Czechoslovakia:
Prof. Vladimír Tichý
Phone: (00422) 25 80 23, Fax: (00422) 25 73 59.

General Editor of all the IDG Co., Czechoslovakia journals:
Ing. Vítězslav Jelínek
Phone: (00422) 25 32 17.

Price Information:
Subscription rate 399 US $ per annum.
One issue price: 66.50 US $.
Subscription adress: IDG Co., Czechoslovakia, 120 00 Prague 2, Blanická 16, Czechoslovakia

Advertisement: Ms. M. Váňová, Ms. Ing. H. Vančurová,
IDG Co., Czechoslovakia, 120 00 Prague 2,
Blanická 16
Phone: (00422) 25 80 23, Fax: (00422) 25 73 59.

For the cover picture see p. 76

Scanning the Issue

Editorial

The preparation of this issue — the second in the up to now short history of our Journal — started almost concurently with the first one.

We try here to continue in presenting to the readers some interesting views on the field of neurocomputing and neuroscience. We also would like to have here a certain balance between the contributions from East and West, and also between theory and applications.

We continue in the series of Tutorials with the second part of J. Hořejš's paper on paradigm development. We present here the record of the Panel discussion given at the International Symposium on Neural Networks and Neurocomputing NEURONET'90 and we devote also the corresponding space to information on new books and recently published papers. We inform the reader about some interesting meetings, seminars, symposiums, conferences and exhibitions. In the section „Neurocomputer Companies" we introduce a survey of neuro-oriented companies which were present at the CeBIT fair, Hannover (Germany), March 13—20, 1991.

Mirko Novák
Editor — in — Chief

UNCERTAINTY AND INFORMATION: THE EMERGING PARADIGMS

M. M. Gupta*)

Abstract:

In this paper, we describe some aspects of information and its cognate the uncertainty from the design of perspectives of intelligent systems. The discussion is centered around statistical uncertainty and cognitive uncertainty, an important class of uncertainty that arises from human thinking and cognition process. Also, we discuss how these two uncertainties can help us in the design of new class of sensors and intelligent systems.

1. Introduction

The world around us is full of uncertainties: for example, the uncertainties caused by natural weather patterns and the uncertainties in world peace caused by power hungry politicians. In weather patterns, we have a fairly well defined deterministic morphology at the ultra-macroscopic level in terms of what we will face in the winter or summer months, however, the weather is almost uncertain and difficult to predict at the microscopic level. So is the uncertain situation in the Gulf crisis at the writing of these lines. One will also notice the uncertainty embodied in the random turbulence of the blood flow of our own cardio-vascular system, in the excitation patterns of nervous cells, or in the chaotic behaviour of neural cells in the brain. The uncertainty in the random vibrations of a musical string creates music which can resonate the neurons in one's brain or the uncertainty in the random loud noise in our living environment creates annoyance and can damage our hearing.

We humans are shrouded in uncertainties arising from our own thinking, mentation, cognition and perception process as well. Here, we present a few examples of uncertainties arising from our thinking and cognition process: *this is a beautiful spring rose full of pleasant fragrance, you are very kind to me; music is very pleasant; summers are pleasant and winters are extremely cold in Saskatchewan, and so on, so forth.*

Uncertainty is a blanket which tightly shrouds our environment and our thinking process. We humans are capable of, to a certain extent, perceiving and uncertain phenomenon, extracting some useful information, and attaching a meaning to this information. The perception of this information is very useful. For example, in the case of some perceiving of danger, we attempt to take a defensive action. As we grow, with experience, we develop our own cognitive faculty to extract useful information from the uncertainties in our environment, and make use of this information in our future actions and decision making tasks.

*) Madan M. Gupta
Inteligent Systems Research Laboratory
(Center for Excellence on Neuro-Vision Research)
College of Engineering
University of Saskatchewan
Saskatoon, Saskatchewan
Canada S7N 0W0

Uncertainty is an inherent phenomenon in our universe and in our lives which stands continuously open to our gaze. To some, it may become a cause of anxiety, but to scientists, it becomes a chapter full of challenges. Scientists attempt to comprehend the language of this uncertainty through the mathematical tools, but still these mathematical tools are incomplete.

To some scientists, uncertainty evokes the notion of probability; that is the very reason morphology of uncertainty has been very dull and dry. The theory of probability is unable to describe the beauty of music emating from the random vibrations of a string, or of the scene of a snow clad mountain or the fragrance of a spring rose. The morphology of uncertainty in the random vibrations of a string is probabilistic, however, the uncertainty associated with the perception of its sound is not probabilistic.

Thus, uncertainty may arise from physical phenomena which, in general, are governed by physical laws such as the laws of electromagnetics, laws of motion, and laws of electrical current flows. Also, uncertainty may arise through the process of human cognition and perception; *this red spring rose is beautiful and full of pleasing fragrance*. This cognitive uncertainty is associated with human perception and with the minds of other intelligent biological species. This is the uncertainty which we can *fell*, but which does not have any shapes or bounds. But the strength of this amorphous uncertainty lies in that it can interact with our cognitive process during intelligent decision making tasks. During the past, to mathematicians and scientists, uncertainty has always evoked the throughts of a probabilistic type of uncertainty and they have disdained the challenges of understanding the amorphous (cognitive) uncertainty. It is only recently that with an increasing interest in the development of intelligent autonomous systems, scientists and mathematicians have directed their efforts to devise theories to give some understanding to this amorphous uncertainty.

The mathematics of cognitive uncertainty formalizes the structure of uncertainty arising from the process of thinking, mentation, and perception. There is one important idealization involved: unlike probabilistic uncertainty, cognitive uncertainty does not have absolute measurements, rather, it is relative and context dependant. *"Today, the weather is warm"* evokes two different temperatures in our minds for the months of January and June in Saskatoon, or for two different places, (Saskatoon and New Delhi, for example) but in the same month (January, for example). The new mathematics of cognitive uncertainty may play an important role in the development of autonomous intelligent systems, just as the mathematics of probability theory has played an important role in the understanding of some natural phenomena associated with quantum mechanics, turbulent water flow, and in forecasting of the uncertain weather patterns or random changes in the stock market.

There are emerging paradigms for uncertainty and information. Perhaps, the most convincing arguement in favor of the study of cognitive uncertainty lies in the extraction of amount of information that is embedded in this type of uncertainty.

2. Probabilistic and Cognitive Uncertainties

There are various classes of uncertainties, however, for the discussion purposes, here we classify these uncertainties into two broad categories: the *probabilistic* and the *cognitive* uncertainties.

The *probabilistic* type of uncertainty deals with the information or phenomena which arise from the random behaviour of physical systems. The pervasiveness of this type of uncertainty can be witnessed in the random vibrations of a machine, randomness of a message, random fluctuations of electrons in a magnetic field, diffusion of gases in a thermal field, random electrical activities of the cardiac muscles, uncertain fluctuations in the weather pattern and the turbulent blood flow through a damaged cardiac valve. The probabilistic type of uncertainty has been studied for centuries, and we have a very rich statistical theory to characterize such random phenomena. The calculus of mean and variance is very rich in this respect, and is being used very widely.

The *cognitive* uncertainty, unlike the probabilistic one, is the uncertainty that deals with phenomena arising from human thinking, reasoning, cognition and perception processes, or cognitive information in general. This is subject which has been either neglected or taken very lightly. The cognition and perception of the physical environment through our natural sensors (eyes, ears, nose, etc.), the perception of pain and other similar biological events through our nervous system and neural networks deserve special attention. The *"perception phenomenon"* associated with these processes are full of *"uncertainties"* and cannot be characterized by conventional statistical theory. We can feel pain: *"the back is very painful", but this pain can be neither measured nor characterized using statistical theory. Similarly, we express our perception linguistically, "this red flower is just beautiful and is full of pleasing fragrance"*. This corresponds to the *"perception"* of our physical environment where *"red"* and *"beautiful"* describe the visual perception, whereas *"pleasing fragrance"* describes the perception of smell. Again, we cannot characterize these perceptions using the strength of the statistical theory.

The cognitive uncertainty and its cognate, the cognitive information, involve the activities of the natural neural networks. To non-scientists, it may seem strange that such "familiar" notions have recently become the focus of intense research. But it is the "ignorance" of these notions, and their possible technological applications in intelligent man-made systems, and not

"familiarity" with them which has forced scientists to conduct research in the field of *cognitive uncertainty* and *cognitive information.*

The development of the human cognitive process and the perception of his environment starts taking shape with the development of imaginative power in a baby's brain. A baby in the cradle can recognize the human face long before it is conscious of any visual physical attributes of humans or its environment.

In spite of the richness of conventional statistical mathematical methods, they are very often thought to be dry and cold. One reason lies in this inability to describe the beauty of white mountains, blue lakes, the rising sun, the full moon, or the richness of the fragrance of a spring flower. No doubt, one can estimate the volume of snow or the heights of the mountains, or the frequencies of vibrating musical strings, but the conventional mathematical methods cannot be used to narrate logically the feelings and the emotions associated with their perceptions.

The study of such formless uncertainties provides us with a scientific challenge. Scientists have started now to think of giving a morphology to this amorphous soft uncertainty. In the past, mathematicians have disdained this challenge and have increasingly chosen to flee from natural mentation by devising theories unrelated to human perception, feelings and emotions.

It was in 1965 when Lotfi A. Zadeh published his first celebrated paper on *Fuzzy Sets* and it is now almost twenty-two years since he first introduced to me this new type of information and uncertainty at the breakfast table in August, 1968 at the IFAC Symposium held at Dubrovnik, Yugoslavia. He showed me the path which leads to somewhat beautiful gardens full of immortal and ever increasing fragrance. Though I was taught the notions of cognition and perception at shool, I was very ignorant about uncertainty and its pervasiveness around these notions. Indeed, this uncertainty has been disdained by scientists and mathematicians.

No one had seen the beauty of these Fuzzy Sets before Professor Lotfi Zadeh, and it was he who showed promise of consolidating this beauty into an organized field with rich theories and promising applications.

Professor Zadeh coined the word Fuzzy Sets. Fuzzy Sets deal with sets of objects or phenomena which are vague and have only soft boundaries. The calculus of fuzzy sets and soft logic is a very promising tool for dealing with cognitive uncertainty (just as statistical theory deals with the probabilistic uncertainty). Indeed, the applications of these fuzzy sets, which once were thought to be dull and dry, can be found in many scientific and scholarly works. It is true that Boole introduced the beautiful notion of binary logic which is so pervasive in our digital world, however, this beauty is naked and without any adornment. Boolean logic is unable to model the human cognition and thinking process. This is the very reason that no one today is indifferent to the soft logic of fuzzy sets. In fact, many view their first encounter with the fuzzy logic as a totally new and exciting experience in their scientific life.

From the purely mathematical view point, the evolution of the theory of soft (fuzzy) logic is very exciting but complex. Many scientific theories start by borrowing notions from the already developed areas of mathematics, but in this case, Professor Zadeh introduced the basic notion of *"vagueness"* having no sharp morphology and which is so common in human thought processes, and this notion of vagueness is modelled using the notion of *graded membership.*

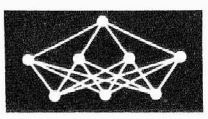
Indeed, Professor Zadeh laid the foundation of fuzzy mathematics on a very robust rock. It now serves the needs of many existing scientific desciplines, but equally important is that many new disciplines, such as the study of fuzzy neural networks, and fuzzy chaos, have started arising around these mathematics. Thus, these mathematics have united several noble (both old and new) narrow streams of scientific disciplines into one while, at the same time, instilling life into several other streams that have been dormant.

Ever since Aristotle, the science of logic has followed a narrow and abstract path through a wilderness of irrelevances and paradoxes incompatible with human logic. Real world is far away from binary logic; computers are asked to solve human-like real world problems but the computer logic is too artificial to handle such problems. A cognitive machine that can appreciate the hedges of truth will appreciate the human-like soft logic. There have been several attempts to use fuzzy logic in the design of such a cognitive machine.

3. Mentation, Cognition and Soft (Fuzzy) Logic

In biomedical engineering, we apply the principles of the natural sciences and engineering to the benefit of the health sciences. Here, we shall take an *inverse biomedical engineering* [(biomedical engineering)$^{-1}$] approach and shall try to apply the biological principles to the solution of some engineering problems. In particular, engineers are investigating the problem of creating intelligence in a robotic system. The thought of the creation of intelligence on a silicon chip (machine) creates some strange feelings in our minds.

Intelligence implies the ability to think, reason, learn and memorize, or, in general, it refers to the human mentation and cognition process. One of the most important frontiers of science, is understanding the biological basis of mentation and cognition: how we think, reason, learn, remember, perceive and act. I still cannot understand how the brain can perceive the dangerous driving situation and act instantaneously while it might take several seconds to multiply two three-digit numbers. How do the genes contribute to



the process of mentation and cognition and how do they develop with the environment?

Here, we have two computational tools: the carbon based organic brain which has existed in humans and animals for several billions of years, and the silicon based modern computers which have evolved only over the last three decades. Recent technological advances in computer hardware have made it possible to carry a very powerful computer in a briefcase which is ultra fast and efficient for numerical computations. However, the *"cognitive information"*, the information which our natural sensors acquire, is not numerical, but the "mentation process" can process such information very efficiently and act upon it accordingly. The modern day computers fail to process such cognitive information.

The fact that the human mentation and cognition process is so marvelously efficient and effective, poses a question for scientists and engineers: *Can some of the functions and attributes of the human sensory system, mentation and cognitive processor, and motor neurons be emulated in a robotic system?*

For such an emulation process, it is necessary to understand the biological and physiological functions of the brain. It is a difficult question to answer. However, it is felt that if we examine some of the "mathematical aspects" of our thinking process and "hardware aspects" of the "neurons", the principle element of the brain, we may succeed to some extent in our emulation process.

The mentation and cognitive activity of the brain, unlike the computational function of the binary computer, is based upon the relative grades of information acquired by the natural sensory system. The conventional mathematical tools, whether deterministic or probabilistic, are based upon some absolute measure of the information. Our natural sensors acquire information in the form of relative grades rather than in absolute numbers. The perception and "action" of the cognitive process also appear in the form of relative grades. While driving on an icy road, for example, we perceive the driving environment in a relatively graded sense and act accordingly.

The mentation and cognitive process thus acts upon the graded information. Information may appear in a numerical form (temperature of the body is 38.4 °C), however, during the process of cognition, we perceive this temperature as near normal, in the form of relative grades. Thus, the cognitive process acts upon the different forms of information and this leads to "formless" uncertainty: *temperature is near normal.*

The theory of fuzzy logic is based upon the notion of relative graded membership and so is the function of the mentation and cognitive process. In the past, studies of cognitive uncertainty and its cognate, the cognitive information, were hindered by the lack of suitable tools for modeling such information. However, with the introduction of the theory of fuzzy logic, it is possible now to expand studies in this important field of cognitive information, neural networks, and cognitive-neural computing tools.

My own laboratory is heavily committed to studies in the field of cognitive information processing, cognitive vision fields, vision perception, neuro-vision research, cognitive-neural computing tools, and cognitive feedback controllers with promising applications to intelligent robotic systems and medical imaging.

4. Biological Basis for Cognition

The genetic control, sensory, cognitive and perception capabilities of biological systems provide an interesting challenge to both engineers and physical scientists in order to learn from these processes and emulate their robust behavior on computers for their engineering and medical applications. For example, the information processing and information extraction capabilities of natural sensors such as vision, tactile, olfactory and auditory along with the cognitive and perception capabilities of the brain provide many challenging examples for the development of intelligent sensors and devices.

The information extraction of our natural sensors and the cognitive functions of the brain are based upon, as well known, on some aggregate properties of attributes in a sensory field. For example, the various attributes associated with the vision such as color, depth, edges, gray-levels, create a vision field. In particular, the attribute color creates a color field with a continuous distribution of color with various intensity levels. Similar is the case with depth and gray-intensity fields. The vision sensory system (consisting of retinal receptors and neural networks at various levels) interacts with these fields in an aggregate manner and extracts information regarding colors, depths, or edges. It must be emphasized here, however, that the perception of these attributes is not in an absolute sense.

The most powerful binary computer of today, however, cannot process the cognitive vision fields or cognitive information, which reflect the qualitative information that arises from human thinking, cognition and perception.

The fact that the natural sensors and mentation and cognitive processes are marvelously so efficient and effective poses a question for scientists and engineers: "can some of the function and attributes of the human sensory system, mentation and the cognitive process as well as that of the motor neurons be emulated on a machine?". This is a difficult questions to answer. However, it is felt that if we examine some of the mathematical aspects of the natural sensory system, human thinking process, and "hardware aspects" of the "neurons" which are the principle computing element in the brain, we may succeed to some extent in the emulation process of these marvelous attributes.

The mentation and cognitive activity of the brain, unlike the computational functions of the binary com-

puter, is based upon the notions of aggregation and relative grades of information that is acquired by the natural sensory system. The perception and action of these cognitive processes also appear to be in the form of relative grades. The physical attributes of a vision field may be measured precisely using physical means, however, during the process of perception, it appears in the form of relative grades, i.e., "the color is yellowish". This introduces a type of uncertainty that can be called a "formless uncertainty". However, we humans are able to convey a message at a greater speed and efficiency using this type of formless uncertainty as a formal basis. Recently, scientists have started to give a morphology to this amorphous uncertainty. In the past, the mathematicians have looked with disdain at this challenge and have chosen to avoid the modelling of mentation by devising theories that are unrelated to human perception and cognitive processes.

The advances in the framework of neural networks and the mathematics of fuzzy logic will lead toward the development of intelligent sensors and systems for engineering and medical applications. They will also lead to new studies in the important fields of cognitive information, neural networks, and cognitive-neural computing tools.

If we want to emulate some of the cognitive functions (learning, remembering, reasoning, intelligence and perceiving, etc.) of humans in a machine, we must generalize the definition of information and develop new mathematical tools and hardwares.

Indeed, biological processes have much to offer to system scientists and mathematicians to solve many practical problems in the world we live in today.

5. Perspectives

Recent progress in information-based technology has significantly broadened the capabilities and application of computers. Today's computers are merely being used for the storage and processing of numerical data (hard uncertainty and hard information). Should we not re-examine the functions of these computing tools in view of the increasing interests in subjects such as knowledge-based systems, expert systems and intelligent robotic systems and for solving problems related to decision and control? Human mentation acts upon cognitive information and the cognitive information is characterized by using relative grades: *"Although it is snowing, it is not very cold"*. Human mentation and cognition function by using fresh information (acquired from the environment by our natural sensors) and the information (experience, knowledge-base) stored in the biological memory.

Shannon's definition of "information" was based upon certain physical measurements of random activities in systems, in particular, in communication channels. This definition of information was restricted only to a class of information arising from physical systems.

If we want to emulate some of the cognitive functions (learning, remembering, reasoning, intelligence and perceiving, etc.) of humans in a machine, we have to generalize the definition of information and to develop new mathematical tools and hardware. These new mathematical tools and hardware must deal with the simulation and processing of cognitive information and soft logic.

Some of the nebulous attributes of the vision field, for example, can be emulated using the theory based upon fuzzy mathematics and fuzzy neurons. Many new notions, although still at an early stage, are springing up around the mathematics of fuzzy neural networks and, hopefully, we will be able to nurture some interesting studies in the not too distant future.

Bibliography

[1] R. E. Bellman, and L. A. Zadeh, "Decision Making in a Fuzzy Environment", *Management Science*, Vol. 17, 1970, B. 141—B. 164.

[2] M. Black, "Vagueness: An Exercise in Logical Analysis", *Philosophy of Science*, Vol. 4, 1937, 427—455.

[3] L. Brillouin, "Science and Information Theory", *Academic Press*, New York, (1956).

[4] R. C. Conant, "Law of Information which Govern Systems", *IEEE Trans. Systems, Man. and Cybernetics*, Vol. 6, 1976, 334—338.

[5] M. Conard, "The Lure of Molecular Computing", *IEEE Spectrum*, Vol. 23, No. 10, Oct. 1986, pp. 55—60.

[6] I. R. Goodman and H. T. Nguyen, "Uncertainty Models for Knowledge-Based Systems", *North-Holland*, New York, 1985.

[7] M. M. Gupta, "Fuzzy Automata and Decision Processes: The First Decade", *Sixth Trienniel World IFAC Congress*, Boston, Cambridge, August 24—30, 1975.

[8] M. M. Gupta, "On the Cognitive Computing: Perspective", in Fuzzy Computing: Theory, Hardware and Applications, *North Holland*, 1988.

[9] M. M. Gupta, "Cognition, Perception and Uncertainty", in Fuzzy Computing and Theory, Hardware and Application, *North Holland*, 1988, pp. 7—10.

[10] M. M. Gupta and G. K. Knopf, "The Percept: A Neural Model for Computer Vision", *IEEE Annual International Conference on Neural Networks*, San Diego, July 24—27, 1988, pp. 1—22.

[11] M. M. Gupta, "Fuzzy Neural Network in Computer Vision", *International Joint Conference on Neural Network*, Washington, June 18—22, 1989. Session: Vision pp. v. 1—v. 2.

[12] M. M. Gupta and G. K. Knopf, "Machine Vision with the Framework of Collective Neural Assemblies", *SPIE Conference on Intelligent Robots and Computer Vision*, November 5—19, 1989, Philadelphia, Paper # 1192—55.

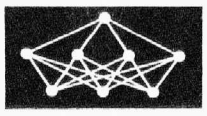
[13] M. M. Gupta and G. K. Knopf, "Theory of Edge Perception for Computer Vision Feedback Control", *Journal of Intelligent and Robotics Systems*, Vol. 2, 1989, pp. 123—151.

[14] M. M. Gupta, "Biological Basis for Computer Vision: Some Perspectives, *SPIES Conf. on Intelligent Robots and Computer Vision*", Nov. 5—10, 1989, Philadelphia, Paper # 1192—49, pp. 811—823.

[15] M. M. Gupta, "Neuro-Morphology of Biological Vision and Artificial Vision Systems", *Canadian Biomedical Engineering Conference*, Winnipeg, June, 1990.

[16] A. Kaufmann, "Introduction to the Theory fo Fuzzy Subsets", Vol. 1, *Academic Press*, New York, 1975.

[17] A. Kaufmann and M. M. Gupta, "Introduction to Fuzzy Arithmetic: Theory and Applications", *Van Nostrand Reinhold*, New York, 1985.



[18] M. M. Gupta, and G. K. Knopf, "Fuzzy Neural Network Approach to Control Systems", *Proceedings of the Int. Symposium on Uncertainty Modelling and Analysis* 1990, College Park, pp. 483—488.
[19] M. M. Gupta and J. Qi, "On Fuzzy Neuron Models", *IEEE Trans. Systems Man. and Cybernetics*, (Under Review), 1991.
[20] M. M. Gupta, "Information, Uncertainty and Intelligence" (Hard Logic to Soft Logic), *Proceedings of the Int. Symposium on Uncertainty Modelling and Analysis* 1990, College Park, pp. 614—618.
[21] G. J. Klir, "Where Do We Stand on Measures of Uncertainty, Ambiguity, Fuzziness and the Like", *Fuzzy Sets and Systems, Special Issue on Measure of Uncertainty*, Vol. 24, No. 2, November 1977, pp. 141—160.
[22] K. Kornwachs and W. von Lucadou, "Pragmatic Information as a Nonclassical Concept to Describe Cognitive Processes", *Cognitive Systems*, I, 1985, 79—84.
[23] F. M. Reza, "An Introduction to Information Theory", *McGraw-Hill*, New York, 1961.
[24] L. A. Zadeh, "Fuzzy Sets", *Information and Control*, Vol. 8, 1965, 338—353.

Literature Survey

The literature on neuroscience increases last few years extremely fast. At present some estimations of more than 20000 existing papers, conference and symposium talks, books and research reports are made. Evidently it is not possible to inform the readers about all the interesting publications, which currently appear. However, we would like to use the existence of the computer oriented Scientific Information System of the Institute of Computer and Information Science in Prague for to present here almost regularly the short survey of the last year records of this base.

Of course, the readers are asked for to be so kind and inform the Editors or the Institute about any publication, which they recommend to insert in this literature survey.

Abbott L. F.: Learning in neural network memories
NETWORK Vol. 1, 1990, No. 1, pp. 105—122
Key words: models of nets.

Abeles M., Vaadia E., Bergman H.: Firing Patterns of Single Units in the Prefrontal Cortex and Neural Networks Models
NETWORK, Vol. 1, 1990, No. 1, pp. 13—25
Key words: neurophysiology.

Adams J. L.: A Complementarity Mechanism for Enhanced Pattern Processing
Neural Computation, Vol. 2, 1990, No. 1, pp. 58—70
Abstract: The parallel ON- and OFF-center signals flowing from retina to brain suggest the operation of a complementarity mechanism. In the proposed mechanism, inhibition and excitation, both feedforward, coequally compete within each hierarchical level to discriminate patterns. A computer model tests complementarity in the context of an adaptive, self-regulating system.

Amit D. J., Parisi G., Nicolis S.: Neural potentials as stimuli for attractor neural network
NETWORK, Vol. 1, 1990, No. 1, pp. 75—88
Key words: models of nets

Anonym: A Window on the Mind
Time, 1990, No. 5/14, pp. 56
Key words: cultured neurons; child brain; laboratory culture.
Abstract: Short description of the results obtained at the John Hophins School of Medicine, Baltimore, Md. with laboratory cultivation of the culture of neurons from the 18-month old girl.

Barkan O., Smith, W. R., Persky G.: Design of Coupling Resistor Networks for Neural Network Hardware
IEEE Transactions on Circuits and Systems Vol. 37, 1990, No. 6, pp. 756—765
Abstract: Each coupling conductance no longer corresponds to a single coupling weight. Rather, each coupling conductance influences all the weights and therefore must be determined from all the desired weight values. In this paper, we present design equations for choosing appropriate coupling resistor values for use in conjunction with practical op amp neurons. We also give illustrated design examples.

Beneš J.: On Neural Networks
Kybernetika Vol 26, 1990, No. 3, pp. 232—246
Abstract: Typical neural network models representing abroad survey of actual trends are considered in an unifying way as Complexes subject to specific control systems, called Formators which may have a two-level arrangement and incorporate human operators. The importance of injected random noise in the process of organization is stressed. Potential applications of neural networks include their use as parts of formators for pattern classification in situational control. Prospective research areas are stated. One possible trend is to bring closer together the theory of neural networks and that of generalized cellular automata.

Braitenberg V.: Reading the structure of brain
NETWORK, Vol. 1, 1990, No. 1, pp. 1—11
Key words: neurophysiology.

Carpenter G. A., Grossberg S.: ART 3: Hierarchical Search Using Chemical Transmitters in Self-Organizing Pattern Recognition Architectures
Neural Networks, Vol. 3, 1990, No. 2, pp. 129—152
Key words: neural network; pattern recognition; adaptive resonance theory; transmitter; modulator; synapse; competition.
Abstract: A model to implement parallel search of compressed or distributed pattern recognition codes in a neural network hierarchy is introduced.

Clark D. M., Ravishankar K.: A Convergence Theorem for Grossberg Learning
Neural Networks, Vol. 3, 1990, No. 1, pp. 87—92
Key words: learning rules; long-term memory; linear threshold units.
Abstract: The prove of a convergence theorem for Grossberg's learning rule. This rule is used to update weights leading into a single processing unit by randomly choosing training sequences $x(x_0, x_1, x_2 \ldots)$ from a finite set Y of training patterns.

PATTERN RECOGNITION WITH HOMOGENEOUS AND SPACE-VARIANT NEURAL LAYERS**)

*H. Marko**)*

Abstract:
In the present article an attempt is made to understand pattern recognition of simple symbols (e.g. alphanumerical letters) by use of a system of homogeneous layers constructed in accordance with known properties of the visual system.

1. Introduction

As a mathematical tool to describe signal transmission and signal processing in layered neuronal systems a "System Theory of Homogeneous Layers" was developed by the author (Marko, 1969). Since then this theory has been applied to the human visual system to investigate signal detection for various stimulus patterns (Marko, 1981) and also for more technical problems of picture processing and pattern recognition (Platzer and Etschberger, 1972; Marko and Giebel, 1970).

In the case of multidimensional linear systems a considerable insight into system behaviour can be reached through the use of Fourier Transformation methods leading to multidimensional spectra. Nonlinear systems such as neural assemblies with thresholds can also be treated with this theory if for instance they have linear and homogeneous subpart receptive fields.

Recently the author proposed this theory to investigate stability conditions in layered systems with reverberating pathways leading to multistable states suitable for classifying perceptual sensations (Marko, 1984). This led to a space variant system formed by a learning process which again could be decomposed in a number of homogeneous systems.

In biological cybernetics the concept of modelling systems by a computable model or by the way of simulation plays an important role. To solve a certain task such as pattern recognition this model has to be optimized or best adapted to this task. This teleological aim seems to be controversial with biology in a Darwinian sense, where development is performed by the selection process. However, evolution theory shows that optimal or nearly optimal conditions develop as a consequence of survival.

Nevertheless, the biological constraints of such an optimization have to be maintained. Those are for the present problem mainly:

1. the system is to be composed of neurons, i.e. elements capable of summation of signals with an effect of thresholding;
2. the initial (hereditary) construction information for the system should be reasonably small (this is done by the structural and functional regularity of neuron layers mostly found and leading to the quality of homogeneity);
3. the system has to be adaptive to environmental experience in the sense of unsupervised learning or self-organization. In other words: there is no teacher or instructor to optimize the system for a specified task.

To our present knowledge the third condition is realized in a neuronal system by a simple mechanism proposed by Hebb already in 1949. The logical table of this mechanism is shown in *Fig. 1* together with a slightly modified version by Singer (see page 5 in this book).

The size of synaptic modification Δ is relevant for the plasticity of the system. Especially in more peripheral layers Δ has a finite value only for a certain adult age and then drops to zero with the result that the learned connections are irreversibly maintained.

A	C	M
1	1	$+\Delta$
0	1	0
1	0	0
0	0	0

a

A	C	M
1	1	$+\Delta$
0	1	$-\Delta$
1	0	0
0	0	0

b

Fig. 1. Logical table of
a) Hebb's mechanism, and (b) its modified version by Singer;
A: afferent signal (1 present, 0 absent), C: state of the neuron cell (1 firing, 0 silent), M: modification of synaptic transmission by a small amount of Δ (+ Δ increasing, — Δ decreasing, 0 no change).

*) Prof. Dr. Hans Marko, Ing. E.h.
Lehrstuhl f. Nachrichtentechnik
Institut f. Informationtechnik, TU München
Postfach 202420
D-8000 München 2, BRD
**) Presented at: Process in Structures for Perception and Action
DFG Deutsche Forschungsgemeinschaft, 1988

2. A General Scheme for Neuronal Signal Processing

Before concentrating on the problem of pattern recognition a hypothetical scheme of neuronal performance as a whole should be considered (see *Fig. 2*).

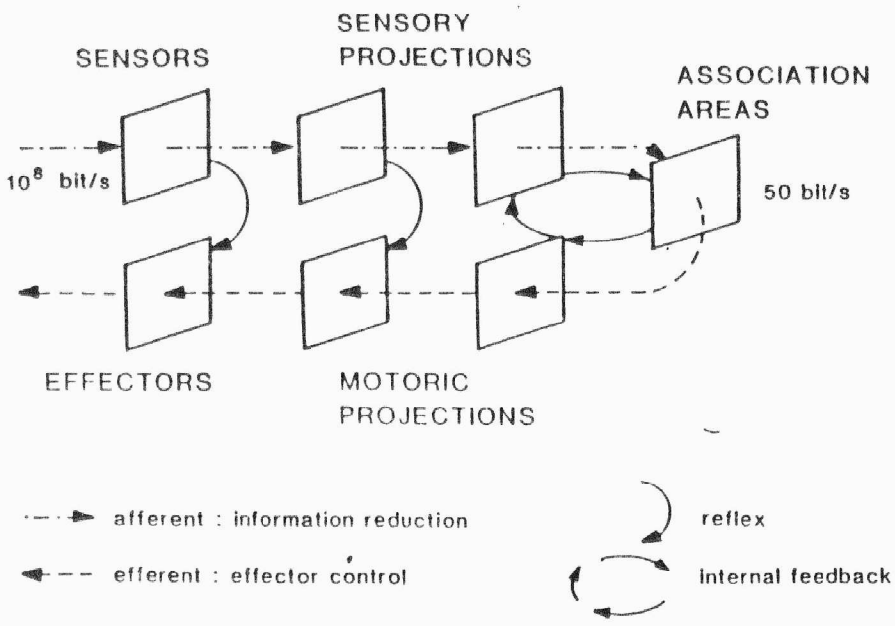


Fig. 2. Scheme of neuronal performance as a whole showing the afferent and efferent pathway and the internal interconnections (10^8 bit/s: estimate of sensory information flow mainly determined for the human visual system, 50 bit/s: upper limit of conscious information flow).

Of course, this scheme does not aim to explain the rather complicated neuronal connection scheme in detail and any layer shown here is only a representative of many parallel or cascaded neuronal areas.

However, some general considerations about the principal functioning of the system can be stated using this scheme:

1. The phylogenetic development proceeded from the left to the right, i.e. the receptors/effectors and the midbrain projections developed first before the cortical layers were added. This, of course, implies a further important constraint for the whole system. In other words: in biological systems a bottom up rather than a top down algorithm is realized.
2. The connection of the afferent to the efferent pathway is multifold. It includes direct and reverberating connections as well.
3. The sensory projection layers act as feature detectors (preserving retinotopy in the visual system). They are formed by experience using Hebb's conditioning algorithm. A possible structure for this performance will be discussed in the next section.
4. In order that the cortical layers may interpret the sensory signals correctly the proliferation process of the more peripheral layers should be completed before the beginning of a similar process within the cortical layers. Thus the establishment of learned specification (proliferation of the afferent branch) proceeds from the left to the right.
5. The plasticity or the ability to produce an irreversible (or long lasting) change in the interconnecting scheme is controlled by an unspecific instance (e.g., reticular system or amygdala and hippocampus to recent findings. This causes to change the size of Δ according to *Fig. 1*. The central control of learning was omitted in *Fig. 2*).
6. There are reverberating connections especially in the cortical area which can lead to instabilities or maintained oscillations. The preferred spatio-temporal patterns of such oscillations will be produced by neuronal assemblies strongly interconnected due to Hebb's principle and more or less correlated to specific sensory signals.
7. Consequently, the perception process is not purely passive (or causal) but partly active. This means that the corresponding areas are not only excited by the sensory signals but also by the reverberating processes. This self-reflexive action of the brain is also supported by the finding that only 1 % of the cortical connections are thalamic (sensory) afferents, while the larger part issues from other cortical regions (Braitenberg, 1978).
8. As a whole the system is likely to produce a stochastic process which tends to predict the incoming sensory signals, so that the incoming information (the prediction error) is minimized. Such a performance would highly increase the survival chance in a varying environment.
9. The information produced and delivered within the efferent pathway is partly supplied by the internal reverberating process and partly by the incoming sensory signals. This corresponds, in terms of information theory, to a dependent information source (Marko, 1966, 1983).

3. The Formation of Sensory Layers as Feature Detectors

In a pattern recognition system the first processing stage consists of feature detectors. Feature detectors are used to extract relevant information from the sensory signal suitable for the next stage, the classifier. In the visual system of vertebrates feature detectors are found in the striate cortex (area 17, 18, 19).

The main functional units are directional and/or orientational filters found by Hubel and Wiesel (1962) and arranged in a columnar structure (see also Singer's contribution to this book). They are to be found in area 17 below and above the input layer (layer 4). From psychophysical detection experiments the existence of orientationally selective filters is also evident. They are described by elliptical receptive fields (Marko, 1981). The eccentricity parameter ε of these fields is suspected to be in the order of 2 to 8 (see Elsner, in this book). More complex feature detectors (complex cells) and motion sensitive cells have been found in the higher stages of striate cortex. A very important constraint expressing some functional and structural regularity that has always been observed, namely the retinotopy, must be considered. This means that cells having equal functional properties are arranged as

a homogeneous system coupled to the original picture or the retinal excitation respectively. Cells within the same neuronal layer but with different functional properties are arranged in an interleaving fashion, therefore leading to the columnar structure. The distance of such columns, with the same functional property, is about 0,8 mm or the cortex surface (the cortex having a thickness of about 2,5 mm). Consequently, a "hypercolumn" which includes all functional properties forms a cylinder with a diameter of 0,8 mm and a height of 2,5 mm, having a volume of about 1 mm^3. It contains about 10^5 neurons with about $2 \cdot 10^8$ synapses.

According to Braitenberg (1978 and personal message) the mean lengts of inhibitory connections correspond to a radius of 0,2 mm (estimated) and that of excitatory connections to a much larger radius of about 1 mm. It should be noted that $^3/_4$ of the cortical neurons are pyramidal cells and $^1/_4$ are stellate cells, the latter probably being responsible for the inhibition of the pyramidal cells. The lateral inhibition found within a small area and the excitatory connections extending over a much larger area lead the author to propose the NIFE structure (NIFE: near inhibition, far excitation) as a model to explain the columnar and retinotopic arrangement of neurons with the same functional property within the cortex.

Given a strong circular inhibition between adjacent neurons, then, according to *Fig. 3*, only sufficiently distant neurons could be excited simultaneously. They form a regular hexagonal pattern. The intermediate neurons are inhibited and stay inactive despite some excitation from outside. For an unspecific or diffused excitation of the whole layer it is by hazard which of the interleaved hexagonal patterns of neurons may be excited. The the whole layer forms a multistable system (a "multiflop"), the states being characterized by the common excited neurons in a regular structure. Assuming now that the excitatory coupling between commonly excited neurons (that means over a larger area) is modified by a learning process using Hebb's lagorithm, this class of neurons may act as feature detectors. For instance, if lines of a certain orientation are presented, the first (by chance) excited class will establish supporting connections within this orientation. Subsequently they will more likely to be excited in all case where patterns containing lines or edges are presented. If the Hebb-Singer algorithm is used (see *Fig. 1*) the excitatory connections perpendicular to the orientations presented will diminish. This leads to the formation of an elliptical receptive field needed for orientational filters.*)

If lines or edges of different orientations are presented next, another neuronal population (with its original circular receptive field) will be preferred and

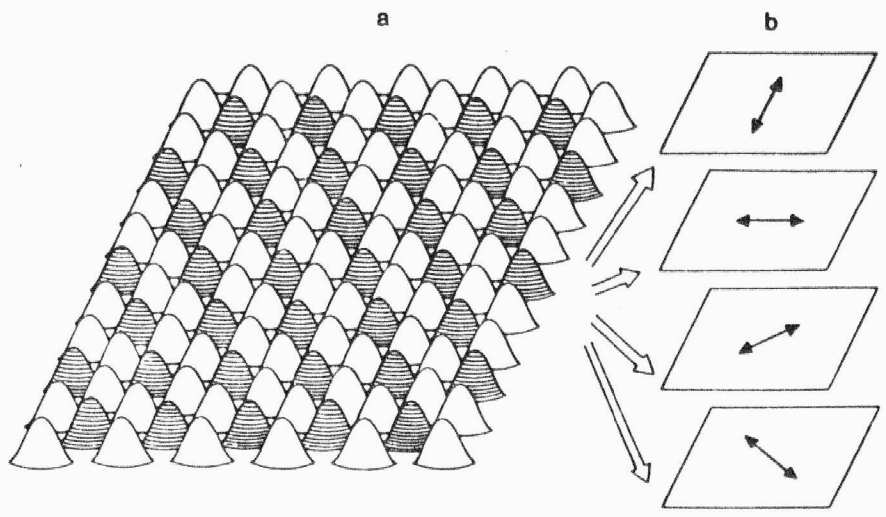


Fig. 3. NIFE — structure of neuron layers (NIFE: near inhibition, far excitation); a) space variant layer with double space periodicy (a commonly excited substructure forming a hexagonal pattern is indicated), (b) decomposition of the neuronal layer shown in (a) into a number of homogeneous layers after completion of the learning process.

excited and adapted with the learning algorithm to the new orientation. In this way the functional specification of the interleaved neuron populations is formed by experience. As the number of different populations is limited, only the most frequently occuring patterns have a chance to be learned. These seem to be lines or edges (which are also transformed to lines via spatial differentiation in the afferent sensory channel).

It should be noted that the number of homogeneous substructures depends on the area of the strong inhibition. For *Fig. 3* it was assumed that only the adjacent neurons were strongly inhibited leading to a number of 4 substructures. If the inhibition area extends to the next adjacent neurons 7 structures are possible and so on.

A calculation by Neumann (1981) shows that the NIFE structure is very resistant to parameter variations. *Fig. 4* shows the distance between commonly excitable neurons assuming Gaussian inhibition $a \cdot g(x)$ with radius x_0 and strength a and common uniform excitation with strength E.

Activity $\overline{A}$ of any neuron within the layer is assumed to be.

$$\overline{A} = A \text{ if } E - \Sigma I > S$$
$$O \text{ if } E - \Sigma I < S$$

where S is a threshold and ΣI is the sum of inhibition potentials of the other active neurons. The inhibitory potential of one active neuron to another neuron with distance x is assumed to be

$$I = A \cdot a \cdot g(x) \text{ with } g(x) = e^{-x^2/x_o^2}$$

Selfinhibition is not admitted, i.e., $g(x) = 0$ for $x = 0$. The initial unlearned state is considered to be without any excitatory connections. The maximum distance in *Fig. 4* corresponds to a condition in which an intermediate neuron cannot be excited due to inhibition of the neighbouring excited neurons. The minimum distance corresponds to a condition in which the

*) It might be noted that the spatial bandpass characteristic of the visual channel is necessary for this feature forming process because the spatial differentiation assumes the simultaneous excitation only along a line if an edge is presented.

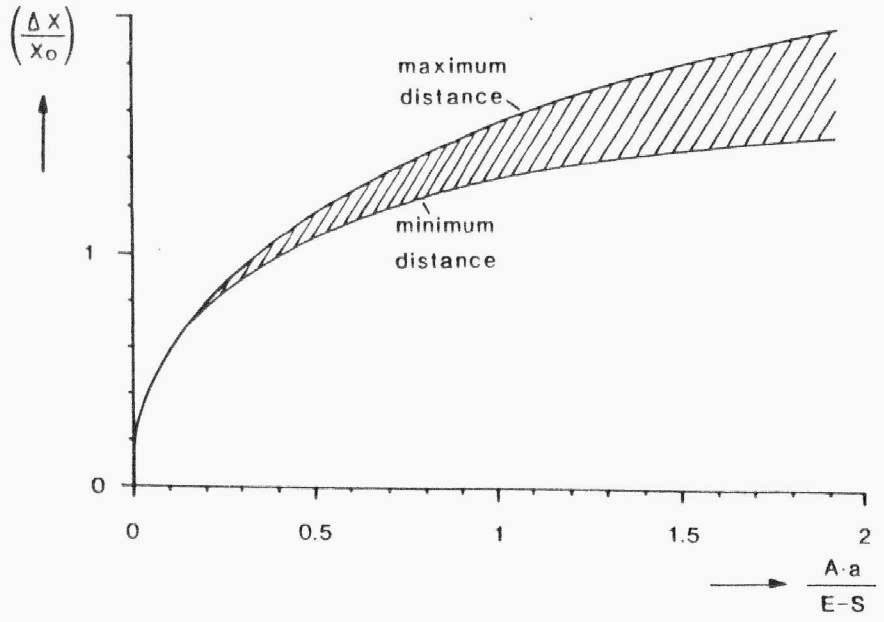


Fig. 4. NIFE-structure. Distance of commonly excitable neurons in a homogeneous layer structure with a Gaussian inhibitory surrounding.

mutual inhibition of commonly active neurons is smaller than the excitation.

Obviously the distance ΔX between commonly excitable neurons (or the diameter of the hypercolumn) is very stable against parameter variations.**)

After the interleaved hexagonal units are organized by creation or modification of excitatory connections according to a learning process, the layer as a whole is no longer homogeneous. The learning process has altered the original homogeneous structure into a space-variant structure with a double-periodic space variance. This, however, can be decomposed into a number of again homogeneous layers as shown in *Fig. 3*, now having different receptive fields. (For a mathematical treatment the excitation function which is present only at discrete points has to be smoothed over the whole area according to the spatial sampling theorem.)

4. The Recognition of Simple Patterns with a Homogeneous Layer System

Recognition means classification of a set of different patterns into a number of specified classes. Here we used handprinted letters and figures (alphanumeric characters) written by many people in a highly varied manner (a total of 36 classes). As a processing system we used a system of homogeneous layers with feature detectors according to those found in the visual system. The scheme is shown in *Fig. 5*. In the first stage four orientational filters have been used. In a successive stage more complex features are formed, such as crossing points, branching points, end points etc.

**) It might be of interest that a similar hexagonal array for the structural growth (in constrast to the function) has been obtained by Meinhardt (1982, 1986) using lateral inhibition accomplished by a diffusion process. It seems that the same mathematical scheme of a homogeneous coupled layer is valid.

From the feature extracting process about 500 single points belonging to all layers (except the picture plane) are finally obtained. In the learning phase of the system only those points exceeding threshold are further regarded. With the incremental learning procedure the coupling coefficients to the corresponding output layer in the last stage were raised, and to the other layers they were lowered in small steps until the right classification was obtained with a certain security margin. This procedure converged after about 1 000 presented patterns written by about 20 persons. This means that a complete separability of the used pattern set was achieved. More detailed information on the system design can be found in Marko and Giebel (1970) and Giebel (1975).

The recognition process runs as follows: The input is provided by writing with a light pen or by a scanning process of the handprinted letter. Next the picture is framed with 16×24 picture elements. Then filtering in four directions (horizontal, vertical, and two diagonal directions) is performed. Only elementary points with a signal above the threshold are indicated. Higher features like angles, curvatures, and endpoints or crossing points and branching points are extracted. For these higher features, the framme might be much broader than for the directional filters. Finally, weighting according to the learned scheme is done and the result is indicated.

The performance of this system may be judged by the improvement of the error rate due to the multistage operation of the system. First it should be stated that the error rate for the set of handwritten characters used in the learning phase of the system was zero; i.e., the learning process converged totally, leading to complete separability. When another set of handwritten characters not used in the learning phase but written by the same writers was presented, a recognition

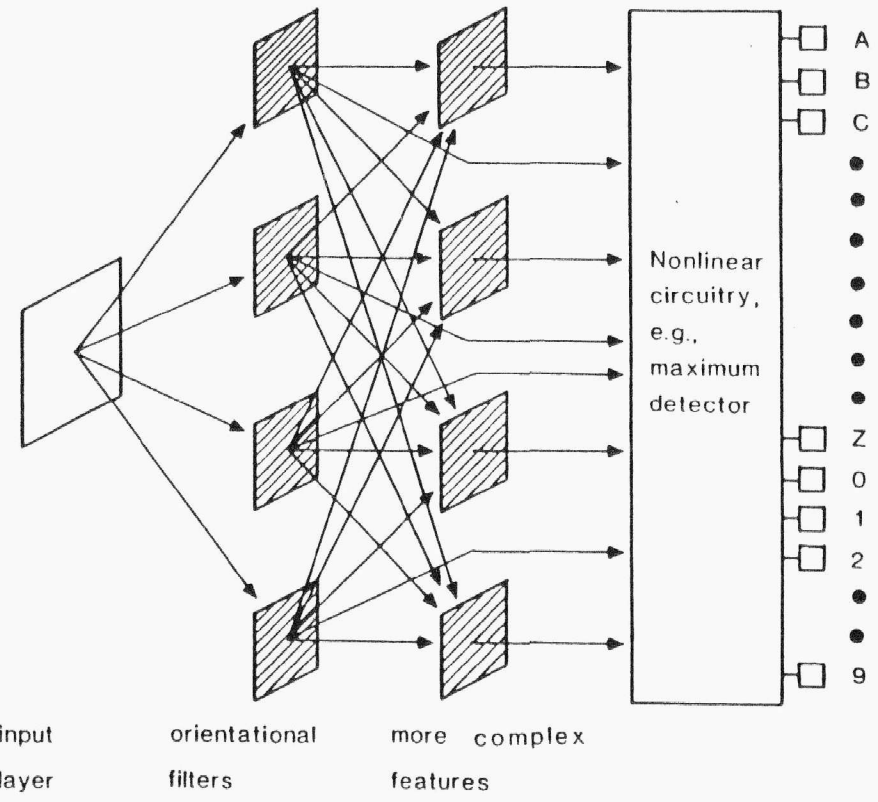


Fig. 5. System of homogeneous layers used in pattern recognition for the classification of hand-printed characters.

error of 3 percent was found. The same character set would lead to an error of 20 percent with one-stage operation only, which corresponds to the Bayesian classifier applied to the image plane. With four directional filters the error rate would drop to 10 percent, and finally to 3 percent with all the features used. It seems therefore that the hierarchical structure of the homogeneous layer system is especially suitable for recognizing new patterns not used in the learning process. This ability of generalization common to living creatures seems to be an advantage of the biological approach, which uses features similar to those encountered in the nervous system.

The contribution of the different features to the final quantity to be maximized by the learning procedure was measured by the average product of the absolute weights and the frequency of using the corresponding features. The result is as follows: 21 percent for the horizontal direction, 20 percent for the vertical direction, 9 percent for the positive diagonal slope, 14 percent for the negative diagonal slope, 4 percent for the curvatures, 11 percent for ending points, 8 percent for right angles, 1 percent for intersection, and 19 percent for other topological features. These figures are mean values taken over all 36 pattern classes.

In order to evaluate the sensitivity of the system to pattern variations the rotation of patterns has been investigated by Tilgner (1982). His results are shown in the appendix.

The comparison between man and this system of the sensitivity under rotations reveals similarity. As compared to the Bayes-classification applied to the pixels of the image plane without feature layers the homogeneous system shows a considerable improvement. It might be concluded that a homogeneous system using many feature filters in parallel provides the invariance capability necessary to recognize highly variable patterns.

5. Discussion

The system described above is of course a simplified version of the pattern recognition abilities of the visual system. Firstly, in reality there are many more feature filters than those used in the simulation. For instance, in the striate cortex not only orientational filters but also receptive fields of different size leading to feature filters of different spatial frequencies were found. According to psychophysical findings different orientational filters with different spatial frequencies have to be assumed (see Elsner, page 79). Secondly, more stages are certainly involved, especially if word recognition is required. They could work principally in the same manner, using the NIFE structure with its learning capabilities. Thirdly, the technical recognition system described here works not sequentially but in a parallel mode. Time is only needed to build up the excitations but not provide a sequence of different excitations.

In the real system a time sequence is generated as a consequence of reverberating connections shown in *Fig. 2*. It is a spatio-temporal pattern of excited neuron assemblies, correlated with the sensory input or its extracted features, respectively. This correlation is established by learning based on Hebb's principle. In this way the process will predict the most probable input even if the real input fails to exist, or if the sensory signals are disturbed or partly omitted. Perception is then an active process synchronized by sensory signals. Synchronization here is equivalent to the recall of information in an associative memory.

Generally, perception is an internal process correlated with the outer world in a variable way. The degree of this correlation or synchronization may determine whether the process is more perceptive (reception of information) or more reflective (production of information). Thus, a meaningful information exchange between the living being and its environment is possible due to the highly interconected layers of the adaptive neo-cortex.

APPENDIX

MAN-MACHINE COMPARISON OF ROTATIONAL INVARIANT CHARACTER RECOGNITION

Ralph D. Tilgner)*

1. Introduction

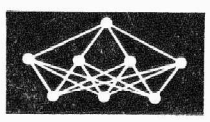
In recent years, automatic pattern recognition has reached a high level in the field of character recognition. Problems due to pattern variations by projective transformations (size, translation or rotation) still remain, even though a variety of preprocessing techniques were developed to improve recognition performance under these transformations (Nagy and Tuong, 1970; Marko, 1973; Güdesen, 1976).

This contrasts to the high capability for invariant recognition of the human visual system. In the following, a recognition system which was proposed in earlier works as an analog model to some structures of the human visual system (Marko and Giebel, 1970; Marko, 1974) is inspected to study the effect of pattern rotation.

2. Data Base

Patterns were drawn from a standard data base which provides 36 classes of unnormalized alphanum-

*) Presented on 6th ICPR



eric handprinted characters, about 3 500 each class, quantized to 32×40 binary elements (Krause and Bleichrodt, 1973; Suen, 1980). For our investigation the number of classes was reduced to 10 (only numerals), each class containing 2 000 numerals, whereby the first 1 000 of each class represents the training set and the second 1 000, the test set. To focus on rotation as a source of classification errors, heavily degraded numerals were eliminated.

In a second step the numerals were rotated by varying a test angle R_t. Test angles R_t were determined with an increment of 5° within a rotation range RR by means of an equally distributed pseudo random sequence. Rotations were performed around the center of gravity. To avoid pattern degradations the used sin/cos transformation was calculated with floating point precision within the 32×40 quantization of the original data base. The rotated numerals then were reduced to 16×16 binary elements with a best fit normalization of size (see *Fig. A.1* for examples).

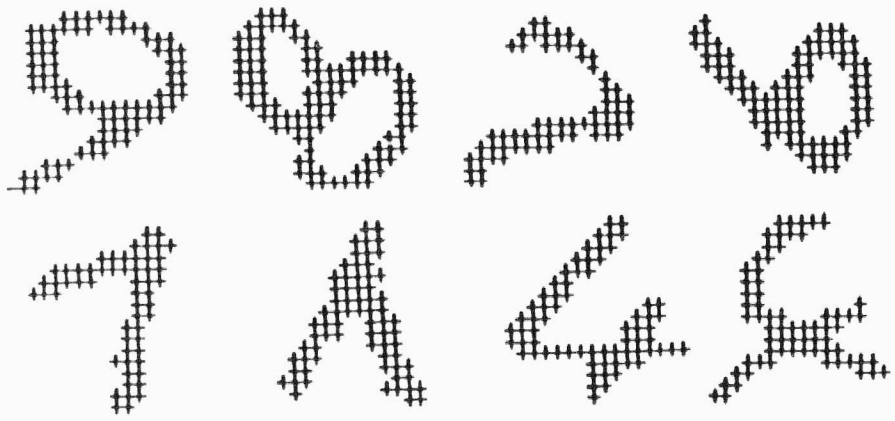


Fig. A.1. Sample of numerals of data set RR60 (upper row), examples of form ambiguities after normalization by second-order moments (lower row).

In this way 4 versions of training and test sets with the following embedded ranges of rotations were established:

$RR0$: $R_t = 0°$ $RR30$: $-30° \leq R_t \leq +30°$

$RR60$: $-60° \leq R_t \leq +60°$ $RR90$: $-90° \leq R_t \leq +90°$

In all cases writer-specific pattern orientations were left unconsidered.

3. Recognition Processes

A. Bayesian Classifier

As a statistical reference a well known simple Bayesian decision criterion was used. Sampling a pattern $s(x, y)$ with a spatial quantization of 16×16 elements leads to a pattern vector $\bar{s}$ in signal space with $N = \dim(\bar{s}) = 256$ components. If components of $\bar{s}$ are assumed: I.) to be binary, and II.) to be statistically independent (whereby the latter obviously is not true for patterns like numerals), the following decision criterion e^c can be derived:

$$e^c = \sum_{i=1}^{N} s_i \log\left(p_i^c(1)\right) + \sum_{i=1}^{N} (1 - s_i) \log\left(1 - p_i^c(1)\right)$$

$p_i^c(1)$ denotes the conditional probability for the i-th component of $\bar{s}$ to be equal to 1 if $\bar{s}$ is a member of class c. An unknown pattern $\bar{s}$ is then assigned to a class c according to the maximum of e^c (Ullman, 1973). Four different estimations of $\bar{p}^c(l)$ were calculated from the learning portion of the 4 versions of rotated data sets described in 2.

B. Homogeneous Layer System

Fig. A.2 shows a scheme of the simulated system designed for character recognition. The input activity of the sensor layer in stage 0, representing the quantized input pattern $s(x, y)$, is directed to the first processing stage. This stage provides four independent directional filters extracting horizontal, vertical and oblique line elements. These directional filter processes are performed by homogeneous spatial convolutions with masks of 3×3 elements shown in *Fig. A.2*. A subsequent threshold operation reduces the continuous convolution result to binary values.

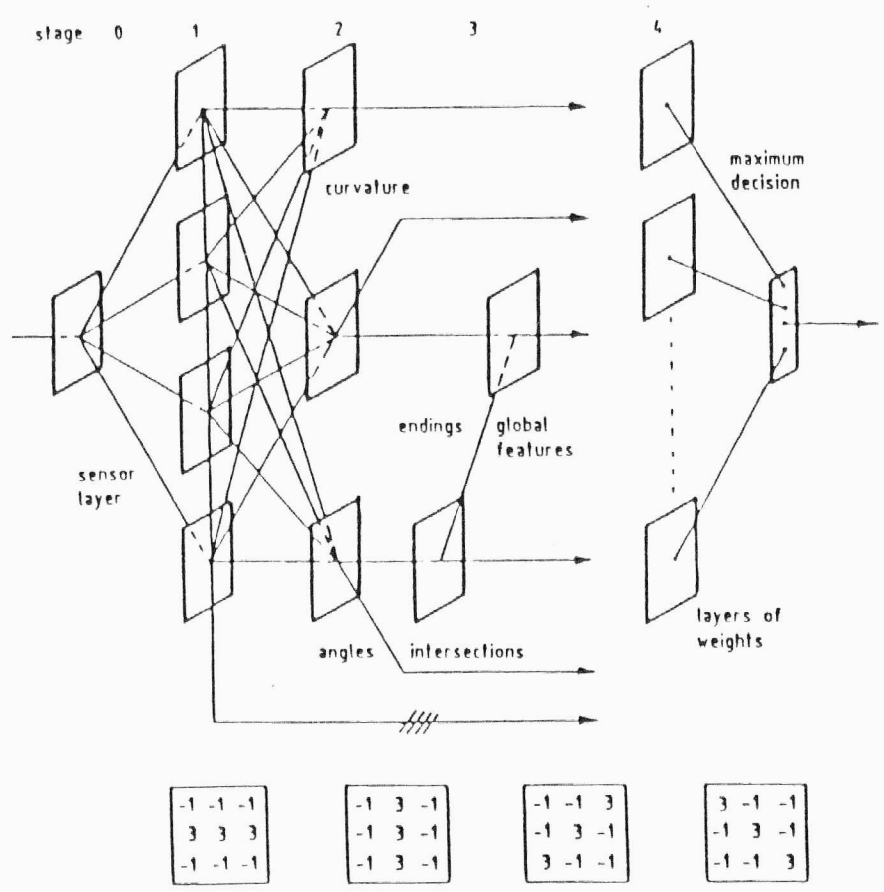


Fig. A.2. Structure of the hierarchical layer system and 3×3 convolution masks applied for line filtering in stage 1 (used threshold: 4).

By logically combining 4 neighboured line elements to 1 element within each layer, the quantization is reduced by a factor 4. The resulting elements then are interpreted as components of the feature vector $\overline{f1}$. Hereby the number of components of $\overline{f1}$ equals that of the sensor layer, but components of $\overline{f1}$ are influenced by an area of 4×4 elements of the sensor layer.

The activity of the following stage 2 is calculated in a similar way by homogeneous filtering of the output

of stage 1. In our application 5 filter processes lead to features representing curvatures, right angles and endings of 3 different orientations. Again by logically combining 4 neighboured elements to 1 within each layer the dimension of the resulting feature vector $\overline{f2}$ is reduced to dim $(\overline{f2}s) = 5/4$ dim $(\overline{f1})$, whereby one component now covers an area of 12×12 elements of the sensor layer. Finally, the activity of stage 3 is derived by combining angular features of stage 2 to 6 different types of intersections. Combinations of endings and these intersections without any regard to their local position yield some global features.

With respect to the 16×16 quantization of the used data sets, a combined feature vector $\bar{f}$ composed of the features $\overline{f1}$, $\overline{f2}$, $\overline{f3}$ consists of dim$(\bar{f}) = 256 + 80 + 56 = 392$ elements.

To categorize an unknown pattern, the last stage provides a set of class-specific layers which represents weights $\bar{w}^c$ for every component of $\bar{f}$. A decision criterion e^c can be derived by the dot product:

$$e^c = \bar{w}^c \cdot \bar{f}$$

The unknown pattern then is assumed as member of class c according to the maximum of e^c. Weight vectors $\bar{w}^c$ were determined during the training phase by an iterative fixed increment procedure with dead zones (Ullman, 1973; Rosenblatt, 1962).

This system with a slightly increased quantization of the input stage (16×24) leading to a feature vector of dim$(\bar{f}) = 579$ components, yields recognition performance comparable to other systems (Krause and Bleichrodt, 1973; Suen, 1980). *Tab. A.1* lists substitution rates for the 10 class (numerals) and 36 class (alphanumerics) problem obtained for the original data base described in 2. Training sets consist of 2 000 characters each class, test sets of 1 000/class. An implementation on a PDP 11/45 minicomputer, mainly written in FORTRAN (only the convolutions of stage 1 and the dot products of the last stage were written in machine language), requires about 9 sec to recognize one character.

Table A.1. Substitution rates of the homogeneous layer system, no rejection was permitted.

data set	substitution rate/% 36 classes	10 classes
training set	1.9	.09
test set	3.1	.7

C. Normalization of Rotation by Second-order Moments

Higher moments of a given pattern $s(x, y)$ were proposed for mapping this pattern into a normalized form (Hu, 1962; Amari, 1978; Teague, 1980). To compensate an assumed misorientation of $s(x, y)$ second-order moments yield the orientation of the principal axis of $s(x, y)$, which can be used as a compensative angle R_n to rotate $s(x, y)$ into a normalized position:

$$\text{with } g_{11} = \int\int s(x, y)xy\, dxdy,$$
$$g_{02} = \int\int s(x, y)y^2\, dxdy,\ g_{20} = \int\int s(x, y)x^2\, dxdy$$
$$\text{follows: } R_n = 1/2 \tan (2g_{11}/(g_{20} - g_{02}))$$

To prove the efficiency of this rotation normalization done by a preprocess in signal space, new versions of 4 data sets with rotate (by R_l) and then rotationnormalized numerals were calculated and then processed by the layer system. First test angles were applied within the ranges $RR0$, $RR30$, $RR60$ and $RR90$ and then compensated by $-R_n$. Rotations by R_l and by $-R_n$ were performed with floating point precision within the 32×40 quantized signal space. Resulting patterns then were reduced to 16×16 quantization constituting the rotation compensated data sets.

D. Human Visual System

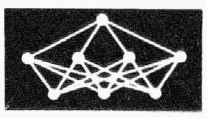
To obtain comparable information about the performance of the human visual system, recognition experiments with 3 subjects were accomplished. In 4 sessions subsets of the 4 rotated data sets described in 2. had to be classified. Of course, subsets were extracted from the test set portion, each subset containing 3 000 numerals (300/class). Numerals were presented in a random sequence at the center of a HP 1310A display with a size of 2.5×2.5 cm. Presentation time was 600 msec, viewing distance 1.5 m, subject's head was fixed in vertical position by a headrest.

4. Comparison of Recognition Performance

Fig. A.3a compares substition rates p_s obtained for the test set portion of the 4 rotated data sets.

Performance of the Bayesian classifier is strongly dependent on pattern rotation with a nearly linear increase of errors starting with $p_s(RR0) = 4.39\%$ to $p_s(RR90) = 24.8\%$. The layer system copes much better with rotation and approaches the performance of the human visual system, though substitution rates of the subjects are better for all examined rotations.

The layer system combined with the normalization preprocess using second-order moments achieves even better results than the subjects, if R_l varies within $RR30$, $RR60$ and $RR90$, but for small rotations ($RR0$, only writer specific variations) the normalization process leads to an increase of errors. This is mainly caused by a weak correlation between the calculated orientation of the principal axis and the actual vertical position of numerals 1, 4, 6, 9 (see examples in *Fig. A.1*). In *Fig. A.3b* the fraction of performance related to each stage of the layer system is studied. If classification is carried out in signal space, dependence on rotation comes near to that of the Bayesian classifier,



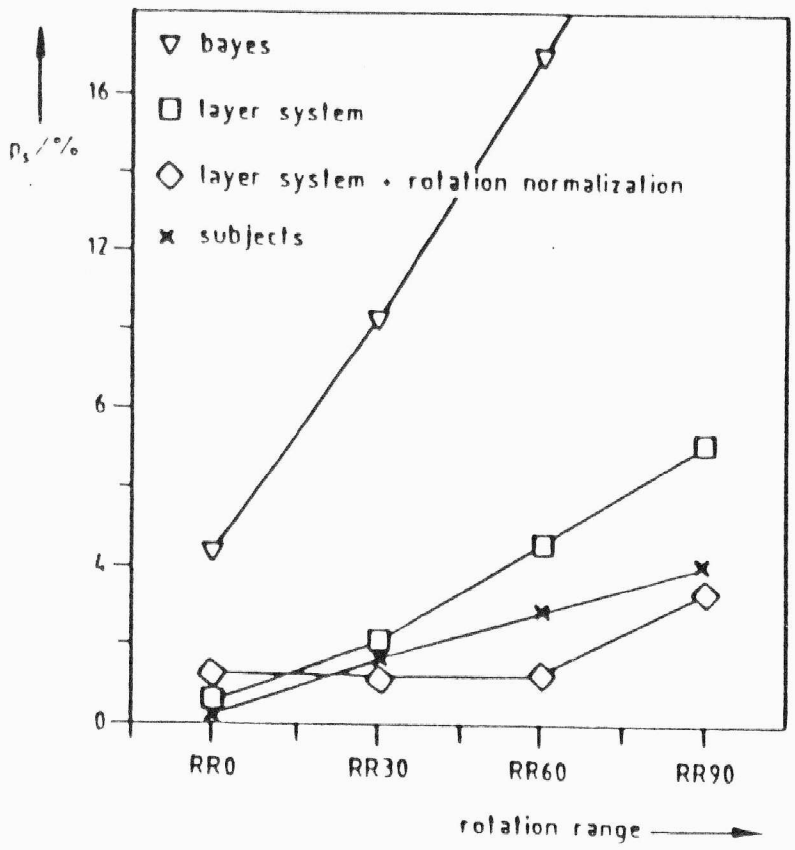


Fig. A.3. Substitution rates p_s a function of rotation range, no rejection was permitted (missing values: a) $p_s(\nabla, RR90) = 25.4\%$, b) $p_s(x, RR90) = 27.9\%$).

though a little better results are given for $RR0-RR60$ but worse for $RR90$. Using the line features of stage 1, an evident increase of performance for every rotation range can be achieved. Including the features of stage 2, substitution rates once more can be reduced for $RR30-RR90$; the performance comes near to that of the complete system except for $RR90$. Of course, in this study separate learning procedures were performed for every subsystem.

5. Conclusions

In contrast to the Bayesian classifier which neglects statistical dependencies of neighboured pixels, stepwise linear combinations of neighboured pixels, as done by the applied convolutions of the layer system, followed by a subsequent threshold operation transfer these dependencies into feature space. This concept leads to good results compared to those of subjects and verifies in some respects the hierarchical layer system as an analog model for the visual recognition process.

The inspected normalization process yields an improvement only for large rotations. Additional errors must be taken into account for small variations of rotation. Thus, to approach rotational-invariant recognition such a normalization in signal space cannot be assumed as a general preprocess.

References

[1] Amari S. (1978) Feature spaces which admit and detect invariant signal transformations. In *Proc. of the 4th Int. Joint Conf. on Pattern Recognition* (Tomaru Y., ed.), pp. 452—456. Kyoto University Press, Kyoto.

[2] Braitenberg V. (1978) Cortical architectonics: general and areal. In *Architectonics of the Cerebral Cortex* (Brazier A. B. and Petsche H., eds.), pp. 443—465. Raven Press, New York.

[3] *Giebel H. (1975) Optimierung mustererkennender Schichtstrukturen durch Merkmalssynthese. *Nachrichtentech. Z.* **28,** 413—417.

[4] Güdesen A. (1976) Quantitative analysis of preprocessing techniques for the recognition of handprinted characters. *Patt. Recogn.* **8,** 219—227.

[5] Hebb D. O. (1949) *The organization of behaviour.* Wiley, New York.

[6] Hu M. (1962) Visual pattern recognition by moment invariants. *IRE Trans.* **IT-8,** 179—187.

[7] Hubel D. H. and Wiesel T. N. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual vortex. *J. Physiol. Lond.* **160,** 106—154.

[8] Krause P. and Bleichrodt H. (1973) Experiments on direct input and recognition of handwritten digits and handwritten letters with computers (translated by the Post Office Res. Dept., London). From *IITB-Mitteilungen 1972/73*, pp. 9—15. Fraunhofer-Institut, Karlsruhe.

[9] Marko H. (1966) Die Theorie der bidirektionalen Kommunikation und ihre Anwendung auf die Nachrichtenübermittlung zwischen Menschen (Subjektive Information). *Kybernetik* **3,** 128—136.

[10] *Marko H. (1969) Die Systemtheorie der homogenen Schichten. *Kybernetik* **5,** 221—240.

[11] *Marko H. (1973) Space distortion and decomposition theory — A new approach to pattern recognition in vision. *Kybernetik* **13,** 132—143.

[12] *Marko H. (1974) A biological approach to pattern recognition. *IEEE Trans.* **SMC-4,** 34—39.

[13] *Marko H. (1981) The z-model — a proposal for spatial and temporal modeling of visual threshold perception. *Biol. Cybern.* **39,** 111—123.

[14] *Marko H. (1983) Information theory and communication theory. In *Biohysics* (Hoppe W., Lohmann W., Markl H., Ziegler H., eds.), pp. 788—794. Springer, Berlin, Heidelberg.

[15] *Marko H. (1984) Grunlagen der Information und Kommunikation. In *Information und Kommunikation — naturwissenschaftliche, medizinische und technische Aspekte* (Karlson P., Bettendorf G., Marko H., Sachsenmaier W., Schneider D., Staab H. A., Gibian H., eds.) pp. 71—86. Wissenschaftliche Verlagsgesellschaft mbH, Stuttgart.

[16] *Marko H. and Giebel H. (1970) Recognition of handwritten characters with a system of homogeneous layers. *Nachrichtentech. Z.* **23,** 455—459.

[17] Meinhardt H. (1982) **Models of biological pattern formation.** Academic Press, London.

[18] Meinhardt H. (1986) Formation of symmetric and asymmetric structures during development of higher organisms. *Comp. Math. with Appls.* **12B,** 419—433.
[19] Nagy G. and Tuong N. (1970) Normalization techniques for hand-printed numerals. *Com. A.C.M.* **13,** 475—481.
[20] Neumann G. (1981) Zur Organisation der intracorticalen Verschaltung in der Sehrinde. Doctoral Thesis, Technische Universität München.
[21] *Platzer H. and Etschberger K. (1972) Fouriertransformation zweidimensionaler Signale. *Laser und Elektro-Optik* **4,** 39—45, 43—49.
[22] Rosenblatt F. (1962) *Principles of neurodynamics.* Spartan Books, Washington/DC.
[23] Suen C. H. (1980) Automatic recognition of handprinted characters — state of the art. **Proc. IEEE 68,** 469—487.
[24] Teague M. R. (1980) Image analysis via the general theory of moments. *J. Opt. Soc. Am.* **70,** 920—930.
[25] *Tilgner R. D. (1982) Untersuchungen zur Rotationsinvarianz im visuellen System und Vergleich zum technischen Zeichenerkennungssystem. Doctoral Thesis, Technische Universität München.
[26] Ullman J. R. (1973) *Pattern recognition techniques.* Butterworth, London.

* Publications from research supported by the Deutsche Forschungsgemeinschaft through the Sonderforschungsbereich "Kybernetik".

Neurocomputer Companies

Artificial neural networks extend in last few years from laboratories into commerce and industry. At present there are known several tenths of companies producing and seling the neuro-software and/or -hardware tools and services. In this section of our Journal we shall inform the readers substantially about the adresses of some of these companies.

Ab Tech Corporation
700 Harris Street
Charlestonville VA 22901
Tel.: (804) 977 0686

Abbot, Foster & Hausermann
44 Montgomery, Fifth Floor
San Francisco CA 94014, USA
Tel.: (415) 955-271

AI Ware, Inc.
11000 Cedar Ave. , Suite 212
Cleveland OH 44106, USA
Tel.: (216) 421-2380

American Interface Corporation P. O. Box 297
Zurich 8027, Switzerland

California Scientific Software
160 East Montecito, Suite E
Sierra Madre CA 91204, USA

Cognitive Software, Inc.
703 East 30th St.
Indianapolis IN 46205, USA

DAIR Computer Systems
3440 Kenneth Dr.
Paolo Alto CA 94303, USA

Excalibur Technologies
2300 Buena Vista SE
Albuquerque NM 87106, USA

Micro Devices
5695 Beggs Rd.
Orlando FL 32810, USA

Nestor, Inc.
1 Richmond Sq.
Providence RI 02906, USA

Neural Systems Incorporated
2827 West 43rd Avenue
Vancouver, British Columbia V6N 3H9, Canada
Fax: (604) 263-3667

NeuralWare, Inc.
103 Buckskin Court
Sewickley PA 15143, USA

Neurix, Inc.
1 Kendall Sq. , Suite 2200
Cambridge MA 02139, USA

Olmsted & Watkins
2411 East Valley Pkwy. , Suite 294
Escondido CA 92025, USA

Oxford Computer
39 Old Good Hill Rd.
Oxford CT 06483, USA

SAIC
Mall Stop 71, 10260 Campus Point Drive
San Diego CA 92121, USA

Syntonics Systems, Inc.
20790 Northwest Quail Hollow Dr.
Portland OR 97229, USA
Tel.: (503) 293-8167

TRW Military Electronics & Avionics Div.
One Rancho Carmel
San Diego CA 92128, USA
Tel.: (619) 592-3482

In this section we present a short overlook on the companies which were present at the CeBIT'91Hannover fair, March 13 — 20, 1991 and which have expressed their activity in neurocomputing and neurocomputers. These are:

Adaptive Solutions
USA, Beaverton, OR 970006

adcomp Datensysteme
Germany, D-W 8025, Unterhaching

AGFA-GEVAERT
Germany, D-W 5090, Leverkusen 1

Apple Computer
Germany, D-W 8000, Muenchen 45

Atlantic Money Systems
USA, Miami,Florida 33169

BCT
Germany, D-W 7990, Friedrichshafen 1
Berthold
Germany, D-W 1000, Berlin 46

CIRRUS TECHNOLOGY
USA, 21046 Columbia,MD

debis Systemhaus
Germany, D-W 7000, Stuttgart 80

DIGITHRUST
Germany, D-W 8500, Nuernberg

DSM Computer Systeme
Germany, D-W 8000, Muenchen 2

Elettronica San Giorgio
Italy, I-16154, Genova

Expert Informatik, GmbH
Germany, D-W 7770, Ueberlingen,Hafenstrasse 10

heddier electronic
Germany, D-W 4420, Coesfeld 2

Hewlett-Packard
Germany, D-W 6380, Bad Homburg v.d.H.

Image Recognition (I.R.I.S.)
Belgium, B-1348, Louvain la Neuve

Industronics
Germany, D-W 6980, Wertheim

INTEGRATA
Germany, D-W7400, Tuebingen am Neckar 1, Schleifmuehleweg 68

I.T.C.
Germany, D-W 8050, Freising

Kleindienst Datentechnik
Germany, D-W 8900, Augsburg 11

Kleindienst Datatechnik AB
Sweden, 161 02 Bromma, Aplvaegen 10,Box 20117

KPMG Deutsche Treuhand Gruppe
Germany, D-W 6000, Frankfurt am Main 26

Lincoln,A.J.
USA, Concord MA 01742

LSI LOGIC, GmbH
Germany, D-W 8000, Muenchen 81, Arabellstrasse 33

Mannesmann Scangraphic GmbH
Germany, D-W 2000, Wedel / Holst.

MDS-Deutschland
Germany, D-W 5000, Koeln 30

MICROTEK Electronics Europe
Germany, D-W 4000, Duesseldorf 11

NEOS International
Germany, D-W 2000, Hamburg 26

Neuro Informatik
Geselschaft fuer Entwicklung und Anwendung Neuronaler Netze mbH
Germany, D-W 1000, Berlin 41, Roennebergstrasse 5 A

NIKEX TRADING
Hungary, H-1809, Budapest

OPTO-TECH
Germany, D-W 8122, Penzberg

PARSYTEC
Germany, D-W 5100, Aachen, Juelicher Strasse 338

Peters
Germany, D-W 2000, Hamburg 50

PE-VON
Taiwan, Taipei

PROFILE
Germany, D-W 6200, Wiesbaden

Rexroth electronic
Germany, D-W 8770, Lohr am Main

Siemens Nixdorf
Germany, D-W 4790, Paderborn

Symbolics
Germany, D-W 6236, Eschborn/Taunus, Mergenthalerallee 77-81

SYSTEM CONSULT
Germany, D-W 1000, Berlin 38

SZKI RECOGNITA AG
Hungary,H-1251, Budapest

Technische Fachhochschule Berlin
Germany, D-W 1000, Berlin 65

Technische Universitaet Braunschweig
Germany, D-W 3300, Braunschweig

Universitaet Erlangen-Nuernberg
Germany, D-W 8520, Erlangen

Victor Technologies GmbH
Germany, D-W 6070, Langen

Xionics
Great Britain, London N3 1HG

NEURAL NETWORK LEARNING WITH RESPECT TO SENSITIVITY TO WEIGHT ERRORS

P. Růžička)*

Abstract

This paper deals with the problem of neural network learning to get the most convenient "configuration". By the configuration is meant the vector of synaptic weights and thresholds of formal neurons creating the network. In the configuration design, we respect the complexity of technical realization of the network and we consider both the possible errors in keeping precise the designed configuration during the realization and fluctuations of the configuration during the net exploitation. To achieve this we introduce a cumulative loss function of the network which expresses the loss evoked by unprecise learning. The network learns through the optimization of sensitivity of the cumulative loss to large changes of configuration, the sensitivity to large changes being constructed on the basis of differentiating linear integral parametric operators of derivatives estimation. The possibilities of such an approach are demonstrated by an example.

I. Introduction

The modeling of biological neural networks and the design of artificial neural networks have, up till now, been based on the nominal values of their parameters (synaptic weights, neuron activity thresholds, etc.), which we will call *the configuration.* However, both in the learning procedure and in the activation period of a neural network's operation, many deviations of the actual configuration parameters values appear. These deviations are caused by the influence of the network environment activity, by the network aging, and, in artificial neural networks, also by manufacturing imperfections. The negative results of all these deviations consist in changes of the network's functional properties, leading in certain cases to their total degradation.

In the methods developed up till now for neural networks analysis and synthesis, the problems of configuration parameters and structure deviations have not been taken into account systematically. Though research in this respect is important for the modeling of various pathological stages of biological neural networks (epileptic situations, memory degeneration), it is especially important for the design of artificial neural networks where the influence of technology imperfections, material aging and environmental changes cause unpredictable deviations of parameters values from the nominal configuration calculated in the process of the design, resulting in a decrease of production yield and reliability. Recently some attempts have appeared for evaluating the sensitivity of the neural network functions to changes of configuration parameters for the preliminary neural network design (see Davis, 1989, Stevenson, 1990). It is possible to attack these problems more fundamentally by exploiting the theory of system tolerances and sensitivity. We show how the negative results of deviations of configuration parameters can be limited or excluded by respecting these deviations in the synthesis procedure of the neural network. To do this we minimize the sensitivity of the input-output function realized by the neural network to the changes of the configuration parameter values using convenient learning methods.

The methods of optimal tolerancing and sensitivity minimization were developed for electronic circuits design, e. g. Bandler, 1980, Bode, 1951, Buttler, 1971, Director, 1977 and 1978, Géher, 1971, Opalski, 1979, Strazs, 1980, Thach, 1988, but they also seem to be very useful for neural net configuration analysis and synthesis. They deal with the design of parameters of systems to assure an optimal performance of the systems. It is supposed that the behavior of the system is described by a vector of *system functions*

$$f: \mathcal{R}^r \to \mathcal{R}^v \tag{1}$$

defined on the r-dimensional real space $\mathcal{R}^r$ of system *parameters* x. The demands on the system behavior are expressed by the system of inequalities

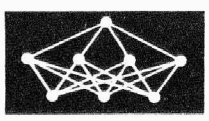
$$f(x) \leq 0. \tag{2}$$

When solving this system of inequalities we obtain an acceptable vector of parameters $x \in \mathcal{R}^r$. The set of all the acceptable vectors of parameters, for which the system fulfills the demands, is called a *region of acceptability*

$$R_a = \{ x \in \mathcal{R}^r \mid f(x) \leq 0 \}. \tag{3}$$

We are usually not satisfied with an arbitrary acceptable solution $x \in R_a$ and we are therefore looking for a vector of parameters that is the „most convenient"

*) Dr. Pavel Růžička
Institute of Computer and Information Science, Czechoslovak Academy of Sciences, 182 07 Prague, Czechoslovakia



in some respect. The following min-max problem is often solved to find some optimal solution

$$\min_{x \in \mathcal{R}^r} \max_{s \in \hat{v}} f_s(x)$$

($\hat{v}$ will be used for a set of positive integers up through v, $\hat{v} = \{1, 2, \ldots, v\}$). However, if the system functions are not convex this task has generally not a unique solution and we can put more conditions on the „convenient" solution. If we need to consider the effect of possible deviations of parameters we can incorporate demands on the sensitivity of system performance to these deviations into the task of system design. The concept of sensitivity in this sense was first formalized because of needs in electronic circuits design. If the changes of parameters are small one works with a *differential sensitivity of system function* that is simply defined as the first partial derivatives of the system function with respect to the parameters or alternatively, *a relative first order differential sensitivity of the function* f_s is used (see Bode, 1951 or Géher 1971)

$${}^1\mathcal{S}_i^s(x) = \frac{\partial \ln f_s(x)}{\partial \ln x_i} = \frac{x_i \partial f_s(x)}{f_s(x)\, \partial x_i}$$

for $i \in \hat{r}$, $s \in \hat{v}$. The differential sensitivies of higher orders are defined analogicaly by using derivatives of higher orders. Then some conveniently constructed scalar measure of the first order differential sensitivities is minimized for one to get an advantageous vector of parameters. Very often, respect to sizable fluctuation of parameters shall be taken into account and the tools of differential sensitivities, which deal only with local changes of system functions, are insufficient. Therefore several authors introduced a *sensitivity of system functions to large changes of parameters* (e. g. Buttler 1971) but their constructions were mostly very tightly bound with a concrete algorithm of system design. Moreover all the approaches have had a bottle-neck in the assumption of differentiability of system functions.

We will concentrate on neural networks with a fixed function that are first learning to perform the correct function and then are put into use. Neural networks with fixed functions have many meaningful applications that require hardware implementation. Before realizing this technically, it is necessary to look for the configuration (i. e. thresholds of neuron activity, strengths of neuron connections, etc.) by computer. In doing so, we take into account the technological impossibility of maintaining precisely the designed configuration in the process of realization and potential deviations of parameters caused by material aging and environmental changes. Therefore, the correct function of the net shall be ensured both for this configuration and for a relatively large area surrounding it in the configuration space. We get such a „stable" solution by minimizing the „sensitivity" of neural net behavior to large changes of the configuration.

In section II. , we will show the possibility to teach a neural network through a loss function minimization. The loss functions express the loss resulting from imperfect learning to input samples and they can serve as system functions (1) in the design process of the network configuration. In section III. , we will mention averaging linear integral parametric operators (LIPOs) and differentiating LIPOs of derivative estimation that serve to estimate values of functions and their derivatives as an alternative to difference formulas, namely in the case when the values of functions cannot be evaluated precisely. In section IV. , we will prove some properties of these LIPOs that are important for the construction of the sensitivity of system functions to large changes of parameters. In section V, this enables us to construct this sensitivity in a mathematically correct way by utilizing the differentiating LIPOs of gradient estimation even for a system described by system functions which are not differentiable but only e. g. locally integrable. In section VI, we will give an example of teaching a three-layer feedforward network on the basis of this sensitivity optimization.

II. Learning via optimization

For a great deal of the neural net models we can find a function which is being minimized in the process of learning. We are always able to transform this function in such a way that it represents the loss resulting from imperfect learning. Let us consider the existence of such a non-negative *loss function* f as a measure of the net behavior correctness and let this loss function fulfill the following.

Assumption 1.

Let the loss function

$$f: \mathcal{R}^r \times \mathcal{A} \to < 0, +\infty)$$

be at least lower semi-continuous on the *configuration space* $\mathcal{R}^r$ ($r-$ dimensional real space consisting of all possible vectors of weights and thresholds) for each *input* $a \in \mathcal{A}$ and at least continuous on the *input space* $\mathcal{A}$ (environment) for each configuration vector $x \in \mathcal{R}^r$

■

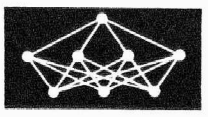
Continuity is necessary for the net to be able to generalize, in other words, to guess the behavior of the input space $\mathcal{A}$ from a small amount of input samples.

Ideal learning could then be viewed as the global minimization of the *cumulative loss*

$$\min_{x \in X \subset \mathcal{R}^r} F(x) \tag{4}$$

$$F(x) = \|f(x,.)\|, \tag{5}$$

where $\|.\|$ is some convenient norm on a space of func-



tions that are defined and at least continuous on the input space $\mathcal{A}$. The lower semi-continuity ensures the existence of a minimum for problem (4) on a bounded subset $X \subset \mathcal{R}^r$, where X represents e. g. the range in which the values of configuration parameters may change. The more we can suppose about the smoothness of functions forming this space, the faster is the learning. See e. g. Poggio, 1990.

If we have the cumulative loss function that is minimized in the process of adaptation common for the neural network model, any adaptation method that is able to minimize it is applicable for us to teach the network. The principal criterion for the selection of a convenient method is its ability to converge to a „good" global minimum of the cumulative loss function.

II. 1. The case of supervised learning

In the case of supervised learning the *input-output function* y realized by the network

$$y(x, .) : \mathcal{A} \to \mathcal{Y}$$

should approximate an unknown map

$$d : \mathcal{A} \to \mathcal{Y}$$

from the input space $\mathcal{A}$ into the *output space* $\mathcal{Y}$. The cumulative loss function is then

$$F(x) = \| y(x, .) - d(x, .) \|.$$

Usually a finite set of patterns $a^s \in \mathcal{A}$ and the corresponding set of prescribed outputs $d^s \in \mathcal{Y}$, $s \in \hat{v}$, are given and thus our theoretical cumulative loss function can be estimated as

$$\tilde{F}(x) = \| \{ \| d^s - y(x, a^s) \|_1 \}_{s=1}^{v} \|_2, \tag{6}$$

where $\| . \|_1$ is a norm on $\mathcal{Y}$, $\| . \|_2$ is some norm on the v-dimensional real space $\mathcal{R}^v$ and $\{ e^s \}_{s=1}^{v}$ is a v-dimensional vector with elements e^s. The learning problem in (4) becomes a common task of discrete approximation

$$\min_{x \in X \subset \mathcal{R}^r} \tilde{F}(x). \tag{7}$$

If the behavior of the environment $\mathcal{A}$ is stochastic, then we will use the mean value of the approximation error as the cumulative loss

$$F(x) = E(\| d - y(x, a) \|) = \\ = \int \| d - y(x, a) \| \mu(a, d)\, d(a, d) \tag{8}$$

for a joint probability density function μ of the vector (a, d). In this case, task (4) becomes a stochastic optimization problem and it can be treated by employing a stochastic approximation method .

We provide a simple stochastic approximation algorithm as an example not only because its applicability to solving problem (8) above but also because we will formulate learning as a task of sensitivity measure minimization in section V, the task being a stochastic optimization problem, and we will use a stochastic approximation algorithm to solve the network design problem in section VI.

Generally, if we have a simple stochastic optimization problem to find a minimum of mean value of a function $g(x, \omega)$ (we use a symbol ω to emphasize that $g(x, \omega)$ is a random variable)

$$\min_{x \in \mathcal{R}^r} g(x),$$

$$g(x) = E(g(x, \omega)),\ g(x, \omega) = g(x) + \gamma(\omega),\ E(\gamma(\omega)) = 0$$

and if we can only obtain realizations of random variables $g(x, \omega)$ and not the mean values $g(x)$ then we can employ the following stochastic approximation algorithm to find a local minimum of the problem above

$$x_{n+1} = x_n - a_n\, Dg(x_n, c_n), \tag{9}$$

where $\{a_n\}_{n \in \mathcal{N}_0}$ is a sequence of positive step lengths, and $\{c_n\}_{n \in \mathcal{N}_0}$ is also a sequence of positive real numbers used for stochastic gradient estimates of the function g ($\mathcal{N}_0$ is the set of all the nonnegative integers). The members of this sequence can be step lengths for differences computations or the averaging parameters of differentiating operators which will be introduced in the next section. The sequences must fulfill

$$\sum_{n=0}^{\infty} a_n = +\infty,\ \lim_{n \to \infty} a_n = 0,\ \lim_{n \to \infty} c_n = 0,$$

$$\sum_{n=0}^{\infty} a_n c_n < +\infty, \quad \sum_{n=\infty}^{\infty} \left(\frac{a_n}{c_n} \right)^2 < +\infty.$$

We suppose the estimate Dg of the gradient ∇g in the form

$$Dg(x_n, c_n) = \nabla g(x_n) + \gamma_n(x_n, c_n) + \hat{\gamma}_n(\omega)$$

in each step $n \in \mathcal{N}_0$. Let the random vector $\hat{\gamma}_n$ have zero mean and let there be valid $\gamma_n(x_n, c_n) = o(c_n^2)$ for the deterministic error of the gradient approximation. Under some noise condition on the sequence $\{\hat{\gamma}_n\}_{n \in \mathcal{N}_0}$ (e. g. if the sequence is a martingale) the iterations x_n either converge to a local minimum of the function g almost surely (i. e. with probability one) or they are unbounded (see Ermoljev, 1976, Katkovnik, 1976, Kushner, 1978 or see Hornik, 1990 for applications in neural networks learning).

II. 2. The case of unsupervised learning

If we also consider a finite set of patterns $a^s \in \mathcal{A}$ used by the network to learn, the network must then

search for the corresponding outputs $d^v \in \mathcal{Y}$, $s \in \hat{v}$, together with the configuration $x \in \mathcal{R}^r$ during the learning. It means that the optimization problem (4) is solved on a larger space in the case of unsupervised learning. The deterministic supervised learning case (7) has its analogy here in the problem

$$\min_{x \in X} \bar{F}(x, d^1, d^2, \ldots, d^v)$$

$$d^1, d^2, \ldots, d^v \in \mathcal{Y}$$

where the parameters $d^1, d^2, \ldots, d^v$ of cumulative loss function (6) are realized as other variables

$$\bar{F}(x, d^1, d^2, \ldots, d^v) = \| \{ \| d^s - y(x, a^s) \|_1 \}_{s=1}^{v} \|_2.$$

The loss function can often be derived from the adaptation rules because they should be related to the gradient of the loss function.

Irrespective of the difficulty of finding a global minimum of F in (4), we know nothing about the quality of the solution obtained by the learning according to (4). Practically, we need not perfect learning. We accept the configuration for which the value of the cumulative loss function is less than some small selected number ε as the solution searched. The set of all such almost optimal configurations we will call the region of acceptability R_a in correspondence with (3)

$$R_a = \{ x \in \mathcal{R}^r \mid F(x) \leq \varepsilon \}. \tag{10}$$

If the region of acceptability is nonempty, then by solving problem (4) we will reach a point from this region. Unfortunately such a solution can be located somewhere near the boundary of the region R_a. If we construct hardware equipment on the base of this solution, there is a great probability that it will perform incorrectly due to the facts mentioned in the introduction. These facts may cause a shift of the original configuration outwards from the region of acceptability during the realization process or during hardware utilization. In order to afford the possibility of imprecise realization of the computed configuration, we would like the configuration to be in the „center" of region R_a. This is also the way how to construct highly reliable hardware which is a problem dealt with by the tolerances and sensitivity theory.

III. Linear integral parametric operators (LIPO) of averaging and differentiation

With the aim of effective estimating values and derivatives of functions, Katkovnik in (Katkovnik, 1976) introduced linear integral parametric operators (LIPO) with a real parameter $c > 0$ defined on the class of Lipschitz continuous real functions in the form of convolution with a kernel h

$$\hat{f}(x, c) = \int_{\mathcal{R}^r} h(u) f(x - cu)\, du. \tag{11}$$

Definition 1.

Operator (11) is called the *averaging operator of degree* ϑ, $\vartheta \geq 0$, with the kernel h^ϑ if the norm condition (12) is valid

$$\int_{\mathcal{R}^r} h(u)\, du = 1, \tag{12}$$

and if

$$\int_{\mathcal{R}^r} h(u)\, u_1^{s_1} \ldots u_r^{s_r}\, du = 0 \tag{13}$$

holds true for all the $s_i \in \mathcal{N}_0$, which fulfill

$$0 < \sum_{i=1}^{r} s_i \leq \vartheta, \quad s_i \geq 0.$$

■

Condition (13) implies that $\hat{q}(x, c) = q(x)$ for any polynomial q of degree maximally ϑ and an arbitrary $c > 0$.

We will always use a carat sign to denote the result $\hat{f}(.\,, c)$ of averaging operator application to a function f. The function $\hat{f}(.\,, c)$ will be called *an averaged function* f.

Katkovnik introduced LIPO of differentiation by the expression

$$\hat{f}^{(l)}_{l_1 \ldots l_r}(x, c) = c^{-l} \int_{\mathcal{R}^r} d_{l_1 \ldots l_r}(u) f(x - cu)\, du, \tag{14}$$

where l_i are nonnegative integers, $l_i \in \mathcal{N}_0$, $i \in \hat{r}$, and

$$l = \sum_{i \in \hat{r}} l_i.$$

We use the symbol $D^l f(x)$ for the derivative of f of order l

$$D^l f(x) = D^l_{l_1 \ldots l_r} f(x) = \frac{\partial^l f(x)}{\partial x_1^{l_1} \ldots \partial x_r^{l_r}}$$

Definition 2.

Operator (14) is called the *differentiating operator of the l-th derivative* $\{D^l_{l_1 \ldots l_r} f(x)\}$ *estimation of degree* ϑ *with the kernel* $d^\vartheta_{l_1 \ldots l_r}$ for $\vartheta \geq l$ if the moment conditions

$$\int_{\mathcal{R}^r} d_{l_1 \ldots l_r}(u)\, u_1^{s_1} \ldots u_r^{s_r}\, du = 0, \tag{15}$$

are valid for any $s_i \in \mathcal{N}_0$, $i \in \hat{r}$, fulfilling

$$0 < \sum_{i=1}^{r} s_i \leq \vartheta,\ s_i \geq 0,\ i \in \hat{r},$$

and if there holds true

$$\int_{\mathcal{R}^r} d_{l_1 \dots l_r}(u)\, u_1^{l_1} \dots u_r^{l_r}\, du = (-1)^l \prod_{i=1}^{r} (l_i)!. \quad (16)$$

■

Operator (14) transforms polynomials q of degree less or equal to ϑ into their exact derivatives $D^l_{l_1 \dots l_r} q$

$$\tilde{q}^{(l)}_{l_1 \dots l_r}(x, c) = \frac{\partial^l q(x)}{\partial x_1^{l_1} \dots \partial x_r^{l_r}}$$

due to the validity of (15) and (16).

We will always use a tilda sign to denote the result $\tilde{f}^{(l)}_{l_1 \dots l_r}(., c)$ of the transformation of a function f by a differentiating operator of the l-th derivate $D^l_{l_1 \dots l_r} f$ estimation.

The parameter c is called an *averaging parameter.* It determines a measure of distance between the function $\hat{f}(., c)$ *or* $\tilde{f}^{(l)}_{l_1 \dots l_r}(., c)$ and the original function f or its derivative $D^l_{l_1 \dots l_r} f$ respectively, and alternatively it determines a measure of "averaging")smoothing; of the original function by considering its behavior in a smaller or larger area surrounding the point x in estimate of $f(x)$ or $D^l_{l_1 \dots l_r} f(x)$ respectively.

An important property of some differentiating operators is their potentiality.

Definition 3.

The differentiating operator of the l-th derivative estimation is called *potential* if an averaged function $\hat{f}(., c)$ can be always found for any function f, any parameter $c > 0$, any integer $t, 0 < t \leqslant l$, and all the $x \in \mathcal{R}^r$ so that

$$\tilde{f}^{(t)}_{l_1 \dots l_r}(x, c) = D^t_{l_1 \dots l_r} \hat{f}(x, c)$$

is valid.

■

In some sense, potentiality renders it possible to neglect a systematic error of function and derivatives values estimates (11) and (14). E. g. , if we use a potential operator of gradient estimate with a fixed parameter $c > 0$ to search for a minimum of function f we will actually find a minimum of function $\hat{f}(., c)$. Because the averaging removes only sudden local changes in the shape of the function f and because it can preserve the global properties substantial for the solved problem (concretely, it removes only shallow local minima) the systematic error appears to be rather positive.

Values of generally complex and multidimensional integral transformations (11) and (14) can be estimated by employing a Monte Carlo method. We select a convenient probability density function p which is positive on the support of the kernel h or $d_{l_1 \dots l_r}$ respectively and we will generate L random samples $u^s \in \mathcal{R}^r$, $s \in \hat{L}$, according to this probability density function. We will then estimate the value $\hat{f}(x, c)$, and thus also the value $\hat{f}(x)$, by using the parametric estimate

$$\hat{f}(x, c, L) = \frac{1}{L} \sum_{s=1}^{L} \frac{h(u^s)}{p(u^s)} f(x - cu^s) \quad (17)$$

and value $\hat{f}^{(l)}_{l_1 \dots l_r}(x, c)$, and thus $D^l_{l_1 \dots l_r} f(x)$, by using the parametric estimate

$$\hat{f}^{(l)}_{l_1 \dots l_r}(x, c, L) = \frac{1}{c^l L} \sum_{s=1}^{L} \frac{d_{l_1 \dots l_r}(u^s)}{p(u^s)} f(x - cu^s). \quad (18)$$

We introduce an artificial randomness into the estimates in such a way. We can use either the same probability density function for estimating both function values (17) and derivatives (18) „in parallel" using the same function values $f(x\text{-}cu^s)$ calculated for the same samples $x\text{-}cu^s$, $s \in \hat{L}$, or we could select various probability density functions in the expressions (17) and (18) to achieve optimal variance of these estimates.

To express the quality of estimates of function value $f(x)$ or derivative $D^l f(x)$ above we use the mean value of the square of the total error of the estimate

$$\Delta^2 = \Delta_o^2 = E\left((\hat{f}(x, c, L) - f(x))^2 \right) =$$

$$= (\hat{f}(x, c) - f(x))^2 + D(\hat{f}(x, c, L)) \quad (19)$$

or

$$\Delta_{l_1 \dots l_r}{}^2 = E\left(\left(\hat{f}^{(l)}_{l_1 \dots l_r}(x, c, L) - D^l_{l_1 \dots l_r} f(x) \right)^2 \right)$$

$$= \left(\tilde{f}^{(l)}_{l_1 \dots l_r}(x, c) - D^l_{l_1 \dots l_r} f(x) \right)^2 + D\left(\tilde{f}^{(l)}_{l_1 \dots l_r}(x, c, L) \right). \quad (20)$$

This measure of error consists of the deterministic part, the square of systematic error of the estimate of function or derivative value for the particular operator (11) or (14), and the variance (denoted by D) of random part, the randomness being the result of Monte Carlo method (17) or (18) application.

The deterministic part, concretely the systematic error, is analyzed by Katkovnik in detail. If the function f has Lipschitz continuous derivatives up to and including the order α and if $l \geq 0$, $\vartheta \geq l$, $\alpha \geq l$ and if we denote $\vartheta^* = \min\{\alpha, \vartheta\}$, then the systematic error of the estimate of 1-th derivative (including $l = 0$) using an operator (11) or (14) of order ϑ is

$$o\left(c^{\vartheta^* + l - l}\right). \quad (21)$$

Obtained approximations $\hat{f}(., c)$ or $\hat{f}^{(l)}_{l_1 \dots l_r}(., c)$ are smooth functions if the kernel of used integral transformation (11) or (14) is smooth.

We will briefly touch the random component of error for the case when we can only obtain noise cor-

rupted values $f(., \omega)$ of the function f with the variance of noise σ^2, i. e.

$$f(x) = E(f(x, \omega)),$$

$$\sigma^2 = D(f(x, \omega)) = E\left((f(x) - f(x, \omega))^2 \right). \quad (22)$$

This situation occurs by learning in random environment (8) when the network shall approximate an unknown map which cannot be evaluated exactly.

Then the variance of the estimate of derivative $D^l_{l_1 \dots l_r} f(x)$, including the case $l = 0$, is

$$D\left(\tilde{f}^{(l)}_{l_1 \dots l_r}(x, c, \omega, L) \right) \leq \frac{\sigma^2}{L c^{2l}} l^2_{l_1 \dots l_r} + o(c^{1-2l}),$$

if the function f is at least continuously differentiable in a neighborhood of the point x (see Katkovnik, 1976). The number

$$l^2_{l_1 \dots l_r} = \int_{\mathcal{R}^r} \frac{d_{l_1 \dots l_r}(u)^2}{p(u)} du \quad (24)$$

is called *an index of exactness* of the averaging or differentiating operator. Thus the random noise is suppressed if the averaging parameter is large and the index of exactness is small.

Let us summarize some advantages of the above way of estimation of function values and derivatives.

If an analytical expression for calculation of derivatives of a function is not known, or if the analytical expression is to complex, one usually uses some difference formula to estimate the necessary derivative. If the function values are corrupted by a random noise then, e. g. by estimating the first partial derivatives using a difference formula, the variance of random error of the estimate is indirectly proportional to the square of the step used for difference calculation. However, kernels of differentiating operators with a small index of exactness can be constructed for one to achieve less variance of noise (24) for the corresponding integral operator as compared with the difference formula when both the same systematic (deterministic) error and the same number of function values L used for the estimate are required. While the necessary number of function values L, used in a difference formula for the first partial derivative estimate with given systematic error, increases linearly with the dimension of problem r, it can stay approximately constant in the case of integral operators (14) with a convenient kernel (see Example 1 in Mathematical Appendix).

If the differentiating operator (14) is potential we can neglect the systematic error of derivative estimate in a certain range due to smoothing properties of the corresponding averaging operator. The averaging parameter c may stay relatively large which positively affects the value of noise (23) of the estimate (18).

If we use a potential differentiating operator of gradient estimate when solving a stochastic optimization problem (9) we need not the sequence of averaging parameters $\{c_n\}_{n \in \mathcal{N}_0}$ to converge to zero. Thus we can reach the maximal order of asymptotic convergence of stochastic approximation algorithms (see Kushner, 1978).

IV. Properties of LIPOs of averaging and differentiating in estimating values and derivatives of nondifferentiable functions

Katkovnik in (Katkovnik, 1976) introduced the operators (11) and (14) on the space of Lipschitz continuous functions and functions with Lipschitz continuous derivatives up to order l respectively. However, kernels of these operators belong practically in the space of quickly decreasing functions. It means that the operators can be defined on a broader class of functions.

Definition 4.

Linear space $\mathcal{Q}$ of all the smooth functions (i. e. functions having all the derivatives) Ψ such that the functions $D^l_{l_1 \dots l_r} \Psi(x)\, q(x)$ are bounded on $\mathcal{R}^r$ for any polynomial q and an arbitrary l-th derivative of Ψ, where $l = \sum_{i=1}^{r} l_i$, $l_i \in \mathcal{N}_0$, $i \in \hat{r}$, will be called a *space of quickly decreasing functions* on $\mathcal{R}^r$.

Linear space $\mathcal{Q}'$ of all the measurable functions f for which a polynomial q exists such that $|f(x)| \leq |q(x)|$ almost everywhere on $\mathcal{R}^r$ will be called *a space of slowly increasing functions* on $\mathcal{R}^r$.

■

If kernels of transformations (11) and (14) are from the space $\mathcal{Q}$ then these transformations have sense, i. e. they exist, for any slowly increasing function $f \in \mathcal{Q}'$. The resulting outputs of the transformations are then smooth functions $\hat{f}(., c)$ and $\hat{f}^{(l)}(., c)$ respectively and the following theorems are valid.

Theorem 1.

Let a kernel h of averaging operator (11) be quickly decreasing, $h \in \mathcal{Q}$, and the function f be a slowly increasing function continuous at the point x. Then the following limit holds true for the averaging operator (11)

$$\lim_{c \to 0+} \hat{f}(x, c) = f(x)$$

■

Theorem 2.

Let the kernel g of potential differentiating operator of gradient estimation (14) be a quickly decreasing function, $g \in \mathcal{Q}$. Let f be a slowly increasing function and let the gradient of this function ∇f exist almost everywhere and be continuous at the point x. Then the differentiating operator of gradient estimate (14) fulfills

$$\lim_{c \to 0+} \nabla \tilde{f}(x, c) = \nabla f(x).$$

■

Let us note that $g = (d_{1\ldots0}, \ldots, d_{0\ldots1})$.

Most generally, the LIPOs of averaging and differentiation could be considered on the space of generalized slowly decreasing functions, see Antosik, 1973 for the definition of this space.

We will briefly concentrate on an important case when the function f is *locally Lipschitz*, i. e. when there exists a constant $U > 0$ for any bounded subset of $\mathcal{R}^r$ that the inequality

$$|f(x) - f(y)| \leq U\|x - y\|$$

is valid for any points x, y from this subset. Then the gradient of function f exists almost everywhere in $\mathcal{R}^r$ and a generalization of gradient can be introduced at the points where the classical gradient does not exist. We utilize Clarke's generalization of the gradient (Clarke, 1983).

If a function f is Lipschitz in the neighborhood of a point x then a generalized directional derivative of f in a direction $v \in \mathcal{R}^r$ can be defined in several ways. Clarke did it by using the expression

$$f^o(x; v) = \lim_{\substack{y \to x \\ t \to 0+}} \sup \frac{f(y + tv) - f(y)}{t}$$

and then he defined the generalized gradient as the set

$$\underline{D}f(x) = \{\xi \in \mathcal{R}^r \mid f^o(x; v) \geq \langle \xi, v \rangle \text{ for an arbitrary } v \in \mathcal{R}^r\}.$$

Basic properties of the generalized gradient are formulated in Theorem 6. in Mathematical Appendix.

We know that the averaged function $\hat{f}(., c)$ defined according to (11) converges uniformly to the original function f, when $c \to 0+$, because f is locally Lipschitz (Katkovnik, 1976) . We would like to know more about the behavior of the gradient estimate $\nabla\tilde{f}(., c)$. We state the results concerning the gradient estimate convergence in the following theorem.

Theorem 3.

Let a function f be locally Lipschitz with a constant $U > 0$ and let g be a kernel of differentiating LIPO of gradient estimation for which a constant M exists such that

$$\int_{\mathcal{R}^r} \|g(u)\| \|u\| \, du \leq M < +\infty. \qquad (25)$$

If the gradient of function f exists at the point x then the following limit holds

$$\lim_{c \to 0+} \nabla\tilde{f}(x, c) = \nabla f(x).$$

If the function f is convex then we have at any point x

$$\lim_{c \to 0} \nabla \tilde{f}(x, c) = \xi,$$

where $\xi \in \underline{D}f(x)$ is an element of the subgradient of f at the point x ($\underline{D}f(x)$ is the generalized gradient introduced for convex functions, which equals Clarke's generalized gradient).

If the differentiating LIPO of gradient estimation is potential and if the kernel of the corresponding LIPO of averaging h is a nonnegative function then there exists an averaging parameter $c_o > 0$ for any positive number $\eta > 0$ and for any point x so that

$$\nabla\tilde{f}(x, c) = \frac{1}{c} \int_{\mathcal{R}^r} g(u) f(x - cu) \, du =$$

$$= \int_{\mathcal{R}^r} h(u) \nabla f(x - cu) \, du \in \underline{D}f(x) + B(0, \eta),$$

(where $B(0, \eta) = \{y \in \mathcal{R}^r \mid \|y\| \leq \eta\}$) for an arbitrary c, $0 < c \leq c_o$.

■

If we solve a stochastic optimization problem (e. g. learning task in a stochastic environment (8)) where the goal function is only locally Lipschitz (e. g. due to using the maximum norm in (5)) and if we apply the stochastic approximation algorithm (9) we must estimate the gradient of the nondifferentiable function by using a potential differentiating LIPO of the gradient estimation. We can use the sequence of averaging parameters $\{c_n\}_{n \in \mathcal{N}_0}$ which does not converge to zero then the claim of this theorem assures the almost sure convergence of the algorithm (9) to an approximate solution of the problem.

V. Differentiating LIPO of gradient estimation as a sensitivity of system functions to large changes of parameters

V. 1. Construction of the sensitivity to large changes of parameters

The result of transformation of a function f by an averaging LIPO is a more or less smoothed course of this function, the smoothing depending on the value of the averaging parameter c. In other words, the value of averaged function $\hat{f}(x, c)$ represents an average behavior of the function f in an area surrounding the point x. Moreover, if we have a potential differentiating LIPO of gradient estimation with the kernel $g = \nabla h$ then the property

$$\nabla\tilde{f}(x, c) = \nabla\hat{f}(x, c),$$

(which means that the smoothed ("average") gradient of the original function equals to the gradient of the averaged function $\hat{f}$) leads us to introduce a sensitivity of behavior of a system described by a vector of system functions (1) $f: \mathcal{R}^r \to \mathcal{R}^v$ to large changes of parameters by employing this potential differentiating LIPO of gradient estimation.

Now we have a basis for construction of various vector functions which can characterize a sensitivity of system functions to changes of parameters. We can hardly find a unique general construction of sensitivity of system functions to large changes of system parameters which would be simple enough for numerical evaluation. A reasonable way for us to avoid the antagonism between simplicity of numerical calculation of sensitivity and a universality of the construction of sensitivity is to concentrate on some concrete, but wide enough, class of system functions. We can then also consider more concrete demands on the design of a system.

We will pay our attention to a neural network design. Thus the system functions are the loss functions $f_s = f(., a^s)$, $s \in \hat{v}$. They are nonnegative, which is important for our construction. The simplest way to construct the sensitivity of these system functions to large changes of parameters at the point $x \in \mathcal{R}^r$ is by using a LIPO

$$\nabla \tilde{F}(x, t) = \nabla \hat{F}(x, t) = t^{-1} \int_{\mathcal{R}^r} g(u)\, F(x - \bar{t}u)\, du, \quad (26)$$

where the function F is constructed as a norm of the vector of system functions f_s, $s \in \hat{v}$ (compare with (5)). We use the maximum norm which allows us to concentrate on the worst fulfilled demands on the system

$$F(x) = \max_{s \in \hat{v}} f_s(x), \; \bar{t} = \operatorname{diag}(t), \; t \in \mathcal{R}^r, \; g = \nabla h. \quad (27)$$

The matrix $\bar{t}$ is the diagonal matrix with the diagonal formed by the vector t. In this case, the large changes sensitivity is the gradient of averaged maximum of particular system functions with vector averaging parameter t and, at the same time, it is a „mean" value of the gradient of maximum of the system functions if the system functions are differentiable almost everywhere.

On the basis of results of section IV. we can state the following:

If the kernel of corresponding averaging LIPO is a quickly decreasing function $h \in \mathcal{Q}$, $g = \nabla h$, and if the system functions are slowly increasing (or e. g. at least locally integrable) then the function $\hat{F}(., t)$, and thus also the function $\nabla \hat{F}(., t)$, are smooth. Moreover, if the system functions f_s, $s \in \hat{v}$, are continuously differentiable then the function F is locally Lipschitz and the sensitivity $\nabla \hat{F}(., t)$ converges to ∇F almost everywhere, when $\|t\| \to 0$. Concretely, it converges to $\nabla F(x)$ at the points where only one system function is active, i. e. where $F(x) = f_s(x)$ for just one $s \in \hat{v}$. At other points, $\nabla \hat{F}(x, t)$ converges to the generalized gradient $\underline{D}F(x)$, which equals to the convex hull of gradients of the system functions that are active at the point x (see Clarke, 1983)

$$\underline{D}F(x) = \overline{co}\,\{\nabla f_s(x) \mid s \in \hat{v}, \; F(x) = f_s(x)\},$$

if $h \geq 0$. Thus the large changes sensitivity becomes the differential sensitivity (the sensitivity to small changes) when $\|t\| \to 0$, which is natural and desired.

We need to construct a scalar measure of the large changes sensitivity which we want to minimize in order to find a convenient design of system. A norm of large changes sensitivity is usually taken as the measure

$$S_F(x) = \frac{1}{2} \|\nabla \hat{F}(x, t)\|^2,$$

where $\|.\|$ can be the Euclidean norm. We find a convenient design of the system by minimizing this sensitivity measure for a fixed vector averaging parameter $t \in \mathcal{R}^r$

$$\min_{x \in \mathcal{R}^r} S_F(x).$$

Such a task as above is sometimes called design centering because we move the design into a „center" of the region of acceptability.

From the numerical point of view, it would be easier to solve the problem of minimization of the averaged maximum of system functions

$$\min_{x \in \mathcal{R}^r} \hat{F}(x, t) \quad (29)$$

then task (28). We shall formulate the relationship between these two tasks in the following theorem.

Theorem 4.

Each local minimum x^* of problem (29) is also a global minimum of problem (28).

Each minimum x^* of problem (28) is a local minimum of problem (29) whenever the Hessian of the function $\hat{F}(., t)$ (the matrix of second partial derivatives) $H_{\hat{F}}(x^*, t)$ is positive definite.

■

The proof of the theorem simply follows from the necessary conditions of optimality.

Thus task (29) can be solved instead of task (28) and, in this sense, the averaged function $\hat{F}(., t)$ can be used instead of the measure of large changes sensitivity (26).

If the system functions f_s, $s \in \hat{v}$, are continuously differentiable and $\|t\| \to 0$ then task (29) becomes the classical MIN-MAX problem mentioned in the introduction

$$\min_{x \in \mathcal{R}^r} \max_{s \in \hat{v}} f_s(x).$$

It is more necessary to watch after the area in which the loss in (5) caused by imprecise learning might be unacceptable than after that area where the loss is very near to zero. Also, the near-to-zero values of loss in acceptable areas could unpleasantly compensate the behavior of loss in some more critical parts of the parameter space $\mathcal{R}^r$ in integral (11). Therefore the following construction of large changes sensitivity is sometimes more practical

$$\hat{F}_i(x, t) = \int h(u)\, F_i(x - \bar{t}u)\, du, \qquad (30)$$
$$F_i(x) = \max\left[\varepsilon_{\mathcal{R}^r}, f_1(x), \ldots, f_s(x)\right], \ \varepsilon \geq 0,$$

where ε is appropriately selected and its value can be controlled during the design optimization.

It is also often convenient to solve a problem of minimization of the cost of realization under the condition that the value of the scalar measure of the large changes sensitivity shall be small instead of the large changes sensitivity measure minimization problem (28) or (29)

$$\min_{x \in \mathcal{R}^r} c(x, t) \text{ under the condition } \hat{F}_i(x, t) \leq 0, \qquad (31)$$
$$t \in \overset{\cdot}{\underset{i=1}{\mathrm{X}}} (0, +\infty)$$

where $c(x, t)$ is a cost function decreasing in the variables t_i, $i \in \hat{r}$. The larger will be the optimal values of elements of the vector t the larger will be the area where the system is little sensitive to changes of parameters and thus the more stable will be the system in the area surrounding the designed parameters.

V. 2. Construction of sensitivity to large changes respecting the real distribution of errors

The definitions of LIPOs (11) and (14) allow us to consider a random dispersion of parameters arising in the technological process of the system production in the process of system design. We can incorporate our knowledge about this dispersion into the construction of the kernel h of the operator (11).

Let $p(., \mu, \sum)$ be a probability density function defined on $\mathcal{R}^r$ for which a mean value μ and a covariance matrix $\sum = \sum(t)$ exist. The vector t is e. g. a technological parameter representing prescribed precision of production. If integral (32) exists (e. g. if F is bounded and measurable)

$$\hat{F}(x, t) = \int_{\mathcal{R}^r} p(u, 0, \sum(\mathbf{1}))\, F(x - \bar{t}u)\, du \qquad (32)$$

it represents the value of the averaged function $\hat{F}(x, t)$ with vector averaging parameter t according to Definition 1. The vector $\mathbf{1} = (1, 1, \ldots, 1)$. So-constructed operator has the order $\vartheta = 1$.

Theorem 5.

Let the gradient $\nabla_u p(u, \mu, \sum)$ of p with respect to u exist for almost all u and for any μ, Σ. Let the functions $\| \nabla_M p \|$ and $\| \nabla_M p \| \| u \|$ be integrable as functions the vector u. Then $\nabla_u p(u, 0, \Sigma)$ is the kernel of LIPO of gradient estimation (14) and this operator is potential. of

■

Therefore the introduced large changes sensitivity with an appropriately constructed kernel can respect the real distribution of errors.

V. 3. The relationship of sensitivity optimization and optimal design of tolerances

When constructing the large change sensitivity by applying the LIPOs with a kernel derived from the density probability function of deviations of system parameters from prescribed values, we easily find the relationship of sensitivity optimization problems to problems solved in the theory of tolerances (e. g. Bandler, 1980, Director, 1977 and 1978, Géher, 1971, Opalski, 1979, Strazs, 1980, Thach, 1988).

Let χ_{R_a} be the characteristic function of the region of acceptability

$$R_a = \left\{ x \in \mathcal{R}^r \mid \max_{s \in \hat{s}} f_s(x) = F(x) \leq 0 \right\},$$
$$\chi_{R_a}(x) = \begin{cases} 1 \text{ for } x \in R_a \\ 0 \text{ for } x \notin R_a \end{cases}$$

and let $p(., x, \sum(t))$ be the probability density function of a random vector v with the mean value x and with the covariance matrix $\sum(t)$ describing the distribution of values of parameters of produced systems v around the (prescribed) nominal design x when the tolerances of parameters t are prescribed. Assume that this probability density function is differentiable almost everywhere, i. e. that the assumptions of Theorem 5 are valid. If we use the transformation $v = x + \sum(t)\, u$ then we obtain a random vector u with a probability density function $p(., 0, \sum(1))$. We can calculate the theoretical yield of production, i. e. the probability that a produced system will be acceptable

$$Y(x, t) = \int_{\mathcal{R}^r} p(v, x, \sum(t))\, \chi_{R_a}(v)\, dv =$$
$$= \int_{\mathcal{R}^r} p(u, 0, \sum(\mathbf{1}))\, \chi_{R_a}(x - \bar{t}u)\, du = \hat{\chi}_{R_a}(x, t),$$

which is actually the averaged characteristic function χ_{R_a}. Thus the problem of minimization of abortions (which is equivalent to the yield maximization)

$$\min_{x \in \mathcal{R}^r} Z(x, t),$$
$$Z(x, t) = \int_{\mathcal{R}^r} p(u, 0, \sum(1))\, \chi_{\neg R_a}(x - \bar{t}u)\, du \qquad (33)$$

($\neg R_a$ is the complement of R_a to $\mathcal{R}^r$) is evidently an analog of the problem (29). Problems (33) and (29) be-

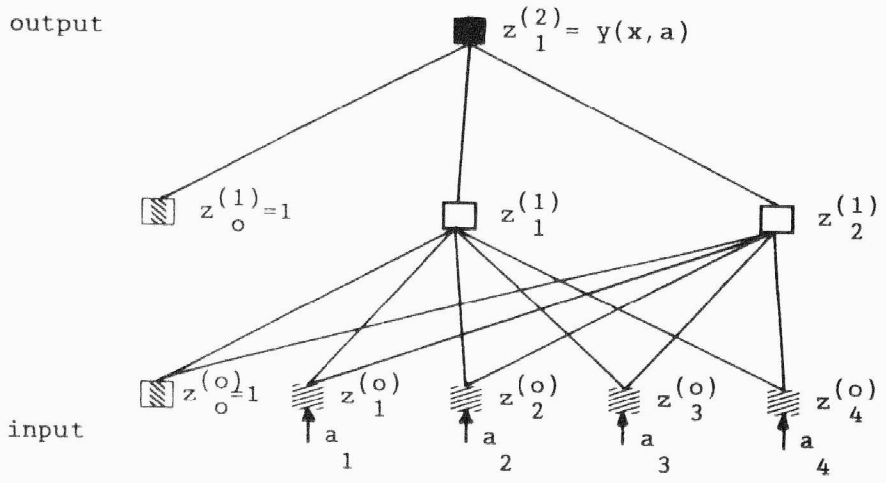
long to tasks of so called design centering. The concept of the center of the region of acceptability may naturally vary case to case. Both problems are tasks of stochastic programming because the functions being minimized are not given exactly. We can only evaluate random estimates of their values in some algorithmic way. All the functions $\hat{F}(., t)$, $\nabla\hat{F}(., t)$, $Y(., t)$ and $\nabla Y(., t)$ must be generally estimated by employing a Monte Carlo method. The same numerical method is thus applicable for solving problems (29) and (33).

VI. Example of the optimal configuration design

We will be looking for the optimal configuration of the three layer feedforward net (perceptron) for the binary symmetry recognition, i. e. for the recognition of the following relation among the input vector elements

$$a_j = a_{N+1-j}, \mathrm{j} = 1, 2, \ldots, [\mathrm{N}/2].$$

The input space $\mathcal{A}$ consists of all binary vectors of dimension N. The topology of the net is as follows. There are five input neurons $l_0 = 5$, three hidden layer neurons $l_1 = 3$ and one output neuron $l_2 = 1$. The first neurons in the first and hidden layers are utilized to set the thresholds of the neurons from the higher layers. The net should learn all possible input patterns from $\mathcal{A}$ when N = 4, thus $v = 16$. We request the net to indicate the validity of the symmetry relation by value 1 on its output and the case of nonsymmetry by value 0. Thus the requested net outputs for the inputs a^s are $d^s \in \{0, 1\}$, s = 1, . . ., 16.



The three layer PERCEPTRON for the binary symmetry recognition

Fig. 1

The active dynamics of the net are described by the expression

$$z_k^{(i+1)} = \Phi\left[\sum_{j=0}^{l_i} x_{jk}^{(i)} z_j^{(i)}\right] \qquad (34)$$

(see Rumelhart, 1986), where

$z_j^{(i)}$ is the state of the j-th neuron in the layer i,

Φ is the neuron transfer (activation) function of the sigmoid type $\Phi(z) = 1/(1 + exp(-z))$

$x_{jk}^{(i)}$ is the weight of the connection between the j-th neuron in the layer i and k-th neuron in the layer $(i + 1)$, $i = 0, 1$, $j = 0, 1, \ldots, l_i$,

$z_j^{(o)} = a_j$ for the imput layer, $j = 1, \ldots, l_o$,

$z_o^{(i)} = 1$ for $i = 0, 1$,

$y(x, a) = z_1^{(2)}$ is the net output for our case

The optimal design of the weights was found solving the following optimization problem of the same type as (31), instead of the classical back-propagation algorithm

$$\min_{x \in \mathcal{R}^r,\ t \in (0, +\infty)^r} c(x, t)$$
$$\hat{F}(x, t) \le \varepsilon,$$

where the cost c is given by

$$c(x, t) = \sum_{i=1}^{13} \frac{1}{t_i}$$

and the sensitivity measure $\hat{F}$ is defined as the averaged cumulative loss function (6) according to Theorem 5

$$\hat{F}(x, t) = \int h(u, 0, \tau(\mathbf{1}))\, F(x - tu)\, du = \int h(u, x, \tau(t))\, F(u)\, du.$$

The function

$$F(x) = \max_{s = 1, \ldots, 16} |y(x, a^s) - d^s|$$

is used as the cumulative loss function. The kernel of the averaging LIPO $h(u, x, \tau(\mathbf{1}))$ is the probability density function of the Gaussian distribution without correlations, with the vector of mean values x and with the standard deviations

$$\tau_i(t) = \frac{1}{100\sqrt{3}}\, t_i |x_i|.$$

Thus the LIPO used here is the same as that given in Example 1 in Mathematical Appendix. The vector t is the vector of tolerances, which will be prescribed for the realization of the net. The numbers d_s, s=1, 2, . . ., 16, are the requested outputs for the inputs a^s. The symbol Y_k, k = 1, 2, in Table 1

$$Y_k = Y_k(x, t) = \int_{R_{ak}} h(u, x, \tau(t))\, du$$

weight (threshold)	Solution		
	optimal nominal value	optimal tolerance [%]	comparat. tolerance [%]
$x_{01}^{(o)}$	-3.927	16.0	20.0
$x_{11}^{(o)}$	-21.462	10.0	12.0
$x_{21}^{(o)}$	-10.309	15.0	18.0
$x_{31}^{(o)}$	10.770	15.0	18.0
$x_{41}^{(o)}$	19.902	10.0	12.0
$x_{02}^{(o)}$	-7.766	15.0	18.0
$x_{12}^{(o)}$	23.667	10.0	12.0
$x_{22}^{(o)}$	13.105	13.0	16.0
$x_{32}^{(o)}$	-11.641	14.0	17.0
$x_{42}^{(o)}$	-23.215	10.0	12.0
$x_{01}^{(1)}$	11.659	16.0	20.0
$x_{11}^{(1)}$	-23.016	16.0	20.0
$x_{21}^{(1)}$	-22.101	16.0	20.0
cost		0.9984	0.8213
Y1		0.9062	0.7808
Y2		0.9497	0.8628
$\hat{F}$		0.0502	0.1342

The optimal design of the three layer PERCEPTRON for the binary symmetry recognition

Tab. 1

is used to denote the yield (the relative number of the acceptable net realizations) for the regions of acceptability R_{a_k}, $k = 1, 2$

$$R_{a_k} = \left\{ x \in \mathcal{R}^{13} \mid F(x) \leq \varepsilon_k \right\}, \varepsilon_1 = 0.05, \varepsilon_2 = 0.4.$$

The yield is one of the possible measures of the design quality.

The exact penalty method (see Ermoljev, 1976) was used to transform the problem to an unconstrained optimization problem and then stochastic approximation algorithm (9) was employed to get the solution. The number of the net output evaluations to get the optimal solution in Table 1 was approximately 5, 000. The tolerances in the second column of Table 1 are introduced only for comparison with those optimal ones.

We can roughly conclude that the errors arising in the process of this net realization will not cause an incorrect behavior of the realized hardware with a probability greater then Y_2, if they are in the range

$$t_i^* x_i^* / 100, \; i = 1, 2, \ldots, 13,$$

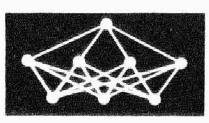
around the optimal nominal values x^* for the optimal vector of tolerances t^*.

The solution that we get by solving problem (31) possesses other interesting properties. If we calculate the gradient of the cumulative loss function according to the weights and compare it with the gradient according to the net inputs we will see that if the former is near to zero then the later is also near to zero. Thus the solution that is little sensitive to weight changes is also little sensitive to changes of input values. We can teach the net using precise samples $a^s \in \mathcal{A}$, $s \in \hat{v}$, despite the fact that the actual inputs from $\mathcal{A}$ activating the net in the process of an exploitation may be noise corrupted.

VII. CONCLUSION

Thus the potential differentiating operators of gradient estimation and corresponding averaging operators allow us to introduce simple criteria of sensitivity of a system to changes of its parameters and corresponding scalar measures of this sensitivity and to consider this sensitivity only with respect to selected parameters and selected system functions. This construction of sensitivity both suits the technical demands for design and fits with a natural idea of using a mean value of gradients of active system functions in a neighborhood of the nominal design and of utilizing the distribution of parameters that corresponds to the production technology when keeping the prescribed nominal parameters values and their tolerances. Also, the necessary range of parameters determined by the tolerances and correlations among parameters can be respected by this construction. Numerical calculations of the sensitivity and the sensitivity measure can be performed simply and also in parallel if it is necessary. If we are optimizing the sensitivity of system whose system functions are given stochastically, the construction based on LIPOs suppresses the random noise by which the values of system functions are corrupted. The so-constructed sensitivity to large changes of parameters converges, when the range of changes goes to zero, to the differential sensitivity. The sensitivity and the measure of sensitivity are smooth functions of parameters and generally, depending on the kernel of the used operator, they are more smooth then the original system functions due to the regularization property of the operators (11) and (14). The construction of sensitivity can be correctly used even if the system function are nondifferentiable in comparison to other approaches.

By using a potential averaging LIPO to construct the sensitivity measure of cumulative loss function (5), we take off non-characteristic local minima of the cumulative loss function F which make the teaching of the net by solving the problem in (4) very difficult. In



other words, using the sensitivity measure $\hat{F}(., t)$, instead of the cumulative loss function F, we easily reach the global solution of the problem given by (4) and we avoid the local minima of F at which the net does not generate the correct outputs. The problem of existence of such points is very serious and unavoidable because of the necessity for the net function as a function of configuration, to have a complex and quickly changing shape to be able to approximate a fairly wide class of functions.

References

[1] P. Antosik, J. Mikusinski, R. Sikorski: Theory of Distributions. The Sequential Approach, Amsterdam: Elsevier Scientific Publishing Corporation, 1973.

[2] J. W. Bandler, H. L. Abdel-Malek: Yield Optimization for Arbitrary Statistical Distributions: Part I-Theory, Part II-Implementation, IEEE Trans. on Circuits and Systems, **CAS-27**, no. 4, (1980), pp. 245-263.

[3] H. W. Bode: Network Analysis and Feedback Amplifier Design, New York: Van Nostrand, 1951.

[4] E. M. Buttler: Large Change Sensitivities for Statistical Design, Bell System Technical Journal, **50**, no. 4, (1971), pp. 1209-1224.

[5] F. H. Clarke: Optimization and Nonsmooth Analysis, New York: John Willey and sons, 1983.

[6] G. W. Davis: Sensitivity Analysis in Neural Net Solutions, IEEE Trans. on Systems, Man and Cybernetics, **SMC-19**, no. 5, (1989), pp. 1078 — 1082.

[7] S. W. Director, G. D. Hachtel: The Simplicial Approximation Approach to Design Centering, IEEE Trans. on Circuits and Systems, **CAS — 24**, no. 7, (1977), pp. 363 — 372.

[8] S. W. Director, G. D. Hachtel, L. M. Vidigal: Computationally Efficient Yield Estimation Procedures Based on Simplicial Approximation, IEEE Trans. on Circuits and Systems, **CAS-25**, no. 3, (1978), pp. 121-130.

[9] J. M. Ermoljev: Stochastic Programming Methods, Moscow: Nauka, 1976. (in Russian)

[10] K. Géher: Theory of Network Tolerances, Budapest: Akademiai Kiado, 1971.

[11] K. Hornik, M. Stinchcombe, H. White: Universal Approximation of an Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks, Neural Networks, **3**, no. 5, (1990), pp. 551-560.

[12] V. J. Katkovnik : Linear Estimates and Stochastic Optimization Problems, Moscow: Nauka, 1976. (in Russian)

[13] H. J. Kushner, D. S. Clark: Stochastic Approximation Methods for Constrained and Unconstrained Systems, New York: Springer-Verlag, 1978.

[14] L. Opalski, M. A. Styblinski, J. Ogrodzki: An Orthogonal Search Approximation to Acceptability Regions and its Application to Tolerance Problems, Proc. of the SPACECAD 79, Bologna, Italy, September 1979, (1979), pp. 163-167.

[15] T. Poggio, F. Girosi: Networks for Approximation and Learning, Proceedings of the IEEE, Special Issue on Neural Networks, I: theory & modeling, September 1990, (1990), pp. 1481 — 1497.

[16] E. Rumelhart, G. E. Hinton, R. J. Williams: Learning Representations by Back-Propagating Errors, Nature, **323**, No. 9, (1986), pp. 533-536.

[17] M. Stevenson, R. Winter, B. Widrow: Sensitivity of Feedforward Neural Networks to Weight Errors, IEEE Trans. on Neural Networks, **NN-1**, no. 1, (1990), pp 71 — 80.

[18] W. Strazs, M. A. Styblinski: A Second Derivative Monte Carlo Optimization of the Production Yield, Proc. of the 1980 ECCTD, Warsaw, Poland, September 1980, (1980), pp. 121-131.

[19] P. T. Thach: The Design Centering Problem as a D. C. Programming Problem, Mathematical Programming, **41**, (1988), pp. 229-248.

MATHEMATICAL APPENDIX

Example 1:

By employing Hermitte polynomials p_i^o, p_i^1, p_i^2, $i \in \hat{r}$, with the weights ρ_i

$$\rho_i(u_i) = \exp\left(-u_i^2/2\right), \quad -\infty < u_i < +\infty, \quad i \in \hat{r},$$

$$p_i^o(u_i) = (2\Pi)^{-\frac{1}{4}}, \; p_i^1(u_i) = (2\Pi)^{-\frac{1}{4}} u_i, \; p_i^2(u_i) =$$

$$= (8\Pi)^{-\frac{1}{4}} \left(u^2{}_i - 1\right)$$

and by applying the method of local approximation designed by Katkovnik, 1976, we get the averaging operators with the kernels

$$h^o(u) = h^1(\mathrm{u}) = (2\Pi)^{-r/2} \, \mathrm{e}^{-\|u\|^2/2}, \tag{35}$$

$$h^2(u) = h^3(u) = \left(1 + \frac{r}{2} - \frac{\|u\|^2}{2}\right) \mathrm{h}^o(u) \tag{36}$$

and the corresponding differentiating operators of first derivatives estimation with the kernels

$$g_i^1(u) = g_i^2(u) = -u_i h^o(u), \; i \in \hat{r}, \tag{37}$$

$$g_i^3(u) = g_i^4(u) = \left(2 + \frac{r}{2} - \frac{\|u\|^2}{2}\right) \mathrm{g}_i^1(\mathrm{u}) =$$

$$= -u_i \left(2 + \frac{r}{2} - \frac{\|u\|^2}{2}\right) \mathrm{h}^{\mathrm{o}}(\mathrm{u}), \; \mathrm{i} \in \hat{r}, \tag{38}$$

and, finally, the corresponding differentiating operators of second order derivatives estimation with the kernels

$$w_{ij}^2(u) = w_{ij}^3(u) = u_i \, u_j \, h^o(u), \; i \neq j,$$

$$w_{ii}^2(u) = w_{ii}^3(u) = \left(u_i^2 - 1\right) \mathrm{h}^{\mathrm{o}}(\mathrm{u}), \tag{39}$$

$$\textit{for } i, j \in \hat{r}$$

$$w_{ij}^4(u) = w_{ij}^5(u) = u_i \, u_j \left[3 + \frac{r}{2} - \frac{\|u\|^2}{2}\right] \mathrm{h}^{\mathrm{o}}(\mathrm{u}),$$

$$\mathrm{i} \neq j,$$

$$w_{ii}^4(u) = w_{ii}^5(u) = \left[\left(3 + \frac{r}{2} - \frac{1}{2}\|u\|^2\right)\right.$$

$$\left.\left(\mathrm{u}_i^2 - 1\right) + 1\right] \mathrm{h}^{\mathrm{o}}(\mathrm{u})$$

$$\textit{for } i, j \in \hat{r}. \tag{40}$$

The upper index by the kernel denotation is the order ϑ of the resulting operator and thus the maximal order of polynomials for which the operators give the exact values or the exact derivatives. Although the method of local approximation does not generally guarantee

the constructed differentiating operators to be potential, in this case, the operators with kernels (37) to (40) are potential.

If the probability density function p of normal distribution is used for generating vectors u, i. e.

$$p(u) = (2\Pi)^{-1/2} \, e^{-\|u\|^2/2},$$

then we obtain the following parametric estimates of values and derivatives of a function f with the aid of its values at L points u^s, $s \in \hat{L}$, generated using the probability density function p

$$^1\hat{f}(x, c, L) = \frac{1}{L} \sum_{s=1}^{L} f(x + cu^s), \tag{41}$$

$$^3\hat{f}(x, c, L) = \frac{1}{L} \sum_{s=1}^{L} \left[1 + \frac{r}{2} \frac{\|u^s\|^2}{2} \right] f(x + cu^s), \tag{42}$$

$$^2\nabla\tilde{f}(x, c, L) = \frac{(-1)}{cL} \sum_{s=1}^{L} u^s f(x + cu^s), \tag{43}$$

$$^4\nabla\tilde{f}(x, c, L) = \frac{(-1)}{cL} \sum_{s=1}^{L} u^s \left[2 + \frac{r}{2} - \frac{\|u^s\|^2}{2} \right] f(x + cu^s), \tag{44}$$

$$^3\tilde{f}''_{ij}(x, c, L) = \frac{1}{c^2 L} \sum_{s=1}^{L} u_i^s u_j^s f(x + cu^s), \quad i \neq j, \quad i, j \in \hat{r} \tag{45}$$

$$^3\tilde{f}''_{ii}(x, c, L) = \frac{1}{c^2 L} \sum_{s=1}^{L} [(u_i^s)^2 - 1] f(x + cu^s),$$

$$^5\tilde{f}''_{ij}(x, c, L) = \frac{1}{c^2 L} \sum_{s=1}^{L} u_i^s u_j^s \left[3 + \frac{r}{2} - \frac{\|u^s\|^2}{2} \right] f(x + cu^s), \quad i \neq j, \; i, j \in \hat{r} \tag{46}$$

$$^5\tilde{f}''_{ii}(x, c, L) = \frac{1}{c^2 L} \sum_{s=1}^{L} \left[\left(3 + \frac{r}{2} - \frac{1}{2} \|u^s\|^2 \right) ((u_i^s)^2 - 1) + 1 \right] f(x + cu^s).$$

The left upper index denotes the order. The indexes of exactness of the particular operators (see (24)) are in Table 2.

Estimate	(41)	(42)	(43)	(44)	(45) i≠j	(45) i=j	(46) i≠j	(46) i=j
Index of exactness	1	$1+\frac{r}{2}$	1	$2+\frac{r}{2}$	1	2	$3+\frac{r}{2}$	7+r

Tab. 2

Let us compare the accuracy of estimates (43) to (46) and the estimates based on the symmetric difference formulas when the first and second derivatives are estimated in parallel and the function f is given stochastically, i. e. we can only obtain its values in a form corrupted by a random noise with a variance σ. The symmetric difference formulas for

$$f'_i(x, c, \omega) = \frac{1}{2c} [f(x + ce^i, \omega) - f(x - ce^i, \omega)], \tag{47}$$

$$f''_{ij}(x, c, \omega) = \frac{1}{4c^2} [f(x + ce^i + ce^j, \omega) - f(x + ce^i - ce^j, \omega) - f(x - ce^i + ce^j, \omega) + f(x - ce^i - ce^j, \omega)] \tag{48}$$

require the function f to be calculated at $2r^2 + 2r + 1$ points.

Let there exist derivatives of the function f up to and including the fourth order and let them be bounded. The mean values of the square of the errors (19), (20) for the estimates (47), (48) are

$$\Delta^2_i = o_1(c^4) + \left[\frac{\sigma^2}{2c^2} + o_2(c^{-1}) \right],$$

$$\Delta^2_{ij} = o_{01}(c^6) + \left[\frac{\sigma^2}{4c^4} + o_{02}(c^{-3}) \right],$$

where the members $o_1(c^4)$ and $o_{01}(c^6)$ represent the square of systematic error (21). On the other hand, if we use the parametric estimates (43) to (46) for $L = 2(r^2 + r)$ we have the mean value of the square of the error for (43)

$$\Delta^2_i = o_{11}(c^4) + \left[\frac{\sigma^2}{2(r^2 + r)c^2} + o_{12}(c^{-1}) \right],$$

for (44)

$$\Delta^2_i = o_{21}(c^8) + \left[\frac{\left(\frac{r}{2} + 2\right)\sigma^2}{2(r^2 + r)c^2} + o_{22}(c^{-1}) \right],$$

for (45)

$$\Delta^2_{ij} = o_{31}(c^4) + \left[\frac{\sigma^2}{2(r^2 + r)c^4} + o_{32}(c^{-3}) \right], \; i \neq j,$$

$$\Delta^2_{ii} = o_{41}(c^4) + \left[\frac{\sigma^2}{(r^2 + r)c^4} + o_{42}(c^{-3}) \right],$$

and finally for estimate (46)

$$\Delta^2_{ij} = o_{51}(c^8) + \left[\frac{\sigma^2 \left(\frac{r}{2} + 3\right)}{2(r^2 + r)c^4} + o_{52}(c^{-3}) \right], \; i \neq j,$$

$$\Delta_{ii}^2 = o_{61}(c^8) + \left[\frac{\sigma^2 (r+7)}{2(r^2+r)c^4} + o_{62}(c^{-3}) \right].$$

We can conclude that the use of estimates (43) to (46) is evidently more advantageous in the case of small values of the averaging parameter c than the use of the estimates based on symmetric difference formulas. When the parametric estimate of gradient (43) is used the required number of values of the function f necessary for us to reach a demanded value of the variance of the estimate error (20) does not grow with the dimension of the problem r. Although it grows linearly with the dimension r when the difference formula (47) is applied.

■

Theorem 6. (Clarke , 1983)

Let a function f be Lipschitz in a neighborhood of a point x with a constant U. Then
1) $\underline{D} f(x)$ is nonempty, convex and closed subset of $\mathscr{R}^r$ and

$$\|\xi\| \leq U \text{ for any } \xi \in \underline{D}f(x).$$

2) The equation

$$f^o(x; v) = \max_{\xi \in Df(x)} <\xi, v>$$

is valid for any $v \in \mathscr{R}^r$.

3) If the function f is locally Lipschitz in $\mathscr{R}^r$ then the point-to-set map $\underline{D}f$ is an upper semi-continuous map from $\mathscr{R}^r$ into the system of all the subsets of $\mathscr{R}^r$ i. e.

$$(\forall \varepsilon > 0)\,(\exists \delta > 0)\,(\forall y, \|x - y\| \leq \delta)$$
$$\underline{D} f(y) \subset \underline{D}f(x) + B(0, \varepsilon),$$

where

$$B(0, \varepsilon) = \{ y \in \mathscr{R}^r \mid \|y\| \leq \varepsilon \}.$$

■

Proof of **Theorem 1 :**

Because $h \in \mathscr{Q}$ and $f \in \mathscr{Q}'$ integral (11) exists and a constant $K > 0$ can be found so that

$$\int_{\mathscr{R}^r} |h(u)|\, du \leq K < +\infty.$$

According to Definition 1 we have

$$\hat{f}(x, c) - f(x) = \int_{\mathscr{R}^r} h(u)\,(f(x - cu) - f(x))\, du.$$

If we have a ball $B(0, R)$ with the center at the origin and with a great enough radius R in relation to the preselected $\varepsilon > 0$ then

$$\int_{\mathscr{R}^r - B(0, R)} |h(u)|\, |f(x - cu)|\, du \leq \frac{\varepsilon}{3},$$

$$\int_{\mathscr{R}^r - B(0, R)} |h(u)|\, |f(x)|\, du \leq \frac{\varepsilon}{3}$$

due to the existence of these integrals over the whole space $\mathscr{R}^r$. Owing to the continuity of the function f at the point x, a number $\delta > 0$ can be found so that

$$|f(x) - f(y)| \leq \frac{\varepsilon}{3K}$$

for any $y \in \mathscr{R}^r$, $\|x - y\| < \delta$. If we set $c < \delta / R$ we obtain

$$|\hat{f}(x, c) - f(x)| \leq \int_{B(0, R)} |h(u)|\, |f(x - cu) - f(x)|\, du +$$
$$+ \int_{\mathscr{R}^r - B(0,)} |h(u)|\, |fx - cu)|\, du +$$
$$+ \int_{\mathscr{R}^r - B(0, R)} |h(u)|\, |f(x)|\, du \leq \varepsilon.$$

Because ε was arbitrary the theorem is proved.

■

Proof of Theorem 2:

Because the differentiating LIPO of gradient estimation is potential and the kernel of the corresponding averaging operator is h, where $g = \nabla h$, we can proceed in the same way as in the proof of Theorem 1 with

$$\nabla \tilde{f}(x, c) - \nabla f(x) = \int_{\mathscr{R}^r} h(u)\, \nabla f(x - cu)\, du.$$

■

Proof of Theorem 3 :

If the gradient $\nabla f(x)$ at the point, $x \in \mathscr{R}^r$ exists we can write according to Definition 2

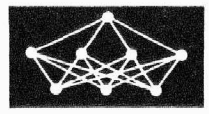
$$\lim_{c \to 0} \nabla \tilde{f}(x, c) = \lim_{c \to 0+} \frac{1}{c} \int_{\mathscr{R}^r} g(u)\, f(x - cu)\, du =$$
$$= - \lim_{c \to 0+} \int_{\mathscr{R}^r} g(u)\, \frac{f(x) - f(x - cu)}{c}\, du. \qquad (49)$$

Because the function f is Lipschitz the inequality

$$\left| \frac{1}{c} (f(x) - f(x - cu)) \right| \leq U \|u\| \qquad (50)$$

holds true. Due to the regularity of f (the existence of $\nabla f(x)$) at the point x, the existence of classical first derivatives $f(x;u)$ of f at the point x in any direction $u \in \mathscr{R}^r$ is assured

$$\lim_{c \to 0+} \frac{f(x) - f(c - cu)}{c} = f(x; u) = \nabla f(x)^T u. \qquad (51)$$



According to Lebesgue's theorem and due to the existence of majorant (25), we can change the limit and the integration in (49). Then we have

$$\lim_{c \to 0+} \nabla \tilde{f}(x, c) = - \int_{\mathcal{R}^r} g(u) \nabla f(x)^T u \, du = \nabla f(x) \tag{52}$$

by utilizing (15) and (16).

The limit (51) does not generally exist at the points where the function f is nondifferentiable. But e. g. the convexity of f in a neighborhood of the point x is sufficient for the existence of it because

$$\lim_{c \to 0+} \frac{f(x) - f(x - cu)}{c} = f(x; u) = \xi^T u,$$

where ξ is an element of the subgradient of f at the point x.

We want to prove a weaker result for the nonconvex case that claims that $\nabla \tilde{f}(x, c)$ approximates a point from the generalized gradient $\underline{D}f(x)$ if the averaging parameter c is small enough. We will use the following two lemmas.

Lemma 1: Separability of convex sets (Clarke, 1983,Ermoljev, 1976)

Let S be a closed convex subset of $\mathcal{R}^r$. Then for each $y \in S$ a vector $a \in \mathcal{R}^r, a \neq 0$, and a number $\varepsilon > 0$ exist such that

$$< a, x > \leq < a, y > - \varepsilon \tag{53}$$

for an arbitrary $x \in S$.

Lemma 2.

Let $h : \mathcal{R}^r \to \mathcal{R}$ be a nonnegative function and let

$$\int_{\mathcal{R}^r} h(u) \, du = 1.$$

Then for an arbitrary vector function $f : \mathcal{R}^r \to S$, where S is a convex, closed and bounded subset of $\mathcal{R}^r$ the following

$$\int_{\mathcal{R}^r} h(u) f(u) \, du \in S$$

is valid.

Proof of Lemma 2:

Let us denote

$$y = \int_{\mathcal{R}^r} h(u) f(u) \, du$$

and let $y \in S$. Claim (53) of Lemma 1. can be employed. Thus we have

$$< a, y > = \int_{\mathcal{R}^r} h(u) a^T f(u) \, du \leq \int_{\mathcal{R}^r} h(u) \left(a^T y - \varepsilon \right) du = < a, y > - \varepsilon.$$

We obtain a contradiction therefore Lemma 2 must be valid.

■

Evidently, there exists a ball with the center at the point 0 and with a radius $K > 0$ so that

$$\int_{B(0, K)} h(u) \, du = C \geq 1 - \varepsilon.$$

In the opposite case, the sequence $\{K_n\}$, $K_n \to \infty$, could be constructed so that the inequality

$$\int_{B(0, K_n)} h(u) \, du < 1 - \varepsilon$$

is valid and then we would obtain a contradiction.

$$1 = \int_{\mathcal{R}^r} h(u) \, du = \lim_{n \to \infty} \int_{B(0, K_n)} h(u) \, du \leq 1 - \varepsilon.$$

Due to the upper semi-continuity of the generalized gradient $\underline{D}f$ there exists a number $\delta < 0$ for which

$$(\forall y, \| x - y \| \leq \delta) \, \underline{D}f(y) \subset \underline{D}f(x) + B(0, \varepsilon)$$

holds true.

Let us set $c < \delta \; / \; \max \{1, K\}$. Then $\nabla f(x - cu) \in \underline{D}f(x) + B(0, \varepsilon)$ for almost all the $u \in B(0, K)$. Because $\underline{D}f(x) + B(0, \varepsilon)$ is a convex, closed and bounded set Lemma 2 is valid and thus

$$\frac{1}{C} \int_{B(0, K)} h(u) \nabla f(x - cu) \, du \in \underline{D}f(x) + B(0, \varepsilon). \tag{54}$$

Since f is locally Lipschitz $\| \nabla f(y) \| \leq U$ holds true for all the $y \in \mathcal{R}^r$ (see Theorem 5). Therefore we have

$$\| \int_{\mathcal{R}^r - B(O, K)} h(u) \nabla f(x - cu) \, du \| \leq \varepsilon U$$

and also

$$\int_{B(O, K)} h(u) \nabla f(x - cu) \, du \in \underline{D}f(x) + B(0, \varepsilon(1 + U + \varepsilon))$$

which give

$$\nabla \tilde{f}(x, c) \in \underline{D}f(x) + B(0, \varepsilon(1 + 2U + \varepsilon)).$$

Thus we can find η according to the statement of this theorem.

Let us note that the nonexistence of $\nabla f(x + cu)$ on the set of zero Lebesgue's measure does not effect the existence of integral (54) and its value.

■

Proof of Theorem 5 :

According to the assumptions of the theorem, we have

$$\frac{\partial p}{\partial u_i} = \lim_{\Delta \to 0+} \frac{p(u + \Delta e^i, 0, \sum) - p(u, 0, \sum)}{\Delta} \quad (55)$$

almost everywhere and also, because p is a density probability function, we have

$$p(u + \Delta e^i, 0, \sum) = p(u, -\Delta e^i, \sum),$$

and therefore (56) and (57) holds true

$$\int_{\mathcal{R}^r} \frac{p(u + \Delta e^i, 0, \sum) - p(u, 0, \sum)}{\Delta} du = 0, \quad (56)$$

$$\int_{\mathcal{R}^r} u_j \frac{p(u + \Delta e^i, 0, \sum)}{\Delta} du = \begin{cases} -1 \text{ for } i = j \\ 0 \text{ for } i \neq j \end{cases}. \quad (57)$$

The limit $\Delta \to 0+$ in expressions (56) and (57) can be performed because the conditions of Lebesgue's theorem are valid. Thus we proved the conditions of Definition 2. The potentiality of the constructed operator is evident.

■

Theorem 5 is also applicable when LIPOs are constructed on the basis of generalized function theory. The gradient and the limits of expressions (56) and (57) are then performed in the generalized sense because the scalar product is the regular operation, see Antosik, 1973.

Example 2.

Let us consider the density probability function of the uniform distribution

$$p\left(x, 0, \frac{1}{3} I\right) = \frac{1}{2^r} \prod_{j=1}^{r} [\Theta(x_j + 1) - \Theta(x_j - 1)],$$

($I \in \mathcal{R}^{rxr}$ is the unit matrix). Then we obtain the kernel of potential LIPO of gradient estimation with i-th component

$$g_i\left(x, 0, \frac{1}{3} I\right) = \frac{1}{2^r} \left\{ \prod_{j=1}^{r} [\Theta(x_j + 1) - \Theta(x_j - 1)] \right\}$$
$$[\delta(x_i + 1) - \delta(x_i - 1)],$$

where

$$\Theta(x_j) = \begin{cases} 0 & \text{for } x_j < 0 \\ 1 & \text{for } x_j > 0 \end{cases}$$

and $\delta(x_j)$ is the Dirack's δ — function.

■

Books alert

The following books can be interesting for the readers of our Journal

An Introduction to Neural and Electronic Networks. Ed. S. F. Zornetzer, J. L. Davis, and C. Lau. -San Diego, CA: Academic, 1990, 493 pp. , bound, 89. 95; paper, 39. 95, ISBN 0-12-781881-2.

Mind, Brain and the Quantum The Compound 'I". M. Lockwood. -Oxford, Basil Blackwell, 1989, 365 pp. , 29. 95. ISBN 0-631-16183.

Modeling Brain Function, The World of Attractor Neural Networks. D. J. Amit. -New York, Cambridge University Press, 1989, 504 pp., ISBN 0-521-36100-1.

Neural Computing — An Introduction. R. Beale, T. Jackson. -Bristol, Adam Hilger, 1990, 256 pp., ISBN 0-852-7462-2.

Neural and Concurrent Real-Time Systems:The Sixth Generation. B. Soucek. -New York, John Wiley & Sons, 1989, 389 pp.,

Neural Computing Architectures, the Design of Brain-Like Machines. Ed. I. Aleksander. -Cambridge, MA: MIT Press, 1989, 401 pp., ISBN 0-262-01110-7.

Neural Networks — Computers with Intuition. S. Brunak, B. Lautrup. -London, Academic Press 1990. 180 pp.

Both specialists and laymen will enjoy reading this book. Using a lively, non-technical style and images from everyday life, the authors present the basic principles behind computing and computers. The focus is on those aspects of computation that concern networks of numerous small computational units, be they biological neural networks or artifical electronic devices.

Naturally Intelligent Systems. M. Caudill, C. Butler. -Cambridge, MA: MIT Press, 1989, 304 pp. ISBN 0-262-03156-6.

The book is divided into three parts. After the introductory chapters, the first part deals with associative memories, the second with learning and memory, and the third with multilayered networks. A final set of chapters describes some implementations and applications.

Neurobiology of Learning and Memory — Reprint Volume. Ed. G. Shaw, J. McGaugh and S. Rose. -London, World Scientific 1990. 850 pp.

New Developments in Neural Computing. J. G. Taylor, C. L. T. Mannion. -Bristol, Adam Hilger, 1989, 264 pp., ISBN 0-85274-193-6.

New Developments in Neural Computing presents new information from researchers from all over the world, enabling workers to have access to the most up-to-date results.

Recursive Neural Networks for Asociative Memory. Y. Kamp, M. Hasler. -London, John Wiley & Sons, 1990, 216 pp., ISBN 0-47192-866-6.

Neural networks have received an upsurge of interest from a broad sector of the scientific community ranging from neurobiology to electronics and computer science and learning to a wealth of engineering applications in speech and image processing.

LIMIT INFORMATIONAL CHARACTERISTICS OF NEURAL NETWORKS CAPABLE OF ASSOCIATIVE LEARNING BASED ON HEBBIAN PLASTICITY.

A. A. Frolov)*

Abstract:

The capability of associative learning is one of the main properties of the brain. We share the idea (Palm, 1982; Kohonen, 1984) that design of devices modelling behavior of some biological organism as a whole can be based on associative memory mechanism. This idea is related to the one of Pavlov: that adapted animal behavior is based on the conditioning ability. A lot of experimental data on neurophysiology of associative learning has been accumulated since Pavlov. Associative memory models have been developed simultaneously to generalize experimental data and to create the basis for further experiments (Rosenblatt, 1959; Konorsky, 1970; Hebb, 1949; Steinbuch, 1961; Willshaw et al. , 1969; Briendley, 1969; Marr, 1969, 1970, 1971; Palm, 1981, 1982; Kohonen, 1980, 1984; Hopfield, 1982, 1984 etc.). As a result of experimental and theoretical research, the following common understanding of learning and memory problems in the nervous system has been reached.

1. Introduction

It is considered that the activity of the nervous system or any part of it may be described by vector A, whose components A_i are activities of individual neurons. A_i is the characteristic of neuron's instantaneous frequency of action potentials or probability of its excitation. Current activity of each neuron primarily depends on the previous activity of other neurons which influence the former through synaptic connections. The level of such influence is determined by synaptic weights. The activity of the input neurons is also determined by external signals. So, current activity of the neural network is determined by its previous activity history, current weights of synaptic connections and current pattern of external signals. Each external event is coded by a certain activity vector A or a sequence of such vectors, and retrieval of this event from the memory corresponds to the setting of nervous system activity pattern close to the storage one. Storage in the memory is based on modification of plastic elements of the nervous system. Synapses are ordinarily considered to be such elements. But in some of ours papers (Frolov, Murav'ev, 1987, 1988 a, b etc.), as well as in articles by neuron as a whole, they are considered to be memory elements. It is supposed that dynamics of the plastic elements modification is „localized" (Braitenberg, 1978), that is, the change of each memory element depends only on its current plasticity state and the current neural activity in its location point (and, possibly, on some signals which modulate the general level of the whole network's plasticity). Hebbian plasticity (Hebb, 1949) is the most popular for memory modelling among the types of synaptic plasticity which satisfy the condition of plasticity localization.

According to the Hebbian rule, the synaptic weight W_{ij} between i-th and j-th neurons depends on the correlation between the activities A_i and A_j of the post— and presynaptic neurons. Here we consider three types of Hebbian plasticity: one gradual and two binary ones. For convenience we introduce a vector S which we call modification state vector. Each component of this vector is definitively related to the weight of one of modifiable synapses. For gradual Hebbian plasticity we suppose:

$$\Delta s = cA_iA_j, \tag{1}$$

where s is the state of modification of the synapse between i-th and j-th neurons (initial value of s is assumed to be zero), c is the rate of learning which can vary from 0 for unimportant events which need not be stored to some maximal value for the most important events which must be stored after the single occurrence. Here it is assumed that $c=(0, 1)$: $c=1$ for learning, $c=0$ for retrieval. For binary plasticity we assume

$$\Delta s = c(1 - s)\, A_i\, A_j \tag{2}$$

where again $c=(0, 1)$, initially $s=0$ and neuron's activity is assumed to be binary: $A_i=(0, 1)$. After the learning is finished, in the case of gradual plasticity we obtain

$$s = \sum_{k=1,\,L} A_i^k\, A_j^k \tag{3}$$

where L is the total number of stored events, while in the binary case s is equal to unity if at least for a single stored event the activity of the presynaptic

*) Prof. A. A. Frolov
Institute of Higher Nervous Activity and Neurophysiology USSR Academy of Sciences
Butlerova 5a
142 292 Moscow, USSR

neuron has occurred simultaneously with the activity of the postsynaptic neuron.

For the binary plasticity we distinguish two types of synapses. A synapse with zero weight before modification (that is for $s=0$) is called the Hebbian synapse. A synapse with weight after modification (that is for $s=1$) is called Albus synapse. These definitions originate from papers by Hebb (1949) and Albus (1972) where these types of synapses have been first mentioned. Ordinarily the weight of a binary synapse in a nonzero state is assumed to be $+1$ or -1.

2. Main notions and definitions.

A neural system is called heteroassociative memory if it performs the following functions:

a) In the learning mode ($c=1$) L pairs of vectors (X_k, Y_k) belonging to X^{n_x} and Y^{n_y}, where X^{n_y} and Y^{n_y} are vector sets having dimensions respectively n_x and n_y, are consequently presented to the system to be stored. Vectors X_k and Y_k are called templates.

b) In the retrieval mode ($c=0$) any vectors $X' \in X^{n_x}$ are consequently presented to the system. All those belonging to the given templates vicinities are called „familiar" and others are called „novel". The system recognizes familiar and novel among $\{X'\}$ *by some decision rule. If* X' is recognized as familiar then a vector $Y' \in Y^{n_y}$ is reproduced at the system output. If recognition is correct, then Y' contains the information $I(Y_k, Y')$ about the template Y_k coupled to X_k to which X' is close.

A neural system is called autoassociative memory if it performs the following functions:

a) In the learning mode ($c=1$) L vectors $X_k \in X^{n_x}$ are consequently represented to the system to be stored.

b) In the retrieval mode ($c=0$) the system recognizes familiar and novel among vectors $X' \in X^{n_x}$ presented to it. If some vector X' has been recognized as familiar the system output will reproduce a vector $X'' \in X^{n_x}$. If the recognition has been correct, the vector X'' contains the information about the template X_k to which X' is close. This information is additional to the one contained in the input vector X'. As a rule, $I(X'', X) > I(X', X_k)$. Then the autoassociative memory fulfills the correction of X'. However, if $I(X_k,(X',X'')) > I(X_k,X')$, then some additional information about X_k can be extracted from X'' any case.

In the following, as a rule (and always for binary placticity) it is assumed that templates are binary vectors belonging to sets $B_l^{n_x}$ and $B_l^{n_y}$, where B_l^n is a set of vectors containing l units and n-l nulls. Additionally, it is assumed that templates are chosen equiprobably and independently of each other.

Heteroassociative memory simulates the procedure of behavioral classical conditioning. Autoassociative memory simulates the development of „local conditioned reflexes", "neural models of stimuli" or „memory engrammes". The recognition function of associative memory corresponds to the one postulated for the nervous system as „novelty detection" (Vinogradova, 1975).

For a neural network performing functions of autoassociative memory input and output layers consist of equal numbers of neurons which areequivalent as informational units. Therefore there exist a lot of autoassociative memory models in which input and output layers are combined into a single layer, considered either input or output in different moments.

Autoassociative memory may operate in single-step or multi-step modes. In the second case its output layer is connected with its input one, and the retrieval of the template stored in the memory takes place step-by-step in portions. As a rule there exists a stable state for each of stored template to which the network activity converges as a result of this excitatory reverberation, and the network may be considered as a dynamic system with a lot of stable states (Little, 1974; Hopfield, 1982). If the output layer is combined with the input one. The network need not any additional connections to reproduce the multi-step mode.

During retrieval each of the stored templates is decoded separately, that is, information about other templates which has been extracted from the memory, is not being used. Such decoding is called simple unlike the complex one, making use of this information (Dunin-Barkovski, 1978). It will be shown below that one of the main sources of information losses in the neworks performing functions of associative memory is the use of simple decoding instead of complex one. Changes of the memory elements' modification states caused by the storage of other templates produce background noise and prevent retrieval of the given template. It is the interference of traces of different templates stored that results in information losses.

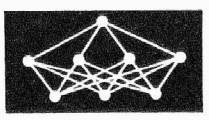
According to the Shannon theorem, the maximal information which can be extracted from the network using any manner of complex decoding is defined by

$$I(S, \hat{X}) = \sum P(S, \hat{X}) \lg (\mathrm{P}(\mathrm{S}, \hat{\mathrm{X}})/P(\mathrm{S})/P(\hat{\mathrm{X}})) = \mathrm{H}(\hat{\mathrm{H}}) - H(\hat{\mathrm{X}}/\mathrm{S}) = H(\mathrm{S}) \tag{4}$$

for autoassociative memory, while for heteroassociative one

$$I(S, (\hat{X}, \hat{Y})) = \Sigma P(S, (\hat{X}, \hat{Y})) \lg (\mathrm{P}(\mathrm{S}, (\hat{\mathrm{X}}, \hat{\mathrm{Y}}))/P(\mathrm{S})/P(\hat{\mathrm{X}}, \hat{\mathrm{Y}})) = H(\mathrm{X}, \mathrm{Y}) - H(\mathrm{X}, \mathrm{Y}/S) = H(\mathrm{S}), \tag{5}$$

where S is the modification state vector of the whole network, $\hat{X}$ and $\hat{X}$, $\hat{Y}$ are full sets of templates and their pairs, P is the probability of their joint distribution,



$$H(V) = \sum P(V) \lg P(V)$$

is the entrophy of a random variable V.

In these equations it is taken into account that for rules of modification given by the formulae (1), (2) the state of modification is completely determined by the stored templates; therefore $H(S/\hat{X}) = H(S/(\hat{X}, \hat{Y})) = 0$.

Maximum of information which may be extracted from the network using any manner of simple decoding is defined by $I = LJ$, where J is maximal information extracted from the network for individual template or their pair, that is

$$J = I(S, X_k) = \sum P(S, X_k) \lg (\mathrm{P}(\mathrm{S}, \mathrm{X_k})/P(\mathrm{X_k})) = \\ = H(\mathrm{X_k}) - H(\mathrm{X_k/S}) = H(\mathrm{S/X_k}) \qquad (6)$$

for autoassociative memory, while for heteroassociative one

$$J = I(S, (X_k, Y_k)) = \sum P(S, (X_k, Y_k)) \\ \lg (\mathrm{P(S, (X_k, Y_k))}/P(\mathrm{S})/P(\mathrm{X_k, Y_k}) = \\ = H(\mathrm{X_k, Y_k}) - H(\mathrm{X_k, Y_k/S}) = H(\mathrm{S}) - H(\mathrm{S/X_k, Y_k}) \qquad (7)$$

Here it is assumed that information quantities extracted from the network are equal for all templates, so k is an arbitrary template number.

If for one of modification states the plastic synapses have zero weights, then the reason of „why does not one neuron react to excitation of another one" cannot be established by testing network's reactivity: it can be either due to absence of the connection between them or the zero weight of such connection. In this case, network testing permits us to learn the structure of network connections with nonzero synapses only. Therefore, to calculate the maximal information, which can be extracted from the network by complex or simple decoding it is necessary to replace in (4) — (7) vector S by the vector σ which determines modification states of nonzero synapses only. The arising uncertainty about the network structure is the second main source of information losses. For the gradual plasticity only a small part of modifiable synapses are in the zero state, so these information losses may be neglected. For the binary plasticity they depend on the modification state in which the synapses have zero weight, i. e. , if synapses belong to Hebb or Albus type.

For decoding routines mentioned above, i. e. corresponding to associative memory functioning, the maximal information values given by (6), (7) and especially by (4), (5) cannot be achieved. They must be considered only as some reference evaluations similar to the one for a heat engine efficiency given by the second thermodynamics law.

It is evident from equations (4), (5) that information which can be extracted from the network by any manner of decoding cannot exceed $H(S)$ and $H(\hat{X})$ for autoassociative memory or $H(\hat{X}, \hat{Y})$ for heteroassociative one. In its turn $H(S)$ cannot exceed $\lg M$ where M is the total number of different modification states of the network. If all memory elements have identical properties then $\lg M = N \lg K$ where N is the total number of memory elements and K is the number of modification states of each element. These limit values give reason for definitions of the main information characteristics of a neural network : the efficiency coefficient

$$E = I/(N \lg K), \qquad (9)$$

and quality coefficient

$$Q = I/H(X) \text{ or } Q = I/H(\hat{X}, \hat{Y}) \qquad (10)$$

where I is the total information extracted from the memory with help of a given decoding routine. A typical dependence of E and Q on the number of stored templates L is shown in Fig. 1. When this number is relatively small the information extracted from memory is equal to the entrophy of the templates, that is, this information is sufficient to retrieve the templates without any errors or „ideally" as Dunin-Barkowski (1978) put it. When L reaches some critical value, the quality coefficient begins to decrease and efficiency coefficient passes its maximum. The search of this critical value of L or of the maximal value of E (which is basically the same) is the main goal of the informational analysis of neural networks performing the function of associative memory.

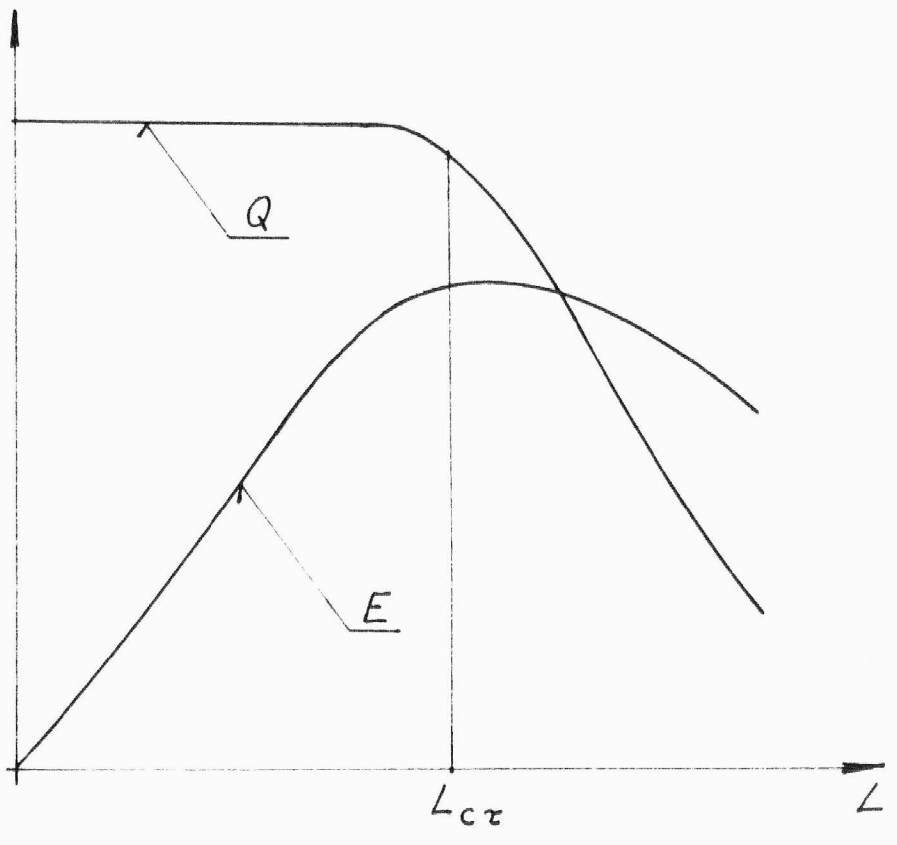


Fig. 1. Typical dependence of the main informational characteristics Q and E on the number of stored templates L.

Here we analyze only two-layered networks containing n_x neurons in the input layer, n_y neurons in the output layer and no hidden neurons. This construction is the simplest for networks, which are able to perform the functions of the associative memory. It

was first analyzed by Steinbuch (1961) for the gradual plasticity and by Willshaw et al. (1969) for the binary one and is called „correlation matrix". For autoassociative memory $n_y = n_x$ and input and output layers may be combined into a single one. Such construction of autoassociative memory has been considered by Hopfield (1982, 1984) and by a lot of his followers (Amit et al. , 1985, 1986, 1987; etc.). In the following it is assumed that each output neuron is connected with m input neurons with connection vector η, and vectors η_i ($i = 1, n_y$) are chosen equiprobably and independently of each other and of the templates from the set $B_m^{n_x}$ (η_{ij}) = 1 if the i-th output neuron is connected with the j-th input one, η_{ij} in the opposite case).

The analysis given below is restricted to evaluation of E from formulae (4)-(7). Therefore we obtain values of this coefficient which can serve as referent ones for any natural decoding routine and helps us to understand the nature of the main sources of information losses. Calculation of informational characteristics of neural networks for a few decoding routines can be found in (Frolov, Murav'ev, 1988, a).

3. The case of a single output neuron.

For a well-designed neural network, the information capacity must have the order of the sum of information capacities of its individual neurons. Thus the calculation of a single neuron information capacity gives a good estimate for the information capacity of the whole network, and, moreover, it enables us to understand the nature of information losses which occur with connection of individual neurons into the integral network.

Gradual plasticity. In the case of gradual plasticity only a small part of synapses are in the zero state. Therefore, we may consider vector η to be completely known. Then from the formulae (4) and (5)

$$I(S, (\hat{X}, \hat{Y})) = H(S) \text{ and } I(S, \hat{X}) = H(S)$$

where y is one of the components of the template Y corresponding to the given output neuron, $\hat{y}$ is the set of components y_k and S is the vector of the modification state for this neuron. From (3)

$$s_j = \sum_{k=1,L} A_{kj} y_k \quad j = 1, \ldots, m, \qquad (11)$$

where A_{kj} (j = 1, . . . , m) are the components of the vector X_k corresponding to input neurons which are linked with the given output neuron. Let components of the templates X_k and the variables y_k be statistically independent and have zero means. Then components of vector S are uncorrelated. Under a sufficiently large value of L one may approximate the distribution of s_j by the normal distribution with zero mean and variance $D = Ld_x d_y$ where d_x, d_y are variances of X_{ki} and Y_{ki}, ignoring statistical dependence of different components of S. Then we may assume

$$H(S) = m\ H(s_j) = \frac{m}{2} \lg (2\pi e d_x d_y L), \qquad (12)$$

where we use the well-known formula (Kolesnik, Poltyrev, 1982) for the entropy of a normally distributed variable. To calculate the number of different states of one modifiable synapse K we may ignore the existence of states with absolute values of s_j which exceed $D^{1/2}$ several times. Therefore, we may assume K to be proportional to $D^{1/2}$. From the formula (11) for sufficiently large L

$$E = \frac{m\ H(s_j)/2}{m \lg K} \cong 1$$

Thus gradual Hebbian plasticity in general may reach maximal possible efficiency. But this is true only for unreal case of complex decoding.

From the formulae (6), (7), the maximal information which can be extracted from the single output neuron for any manner of simple decoding is given by $I = LJ$ where

$$J = H(S) - H(S/X_k, Y_k) \text{ or } J = H(S) - H(S/X_k).$$

But if the template X_k or pair of templates (X_k, Y_k) are known then from the formulae (11) its contributions to vector S are also known therefore $H(S/X_k)$ or $H(S/(X_k, Y_k))$ are equal to the entrophy of the vector of modification state produced by the storing of the other templates. Therefore $H(S/X_k)$ or $H(S/(X_k, Y_k))$ are equal to $H(L\text{-}1)$ where $H(L)$ is given by the formula (12). Then

$$J = H(L) - H(L-1) =$$

$$= \frac{1}{2} m \lg (L/(L-1)) \cong m \lg e/(2L),$$

$$I = LJ = m \lg e/2 \qquad (13)$$

Thus the maximal information which can be extracted from one graduel Hebb synapse does not depend on the number of its gradation states, i. e. , for the gradual Hebbian placticity the passage from complex to simple decoding leads to a substantial decreas of the information capacity of the neural net and a corresponding drop in its efficiency coefficient. Threfore in the following we use for gradual placticity coefficient $E' = I/N$ instead of coefficient E determined by the formula (9). From the formula (13) we get

$$E' = I/m = \lg e/2 \cong 0{,}72 \qquad (14)$$

Binary plasticity. Vector η is known compleatly. For binary templates which are considered for binary plasticity it is evident that for a single neuron $I(S, (\hat{X}, \hat{y}))$ or $I(\hat{S}, \hat{X})$ are equal to $I(S, \tilde{X})$ where $\hat{X}$ is the set of tem-

plates X_k which activations at the input layer coincide with activations of the given neuron at the output layer. Therefore for the binary plasticity we may restrict ourselves to the calculation of $I(S, \tilde{X})$ and use for it formula (4) where we must replace $\hat{X}$ by $\tilde{X}$. Since all synapses are statistically identical one may put $P(S) = P(k)/\ C_m^k$ where k is the number of modified synapses of the given neuron, therefore $I = I' + I''$ where

$$I' = \sum_k P(k) \lg C_m^k \quad I'' = -\sum_k P(k) \lg P(k)$$

If $m \gg 1$ we may approximate the distribution $P(k)$ by a normal one with mean $M(k) = \varkappa m$ and some variance $D(k)$ where $\varkappa$ is the probability of modification of the given synapse after recording of all templates. Then

$$I'' = \frac{1}{2} \lg (2\pi e D(k)).$$

Using Stirling expansion for C_m^k one may put

$$I' = \sum P(k) \left[mh(k/m) - \frac{1}{2} \lg (2\pi k(1 - k/m)) \right].$$

Then expanding the expression between the square brackets in the Tailor series in the vicinity of $k = M(k)$ and neglecting the terms of the order $(1/m)$ one may put

$$I' = mh(\varkappa) - \frac{\lg e}{2m\varkappa(1-\varkappa)} D(k) - \frac{1}{2} \lg (2\pi\varkappa(1-\varkappa) m)$$

Then

$$I = mh(\varkappa) - \frac{1}{2} \frac{D(k)}{m\varkappa(1-\varkappa) \ln 2} + \frac{1}{2} \lg \frac{eD(k)}{\varkappa(1-\varkappa) m} \tag{15}$$

This expression reaches its maximum $mh(\varkappa)$ for $D(\varkappa) = m\varkappa(1-\varkappa)$, i. e. , when the states of modificaition of different synapses are statistically independent and the distribution $P(k)$ is binomial. But it has been shown (Frolov, Murav'ev, 1987) that if $l_x \ll n_x$ then $D(k)$ is of m-th order. Therefore, for $m \gg 1$ two last terms in formula (15) may be neglected relative to the first one and we may put $I = mh(\varkappa)$. Under these conditions one may neglect the presence of statistical dependence of different synapses of a single neuron to evaluate its information capacity. For the binary plasticity $K = 2$ thus from the formula (3) $E = h(\varkappa)$. For $\varkappa = 1/2$ this expression reaches its maximum which is equal to 1. Thus the efficiency of the binary Hebbian plasticity also in general may reach its maximal feasible value.

To evaluate the information capacity of a single neuron under the simple decoding, let its different synapses be assumed to be statistically independent at once. Then one may put $J = mJ'$ where J' is the information extracted from a single synapse. To calculate J', the formula (6) can be rewritten in the following form:

$$J' = I(s_j, \tilde{x}_j) - I(s_j, \tilde{x}_j/x_{kj}), \tag{16}$$

where $I(s_j, x_j)$ are equal to $H(s_j) = \mathrm{h}(\varkappa)$ and x_{kj} is a component of one of the vectors of the set $\tilde{X}$.

For binary plasticity the state s_j under given k-th template or templates pair is compleatly determined only if $\Delta s = 0$ during its storing. In this case

$$I(s_j, \tilde{x}_j/x_{kj}) = I(s_j, \hat{x}_j), \tag{17}$$

where $\hat{x}_j$ is the set of template components $\tilde{x}_j$ without given x_{jk}, i. e. , this information is equal to $h(\varkappa')$ where $\varkappa'$ is the probability of modification of given synapse after the storing of L'-1 templates, where L' is the total number of the templates in $\tilde{X}$. In the opposite case these informations are equal to zero. Since $\Delta s_j = 0$ with the probability 1-q where q is the probability of presynaptic activation during the recording of given template of $\tilde{X}$, then $I(s_j, \tilde{x}_j/x_{kj})$ are equal to $(1-q)\, h(\varkappa')$. Therefore

$$J' = h(\varkappa) - (1-q)\, h(\varkappa').$$

Since templates are assumed to be statistically independent, then

$$\varkappa = 1 - (1-q)^{L'} \text{ and } \varkappa' = 1 - (1-q)^{L'-1}.$$

For $L' \gg 1$ one may put $q \ll 1$, then

$$\varkappa = 1 - \exp(-qL'), \; J' = q \ln(1/\varkappa), \tag{18}$$

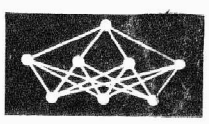
$$E = LJ' = \ln(1/(1-\varkappa)) \ln(1/\varkappa). \tag{19}$$

The same result for efficiency coefficient has been obtained in (Frolov, Murav'ev, 1987, 1988, b) more accurately and in (Dunin-Barkowski, 1978). The maximum of the efficiency coefficient is reached at $\varkappa = 0.5$ and amounts $\ln 2 \cong 0{,}69$, i. e. , the information capacity of the network with binary plasticity is only by 4 per cent less then in the case of gradual plasticity. Such equivalency of gradual and binary plasticity has been noted many times on the computer simulations.

Hebb and Albus synapses. For these types of synapses on decoding not the vector S but the vector ξ is known such that $\xi_j = 1$ if $\eta_j = 1$ and the corresponding synapse has non-zero weight. In the opposite case $\xi_j = 0$, then one must use in the formulae (4) and (6) vector ξ instead of S. Let us ignore as in the previous case the presence of statistical dependence between different synapses of a given neuron. Then for complex decoding we may put $I = nI'$ where

$$I' = H(\xi_j) - H(\xi_j/\tilde{x}_j)$$

is the information which can be extracted from one



component of vector ξ, $H(\xi_i) = h(\varkappa m/n)$ for the Hebb synapses and $H(\xi_i) = h((1-\varkappa)\, m/n)$ for the Albus ones, $\varkappa m/n$ and $(1-\varkappa)\, m/n$ are the probabilities that given input and given output are linked by the synapse with non-zero weight. To calclulate $H(\xi_i/\bar{x}_j)$ let us introduce the binary variable ε which is equal to the modification state of the synapse if given input and output neurons are linked. For given set $\bar{x}_j$ the value of ε is known. If $\varepsilon = 0$ for the Hebb synapses or $\varepsilon = 1$ for the Albus ones then $\xi_i = 0$ in advance and we cannot extract any information from this event. In the opposite cases this information is equal to $h(m/n)$. Therefore for the Hebb synapses

$$I' = h(\varkappa m/n) - \varkappa h(m/n)$$

and for the Albus ones

$$I' = h((1-\varkappa)m/n) - (1-\varkappa)\, h(m/n).$$

The same equations have been obtained more accuratly in (Frolov, Murav'ev, 1987, 1988, b). At $m = n$ for both types of synapse we have

$$I' = h(\varkappa) = h(1-\varkappa), \quad \mathrm{E} = h(\varkappa),$$

which coincides with formulae obtained for the case when vector η is compleatly known. This is natural since for $m = n$ given neuron is linked with all input neurons, i. e. the vector η is known in advance. For $m \ll n$ for the Hebb synapses

$$I' = (m/n)\varkappa \lg(1/\varkappa), \quad E = \varkappa \lg(1/\varkappa) \qquad (20)$$

and for the Albus ones

$$I' = (m/n)(1-\varkappa) \lg(1/(1-\varkappa)),$$
$$E = (1-\varkappa) \lg(1/(1-\varkappa)). \qquad (21)$$

The maximum E is reached for $\varkappa = 1/e$ for the Hebb synapses and for $\varkappa = 1-1/e$ for the Albus ones and amounts $(\lg e)/e \cong 0{,}53$.
It is easy to show that expressions (16) and (17) also remain true for the Hebb and Albus synapses. Then for $m \ll n$ for the Hebb synapses from the formulae (20) one may put

$$J' = (m/n)(\varkappa \lg(1/\varkappa) - (1-q)\varkappa' \lg(1/\varkappa')) \cong$$
$$\cong q(\ln(1/\varkappa) + \varkappa \mp 1)$$
$$E = nL' J'/m = \lg(1/(1-\varkappa))(\ln(1/\varkappa) - \varkappa - 1) \qquad (22)$$

and for the Albus ones

$$J' = (m/n)((1-\varkappa) \lg(1/(1-\varkappa)) -$$
$$-(1-q)(1-\varkappa') \lg(1/(1-\varkappa'))) \cong$$
$$\cong (m/n)\, q(1-\varkappa) \lg 2, \quad \mathrm{E} = (1-\varkappa) \lg(1/(1-\varkappa)).$$

The same expressions were obtained more accuratly in (Frolov, Murav'ev, 1987, 1988, b).
It is interesting to note that for the Albus case there are no information losses on passing from complex to simple decoding. For the Hebb case expression (22) reaches its maximum for $\varkappa = 0.24$ which amounts to 0.26, i. e. on passing from complex to simple decoding the efficiency of the Hebb synapses shows more than two fold decrease.

4. Two-layer networks

To illustrate the nature of the information losses produced by the connection of the individual neurons into the integral network we restrict our analysis by the case of fully connected networks with gradual synaptic plasticity and normally distributed templates with zero means. So the information capacity of the network which will be calculated below must be compared with the capacity of a single neuron given by formula (12). Moreover we consider only autoassociative memory which has been thoroughly investigated by other methods (Hopfield, 1982, 1984; Amit et al., 1985, 1986, 1987). It permits to compare the results of different approaches. The analysis of heteroassociative memory one can find in (Frolov, Murav'ev, 1988, a)
For convenience' sake we introduce the matrix M instead of vector of modification state S. From the formula (3) $M = XX^T$ where X is the matrix which columns are formed by the templates X_k and X^T is the matrix transposed to X. Since for the autoassociative memory $n_x = n_y$ let us denote the number of neurons in the both layers by n. Thus matrix M has n rows and L columns. The maximal information which can be extracted from the network is given by $H(M)$. It is known (Girko, 1980) that for normally distributed X_k and $L < n$

$$P(M) = \frac{|M|^{(L-n-1)/2}\, e^{TrM/(2d_x)}}{(2d_x)^{nL/2} \pi^{n(n-1)/4} \prod_{k=1,n} \Gamma((L+1-k)/2)} \qquad (23)$$

M is the symmetric positively determined matrix, $P(M) = 0$ in the opposite case. In the formula (23) d_x is the dispersion of the template components, TrM is the trace of matrix M and $|M|$ is its determinant. Then

$$H(M) = H_1 + H_2 - H_3 + H_4$$

where

$$H_1 = \frac{1}{2} nL \lg(2d_x) + \frac{n(n+1)}{4} \lg \pi$$

$$H_2 = \sum_{k=1,n} \lg \Gamma((L+1-k)/2)$$

$$H_3 = (\mathrm{L} - n - 1)\,\mathrm{M}\,\{\lg |M|\}$$

$$H_4 = \frac{1}{2d_x} M\{TrM\} = \frac{1}{2} Ln$$

For calculation of H_2 let Γ-function be approximated by the formula of Stirling (Korn, Korn, 1968) and summation be interchanged by integration. Then

$$H_2 \cong \left[n(2L - n) \ln (\mathrm{L}/2)/4 - \right.$$
$$-(\mathrm{L} - n)^2 \ln(1 - (\mathrm{n}/L))/4 -$$
$$\left. - n(\mathrm{L} - n)/2 - n^2/4 \right] \lg e.$$

For calculation of H_3 it may be noted that $|M| + |XX^T| = V_n^2$ where V_n is the volume of the parallelepiped formed by n vectors-rows ξ_i (i=1, n) of the matrix M in the L-dimensional space (Gantmakher, 1966). Let V_l be the volume of the parallelepiped formed by the first l vectors of ξ_i. Then $V_{l+1} + V_l \| \xi^*_{l+1} \|$ where ξ^*_{l+1} is the component of ξ_{l+1} which is orthogonal to the space formed by the the first l vectors ξ_i. Since all ξ_i are statistically independent then $\| \xi^*_{l+1} \|^2 = d_x \chi^2_{L-l}$ where χ^2_{L-l} is the random veriable which has χ-distribution with L-l degrees of freedom. Therefore,

$$M(\ln |M|) = M\left\{ \sum_{i=1,\,n} \ln \| \xi^*_l \|^2 \right) = n \ln d_x + F$$

where

$$F = \sum_{l=L-n+1}^{L} \mathrm{I}(l), \quad \mathrm{I}(l) = M(\ln \chi_l)^2.$$

It is known (Korn, Korn, 1968) that the variable χ^2_l has the following density function

$$(\eta) = \begin{cases} 0 & \text{if } \eta < 0 \\ \dfrac{1}{\Gamma(\lambda/2)2^{1/2}}\, \eta^{(1-2)/2}\, e^{-\eta/2} & \text{if } \eta > 0 \end{cases}$$

that is

$$I(l) = \frac{1}{\Gamma(l/2)2^{1/2}} \int_0 \ln \eta\, \eta^{(1-2)/2}\, e^{-\eta/2}\, d\eta =$$
$$= \psi(l/2) + \ln 2$$

where $\psi(x) = d\ln \Gamma(x)/dx$ and integral value is obtained by (Gradstein, Ryzhik, 1963). Changing summation in formula (23) by integration one may get

$$F = n \ln 2 + 2 \ln \Gamma(L/2) - 2 \ln \Gamma((L - n)/2) \cong$$
$$\cong n \ln(\mathrm{L}/e) - (\mathrm{L} - n) \ln(1 - n/L).$$

Thus

$$M(\ln |M|) \cong n \ln(\mathrm{L}d_x) - (\mathrm{L} - n) \ln(1 - n/L) - n,$$

$$H_3 \cong \frac{1}{2}(\mathrm{L} - n)\mathrm{n} \ln(\mathrm{L}d_x/e) - \frac{1}{2}(\mathrm{L} - n)^2 \lg(1 - n/L),$$

$$H(M) \cong \frac{n^2}{4} \lg(2\pi e d_x^2 \mathrm{L}) + (\mathrm{Ln}/4 - 3n^2/8) \ln e +$$
$$+ \frac{(L - n)^2}{4} \lg(1 - n/L), \qquad (25)$$

$$E' = H(M)/n^2 \cong \frac{1}{4} \lg(2\pi e d_x^2 L) -$$
$$- \left[(3/8 - L/(4n)) \lg e - \right.$$
$$\left. - \frac{1}{4} (\mathrm{L}/n - 1)^2 \lg(1 - n/L) \right]. \qquad (26)$$

The first term in (25) is similar to the expression (12) but is two fold less. This term corressponds to the case of statistically independent modification states of different synapses. Two fold decreasing is explained by the fact that matrix M is symmetrical relative to the main diagonal, therefore, only one half of its element may be considered as indepentent in principle. The same two fold decreasing of efficience coefficient occures evidently for heteroassociative memory when its output layer is linked with the input one by backward connections forming the so called bidirectional memory. Such linking inproves the dynamical properties of the network but dos not enhances its limit information capacity.

The expression between square brackets in (26) reflects the influence of statistical dependence between the elements of a half of matrix M under or below the main diagonal. Dependence of this expression on (L/n) is shown in Fig. 2. It may be seen that this influence is relatively large only if $L \cong n$. As one may expect, this influence results in a slight decrease of efficience coefficient.

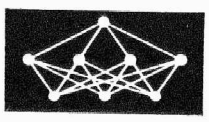
As for the case of a single neuron with gradual plasticity maximal information, which can be extracted from the network about a single template using any routine of simple decoding are given by $J = H(L) - H(L-1)$ where $H(L)$ is given by (23) and $H(L-1)$ is the entrophy of matrix M formed by storing of L-1 templates. Differentiating equation (25) with respect to L one can get

$$J = (1/2)(\mathrm{n} + (\mathrm{L} - n) \ln(1 - n/L)) \lg e.$$

Then for simple decoding

$$E' = LJ/n^2 = \frac{L \lg e}{2n} (1 + (\mathrm{L}/n - 1) \ln(1 - n/L)).$$

Dependence of E' on (L/n) is shown in *Fig. 2*. For increasing of L coefficient E' tends to 0. 36 what is twofold less then its limit value given by formula (14). The



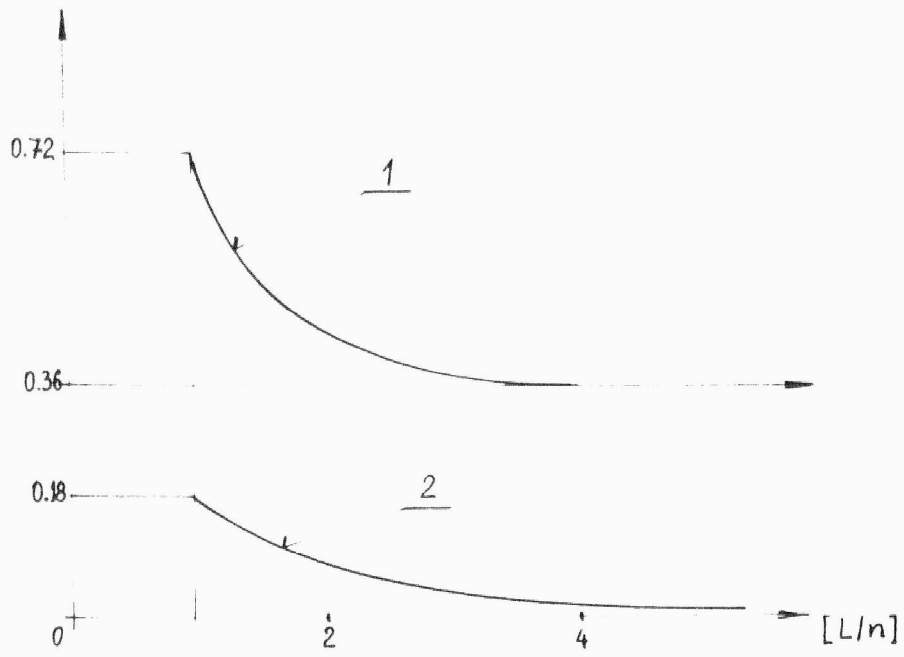


Fig. 2. Efficience coefficient E' for two-layered network performing functions of autoassociative memory 1 — E' for simple decoding; 2 — ΔE' for complex decoding.

reason of this difference is evidently in the symmetricity of matrix *M*. But what is realy interesting is the excess of the efficience coefficient of a half of matrix *M* under the efficience coefficient of a single neuron. Terefore, unlike complex decoding, for simple decoding the efficience coefficient of the whole network may exceed this coefficient for a single neuron. Thus information losses caused by two different reasons (passing from complex to simple decoding and statistical dependence of modification states of different synapses) may be partially compensated.

References

[1] Albus J. S. :A theory of cerebellar function Math. Biosci. 1971. Vol. 10, N 1. P. 25-61.

[2] Amit D. J. , Gutfreund H. , Sompolinsky H. : Spin-glass models of neural networks
Phys. Rev. 1985. Vol. A32. P. 1007.

[3] Amit D. J. , Gutfreund H. , Sompolinsky H. :Annals of Physics. 1986. Vol. 173. P. 30.

[4] Amit D. J. , Gutfreund H. , Sompolinsky H. :Phys. Rev. 1987. Vol. A33. P. 1078.

[5] Braitenberg V. : Cell assemblies in the cerebral cortex: Theoretical approaches to complex systems/Ed. R. Heim, G. Palm/, Berlin , Springer, 1978, p. 171.

[6] Briendley G. S. :Nerve net models of plausible size, that perform many simple learning tasks Proc. Roy. Soc. London , 1969, Vol. 174, p. 193-227.

[7] Dunin-Barkowski V. L. : Information processes in neural structures Moskva, Nauka, 1978. 166 p. (in Russian).

[8] Frolov A. A. , Murav'ev I. P. : Neural models of associative memory Moskva, Nauka, 1987, 160 p. (in Russian).

[9] Frolov A. A. , Murav'ev I. P. :Informational characteristics of neural networks Moskva, Nauka, 1988. 159 p. (in Russian).

[10] Frolov A. A. , Murav'ev I. P. :Informational characteristics of neural and synaptic plastisity Biophysics, 1988, Vol. 33, No. 4, p. 708-716.

[11] Gantmakher F. R. : The theory of matrices. Moskva, Nauka, 1966, 576 p. (in Russian).

[12] Girko V. L. : Thery of random determinants Kiev, Vys. sk. , 1980, 368 p. (in Russian).

[13] Gradshtein I. S. , Ryzhik I. S. : Tables of integrals, sums, series and productions Moskva, Fizmatgiz, 1963. 1110 p. (in Russian).

[14] Hebb D. O. The organisation of behavior J. Wiley, New York, 1949, 335 p.

[15] Hopfield J. J. : Neural network and physical systems with emergent collective computational abilities Proc. Nat. Acad. Sci. USA, 1982, Vol. 79, p. 2554-2558.

[16] Hopfield J. J. : Proc. Nat. Acad. Sci. USA, 1984, Vol. 81, p. 3088-3092.

[17] Kohonen T. : Associative memory Moskva, Mir, 1980, 230 p. (in Russian).

[18] Kohonen T. :Self-Organisation and Associative Memory Springer, Berlin, 1984, 225 p.

[19] Kolesnik V. D. , Poltyrev G. Sh. :Handbook of the theory of information Moskva, Nauka, 1982, 416 p. (in Russian).

[20] Korn G. A. , Korn T. M. :Mathematical handbook (for scientists and engineers) McGraw-Hill, New York, 1968, 830 p.

[21] Konorski J. : Integrative activity of the brain Moskva, Mir, 1970, 412 p. (in Russian).

[22] Little W. A. : The existence of persistent states in the brain Math. Biosci, 1974, Vol. 19, p. 101-120.

[23] Marr D. :A theory of cerebellar cortex J. Physiol. , 1969, Vol. 202, p. 437-470.

[24] Marr D. :A theory of cerebral neocortex Proc. Roy. Soc. , London , 1970, Vol. 176, . p. 161-234.

[25] Marr D. : Simple memory: A theory for archicortex Philos. Trans. Roy. Soc. London, 1971, Vol. 262, p. 23-81.

[26] Palm G. :On associative memory Biol. Cybern. , 1981, Vol. 36, No. 1, p. 19-31.

[27] Rosenblatt F. :Principles of neurodynamics Moskva, Mir. 1965. 400 p. (in Russian).

[28] Steinbuch K. : Die Lern-matrix Kybernetik, 1961, Bd. 1. , p. 36.

[29] Vinogradova O. S. : Hippocamp and memory Moskva, Nauka, 1978, 166 p. (in Russian).

[30] Willshaw D. J. , Buneman O. P. , Longuet-Higgins H. C. : Non-holographic associative memory Nature, 1969, Vol. 222, p. 960-962.

Literature Survey

Greiner B.: Neural Networks: for the Thinking Person
Computing-Canada, Vol. 16, 1990, No. 1, pp. 37—38

Key words: neural networks; research and development; applications; expert systems; parallel processing; artificial intelligence.

Abstract: Neural network technology, with its potential for creating a computer capable of simulating the learning process, holds great potential for research and development over the next decade. Unlike expert systems technology, which requires a complete set of facts and rules in order to function, neural networks can incorporate fuzzy logic and deal with ambiguous real world situations. As so-called 'non-algorithmic function learners', neural networks require no mathematical models. They have such properties as differentiation capability, generalization, fault tolerance and optimization. Essentially a collection of parallel processor with each connection having a weight factor, the neural network can process raw information and responses, generating the 'meanings' associated with learning. They do not replace traditional algorithm-based computing, but apply to different types of problems.

Grossberg S.: The Second Anniversary of Neural Networks (Editorial)
Neural Networks, Vol. 3, 1990, No. 1, pp. 1

A MODEL OF A NEURAL NETWORK WITH SELECTIVE MEMORIZATION AND CHAOTIC BEHAVIOR

Yu. M. Sandler, V. F. Artyushkin)*

Abstract:

In the present paper a generalization of Hopfield model is shown, associated with a break of the specific invariance of the equations of motion (2). Unlike the Hopfield model, the present model can exhibit selectivity in the process of learning (that is, "memorizing" only the patterns of certain kind) and has quasi-stochastic attractors.

1. Introduction

Hopfield has demonstrated [1] that a completely interconnected network of N Mc-Culloch-Pitts neurons, in which each neuron has two states ('on' and 'off'): $\varphi_i(t) = \pm 1$, (where i is the number of neuron in the network, and t is the time), can be described as an Ising spin glass with the Hamiltonian

$$H = -\frac{N}{2} \sum_{i,j} J_{ij}\, \varphi_i\, \varphi_j = -\frac{N}{2}\, \varepsilon(\vec{\varphi}) \tag{1}$$

$$J_{ij} = \frac{1}{N^2} \sum_{s=1}^{P} \mu_s\, \zeta_i^s\, \zeta_j^s$$

where ζ_i^s are the 'frozen' variables which assume the values of (± 1). If the equations of motion for $\varphi_i(t)$ are chosen in the form

$$\varphi_i(t+1) = \operatorname{sign}\left[-\frac{\delta_H}{\delta \varphi_i(t)} \right] = \operatorname{sign}\left[\sum_j J_{ij}\, \varphi_i(t) \right] \tag{2}$$

the vectors $\vec{\zeta}^s = \left(\zeta_1^s, \ldots \zeta_N^s\right)$ turn out to be the stationary states of the neural network.

By now the behavior of the Hopfield model of neural network has been studied sufficiently well. The networks with nonlinear dependence of J_{ij} on $\zeta_i^s\, \zeta_j^s$ are considered in [2]; the neural nets with a more general Hamiltonian are studied in [3], etc.

In the present paper we are going to show that there exists an interesting generalization [4] of the Hopfield model, associated with a break of the specific invariance of the equations of motion (2).

*) Dr. Yu. M. Sandler
Dr. V. F. Artyushkin
Institute for USA and Canada,
Academy of Sciences of the USSR
Moscow 121814 USSR
Khlebny per. 2/3

Unlike the Hopfield model, these networks can exhibit selectivity in the process of learning (that is, 'memorizing' only the certain kinds of patterns) and have quasi-stochastic attractors (which means that for a certain region of the initial states the asymptotic behavior of the network is quasi-chaotic).

2. Description of the Model

The equations (2) are invariant with respect to the scale transformation: $J_{ij} \Rightarrow \lambda J_{ij};\ \lambda > 0$.

This transformation, however, can lead to nontrivial implications, if we treat λ as a function of ε. Then, the equations of motion for $\vec{\varphi}(t)$ take the form

$$\varphi_i(t+1) = \operatorname{sign}\left[\lambda(\varepsilon) \sum_j J_{ij}\, \varphi_j(t) \right]. \tag{3}$$

Such systems can be described by the Hamiltonian

$$H = -\frac{N}{2} F(\varepsilon);$$

$$\varepsilon = \sum_{i,j} J_{ij}\, \varphi_i\, \varphi_j; \tag{4}$$

which leads to the equations of motion similar to (3). On the other hand, the learning rules for this model will be different from well known **Hebbian** learning rules. If the system learns the pattern $\vec{\zeta}^*$ over the time Δt, $(1 \ll \Delta t \ll 2\gamma N^2/F')$, the matrix of connections J_{ij} would change by the value

$$\Delta J_{ij} \cong \frac{\Delta t}{2\gamma} F' \left(\sum_{i \neq j} J_{ij}\, \zeta_i^*\, \zeta_j^* \right) \frac{1}{N^2}\, \zeta_i^*\, \zeta_j^*;\ (\mathrm{i} \neq \mathrm{j}) \tag{5}$$

where — γ is the 'kinetic coefficient' in the equation of motion for the 'slow' (in comparison with $\vec{\varphi}(t)$) variables — J_{ij} and $F' = \partial F/\partial \varepsilon$. Although the form of the matrix J_{ij}, corresponding to the new learning rule (5), is rather similar to (1):

$$J_{ij} \rightarrow J_{ij}^* + \sum_s \mu_s\, \zeta_i^s\, \zeta_j^s \tag{6}$$

there is a fundamental difference, since in (6) the quantities μ_s depend on the entire set of patterns $\{\vec{\zeta}^s\}$, and on the sequence in which they are presented to the system (we shall discuss this circumstance in more detail in Sect. 3).

In this paper we also assume that $N \gg 1$, and $P/N \ll 1$ $(P/N \rightarrow 0$ when $N \rightarrow \infty)$.

3. Learning with Selective Memorization

Let, in the process of learning, the system be presented with a sequence of patterns $\vec{\zeta}^{(1)}, \vec{\zeta}^{(2)}, \ldots \vec{\zeta}^{(n)}$ (it is not required that all patterns in the sequence be different). Unlike [5] we assume that in the process of learning the system is 'frozen' in the state $\vec{\zeta}^{(k)}$ during the time $\Delta t_k = \Delta t$ and its matrix of connections J_{ij} is changed in accordance with (5). Then we have

$$J_{ij}(n) = J_{ij}(0) + \sum_{k=1}^{n} \upsilon(k) \frac{\zeta_i^k \zeta_j^k}{N^2};$$

$$\upsilon(n) = \varkappa f\left[\sum_{i \neq j} \zeta_i^n J_{ij}(0) \zeta_j^n + \sum_{k=1}^{n-1} \upsilon(k) \frac{\left(\vec{\zeta}^k \cdot \vec{\zeta}^n\right)^2}{N^2}\right] \tag{7}$$

In the standard Hopfield model $f \equiv 1$, if any pattern in the sequence is repeated often enough, its basin of attraction will eventually take over almost the entire space of states, displacing all the other images from memory [3].

As seen from (7), in our model this effect can easily be avoided, if we choose $f(\varepsilon)$ rapidly tending to zero with the increasing ε (actually, if $f(\varepsilon) \cong 0$ for $\varepsilon \geq \varepsilon_*$, then $\mu_{\max} \cong \varepsilon_*$). However, a much more intriguing situation is when in the process of learning the neural network will memorize only those images which are close to the present patterns, the reference patterns being not recognized but rather remaining in the 'subconscious'.

Let us illustrate this with a simple example. Assume that

$$J_{ij}(0) = -\eta \frac{\sigma_i \sigma_j}{N^2}$$

$$f(\varepsilon) = \varepsilon^2 \tag{8}$$

Then in the process of relaxation the neural network will be unable to recognize $\vec{\sigma}$ (because the weight of $\vec{\sigma}$ is negative and in the process of relaxation the system goes to some state, which is orthogonal to $\vec{\sigma}$). Now, let us teach the neural network by presenting it alternately with two patterns $\vec{\zeta}^{(1)}$ and $\vec{\zeta}^{(2)}$, whereas

$$\frac{1}{N}(\vec{\sigma} \cdot \vec{\zeta}^{(1)}) \approx 1; \frac{1}{N}(\vec{\sigma} \cdot \vec{\zeta}^{(2)}) \leq \frac{1}{\sqrt{N}} \ll 1$$

Then after $2n$ presentations we get

$$J_{ij}(2n) \cong \frac{\chi(n)}{\varkappa N^2}\left(1 + \frac{\varkappa \eta}{\chi(n)}\right)\left(\zeta_i^{(1)} \zeta_j^{(1)} + \beta^2 \zeta_i^{(2)} \zeta_j^{(2)}\right) - \eta \frac{\sigma_i \sigma_j}{N^2};$$

$$\chi(n) = \chi(n-1) + \chi^2(n-1); \ \chi(o) \approx -\varkappa \eta \ll 1 \tag{9}$$

where $\beta = \frac{1}{N}(\vec{\zeta}^{(1)} \cdot \vec{\zeta}^{(2)}) \approx \frac{1}{N}(\sigma \cdot \vec{\zeta}^{(2)}) \approx \frac{1}{\sqrt{N}}$.

After learning, the neural network will practically be able to recognize only the image $\vec{\zeta}^{(1)}$, since the basin of attractor $\vec{\zeta}^{(2)}$ is exponentially small (because of the smallness of β^2) [3]. The generalization of this example to the more general case is sufficiently obvious and does not lead to new results.

4. Pattern Recognition and Quasi-Stochastic Attractors

Let us consider the model (4) without learning, and with matrix of connections (6), where we have set $J^*_{ij} = 0$. From (2), and (4) we get for $\varphi_i(t)$ (accurate up to o (P/N))

$$\varphi_i(t+1) = \text{sign}\left[f\left(\sum_{S=1}^{P} \mu_s m_s^2\right) \sum_{s=1}^{P} \mu_s m_s \zeta_i^s\right];$$

$$m_s(t) = \frac{1}{N}\sum_{i}^{N} \zeta_i^s \varphi_i(t) = \frac{1}{N}(\vec{\zeta}^s \cdot \vec{\varphi}); \tag{10}$$

$$f(\varepsilon) = F'(\varepsilon);$$

Note that, if not specified otherwise, in the future we assume that we are dealing with the regular version of the asynchronous dynamic behavior [1].

If $F(\varepsilon)$ is a monotonicly increasing function, then sign $[f(\varepsilon)] \equiv 1$. Obviously, in this case (10) is equivalent to (2), and the process of pattern recognition (relaxation of $\vec{\varphi}(t)$ for the model (4) is completely equivalent to the appropriate process in the standard Hopfield model [1]. The qualitative difference in the behavior of these models arises in the case of nonmonotonic $F(\varepsilon)$. Of special interest is the situation when $F(\varepsilon)$ has a maximum at a certain $\varepsilon = \varepsilon_m$.

Depending on the magnitude of μ_q for the given pattern $\vec{\zeta}^n$ we shall have either $\varepsilon(\vec{\varphi} = \vec{\zeta}^q) \leq \varepsilon_m$, or $\varepsilon(\vec{\varphi} = \vec{\zeta}^q) > \varepsilon_m$. If the initial state of the system is in the basin of attraction of $\vec{\zeta}^q$, then in the former case the neural network will go to the stationary state $m_{s=q} = 1$; $m_{s \neq q} = 0$; ($\vec{\varphi}(\infty) = \vec{\zeta}^q$ — that is, the system will recognize the pattern $\vec{\zeta}^q$ *(fig. 1(I))*. In the opposite case the behavior of the neural network will depend on the type of the dynamics — whether it is synchronous or asynchronous. (In the former case the behavior of networks with different dynamics is practically similar).

In the case of synchronous dynamics at $\varepsilon(\vec{\zeta}^q) > \varepsilon_m$ the stationary state $m_{s=q} = 1$, $m_{s \neq q} = 0$ goes over to the cycle with period 2: $m_q(t+2) = -m_q(t=1) = m_q(t)$; $m_{s \neq q} = 0$ — that is, the system performs jumps between the image $\vec{\varphi}(t) = \vec{\zeta}^q$ and its 'negative' $\vec{\varphi}(t+1) = -\vec{\zeta}^q$. Both the form of oscillations and the period do not depend on the magnitude of the difference $[\varepsilon(\vec{\zeta}^q) - \varepsilon_m]$ and the distance between $\vec{\varphi}(0)$ and $\vec{\zeta}^{(q)}$ (of course, provided that $\vec{\zeta}(0)$ is in the basin of attraction of $\vec{\zeta}^q$).

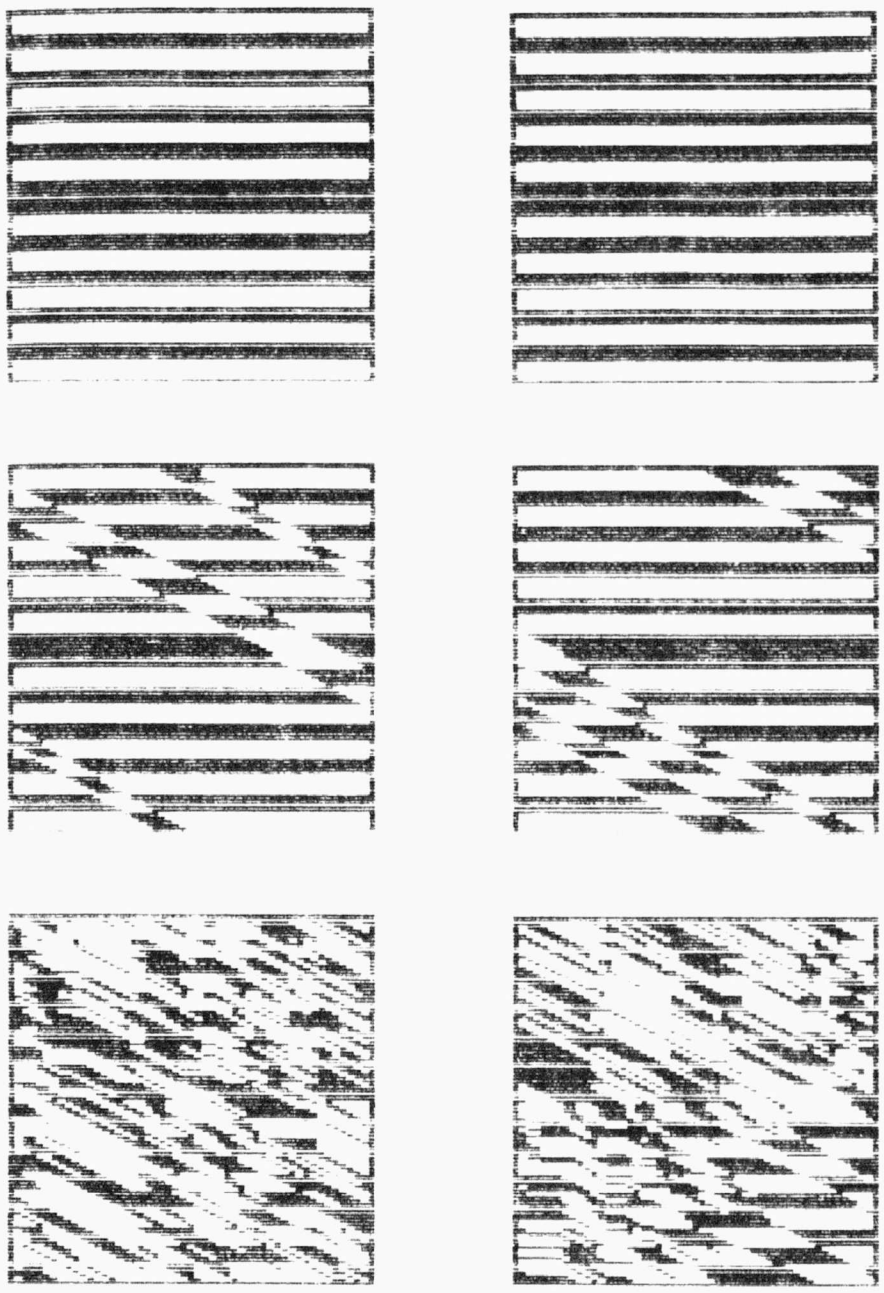


Fig. 1. Dynamics of the neural networks for $N = 100$; $\varepsilon_m = 1$; $0 \le t \le N^2$ (in these drawings every step on the time axis is equal to N steps in the computing simulation), and for different d_0 and $\varepsilon(\vec{\zeta}^q)$.
a) — $d_0/2N = 0{,}06$; b) — $d_0/2N = 0{,}13$
I — $\varepsilon(\vec{\zeta}^q) = 0{,}5$; II — $\varepsilon(\vec{\zeta}^q) = 2$;
III — $\varepsilon(\vec{\zeta}^q) = 20$; IV — $\varepsilon(\vec{\zeta}^q) = 200$.
The black points designate $\varphi_i = 1$ and space the opposite one.

A much more complicated behavior is exhibited by the network with asynchronous dynamics. *Fig. 1—2 (II—IV)* shows the behavior of the neural network and its autocorrelation function

$$C_s(\tau) = \langle m_s(\tau)\, m_s(0)\rangle = \lim_{\kappa\to\infty} \frac{1}{K} \sum_{k=0}^{\kappa-1} \hat{\chi}^s_{k+\tau} \hat{\chi}^s_k; \quad (11)$$

$$\hat{\chi}^s_k = m_s(k) - \lim_{L\to\infty} \frac{1}{L} \sum_{l=0}^{L-1} m_s(1)$$

for different indices s and $\delta_q = \dfrac{\varepsilon\left(\vec{\zeta}^q\right) - \varepsilon_m}{\varepsilon_m}$.

Clearly, the behavior of the neural network depends strongly on the quantity δ_q. For small $\delta_q \ge 0$ the system goes to the state with $m_q = 1 - \Delta_q(t)$; $m_{s\neq q} = \Delta_s(t)$; $(\Delta(t) \ll 1)$. In practice, the neural network slightly oscillates near $\vec{\zeta}^q$, the period of oscillations being much larger than N. The magnitude of Δ_q increases with increasing δ_q (whereas $\Delta_{s\neq q}(t)$ remains small), and the behavior of the neural network becomes more and more chaotic. This may be seen both directly and by watching the correlation function $c_s(\tau)$, which quickly decreases with the increase in τ *(Fig. 2 (III, IV))*.

Another peculiarity of the asynchronous dynamics is associated with the fact that for small δ_q the deviation of $\vec{\varphi}(t)$ from $\vec{\zeta}^q$ depends on the distance between $\vec{\varphi}(0)$ and $\vec{\zeta}^q$: $d_0 = \sum_i |\varphi_i(0) - \zeta^q_i|$ (in the Hamming metric [4]). The larger d_0, the larger $|\Delta_q|$. For large $\delta_q \gg 1$ the behavior of the network does not depend on d_0. A similar situation is observed also with $c_q(\tau)$ and $c_{s\neq q}(\tau)$, which are substantially different for $\delta_q \ll 1$ and are practically the same for $\delta_q \gg 1$, which is in accordance with the fact that for large δ_q the behavior of the system is close to chaotic.

The illustrations of a transition to chaotic behavior are shown on a graphic of the power spectra of the overlaps $m_s(t)$:

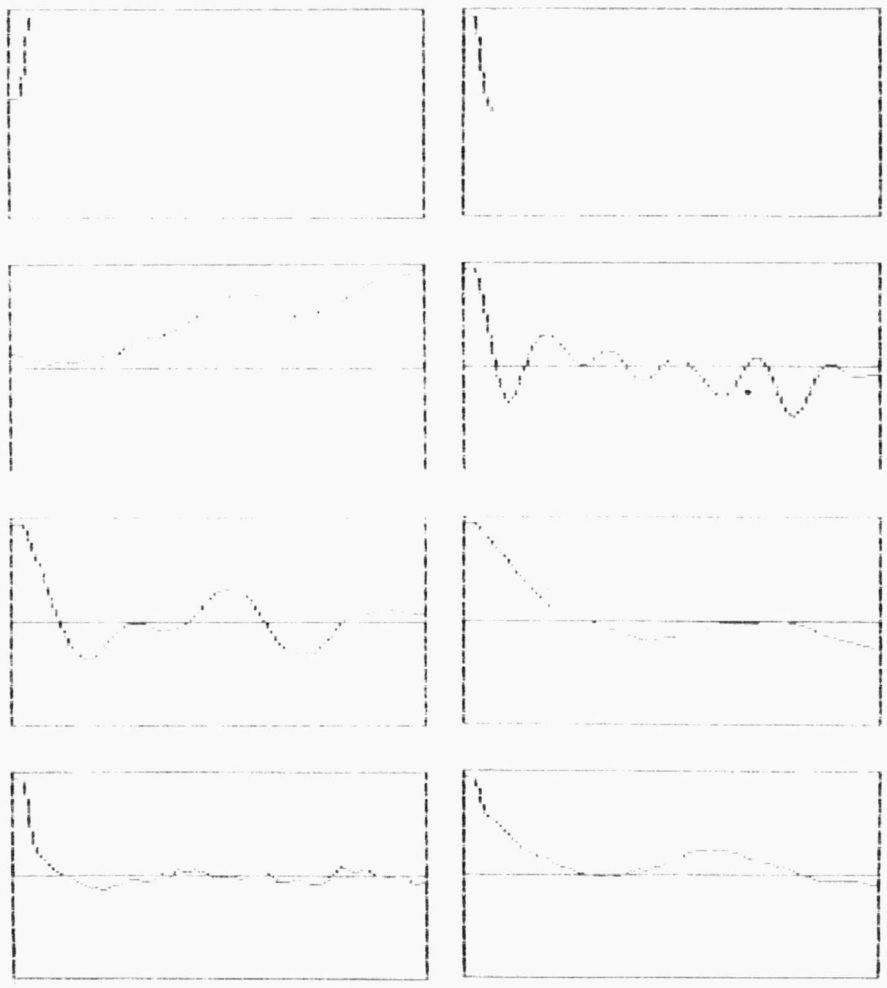


Fig. 2. The correlation funcions of the overlap: $c_s(\tau) = \langle m_s(\tau)\, m_s(0)\rangle$
a) for $s = q$; b) for $s \neq q$
and $0 \le T \le N^2/10$; $N = 100$; $\varepsilon_m = 1$.

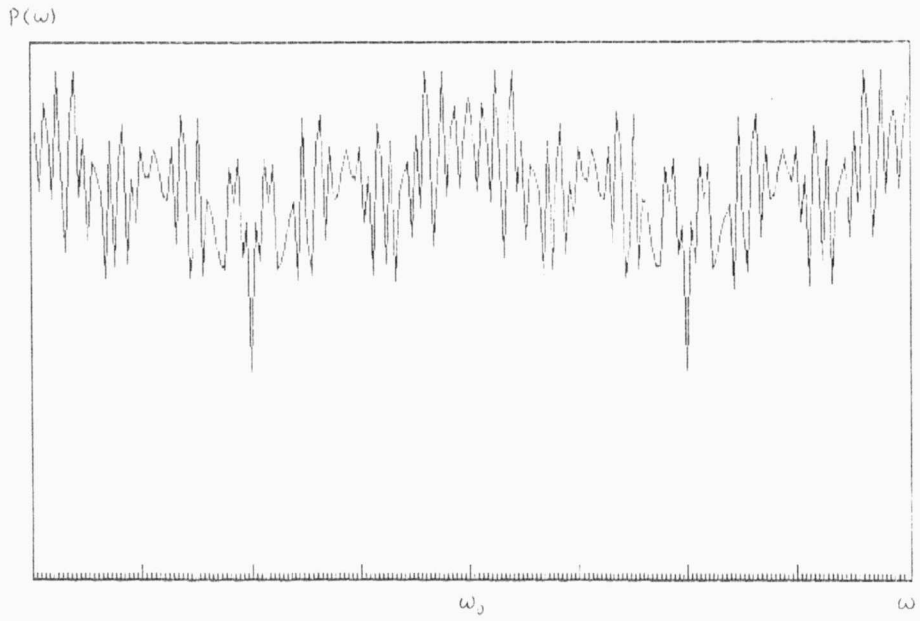


Fig. 3 Power spectrum of the neural net in a oscilation state.

$$P(\omega) = \ln\left[\left|\int m_s(t) \sin \omega t \, dt\right|^2 + \left|\int m_s(t) \cos \omega t \, dt\right|^2\right]$$

In *Fig. 3* we see a typical spectrum of the oscillation behavior with frequencies $\omega_\eta = n\omega_o$, and in *Fig. 4* is seen that for large $\varepsilon(\vec{\zeta}^q) \gg \varepsilon_m$ there is a chaotic spectrum.

It is important that if the neural network has patterns $\{\vec{\zeta}^s\}$ with different signs of δ_s (for instance, $\delta_{s1} > 0$, $\delta_{s2} < 0$), then the attractor $\vec{\zeta}^{s1}$ will be quasi-chaotic, whereas the attractor $\vec{\zeta}^{s2}$ will be a common stationary point.

As a matter of fact, for discrete $\varphi_i(t)$ and finite N the neural network cannot display genuine chaos. This derives from the circumstance that in a system with a finite number of discrete states such k and t will always be found that $\vec{\varphi}(t) \equiv \vec{\varphi}(t + kN)$. Since the equation of motion for $\vec{\varphi}$ are dynamic, it would follow that it is only the cycles that can exist in the system. However, with large N and δ_s the values of $k = k(\delta_s, N)$ become very large, and the behavior of the system for the times $\Delta_t \ll kN$ becomes very close to chaotic. (In systems with continuous variables in the chaotic region the exact equality $\chi = \chi(t + \tau)$ is possible only on the set of measure zero, and therefore they can display genuine chaos.)

5. Neural Network at Finite Temperatures

As indicated in the Introduction, the effects of noise upon the neurons as the random external input can be analyzed in terms of statistical mechanical treatment of the system with the Hamiltonian (8). The probability of the flip $\varphi_i \rightarrow -\varphi_i$ is chosen in the form:

$$P(\varphi_i \rightarrow -\varphi_i) = \left(1 + \exp \frac{\Delta_i H}{T}\right)^{-1}$$

and an overlap m_s in the equlibrium state is

$$\overline{m}_s = \mathrm{sp}\,\{(\vec{\zeta}^s \vec{\varphi})\, P(\vec{\varphi})\}$$

In the limit $N \rightarrow \infty$ the $\overline{m}_s$ can be calculated exactly. This is done most simply by using the method of molecular field, which for the systems with the long-range interaction yields an exact solution (this can be proved by calculating $\overline{m}_s$ by the method of the steepest descent).

Then, for $\overline{m}_s$ we gain:

$$\overline{m}_s = \frac{1}{N} \sum_j^N \zeta_j^s \,\mathrm{th}\left(\frac{f(\bar{\varepsilon})}{T} \sum_{s=1}^p \mu_s \overline{m}_s \zeta_j^s\right); \qquad (14)$$

In order to illustrate the difference between our present model and the Hopfield model, let us consider a simple case with $p = 1$ and $f(\varepsilon) = \varepsilon$. Then the condition of correspondence (14) will take the form

$$\overline{m} = \mathrm{th}\left(\frac{\mu}{T} \overline{m}^3\right) \qquad (15)$$

Hence it follows that there always is a solution $m = 0$. For $T \leq T_* = (3/4)^4 \mu \cong 0{,}315\mu$ this solution becomes metastable, and another solution appears, $m = m(T) \neq 0$, whereas $m(T_*) = (3/4)^{3/2} \cong 0{,}66$; and $m(T_* - \delta_T) \cong m(T_*) + o(\delta_T)$. This means that the transition to the ordered phase in our model is the phase transition of the first order. Note that the critical noise intensity δ_*, which prevents recognition, is much lower than in the Hopfield model:

$$\delta_* / \delta_c^{Hopf} \cong 0{,}56$$

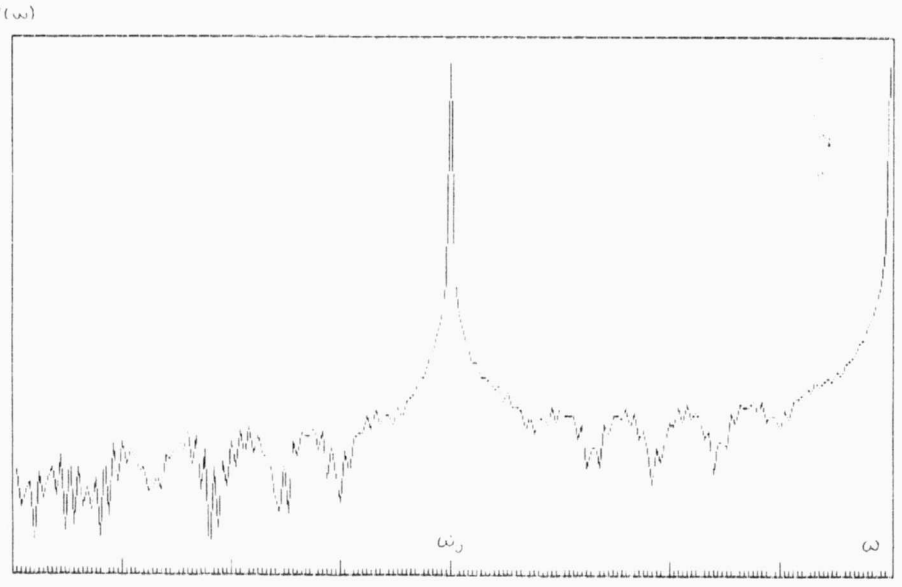


Fig. 4 Power spectrum of the neural net ina quasi-chaotic state.

6. Conclusion

Selective learning and the creation of quasi-chaotic attractors exhibits a clear analogy with the properties of real nervous systems. In particular, the process of selective learning can be interpreted as extracting 'knowledge' from the subconsciousness (compare with 'process of unlearning' in [6]). On the other hand, the inability of a neural network to recognize images for which $\varepsilon(\vec{\zeta}^q) \gg \varepsilon_m$ can be interpreted as temporary amnesia, since by changing ε_m these images can be made comprehensible again.

References

[1] J. J. Hopfield, Proc. Nat. Acad. Sci. USA, v. 79 (1982) 2554; v. 81 (1984) 3088.
[2] H. Sompolínsky, Phys. Rev. A, v. 34 (1986) 2571.
[3] D. Horn, M. Usher, J. Phys. France, v. 49 (1988) 389.
[4] A. A. Vedenov: Modelirovanie elementov Myshleniya. (Moscow, Nauka, 1988) (in Russian); A. A. Vedenov, E. B. Levchenko, JETP Lett. v. 41 (1985) 402.
[5] Y. M. Sandler, V. F. Artyushkin, J. of Nonl. Biol., v. 1 (1990) (in press).
[6] J. J. Hopfield, D. T. Feinstein, R. G. Palmer, Nature, v. 304 (1983) 402.

Book review

Biological Complexity and Information

Proceedings of a Conference on the Amalgation of the Eastern and Western Ways of Thinking

Fuji-Susono, Japan, April 21-24, 1989

Edited by Hiroshi Shimizu
World Scientific, 1990

The book contains papers presented at a small conference attempting an amalgamation of Eastern with Western thinking for a better understanding of information processing. To enable an interaction of an Eastern way of thinking based more on a parallel and global approach and a Western one based more on a serial and analytical approach, scientists as well as philosophers from Japan and Europe and America were invited to this meeting. The papers are divided into the following sections:

Information Dynamics in Biological Systems,
Brain as a Complex System,
Complexity in Information Dynamics,
Toward the Science of Semantic Relations,
Mathematical Expressions of Relations,
Consciousness and Reality.

To illustrate the diversity of topics discussed in the Proceedings, we shall briefly mention some of the presented papers.

J. S. Wicken in „Can Information Be Quantified by Shannon Formalism?" argues against genetic reductionism by showing that the information content of an organism cannot be quantified by complexity analysis of DNA sequences. He stresses that without the special environmental factors that have evolved with living systems, those sequences would not even exist. DNA codes only for the primary structures of proteins, but particularities of coilings, foldings and higher level interactions are left to the physico-chemical milieu, which plays a role of a reader of this genetic information. To quantify the information content of this milieu would require nothing less than the quantification of the ecological-historical context. Organisms'"hardware" and „software" are indissociably integrated. Information content belongs irreducibly to whole systems, not only to their parts. Finally, he warns that science, as a myth-maker of our age, must carefully examine the context in which selfishly directed survival and reproduction get their validation.

R. Suzuki, M. Kawato and Y. Uno, in „A Neural Network Model of Human Motor Skill Development", propose a neural network model of motor skill learning, by which they attempt an explanation of the intuitive or unconscious decision making process. In their model of control of hand movement, a control mechanism added to a feedback control system is implemented as a three-layered neural network. The authors compare computation of a correlation between a desired trajectory and a memorized experience to an intuition.

E. Koerner, H. M. Gross and I. Tsuda, in „Holonic Processing in a Model System of Cortical Processors", criticize homogeneous neural networks models with simple unstructured nodes. They propose considering as nodes instead of single neurons whole minicolumns. They describe a model of columnar dynamics with nodes possessing a complex structure made of heterogeneous elements with three different homogeneous network systems (one of them with locally restricted communications, the other two with global connectivity) communicating vertically in each node. K. Kubota in „Roles of the Prefrontal Cortex on Behaviors, Simple as well as Complex" presents an experimentally inspired hypothesis that small groups of columns in the prefrontal cortex form functional units responsible for different behaviors (like delayed response or delayed alternation).

Y. Tanaka in „The Concept of Reality in Quantum Physics" proposes a new interpretation of quantum reality as a self-projecting organism suggesting the possibility of a synthesis of relativity and complementarity.

R. Rosen in „'Hard' Science and 'Soft' Science" argues that the duality between „hard" or quantitative science and "soft" or qualitative science rests on an entirely false presumption. It is rather a relative question of simplicity versus complexity.

Y. Kajikawa in „Folding the Polyhedra" presents a kind of periodic table describing relationships of all the five Platonic polyhedra and thirteen Archimedean semiregular polyhedra. He proves the existence of some basic states that would have pleased Plato.

Unfortunately, there are a lot of misprints throughout the whole book and also the graphical organization is poorly done (e.g. titles of paragraphs on the last lines of pages, etc.).

Věra Kůrková
Institute of Computer and Information Science,
Czechoslovak Academy of Sciences, Prague

RECORD OF THE PANEL DISCUSSION ON SYMPOSIUM NEURONET '90 HELD IN PRAGUE, CZECHOSLOVAKIA IN SEPTEMBER 1990

W. Eldridge)*

The following discussion is a record of the panel discussion that took place during the International Symposium on Neural Networks and Neural Computing (Neuronet '90) that was held in Prague, Czechoslovakia in September of 1990. The discussion was chaired by Dr. Robert Hecht-Nielsen leading the panel consisting of Dr. Joel Davis from the Office of Naval Research in Virginia, Dr. Lee Giles of NEC computers and the University of Maryland, Dr. Jiří Hořejš of Charles University (Prague), Dr. Vitalij I. Kryukov of the Soviet Academy of Science in Pushchino (Moscow Region), Dr. Hiroyuki Mori of Meiji University (Kawasaki) and Dr. John G. Taylor of King's College (London).

Chair: The goal of our roundtable discussion — we are not going to make formal presentations, except if someone cares to in answer to a point or to make a point, but there is no plan for that. Our goal is to discuss the future of this ensemble of activities that people call Neural Networks, neural computing, neural science, and more importantly, I would like to keep the discussion focused if I could not on the far future, but on the near future. We know there are trends currently in this field. We have heard papers relating to many of these trends during the sessions of this conference. Let me give you some examples. We have for example a great deal of work going on around the world in the area of reinforcement learning — learning where we do not have a specific knowledge of what each processing element should be doing at each time step but where we have some sort of general reinforcement knowledge about the performance of the system, and we have heard very recent interest and interesting research in that area. We have also heard a great deal about oscillator networks, and this is a subject that is attracting a significant amount of attention around the world, both from people studying neurobiology where oscillations have been known for years but where phase-locking has only recently become well-established, and also from the practical end of the subject, where people are looking for ways of building systems that can bind together, at least temporarily, the features of a single object. So this is another trend.

*) W. Eldridge, Institute of Computer and Information Science, Czechoslovak Academy of Sciences, Prague

In the discussions during this week, we have also heard a lot about image processing and image analysis. We heard for example talks on multi-resolution methods where you have say a phobial type of process with a high resolution in the center going to a very low resolution at the periphery, and dynamic gaze shifting models, and other methods of image analysis, such as these R-wave processors that take the object and then allow it to go through a temporal transformation until it reaches the periphery of image the area where a signature is formed — very exciting work, we have heard three or four papers on that — so there are these near-term trends and I would d like to focus the discussion on these trends which we know about and have identified, but where we do not know exactly where they are leading. Other trends include some very exciting new work that you've heard about regarding the use of automata put like a push-down stack in conjunction with a neural network that actually runs the stack and uses it as a resource; this is a concept that appears to have enormous promise. So we have had papers on these trends and so what I would like to do is begin by asking a question and maybe each of you people here can make comments about this, and that is: „What in your area of work, or in your area of interest shall we say, what are the areas that you feel are going to make the most progress in the next two or three years, which of the areas that are attracting the most attention. So in other words, I would like to start by establishing a sort of panorama of areas that are considered exciting, interesting and highly active. So if we can start, we have heard discussions on learning theory and back-propagation theory and so forth, and so perhaps you have a comment —

Hořejš: Well I feell that you seem to be at home among us. It is rather difficult for me to start. You know in Czechoslovakia the position is somehow different than in the West, so we are now keeping track of you, how you are progressing and trying to follow you, I do not mean you personally, but the whole thing. The theory perhaps will be the first such topic in which we might hope that we will be able to contribute something more important than just to see how the train or the world is just leaving us far behind. Maybe then image processing which we have some experience in EEG analysis and EKG analysis and so on, and then perhaps again other image processing to some extent. That is the only thing which I know to say on the behalf of the group which you know already.

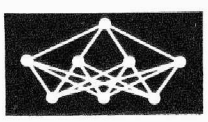
Chair: Thank you very much, and perhaps Professor Taylor would make some comments. I would like to, if possible, focus this more on a world scale in terms of trends. What are the areas that are attracting the most attention and the most intellectual effort, and which of these do you think show the most progress?

Taylor: Of coure this is a very difficult problem because I can look at it from an industrial side, from what will go into new industrial developments, or from the other extreme, what will be important on the theory side. I want to just tell you what experience I had about two weeks ago. The British Association for the Advancement of Science held a meeting in Swansea and we ran a Neural Nets Day there. That day involved an introduction to neural networks, very good lectures from various people, then a description of the industrial developments that are occurring, in particular through the ANNIE project, that's Artificial Neural Nets for the Industry in Europe, a very big project, an ESPRIT project, and the Pygmalion project, which is the development of a software environment for running all the main neural net algorithms we have in a user friendly manner. And these projects, I think, are very important in what they are aiming and what they are achieving to do in industrial penetration. Though the presentation is also from British Telecom, which is funding half the neural net research in the UK, and then there was a presentation from Roger Penrose who's written a book called The Emperor's New Mind. Some of you may have heard of that. Now that book presents the claim that neural nets can never think, can never be made to think, and indeed it presents a much more general claim that anything that is constructed in an algorithmic form will never be made to think and there's some need to have quantum mechanics, even quantum gravity, although that is far beyond the reach of any machines we can make to control it, so we've got really no problems for a long time. Now, I had to follow that, and it seemed to me that that was a challenge that I do see beginning to come into neural networks which I think is very important in a way of strengthening the base of neural networks. We have the fear that what goes up may well come down and it did do that in the early revolution of neural networks that was started by Rosenblatt, by Wilson, by McCullough and Pitts, and it went down when Minsky and Papert had opened up their big gun; the set crumbled, they'd destroyed it. But the situation now I think is becoming different in the sense that I see — to answer the question — that there are the contexts through into neuro-cognition, neuro-philosophy, where it is important to begin to understand in what way one could say model the mentation processes which occur in the developing infant. Now I was shocked to find out, not shocked in a way, not badly, but surprisingly, that infants at the age of one month have a concept already of an object that persists and if they are shown an object that moves across the screen and suddenly disappears, that is surprise. What neural net that we build could do that? And that is the revelation that takes me into temporal processing, temporal sequences. Human neural nets can do that automatically, naturally, already at the age of one month. Now our artificial neural nets are not even at the age of one month as an infant, and I feel that what is the work that is presently going on in temporal sequence storage is very important to be able to begin to model that sort of activity. Now it relates to the work that is coming from computer science in grammar generation, which of course if you are thinking of being able to store a sequence you can generate it and being able to generate these sentences of the grammar with a Chomskyan base structure I feel is again a challenge we have to face up to, but one in which we should try and relate it to what's going on in the infant who at the age of six months again begins to recognize sound, syllables, and so forth.

Well that is one area that I feel is going to be very important: temporal sequences, and the ability to go on and even understand how we may begin to think at the age of a month. The other area I would say is important is stochastic features. We have heard a number of talks in which stochasticity is being brought in. We know from simulated annealing which is a very important feature in some of our algorithms. Now, real neurons are noisy intrinsically. We might say again „What is that noise doing?" It is not something that is necessarily got rid of because it would seem to be strange if we have in survival of the fittest an evolution of something that in fact has to be avoided. So I would wonder whether stochasticity is not something that we should take much more note of, and I see that various groups are involved in that. And the ability to put noise in intrinsically into systems and use it I think will become very important over the next few years. Well, that's a way of getting into hardware, I feel that's the other area, and there are many groups who are working on the hardware aspects. Whoever gets a learning chip that can be used in a broad range of algorithms I feel will have a device that will allow us to properly take off in neural networks. And the message that comes from ANNIE, and I think this is one that we should all recognize in neural networks, the message that they are giving is that in the benchmarking, neural networks are okay, but in most task domains, they are only as good or a little bit better than most standard algorithms. Neural nets are not going to take off until they can take off from Mars in the sense that they are put in hardware devices, into a robot that can move around intelligently. And it would seem to me then that the ultimate is going to be effective hardware devices based on neural net algorithms, but hardware in not great pieces: small chips. And I would say that's the other area I see as the future.

Chair: I hope that Professor Kryukov can add some comments to this subject. What are the trends that you see and which of these do you feel are going to be the most important in the next few years and which will gather the most intellectual activity?

Kryukov: I would like to mention three tendencies at present appearing. One is that neural networks are clearly becoming stated in general theoretical terms. I will give you an example where someone tried to



formulate back-propogation in terms of differential equations. It is a simple method but it helps much to unite the main ideas into simple ideas and more than that I feel that other branches of neural activity must be put into general formulation rather than a good mathematical formulation-for example, the phase transition period that is occurring in mathematical proofs and its going at present in neural networks. We now see phase transitions not like a catastrophe but as a useful property of emerging, spontaneous calculations, as a source of something appearing unexpectedly that cannot be expressed in terms of differential equations, because some situations cannot be described analytically. Nevertheless, they can be observed to exist and used. So I shall answer that differential equations and algorithmic review is only a limited part of our activity. This very famous situation: we calculate, we learn and nevertheless we cannot distinguish how this hidden layer works. Integrate some task. I think it is great in some sense, because we must admit our limited capability to understand in state terms ways of what they do by neural elements. But we can catch the idea and I think this idea of phase transitions, its point is to communicate with many elements as a single entity. It is very important for complex situations like in thinking, moving, and so on.

The next point is this phase transition as a synchronization problem. Of course it is trivial to think that synchronization, and especially local synchronization, is a particular case of phase transition. More interesting is that in behavioral sciences we constantly observe many phase transitions, like going to sleep, like heavy creating, et al. , and if we will be able to explain them for example in terms of a good statistical theory of phase transitions, I think we can understand new data and we can build something interesting for technology.

And my last point is about apprehension computation. Of course it is understandable to select some portion of information from a complex program, is good for present needs, computers and so on. But the more important part is to understand various dispersed numerous data in various sites, in neurophysiology, psychology, physiology and others. And I believe that understanding will come not from a particular piece of data, new data, like particular data, but in attempts to understand the meaning of all the existing data from a rather simple single point of view. Of course it is not ultimate understanding but if you give us something important to believe in, to think in future terms. So I think that an attentional neural computer is not only a technological problem; it is also a general scientific problem to build a model of a rather general nature, of course using all our achievements in present products. And nevertheless to try to understand this numerous data, this unlimited wealth that we have acquired.

Chair: Now Doctor Mori comes from a more practical side of this question, from an industrial application point of view, and I will be very interested, as I am sure you will, in his perspective on what some of the major trends, perhaps in the application domain, will be.

Mori: My area is really the computation of the control data to electric power systems. So far in power system operation I expect neural networks in two areas. One is real-time computation and a high accuracy. That is a clerical central to the system composition. So what I extend to neural networks is there are three points. One point is a learning procedure or algorithm. Right now we have two main-stream algorithms. One is a back-propagation task and the other is a conclude type. But most of them give us a longer, meaner problem (laughter), so we'd like a stochastic algorithm, just like a future value theory based on a theory of random generator. And on the other hand, we would like to extend an aproximate method and to define an algorithm. The second point is a power system is one of a very complicated system. So again we have a large type system. Currently when you talk about neural net application in Japan you get people who ask "How can we solve a large-scale problem?" It is very important to solve that problem with some crispening technique affecting the number against the number. At the same time we need a sort of optimal new neural nets. Mainly in a power system we are interested to get almost the exact data. This means all the data involved, and also all the knowledge of the system function. So even though we construct a neural network, actually we cannot use it immediately because the expert ability of such a network is at the beginning very weak. My last point is in our area, so far the expert system technique is very popular as an AI tecnology, so in the near future we expect integration with the computational numerical algorithms and expert systems and artificial neural networks. That is kind of idea is not so popular right now, but I believe in the near future we can solve difficult problems with such computational techniques.

Chair: Next we have the advantage of someone who lies sort of midway between industrially oriented research and academic research. Dr. Giles is with NEC corporation and runs a research facility. In addition he is a professor at the University of Maryland and has a very strong interest in all matters, academic and practical. What are your views on these things?

Dr. Giles: Well the terrible thing is that I think agree with what most people here have said, so as a consequence I cannot disagree with you.

Chair: We are not here to disagree.

Dr. Giles: Well I always feel that argument is healthy. On the other hand, there are some issues that were not brought up, and I am not maybe the person who should be talking about this, but since no one else has

brought them up, I will bring them up. I think some of the most near-term excitement is going to come from engineers who are using neural nets to solve real problems. Let me give you some examples.

There was recently a workshop at Yale in August on neural networks for control of adaptive systems and robotics, well attended. Next year, there will be a workshop — these are all sponsored by IEEE — the IEEE is an engineering society which in the United States, excuse me, world-wide, has over 210,000 members. That is a large society. And you get these people interested in neural networks, really nice things happen, and the control people and professor Guez will be here tomorrow to give you a talk on the control problems. I think that you will see amazing progress made in areas like that. Next year in Princeton, there is going to be a workshop on neural networks for signal processing and pattern recognition. What that means is that these people — it is not like we are splintering or going our separate ways — it means that these individuals are so excited, they only get together and focus on specific directions and ideas.

So I think that some of the areas that have really nice near-term applications, ones you see people talking about, products coming out, are in speech and signal and image-processing. Larry Jackel at AT&T, Bell Labs, is now trying to sell his chip, in fact successfully selling his speech-processing chip within AT&T itself. So those to me are very exciting directions, and if you want to look at what is coming down the pike, I think there are seven papers — we are looking at this area ourselves — there have been seven papers in neural networks or control of switching systems. Now switching systems are complex structures. They are very complex structures, and there are people who could build chips for switching systems using neural nets, and I will be glad to share these references with you if you're interested. What I am getting at is that this is the same kind of direction that the robotics people are going in. They feel that neural networks for control of massive systems, and you know, it's your only hope for really solving some of these problems because the dynamics ... Crays are dwarfed by these problems. So these are the nice kinds of problems that I think one can look at.

There is one other direction which I think you might be quite interested in. Some people are using neural networks for music composition. There have been a couple of papers out on this already and my company, NEC, is also developing neural nets for doing these kind of compositions in a multi-media structure, in a multi-media format. And even though I do not think that has as much excitement after talking with Dr. Ezhov yesterday, I think there is a lot of potential there. I mean that we should look at neural networks in terms of artistic creations. And that's all I want to say.

Chair: Dr. Joel Davis is in a unique position to comment on these issues because of his leading role in developing this field as director of a large research program. Joel, would you share your comments with us please?

Davis: I think that most of you are sort of aware of the position that I take in my advice with regards to this area, that is I am a very strong supporter and defender of basic biological and cognitive research, and computation models that spring from this biological and cognitive research, and the idea that one can through processes of reverse engineering try and eventually build devices that do the kinds of clever things that biological systems do. I think my predecessors have by and large talked about applications, and I would like to just briefly address for a second some of the basic research issues from that side of the coin. I think the field as a whole, near-term anyway, is very healthy one sees, if you could measure health in a scientific discipline by the production or the coming into existence of new journals, and certainly we see that happening. If you can measure it by success with increased memberships in societies, and actually the proliferation of societies, both in the U. S. and Europe. You can see that as a good indication of health in this field. I wish that — Professor Taylor is too modest. He has a very dynamic role in the formation of new European — I guess it is not so new now — but the major European neural network society, the acronym is JENNIE. And in the States anyway, we see a number of new courses being funded and new programs being funded and coming into existence at major technological institutions like Cal Tech and M. I. T. that attempt to train a new generation of students in these quantitative neural network techniques that then can be applied to, among many possibilities, biological systems. And for those people who are already in the field, there is a proliferation of courses, summer courses typically, in places like Wood's Hole, where neural biologists tend to congregate in the summer (they all go to the shore or they all go to the seacoast, of course), and places like with major meetings, there will be all day courses.

So what I am talking about is what I think is scientifically short-term success, and our charge from Hecht-Nielsen was short-term, but if you allow me to extend that out a little bit more, I think really that this discipline, whether they accept it or not, is in some sense in a contest for support, at least on the basic research level, with what you might call the AI community.

And you can agree with that or disagree with that, but scientifically, from a point of view of providing support for this research, I think that at least from a governmental level, where still most of the money for basic research comes from, this is still a battle that's being fought. How can you win this battle? I think that you need some short-term successes, the kind of successes that Robert Hecht-Nielson and small aggressive companies like his have been able to provide, and some of the things that hopefully Lee

Giles and his collaboration with NEC will be able to provide. I mean you need directed applications, and I think still at this point this field is still regressive, it still needs financial support from government agencies. In the short term, I think it's problematic. I think that in the mid-term and long-term, neural science has a lot to offer this field. I think there are new technologies being developed in the confocal laser scan electron-microscopy, and other technologies such as voltage sensitive dyes, which allow one to spread a voltage sensitive chemical on the surface of a tissue (either in vivo or in vitro) and determine microvoltage changes in real time in biological neural circuits.

And I think another future is new algorithms for multiple micro-electrode recording. I think this is a kind of feedback, a very interesting loop from neural networks to neural biology back to neural networks again, because either algorithms being developed — traditionally if you wanted to look at a real neural circuit, you probed with a number of micro-electrodes simultaneously, and what you tried to do was isolate one cell with one micro-electrode. Well there are techniques being developed that use mathematical analysis, I think one of the terms is "stereotrode", that you basically use mathematical techniques for teasing out a number of cells, so instead of putting in 22 micro-electrodes to record good recordings of 22 different cells, this process becomes much more simple.

Let me just briefly end with where I think the applications for neuroscience are. I think number one is robotics, that is been alluded to. I think there are very interesting cerebellum models that are being produced by people like Jim Houk at Northwestern and others. There is a good background of computational early network models by David Marr and others in the system. Vision, we talked about this before, I think the stuff of Gray and Singer is very exciting because, as I have said in my talk, it is putting the visual system back together again, instead of reducing it to finer and finer levels of analysis. I think attention's been brought up a number of times. I think one problem is when five of us talk about attention at this meeting, we mean five different things. That might even be a good thing that this panel could discuss, or some other panel in the future. How do we define attention, and of course how do you implement it. If you could put what a cognitive psychologist calls attention into a machine, you would have an incredible machine.

One last point to Professor Horejs, and that is — you called yourself a pebble between two, I don't remember exactly, a pebble between two hills, but I think given the past scientific history of this country, and the technological history of this country, maybe the short-term looks bleak, but I think the long term is probably very exciting. I spent my lunch time going to what you call the Mikrobiologický Ústav (the Institute of Microbiology of the Czechoslovak Academy of Sciences) and I saw a new building that will be partly devoted to computer analysis directed at sterology, according to Dr. Ivan Krekule. Sterology is a new computational technique for neuroanatomical analysis. Certainly, this and other kinds of new, computational techniques could be reasonably applied to a neuroscientific analysis of neural networks.

Dr. Hořejš: Thank you for your encouragement, and I would like to add that our effort will be to diminish the gap between us and the world solely symbolically, (moves chair closer . . laughter . .)

Taylor: Well if I could add an insert, because I did not feel that it was necessarily appropriate at this point in the proceedings to talk about joint activities and joint ventures in neural networks, and I feel that they are very important to appreciate the way that the network community is becoming more of a network than it was before. If we can then get a proper network established in Europe, it would be helpful for all of us; I had a very enjoyable lunch with Dr. Novak and his colleagues, and it was very pleasant to sit and look at your beautiful city here, but also to talk about possibilities for the future where I see that the main problem, which is one that will have to be solved I think in the proper way, is the relationship between Eastern and Western Europe. In Western Europe, we have the EC, we have the ECCU, we have of course Margaret Thatcher, but we will not talk about that . . . (laughter) We will talk about the problems that we see with the way that the EC is becoming more of an economic community. It is funding neural networks in an ever-increasing amount, and the problem really is how this relates to Eastern European activity, which in some ways would become more disadvantaged because it will have lesser funding and therefore there will be less going through into the industrial marketplace, and hence the gap will widen. I do not think that is right, you do not think that is right, we have to make sure it does not happen. So there has to be proper representation to the EC, through this country, in Western Europe also, to make sure that there is proper collaboration. And we want, I think, to try and have a proper new network community of neural netters through the whole of Europe, and this includes not only Czechoslovakia, but other East European countries, Russia and so on.

Now I see that there is the bigger problem of total world-wide contacts. I think we are still in our infancy as far as that is concerned. But the point is how to deal with the AI community; clearly we will gain strength by having an international world-wide structure that is strong. We should know how to do it, and in fact we should be able to build a neural network to tell us how to do it, and I leave that as an exercise for the interest of all participants. Thank you.

Giles: I really do not think we are in conflict directly with the AI community. I think they have a lot to offer, and it's true that there's a lot of discontent among AIers in the AI community, and we in neural nets are

fortunate enough to have gotten some of those people who have not been happy with the conceptual and scientific directions that AI has taken. But they have learned an awful lot in AI; they have learned how hard some of their problems are and therefore have benefited us immensely, because we know how hard those problems are now because they worked on them for twenty years and did not solve any of them.

On the other hand, they had some tremendous successes in Expert Systems, and it is important to realize that. It is important to realize that these rule-based systems really are useful models, but they have limitations. If you can make the rule-based systems adaptive, where they create their own rules, then that is going to be very useful for us. I mean, the machine learning community talks about all types of learning, learning by memorization, learning with a teacher, learning by example, learning by discovery, learning by analogy. I mean they have done very useful things in terms of exploring the universe of learning and knowledge, and I think that we have a lot to be gained from what they have done and I think we can help them a whole lot; I think they can help us too. And I think that hybrid systems — you're not going to find neural net stand-alone computation, and I really believe we are not going to displace traditional methods of computation; there are good reasons why traditional methods of computation are so successful and are going to continue to grow exponentially over the next years. But the nice thing is to live with them and to leverage our resources based on what they have done and what they have to offer, i. e. neural net chips, neural net VLSI, etc. So it is fitting into those directions which I think will offer tremendous promise for the growth of neural networks.

Chair: In very much along the lines that Professor Taylor raised, regarding the structure and sort of the world structure of this research, particularly the scientific aspects — I think the industrial aspects of this research are very likely to split into divisions such as individual companies or countries, or whatever, because at least potentially there is a tremendous economic impact that this may have, and I think those benefits will be jealously guarded. But ignoring that, and speaking strictly about the scientific research aspect, the question I have, and I would like everyone to comment on, is how should the meetings in this field be structured over the next few years? We have gone through a period of chaos. We have now a very simple sort of structure where we are starting to see these specialized meetings, that Lee mentioned, emerge and be very successful, and yet we still have large meetings. We now have one IJCNN per year and there seem to be also large continental meetings like a large European meeting and a large Asian meeting, and do any of you have any comments? Maybe as to the current approach? Are there better ways, and in that regard, this issue of countries and researchers who are not able to participate as much — are there answers? We have spoken of Czechoslovakia, but what about Gambia: imagine being a neural net researcher in Gambia or in Paraguay. We feel suddenly very lucky. Okay, maybe we each could comment on that subject. How the meetings are structured —are you happy with the meetings? Do you see the need for more meetings, fewer meetings? How should these things — should the sponsorship change? Or is everybody pretty content? I have to say personally I'm content.

Hořejš: For this one we can end completely exhausted. As for the next one I look forward to it very eagerly, but I do not know where and when. I have no precise idea.

Taylor: Well from the point of view of Western Europe, we are proposing one main meeting a year. There is the ICANN '90 meeting, we had it in Paris a few weeks ago, the next year's will be in Helsinki, it will be organized by David Kohonen, the year after will be in Brighton in the United Kingdom, on September the 4th through 7th, for those of you interested. We will be circulating to everybody whose mailing address we have, and we are presently developing a mailing list, because that is one of, I think, the crucial things in networking: to get to people the information about what the meetings are. There are constantly meetings throughout the world, but many people in the field do not know, and do not have access necessarily. This is where in Europe we are proposing to get a newsletter out that will be able to inform people. Now the question is whether to, hopefully with our colleagues in Eastern Europe as well, try and develop specialized meetings, because we do know that there are specialized areas that are becoming clearer know. For example, if we look after the mathematical theory, we could have a whole meeting on that. I am holding a meeting in London on coupled neuron oscillators in December. There are all these specialized meetings that should be developed with the associated areas, which can then report back at the international meetings as to how developments are proceeding. That, however, leaves a problem paramount: Paraguay. Leave the problem or gamble it — what do we do? My good colleague A. Salam at Trieste at the International Centre for Theoretical Physics has, through UNESCO, developed a network called Centres or Institutes of Advanced Learning in Third World Countries. Now there is a problem still because in this form we have already had to face up to the fact, mainly that there are too many international conferences already, there are three a year, and if one takes account of other ones as well, such as this very good one here, one would have even more, one would have a half a dozen, even a dozen; and I am not sure I quite agree with Bob in the sense that maybe there are just becoming too many, but it may be that one way of helping the situation has been suggested already, that we run a conference world-wide, using sat-

ellite links so that somebody can lecture say here, through a satellite link that is beamed to an audience in Seattle or in New York or well, Gambia if they have satellite links as well, I am not sure. But we have many problems on that side, but let us assume that wherever there are good centres of communication, one can have involvement.

There are problems about time zones on that. It may have to be in the middle of the night to some audiences and how do they stay awake? Well maybe enough coffee if you are down in Venezuela, but I am not sure it is quite fair to fill them with caffeine to keep them awake. I think we have to face up to that problem. Our students are not able necessarily to be supported by funds to go to all the international conferences. To send them off to the United States for the ISCNN conferences is more difficult. And so it may be an effective way to do it if we can have a go at it. I think it is something for the future we are trying to think of. But if anyone has any further suggestions as to how to enable us to avoid a plethora of conferences, but yet to let people have access to all the conferences that they feel they should attend to keep themselves topped up, I think its very important we try and resolve this.

Chair: Now Dr. Kryukov ran a very successful meeting last summer in Moscow, and being from the Soviet Union has a unique perspective on scientific communication. He is one of the globe-trotters of the Soviet Union; he has been going to many of the most important meetings, in fact all of the important meetings for the last few years. And so maybe your perspective on meetings — are there enough? Are there too many?

Kryukov: I will try to express my personal view. Of course my knowledge is limited, but the general feeling is that small highly-professional meetings must be the basis of other activities. At the same time, I would stress that this activity is very good as a phase transition point, and as you know, at this point every level of structure and organization plays a role. You cannot reduce the situation to a single or two levels only, because they are closely interconnected and play a serious, important role. I will give you an example. Those who are skeptical about micro-tubules (and I was among them) can think that we have the leading role, and here is a situation now closely associated with phase transition. At this point, you cannot limit yourself to past activity, to a remembering level. Of course in the vicinity of this critical point, you can reduce to one level miniscule approximations for example. Nevertheless, I believe that some situations will require simultaneous contemplation, simultaneous analysis of different levels, and as for meetings, of course you need both very great meetings, like in America, and, as I said, small professional meetings to cooperate with. As for concrete steps, I am not a professional. I do not know what is best to do, but I think we are self-organizing, and we do almost natural things.

Chair: Thank you. Now Dr. Mori has, I am sure, another opinion. In Japan, the Japanese Neural Network Society has excellent meetings and there are a number of other important technical meetings in Japan that have a strong neural network component. So I think your perspective will be most interesting.

Mori: Right now we have three big institutes focused on neural networks, so that each institute has a neural network briefing every month. So in industry, we are always confused [laughter], because even though we develop a lot of papers, we do not have time to look over everything. So my graduate students always have to deal with that. As far as the neural networks papers are concerned, we actually don't know which one is our best one.

By the way, talking about my area, power, our point of view — it is a very conservative area — so we do not have any international cooperation. I mean the other institutes have a joint contract with Korea, with Shanghai. Fortunately the Institute for Electrical Engineering Society has an international juried committee in the power area. I understand the lack of international communication. I hope that we shall become a member of an international conference in the power area, because graduate students and younger professors do not have the opportunity to exchange ideas with foreign people.

Chair: Thank you, that is a very interesting point. I have had the pleasure of attending a number of neural network conferences in Japan, and I have two problems. One is this problem of sheer volume, the number of papers, and trying to figure out which of them mean something to me, and the other problem is the fact that many of them are in Japanese, which is an impossible barrier for me, having no language skills. Now, Lee Giles has been involved in probably more organizational activities relative to meetings than most people have attended meetings, and so he is in a unique position to comment on this sort of topic, of meetings, their structure and number and so forth.

Giles: I will assume that is a compliment [laughter]. I think there are too many meetings, but I think as Professor Kryukov said, they are self-regulatory. After a while, people stop attending them if there are too many. I mean the attendance will go down, so you find that the meetings will come and go. Meetings require people's efforts, a lot of effort, a lot of time, and after a while it will be decided by the group that puts them on that it is just not worth doing anymore, and it has happened, you know, you can see this in other fields from meetings that just sort of died away. However, it is really hard. I mean I do not have any ideas. I would like to have some body, some formal body make some decisions on endorsing meetings, whether we should have a joint world meeting — I think that would be superb, that is a fantastic idea. And at that time you have to get an agreement from other large

meetings that they probably would join together for this one meeting every year. It would also be very useful to either expand or condense possibly the proceedings. IJCNN is known as the phone book meeting, because you come away with four phone books, and for some people it has even cost them extra on their luggage. We need to do something about that. We recently had this discussion with NIPS, the Neuron for Information Processing Systems, conference held every year in Denver. We receive so many fine papers, and we had to reject so many fine papers, just because we wanted ONE book. We did not want two, you know, we did not want to go to the phone book approach, but on the other hand, the feeling was that we wanted papers of significant length, so if you read them you had more than just a glimpse of what someone was doing, you really knew what he was doing to the point where you might be able to go and duplicate the work yourself. So there was an issue of „Should we expand from eight to ten pages?" I think these are hard issues.

Let me bring up something just very quickly though that I think that people maybe here are not aware of, but in the United States and other places, it is becoming a great way to communicate results or get into good arguments, and that is called E-mail. In these neural network nets, there are two that I am aware of, that I participate in, one is Neuron Digest and it comes out as a large digest of things maybe once a week, and then there is the Connectionist Net, and that is supposed to be neatly after you make your comment, so you bring a comment in, you say „Well I'm . . . " and you do it sort of politely, „I don't agree with so-and-so, these are the reasons . . . We should look at this, this and this . . . " and then there is a flurry of responses, and it is fun to read this mail, because you get to see what people are really thinking about and you find out about the latest trends in research. But be careful, you can also get some bum steers too. We were led down the path of the contiguity problem for about a year because there was a flurry of it on the Connectionist Net. As it ends up, the contiguity problem really is not that interesting. But I think that is a good way to have a good world-wide communication, when everyone can afford to be on the E-mail system.

Chair: I think now Dr. Davis has probably funded more meetings than most of us have been to, and so he also has a unique perspective. I like to think of Joel as a philanthropist and also someone who happens to be a very good scientist and a connoisseur of this research, so I think his perspective will be quite unique.

Davis: Well the difficulty in the U. S. is — here I get the crumbs from the table . . .

Chair: We will let you go first next time . . .

Davis: The best I can do here is say yes I agree that the number of meetings with a free market, where people are free to go to the meetings, either more or less depending on the interest, as well as the need to pay for it, because the bottom line really comes down to the fact that they are very expensive to run and they involve the marshalling of lots of time, effort and money, so if people want them, there will be moremeetings — if people do not want them, there will be less. I think the idea of large international meetings is a good idea. I would also like to re-iterate Lee's point about E-mail. I know here in Czechoslovakia now you certainly have FAX contact with the west. I do not think you have E-mail, or at least if you do, at least it is not very common, but clearly this is coming and clearly the world is going to be linked together with E-mail. And it seems to me up until this point a very positive development, but just as in the West you know, you take a look at your FAX machines and what you see is ten pages worth of advertisements sometimes. I think it is not long before we are going to see the E-mail system cluttered with garbage. But c'est la vie, okay?

My last point is that: what is the purpose of a meeting? I think with modern telecommunications being what they are, or what they soon will be, that the traditional presenting a paper may be a thing of the past. Let me just suggest this to you — I do not know. I think that the rise of the poster session was in some sense a response to this traditional paper method, which as far as I know has been ever since the Royal Academy, a couple of hundred years and maybe before that. I think, to leave you with an idea, the primary purpose of a meeting like this, is the interaction of individuals outside the meeting hall, over coffee, over pivo [Czech for beer] (laughter), whatever it be, and that is where we are going to gain most from. We can take the program, and we can on the plane home or on the train ride home read it at our leisure and think about it. But look into your own selves: what you are really going to remember here is ideas that you discuss for the first time with people that you have not seen before. I think that you can do this with E-mail probably, though it is not as easy.

Chair: Down at the other end . . .

Taylor: Now, with respect to Paraguay which was raised earlier, I am in fact just about to go out there at least just before Christmas, to Paraguay. Now you might say „What are you doing going out to Paraguay?" [laughter] But if anybody has got an answer, I would be most interested to hear it. [more laughter] But seriously, the point is that this is being organized by the EEC, it is funded by the EEC at least, it is a group of neuro-scientists who are going to go out there to try and arrange joint collaborative work in neural models. And that seems to be the sort of thing that actually has not been raised by anybody, which is this, also EEC, idea of cleaning of lavoratories, and this is a way of having what you might say are micro-

meetings. The mini-meetings are the workshops and the macro-meetings are the big conferences, but it is the micro-meetings that are also as important, and in a sense should go on after the macro— and mini-meetings. And it is that approach that I certainly think is developing in the EEC, through the cleaning of lavoratories, through funding of travel. The British Council, I am proud to say, is doing its little bit for this as well. And I feel it is important that we recognize it is true contact between people over a period of time, and not just what has happened here and now say for a few weeks, that is going to get a lot of development done. And I do feel that we should all think about that, for what we may get out of it. And you have got then to push that yourselves to try and get twinning arrangements or whatever with Eastern European, with EEC, with American, with Japanese groups.

Joel Davis: Could I just add a note to that? I agree with you, and I think the Japanese are to be congragulated on the Human Frontiers program which has a strong flavor of directed work with laboratories outside of Japan.

Chair: Now in this research that we are discussing here, neural network research, one of its characteristics is the multidisciplinary flavor. This is not a unique subject, but an unusual subject in the broad range of disciplines that make up the work in this field. I have a question in terms of future trends. We have already seen large numbers of physicists become interested in this field, and more and more neuro-scientists have become interested, and some mathematicians — not many, but some —and other fields as well. One of my questions is: is it maybe time to become proselytizers? In other words — John Hopfield was the originator of this theme, when he made so many talks — he gave talks all over the world to physicists and got many people interested in this field. Is it time for us to proselytize over the next few years to convince some of the top people in related fields to join this field? I know Dr. Kryukov has a young mathematician in his group who I do not believe started off doing this sort of thing, but has been now sort of coopted and drawn into this field and will probably make spectacular contributions. So I think that a question for everyone is: Do we need to be active in our pursuits of the interests and time and talents of people in related fields, or should that be a natural process? Any thoughts on that?

Davis: I think it is absolutely necessary, I think you brought up a good point. A good example with John Hopfield: not only has he gone around the world talking to physicists, and of course he addresses major neuroscience meetings on modelling and neural back-resistance, which I know is a very great interest of his. But you know John carries it one step further: I have to tell you a little story that he told me. That is in July when he was at the Aspen Institute for Physics in Colorado, who else should be in Colorado at the same time but Maggie Thatcher visiting George Bush, and Maggie Thatcher I believe is trained as a chemist, and she asked what was going on at the Aspen Physics Institute, she had heard about it, and in fact John Hopfield and two or three other physicists gave her a briefing on, in John's case, neural networks. So taking what you suggested I think is taking it to the very top. [laughter] I do not know if Dr. Kryukov has Chairman Gorbachev's ear, but it can never hurt to go all the way to the top.

Lee Giles: Well, I agree with Joel, but there are some problems, and let me state the problems as they have been told to me, not as I see them. Institutions — academic institutions tend to be quite conservative, and if a young Ph. D. comes into that institution, at least in the United States, he has six or seven years before he can get tenure. They usually hire only those who they think have a good chance of getting tenure. And if you are going to get tenure, they want you to get tenure in that field, whether it be Computer Sciences, Electrical Engineering, whatever. There is a feeling in some institutions — in quite a few institutions — that if you have a degree, if your Ph. D. , your research, is in neural networks, that is fine, that is good, but you also better be that other person, you better be that physicist, you better be that computer scientist, you better be that electrical engineer, because if you are not, you are not going to get tenure. Because your peers are not going to give you tenure. So it is a hard topic in some ways to encourage young people to do exclusively neural nets. I mean you have got to encourage them to do neural nets, but also be whatever field they are in, unless somehow there is going to be a Department of Neural Networks, or something like that. And I do not know what to do about this, because I have even seen it in my own institute where we have been hiring quite a bit and someone will come in, like in a field like neural nets or in another field like nonlinear dynamics, and the question is „Well, what else are you?" You know, „Are you more than just a neural nets person?" because neural nets is a purely speculative field — it died once, maybe it will die again. These are comments you hear from conservative, but I think very famous physicists and computer scientists and neural scientists, etc. , and I worry about that.

Chair: Professor Mori, could we have your comments on this, in terms of recruiting people from other specialties. In power engineering — Lee was talking about conservative people, very cautious. Power engineering is the height of this.

Mori: Fortunately in Japan, the big companies encourage us to do research on neural net applications. For example, in Tokyo our company provides us with a research grant for application of neural ————nets in

power systems. Talking about the U. S. , the National Science Foundation supports several universities, for example the University of Washington in Seattle has a brief project for signal processing as an application of neural networks, and also Carnegie-Mellon has a project.
In addition, neural networks' potential supports a special workshop on power systems by the neural nets applications. Next year a power systems international forum on application of neural networks in power systems will be held in Seattle. And also the Power Research Institute supports a big research grant. ECT and a power systems company out of San Diego, California have a project, etc.

Chair: Okay, good. Now Dr. Kryukov, as I have said, has already been very effective in recruiting people who work in other fields. The Soviet Union has an enormous supply of brilliant scientists and mathematicians, many of whom I think could be interested in this field. Is this something that you feel should be done? Is it going to be done? How do you feel about this?

Kryukov: From my talk you can understand that I like both local and global directions, for example these particular meetings and this other type, about the multidisciplinary state of neural networks. I think that we have now a unique example of a special role of neural networks, clearly integrating different branches of science into maybe a future science, and I will give only one example. Just recently, not many years ago, psychology and neuro-physiology were quite far apart principally. You can check it from Francis Crick's paper on the brain several years ago. And now there is a growth of specialists in both these areas using the same model for an explanation of their data in terms of each other. Another example is that neural networks through the direction of mind, in some sense, attracts the attention of good specialists. Of course I know that we need their professional skills, their personality. We have a good example: we have at least two Nobel Laureates here. But the point is that neural networks present the unique opportunity to attract the attention of good specialists through not bright, but promising results. For example, this using of Kolmogorov representation theory I am certain will attract the attention of good specialists in analysis and other branches. And I could give many other examples like that, and I believe also in travelling not in extensive sight of this problem, but in intensive. This rare opportunity to propagate and to propose, and maybe I exaggerate, that we are starting a new sort of science.

Chair: And now Professor Taylor has thoughts on this matter, proselyzation.

Taylor: Well I think the problems have been raised already. I think Giles' question as to how to support or advise the young student going into a branch or a subject which may make it difficult for him to be employed on a tenure track, and get tenure at a later day is a very critical one. I do think that we should address ourselves to that. I see that there are actually two different aspects of the way of attacking this problem. One is, and I am delighted to hear that story, I am absolutely delighted that John Hopfield could get at Margeret Thatcher, who really is the government of England. We have to persuade her, and I think John Hopfield is the ideal person to do it, so ...

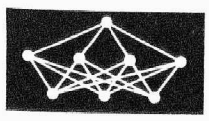
1990 IEEE International Workshop on Cellular Neural Networks and Their Applications CNNA-90, Budapest Dec. 16-19, 1990

The conference on cellular neural networks brought many new approaches to cellular automats and neural networks as well as to their applications especially in picture processing.

The cellular automats and neuron nets are in their principles a little contradictory approaches in the sense that in cellular automats each active cell communicates with tight neighbors only. The two dimensional grid of these cells naturally leads to the application in picture processing. On the other side the standard neural nets are usually fully connected. This way connected is Hopfield net and its modifications, perceptron and many of others, but not all, paradigms of the neural nets. It is spoken about large number of connections, about massive interconnection. The cellular neural networks comprise local interconnection in cellular structures and nonlinearity and adaptivity of the neural nets. Standard approaches in picture processing proceed from small tight neighborhood of each pixel and so they use local matrix of three times three points. The ability to classify in neural nets allows to enlarge this matrix to 5 times 5, 7 times 7 or larger and thus reach very interesting results. In such large matrices it is impossible effectively analyze and evaluate all possible template combinations as in the 3 times 3 matrix. There the notion of nearest template and the ability to classify the patterns in classes seem to be very useful.

The author of this notice is interested in problem of equivalence or similarity in behavior of neural nets having Dr. Levendovsky pointed out a procedure for possible transformation of the fully interconnected net to partially interconnected net.

Rather surprising was the fact, that smaller part of the papers dealing with picture processing dealt also with learning of the net considered. In most of contributions of this kind the optimal template for given task of picture processing is looked for. The ability of learning is most fascinating feature of the neural nets.



This problem was considered in the papers dealing with design of cellular neural nets from the theoretical (Zou et al.) as well as from practical realization in VLSI technology point of view.

As to technical realizations of the neural nets using VLSI technology it was possible to find the analog as well as pure digital approaches. The analog realizations seems to be a little preferred for their larger potential speed and inherent integration process which excludes or diminishes lengthy iterations. It was described many realizations based in one or in the other principle. It was also described several hardware accelerators and hardware realizations of neural nets.

The conference was held in pleasant environment of Budapest. It was good and thoroughly organized and it is possible only to thank to initiator and main organizer of the conference, Dr. T. Roska from CAI of Hungarian Academy of Sciences.

Marcel Jiřina
Institute of Computer
and Information Science
Prague, Czechoslovakia

Book Review

Spectral Analysis in One or Two Dimensions. S. Prasad, R. L. Kashyap (editors),

Proceedings of an Indo-United States Workshop
New Delhi,India,November 27—29, 1989

Oxford and IBH publishing Co.PVT.LTD, New Delhi, Bombay,

Calcutta 1990.
842 p.

The book contains fifty-one invited papers which were presented at an Indo-United States workshop held in New Delhi in November 1989. The collection contains contributions on current and topical problems in the area of the signal processing. The emphasis has been on capturing the important developments currently taking place in various facets of signal processing viz. techniques, algorithms and architectures, and their applications such as array processing, spectral analysis and image processing.

The contributions were organized into six areas of themes:

1. Spectral analysis with multiple nodes. Contained here is a estimation of the directions of arrival of multiple plane waves from data arriving at an array of sensor. It includes the formulation of the problem in terms of artificial neural networks, beam space processing for computational efficiency and robustness and others.

2. Nonlinear and adaptive techniques. Contained here is a significant review paper of techniques associated with the use of higher order statistics in signal processing, the other papers deals with topics on adaptive filtering and control.

3. Multidimensional systems.This consists of papers dealing with modeling and algorithmic issues in image processing, texture classification, computer vision and holographic imaging.

4. Spectral estimation and detection. Included are investigations into high resolution spectral estimation, robust estimation of AR parameters, different methods for a detection of the number of signals in the incoming data and others.

5. Parallel processing. This section contains papers dealing with parallel implementations of various signal processing problems including the solution of the Toeplitz system of equations and image computations.

6. Filtering techniques. Presented here are papers dealing with digital filter design problems as well as with some issues related to numerical robustness and adaptive filtering.

Some of the papers are very useful topical overviews, some contain interesting new and original ideas, the rest are of a more specialized nature and are incomprehensible to an ordinary reader, but, as they are invited papers, all contain an introduction into the problem studied. On the whole it can be said that the book supplies a satisfactory notion of the present level of the field and research directions. We can therefore recommend it to all who take the interest in signal processing. The peoples from neural network society will find there at least three papers which are strongly related to neural networks. These are: **On Parallel Sorting Methods by S. Rao Kumar, Sensor Array Processing with Artificial Neural Networks by D. Goryn & M. Kaveh** and **Deterministic Networks for Image Estimation Using a Penalty Function by A. Rangarajan, R. Challappa & T. Simchony.**

Zdeněk Fabián
Institute of Computer and Information Science,
Czechoslovak Academy of Sciences, Prague

A VIEW ON NEURAL NETWORKS PARADIGMS DEVELOPMENT

(Part 2)

J. Hořejš)*

Here we continue in the tutorial paper concerning the neural network paradigm, which first part was published in the Neural Network World, No. 1, 1991.

3. Adaptation [an exercise in a bit more abstract reasoning]

Now consider what happens if the "neural system" of the chicken from the last section makes a mistake, wrongly considering a hawk $\mathbf{x}^*$ as a farmer due to poor life experience. If it survives, it will perhaps forever remember the difference; in its long term memory LTM, which is in NN models represented by synaptical weights $\mathbf{w}$ (as opposed to *short term memory* STM which may be roughly compared to current activities of neurons, given by the vector $\mathbf{a}$), this memory trace is causes as an adaptation, a proper move of $\mathbf{w}$. The original memory given in *Fig. 7a* by a dashed separating line between squares [hawks] and circles [farmers] should change so as to place the new terrible experience $\mathbf{x}^*$ (full square below the dashed line) in the positive halfspace of the square dangers. The possible resulting full line is shown in *Fig. 7b.*

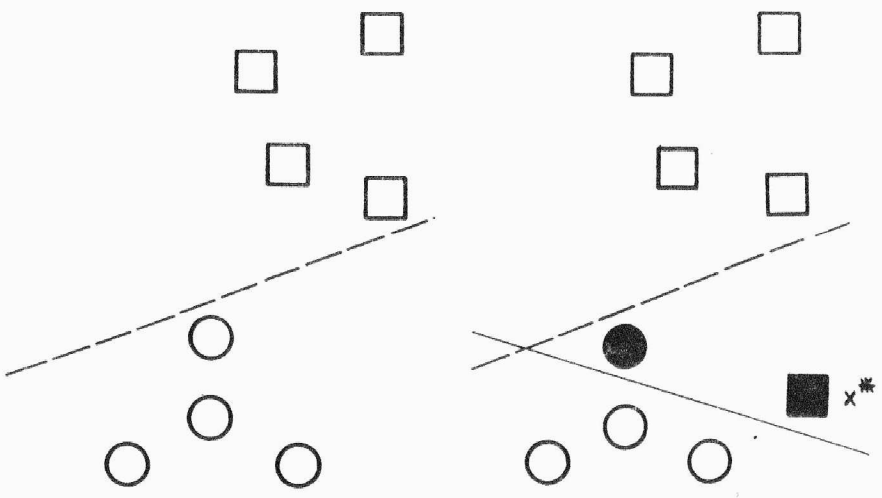


Fig. 7a,b

Rosenblatt invented an algorithm, the so-called *perceptron algorithm,* which handled such simple cases. The algorithm was able to treat many successive corrections of separating lines (separating hyperplanes) as the need arises. It is not quite easy, because if you solve one counter-example like $\mathbf{x}^*$, you may generally introduce others: the full line in Fig. 7b covers the terrible experience of the chicken, but makes it a bit too anxious — even some GOOD messages now cause it to get into a panic (cf. the full circle). So the algorithm should take into account previous cases as well and Rosenblatts' task — successfully completed — was to show that the invented algorithm converges, finally establishing the separating line so that all squares are in one halfspace and all the circles in the other. The idea of the algorithm may be roughly (but inadequately) explained by the formulas

$$\mathbf{w}' = \mathbf{w} + \mathbf{x}^*;$$
$$\mathbf{w}' \,.\, \mathbf{x}^* = (\mathbf{w} + \mathbf{x}^*) \,.\, \mathbf{x}^* = \mathbf{w} \,.\, \mathbf{x}^* + \mathbf{x}^* \,.\, \mathbf{x}^* \geq 0 \quad (4)$$

The first formula exemplifies the adaptation law:the new (full) separating line specified by the vector w' after the occurrence of $\mathbf{x}^* \in$ BAD is simply adapted as a sum of the previous (dashed) line given by w and the terrible input $\mathbf{x}^*$. [Remember that $\mathbf{w}$'s are normal vectors of corresponding lines]. The problem was that $\mathbf{w} \,.\, \mathbf{x}^* \leq 0$, indicating wrongly that $\mathbf{x}^* \in$ GOOD. After making the sum $\mathbf{w} + \mathbf{x}^*$ and under the additional assumptions that $\vartheta = 0$ and all the vectors under consideration are normalized [so that we can sum up apples and oranges], we get among others $\mathbf{x}^* \,.\, \mathbf{x}^* = 1$ which itself exceeds (or is at least in absolute value equal to) $\mathbf{w} \,.\, \mathbf{x}^*$, so the situation changes. The new hawk $\mathbf{x}^*$ is again recognized as a hawk. As already mentioned, the proper algorithm of a perceptron looks somewhat different; having an iterative character, the adaptation law is not so strict as stated in (4), but is of the form $\mathbf{w}' = \mathbf{w} + \alpha \,.\, \mathbf{x}^*$ for some α, $0 < \alpha < 1$ etc.

There is however a good reason why we do not dwell on a detailed description and proof of the perceptron algorithm and it is actually the reason which was pointed out by Minsky and caused the second gap in investigations of NNs. We have already noted that a single perceptron has a limited recognition capability (take any linearly nonseparable subsets, say the XOR problem). Multilayered nets do not suffer from this deficiency, but in the time of the second NNs generation, research had not provided us with any algorithm similar to the perceptron algorithm that would function for multilayered nets and thus for their automatic adaptation and to *learning by examples,* which is one of the most important feature both of living creatures and NNs today. Moreover, Minsky gave reasonable argumentation that it would never be possible to solve this problem successfully. Fortunately enough, he was wrong in this respect and invention of new adaptation algorithms disproved his claim — but this is the question of the "third generation" of NNs, to which we return later on.

Now we recall our chicken toy to formulate the general problem of adaptation and some related concepts. Assume we are given some NN for which it is easy to state active dynamic laws (that govern how the

*) Prof. Dr. Jiří Hořejš, Department of Computer Science, Charles University, 118 00 Prague 1, Malostranské nám. 25, Czechoslovakia

activation of neurons is spread over the net and of which (3) is just an example); for the net it is also easy to specify a mapping from the input (say m-dimensional) space into the output (say n-dimensional) space. The net thus implements a mapping φ: R^m-->R^n, where R is the set or suitable subset of real numbers. Denoting **x** an input vector and **y** an output vector, we can write $\varphi(\mathbf{x}) = \mathbf{y}$. Multilayered nets are good examples of structures which can realize such mappings in a rather unusual yet useful way.

The mapping φ of course depends on the *topology (geometry)* of the net (number of neurons, layers, way they are interconnected, etc). All these things being fixed, φ depends heavily on the synaptic weights assigned to all existing connections. If **w** again denotes the weight "vector" (which may now have a rather strange shape, in multilayered NNs reminiscent rather of a sort of matrix), we should take it as a parameter of the mapping, φ_w. It is not difficult to imagine that the number of components of **w** can highly exceed the dimensions of **x** and **y**. In a specific yet important net of n neurons, where each neuron is interconnected with each other, **w** is usually given by an $n \times n$ matrix having n^2 weights [this net does not belong to the class of multilayered nets!]. Because of the important role of connections, and long term memory (as well as other capabilities to be discussed later on) represented in these highly interconnected nets by synaptic weights **w**, you can often meet the terms "neuronal" and "connectionist" used almost interchangeably.

One way how to use NNs is to use them to implement a "given" mapping. In mathematics, "given" usually means that we are able to specify it analytically (by a formula) or algorithmically (by a program). It is difficult to imagine that our chicken has wired in its "brain" a Pascal program to create the lot of recognizing capabilities necessary for its survival. Even man does not solve all his daily life problems strictly logically on the basis of theories he/she has learned in school. A four-year-old child speaks quite well with its words and yet it has no idea of grammar and its rules. It simply *learns* to understand and to speak by examples, which it observes around it. NNs are trying to mimic to some extent this ability. A mapping to be implemented is mostly "given" by the examples.

Formally we consider a so-called *training set* T as a set of pairs $T = \{ [\mathbf{x}^1, \mathbf{y}^1], [\mathbf{x}^2, \mathbf{y}^2], \ldots, [\mathbf{x}^p, \mathbf{y}^p] \}$, all $\mathbf{x}^i$'s being m-dimensional vectors from the input space (representing external stimuli) and $\mathbf{y}^i$'s being n-dimensional vectors from the output space (representing proper responses) and then try to establish the net N_w [that is weights **w**], which would implement the map φ_w so that

$$\varphi_w(\mathbf{x}^j) = \mathbf{y}^j \text{ for all considered } j\ (1 \leq j \leq p) \quad (5)$$

Fig. 8 depicts abstractly the net $N_w(\varphi_w)$. Taking an arbitrary set of initial weights, **w**0 say, it is improbable that (5) will be satisfied by φ_{w0} (**w** = **w**0, which is often chosen randomly, or from some previous experience

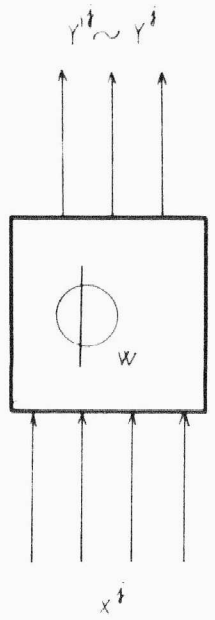


Fig. 8

playing a sort of genetic memory or our knowledge of the problem to solve, like the assumed symmetry of some parts of **w**). Yet in many cases there is a **w*** such that φ_{w^*} satisfies [at least approximately] the equations (5). The process of (successive) changes **w**0-->**w**1--> . . . **w***i* . . . -->**w*** is called *adaptation* of N. Often [but not always] it relies on a *supervisor* j

(teacher) which signal differences between *expected* values $\mathbf{y}^j$ and the actual values $\mathbf{y}'^j$, produced by the net which is not (yet) well adapted [using e. g. weights **w***i* so that $\varphi_{wi}(\mathbf{x}^j) = y'^j \neq \mathbf{y}^j$ for some j]. The announced differences are then utilized by a given adaptation algorithm, which tries to change weights *wi* so as to diminish these differences. Unlike our simple (and not completely discussed) chicken example, most adaptive algorithms change the weights in small steps so as not to miss a satisfactory **w*** nor to make the weights oscillate too much. It follows that the weights should be adapted only slightly and many times; the algorithm of adaptation dynamics then needs also to be exposed to the members of the training set many times to achieve slight changes of **w**. Together with the adaptation algorithm we have to present also a *training strategy* which selects and submits training pairs to the adaptive algorithm in some order (including perhaps random features and a lot of various repetitions).

In a biological system the "teacher" who establishes the differences is a vague mechanism of punishment and rewards ($\mathbf{y}^j$'s are not stated precisely and unambiguously and also the differences between $\mathbf{y}^j$'s and $\mathbf{y}'^j$'s may be difficult to measure). When the error is measured only by some "grades of success", the term *"critic"* is used instead of "teacher": teacher is supposed to be able to confront every result with a correct one, while critic just estimates the general performance in (perhaps also general and fuzzy) terms.

In technical systems on the other hand there are several reasonable formulas which can serve this purpose. Mostly the error function E_w is given by the least mean square criterion,

$$E_w = \sum_j \sum_i (y_i^j - y_i'^j)^2 \quad (6)$$

where the inner sum is taken over all n output neurons ($1 \leq i \leq n$) and the outer sum is taken over all training pairs sets submitted (perhaps repeatedly) during some stage of the adaptation process.

Displaying the value of E_w we can see how the adaptation process behaves. This value should generally decrease until E_w^* is sufficiently small [although rarely zero in the case of input/output vectors composed from real numbers; anyway this would mean that the net perfectly satisfies (5)]. E_w is of course function of many-dimensional **w** (the dimensionality is usually enormous w. r. t. the many neurons used). The graph of E_w (difficult to depict unless we restrict ourselves to dimension 2 or 3) is called the *error function landscape*, often rather bizarre, with many hills, valleys, etc. It is purpose of all adaptation algorithms to find a trajectory over the landscape, which — starting with **w**0 — would ultimately fall down into a (possibly stable) minimum (hole) of the landscape. In the best case we end up with reaching the *global minimum* of E_w^*, but sometimes we have to be satisfied with a not too high *local minimum*. Not necessarily, but very often, we look for a minimum using a sort of gradient method: **w***i* as a point on the landscape rolls down along the steepest descent. *Fig. 9* shows several trajectories on

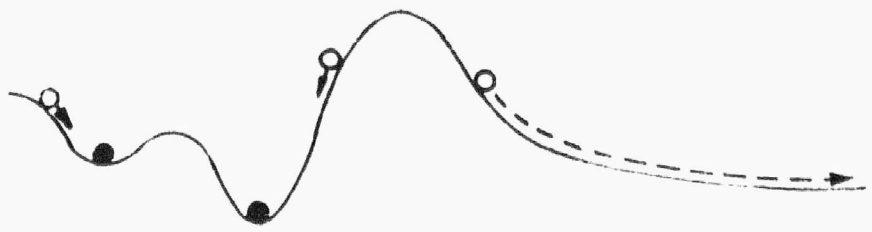


Fig. 9

a rather wild landscape, depending on the initial position of **w**0. In practice we however usually meet a milder climate.

4. Generalization and knowledge extraction

If we again draw an analogy from the animal world, we surely understand the importance of the so called *test set Q*. It is OK when our net has learned well all the pairs from the training set T. Given an inquiry (question) x^i, it answers with an output vector near y^i, provided that $[x^i, y^i] \in T$. It proves a good memorizing capability, but still may lack the ability to respond reasonably to inputs which have not been seen before, say to an input $x^* \in Q$ [Q is thus formed not by pairs, but by singletons from the input space], for which there is no y^* such that $[x^*, y^*] \in T$. In the the situation just described the net simply lacks the intelligence necessary to behave well in a not fully understood environment. It is not able to grasp "inner laws" hidden in examples from T; it is not able to *generalize* its memorized knowledge.

A simple and not too exact example is that of a student who carefully memorized some facts according to given exam questions (first coordinates of pairs from T), but still does not understand the theory behind all of them. The function φ_w mapping questions x^j of the examiner to correct answers Y^j (where $[x^j, y^j] \in T$) may look like that in *Fig. 10a*. If the lecturer confines himself just to the given list of questions, the student can still be graded as excellent. However some additional question x^* can cause him troubles. He is good on training set T, but poor on test set Q.

On the other hand the student from *Fig. 10b* has built a sort of comprehension: he generalized somehow the knowledge hidden in the examples from T and formed an inter- or extrapolation ability from what he learned. It is not possible to say from the figure whether his understanding is correct; actually there are many possible generalizations from T onto Q and a general (meta)theory of generalization has not been successfully developed yet. Yet in many cases, for many *NNs*, many adaptation algorithms and many applications we have good reasons to believe that the *NN* under consideration (together with the adaptation procedure) generalizes correctly. for some examples we are even pretty sure about it. Perhaps the reader will be convinced as well when these two rather abstract sections will be made more concrete later on.

The responses to questions from Q also heavily depend on a sort of *similarity* of T and Q. Sometimes we extend the set T to make it more close to the environment represented by Q (supplying correct answers for questions which, we admit, were too artificial w.r.t. the set T). On the other hand, if the net is exposed too many times to some members of T, it may become (similarly to people) *overtrained* and its generalization as well as other capabilities may decrease.

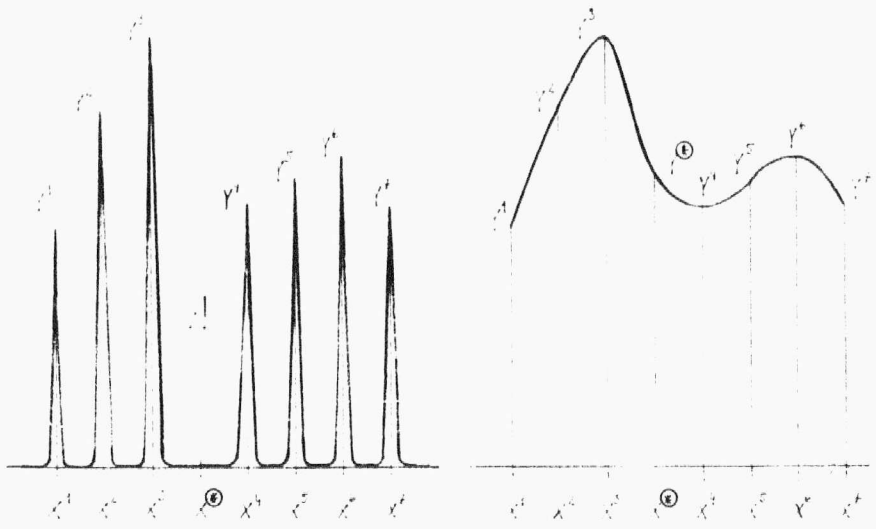


Fig. 10 Students, a, b.

Assume now a suitable net N with good adaptation abilities: *say* a *highly interconnected* multilayer *NN*. By the way, highly interconnected means that the net uses nontrivially (i. e. with weights not constantly 0) many of the possible connections, the amount of which is restricted only by the general topology of the net. Completely interconnected multilayer nets are discussed in the next section.

Now imagine a situation which has indeed been reported. A human-controlled plant works in such

a way that someone follows a few indicators and his job is to set up a few control mechanisms (knobs, gears etc; details are unimportant). The environment is so complicated and implicitly dependent on so many factors (like behavior of customers, their mood, market tendencies, weather etc. etc) that there is neither an analytical nor a stochastical model for relating input information (given by the indicators) to outputs (manipulating control devices), although all pertinent information can be expressed numerically. Yet the human performs after thirty years of service his job excellently. He proves to be able to combine some common reason with a sort of intuition given by the lifetime *experience*, that nobody dares to simulate his behavior by classical computers. However he is going to retire and the company should look for a replacement. They try expert systems, but the attempt fails: the person is not able (although willing) to state exactly the *rules* according to which he behaves.

In this situation a suitable (sufficiently big, highly interconnected, etc) net N may save the plant. For some time (perhaps few months), N is exposed to the same inputs as the man and obtains also copies of his reactions. It successively trains on the enormous training set of situations. From the beginning it itself behaves poorly, but as more training data is processed by an efficient adaptation process, its behavior improves. Finally it is able to compete successfully with the man, who was so long its teacher.

In the very many synaptical weights there appeared to be incorporated the *human knowledge*. Nobody is able to decipher the strange language of large weight matrices, yet the system works. Similarly like in many professions, only an apprentice observing his master for a long time can replace him. In *Fig. 11* we tried to illustrate this special sort o man-machine communication; after some time, the difference δ between a man and machine's responses diminishes.

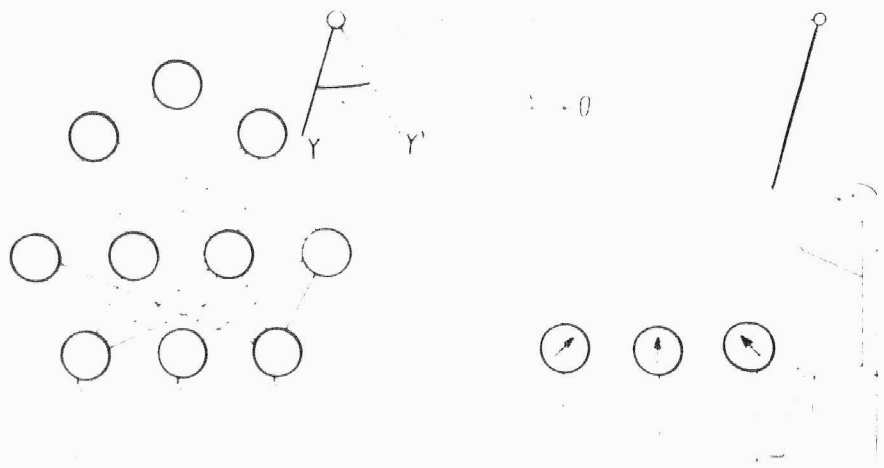


Fig. 11 A man and Neural Network.

There are other examples you can read about. Some guy trained a robot to play ping-pong. Surely there were many technicalities involved and *NNs* were only a part of the whole machine. But the difference between direct learning by observation and a rule (algorithm) obeying device can be again demonstrated. Suppose that you are a beginner in this game and you have a champion to direct you during the play, giving you real time instructions on what to do: now move your right hand forward about four inches, rotating it twenty degrees, lowering yourself ten inches at your knees, . . . Neither the champion would be able to express quickly what he almost *unconsciously* has in his head, nor would you be able to follow his instructions. None of our many languages is suited in many circumstances to express our knowledge completely and pass it to others; and yet we have it. One of the fascinating (although not yet well mastered) features of *NNs* is that they sometimes enable us to *extract knowledge* which is otherwise unaccessible.

Perhaps the time has come when a remark on the style of presentation would fit. As the reader many times has noted, we often use metaphors and parables within a broad range of fantasy: from a one-neuron chicken up to similes between training strategy of *NNs* and students. You can interpret these in a various ways. Because we believe that a description of formal aspects is to a great extent independent of them, you have several possibilities: a rigorous reader can take them as bad jokes or skip them. A more adaptive one can read them for a sort of amusement or mnemonic which helps to remember more formal tricks. Finally an enthusiast can believe that to the true (although oversimplified) isomorphism between structures in the brain and NNs may correspond a slight isomorphism of behavior. It depends on the specific topic, which interpretation is in a given case most suitable. generally it will be a weighted sum of all, but we leave it to the reader to fix the weights.

5. Complete multilayred nets

A *complete multilayered net* (*CMN*) is a multilayered net in which for any pair of adjacent layers, every neuron j in the layer below is connected to every neuron i in the layer above by the connection with a synaptical weight w_{ij}. Let us repeat explicitly that there are no interconnections between neurons in the same layer and the graph of the net is acyclic: if we conveniently order the layers with the input at the bottom of the figures, there are thus no arrows leading back from higher to lower layers. Therefore the activation is always spreading bottom-up and is *"feed-forward"*. In the *activation (working) mode* the information flows in one direction and reaches the output layer by layer in finite time. *Fig. 12a* depicts such a net with an m-dimensional *input layer*, k- and l-dimensional *hidden layers* and an n-dimensional *output layer*. We call it an *m-k-l-n* net. Additional auxiliary *fictive neurons* representing thresholds are shown only if there is some additional reason for it. Sometimes we use an abbreviated graphic picture form *Fig. 12b*. In many cases we restrict ourselves to nets with one k-dimensional hidden layer, an *m-k-n* net. A net 2-2-1 is shown in *Fig. 15*, where moreover some of the neurons are split to be able to speak separately about outputs (after the nonlinearity transfer function S) and net incomes ζ, cf. the last picture of *Fig. 4*.

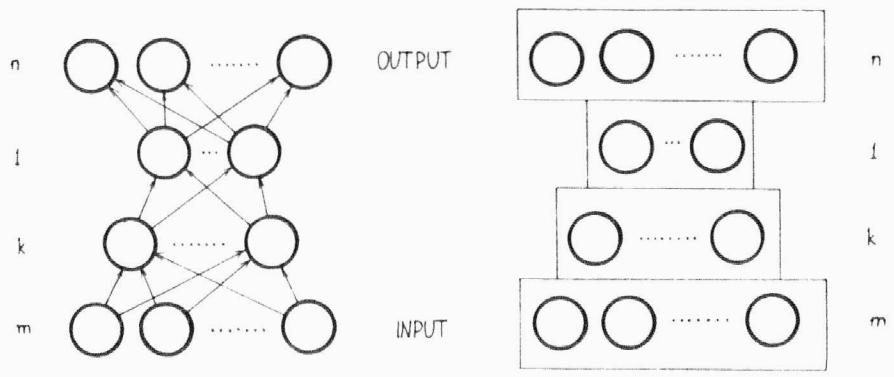


Fig. 12 Multilayred net with rightangles.

We see that in this sort of net there are again many interconnections and we can speak about connectionist systems. Due to so many interconnections in the majority of the nets in use practically, nothing tragic as a rule happens if a relatively smal portion of them become missing (e. g. by setting some of the w_{ij} constantly to 0); such damage occurs often in biological organisms. Although they are not able to recover their axons, if the damage (caused e. g. by an accident or a smal brain stroke — a *lesion*) is not too far reaching, the net is usually able (possibly after some retraining) to replace the damaged connections by circumventing them — from this fact, shared by artificial *NNs* as well, follows a sort of *robustness* or *graceful degradation*. Compare this property with classical computers where a one bit fault, in say a memory check,disables the device totally.

There is one more general advantage of highly interconnected nets. No matter that at present we often simulate their activity on serial machines, they are in fact *inherently parallel* (and more advanced implementations take this fact into account). This helps to explain their efficiency. The real brain works at a frequency of tens to hundreds of hertz, while standard silicon technology is about a billion times faster. The fact that living organisms are still superior in many "intelligent" activities [a two year old child recognizes immediately, among dozen of people, its mother, never seen before in exactly the same dress, hair, smile, etc., a task not yet satisfactorily solved by a supercomputer], stems from the great parallelism involved in the brain's work, where almost every neuron reacts at every moment of time; nobody is lazy, nobody waits. NNS actually also mimic this important feature.

In spite of a high degree of interconnections, informing a given neuron about the situation in a broad area of the net and perhaps the environment, its own actions (computations) are *local*: each neuron decides for itself what to do next. In the majority of cases in life and in most *NNs*, there is no central planning: everybody contributes to the system performance just by being responsible for himself/herself/itself. The *principle of locality* precludes e. g. such things like finding the neuron with maximal output — other than through information delivered directly by the net; no algorithm of finding a maximum (an exercise a beginner in programming), is applied (see section 9 for a neuronal solution to this task and for further comments).

To get a still more general model of *CMNs*, we will now generalize also the nonlinear transfer function *S*. What we need is first to obtain any real numbers (or proper subsets of reals) at the outputs of the neurons [specifically on the output of the net] and second, to be able to formulate more powerful adaptation algorithms. Actually the example of *Fig. 11* and the following one implicitly assume these possibilities.

Instead of binary signum *S*, the so called *hard nonlinearity* because it is not continuous at point 0, we will often take a *sigmoidal transfer function S* which is monotonically increasing and differentiable. the shape of sigmoid reminds one of a "flattened *S*", similarly to *Fig. 13*. Such functions have several plausible properties even from the neurophysiological point of view. When the *inner activation of potential* ζ of a neuron (delivered by its net income) is too high, it loses *discrimination*, while near zero, the sigmoid is almost linear, more sensitive to little changes of inner activation. This may simplify modeling properties like habituation; on the other hand in formal theory of *NNs* we should be aware that admitting too big weights (and thus too big net incomes) can cause problems with accuracy of computation and other problems. Therefore we usually try, wherever possible, to keep weights in reasonable ranges; one way how to achieve that is to perform *normalizations* (dividing weights and perhaps also input and output vectors by their lengths or taking other measures to reduce them systematically] in some stages of computations.

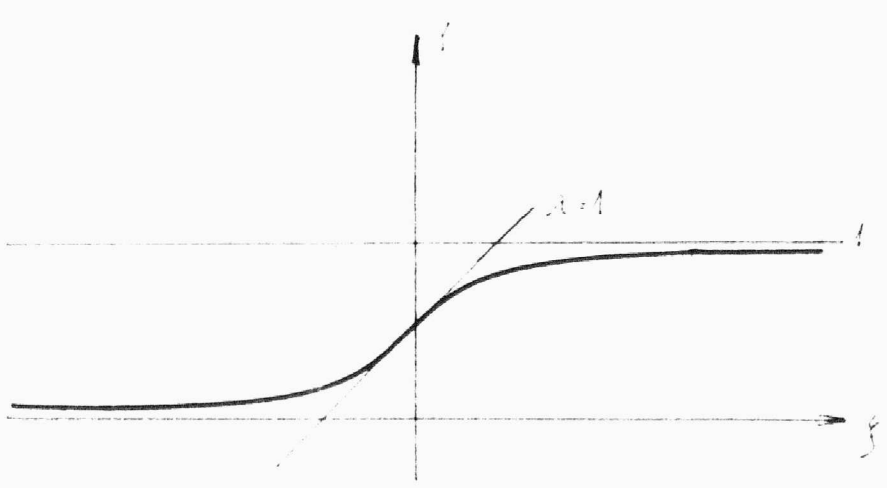


Fig. 13 Sigmoidal transfer function.

There are analytical formulae for sigmoidal *S*, having nice mathematical properties. Unless stated otherwise we shall use the *logistic function*

$$\xi = S(\eta) = (1/(1+exp(-\eta)) \qquad (7)$$

for which it holds

$$d\xi/d\eta = \xi \cdot (1-\xi)\,, \qquad (8)$$

as can be easily calculated.

Another possible formula is e. g.

$$S(\eta) = \tanh(\eta)\,, \qquad (9)$$

which has similar shape. Unlike (7) which ranges bet-

ween 0 and 1, (9) increases from -1 to 1 [both limits in both cases being reached only asymptotically, which has the consequence that the outputs, say in clkassification problems, will never be exact].

Between $X \in [-1, 1]$ and $x \in [0, 1]$, there are of course simple lineasr mappings:

$$X = 2x - 1; \quad x = (X + 1)/2 \tag{10}$$

These allow us to pass also between *binary* (boolean) values {0, 1} to *bipartite* values {−1, 1}, even that in some models there are slight behavioral differences between models using the two different codes.

Consider now a complete multilayered net in which all layers except the bottom one are split (cf. the last drawing of *Fig. 4*); potentials of neurons are denoted by Greek script, while the final output signals are denoted by Roman letters. In *Fig. 14* there is a *CMN* with one hidden layer *m-k-n*, where

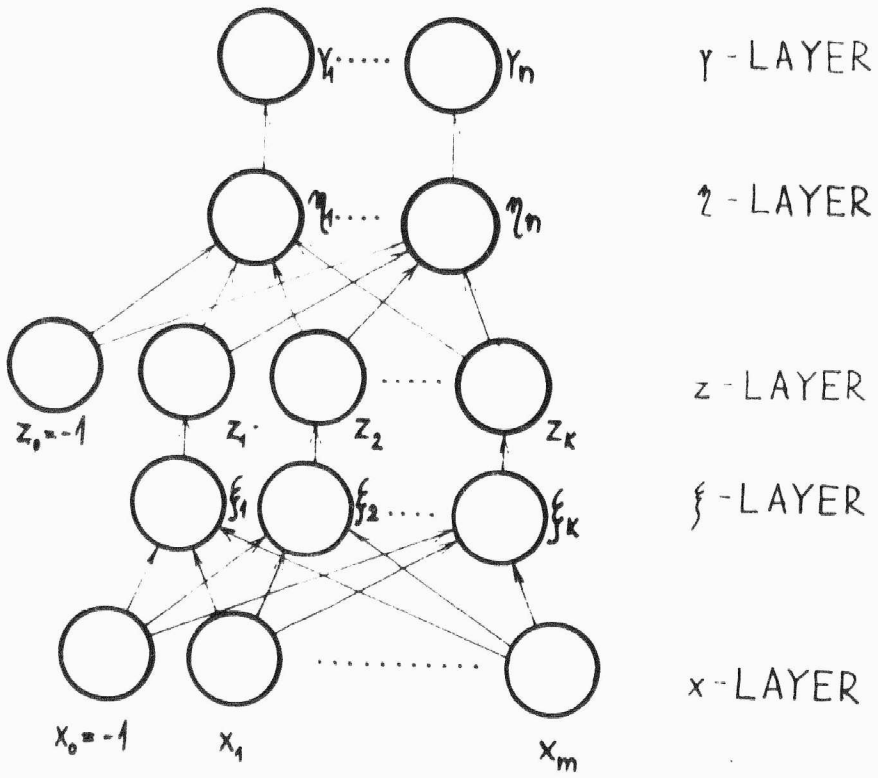


Fig. 14 {x-ξ-z-η-y}

$$\zeta_i = \sum_{j=0}^{m} w_{ij} x_j, \quad z_i = S(\zeta_i), \quad i = 1, 2, \ldots k \tag{11a}$$

$$\eta_i = \sum_{j=0}^{m} v_{ij} z_j, \quad Y_i = S(\eta_i), \quad i = 1, 2, \ldots n \tag{11b}$$

If we arrange the $k.(m + 1)$ weights w_{ij} into a matrix W and similarly $n.(k + 1)$ weights v_{ij} into V and $x = [x_0, x_1, \ldots, x_m]$, $\xi = [\xi_1, \ldots, \xi_k]$, $z = [z_0, z_1, \ldots, z_k]$, $\eta = [\eta_1, \ldots, \eta_n]$, $Y = [Y_1, \ldots, Y_n]$ ($x_0 = z_0 = -1 \; -> \; w_{i0}, z_{i0}$ represent thresholds), the linear parts of the mappings can be written in a matrix form

$$\xi^T = W \cdot x^T, \quad \eta^T = V \cdot z^T, \tag{12}$$

where superscripts T denote transposes. Note that the matrices express cross dependency of all neurons in adjacent layers, while the nonlinear parts of the mappings S are pointwise.

The whole mapping taking input vectors x to output vectors y is thus decomposed into interleaving linear and nonlinear components. The linear parts preserve convexity, parallelism in the case that an input pattern x possesses one and another affine properties, while the nonlinear parts are able to "distort" it converting convex patterns to concave ones, etc.

In *Fig. 15* we present an example of how such a successive mappings may look like when solving the *XOR* problem: first an input square [in coordinates x_1, x_2] is given. For suitable W it is first transformed into a "thin" parallelogram [coordinates ξ_1 ξ_2], which is next taken by the S-transformation into a concave shape [coordinates z_1, z_2] allowing one to place an appropriate separating line. This line forms the coordinate η and the distances of the four given points from it give the next picture, which is then somehow smoothed (due to the shape of (7)) to the final representation on the y output classifying coordinate line, placing one pair of the points near 0 and the other near 1.

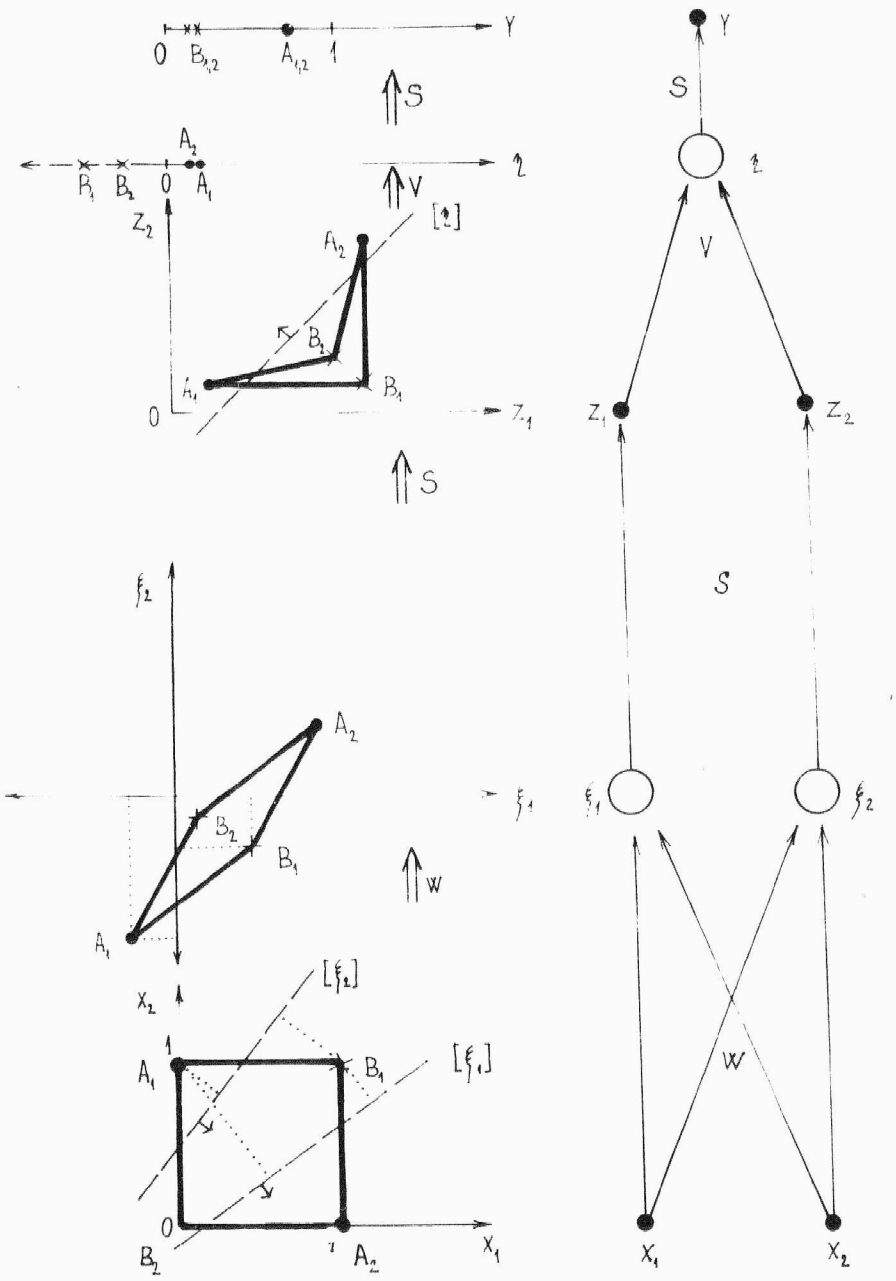


Fig. 15 A distored square.m

While the power of *CMN* relies on the two sorts of mappings, some properties can be deduced from the linear parts alone. Thus e. g. if the matrix W is not full rank (the set of equations (11a) is linearly dependent), several vectors of the input space map onto the same

point in the ζ layer giving thus the same final output. These inputs form a linear hyperspace (whose dimensionality depends on the difference between m and the rank of W) and the same "collapse" of the input space is observed in all parallel hyperspaces. In *Fig. 16* a four dimensional hypercube together with two lower layers of a 4—2—* *CMN* is shown (cf. *Fig. 6*). The equations (11a) for this case read

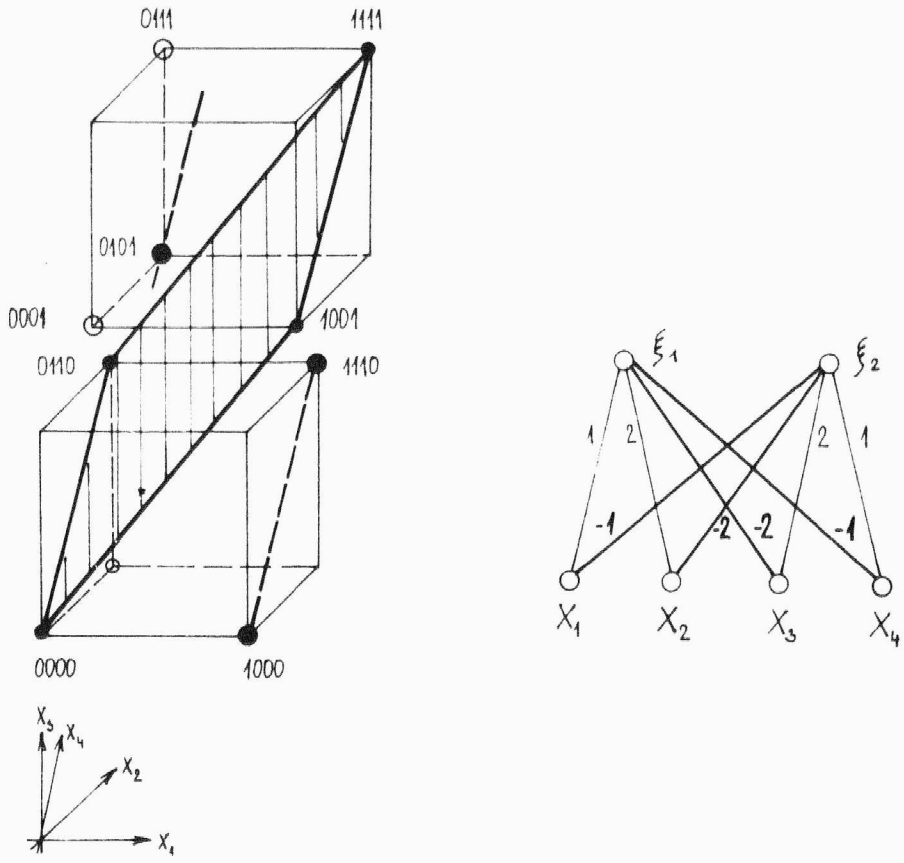


Fig. 16 4—cube.

$$\xi_1 = x_1 + 2x_2 - 2x_3 - x_4 + \vartheta \tag{13}$$

$$\xi_2 = -x_1 - 2x_2 + 2x_3 + x_4 - \vartheta$$

The equations are clearly dependent. If $\vartheta = 0$, the boolean vectors [0000], [0110], [1001], [1111], as well as all others lying in the same 2-dimensional plane, map to the same values of $[\xi_1, \xi_2]$, namely [0, 0]. So do the inputs from the plane given by [0101], [1110] and [1000], all mapped to [1, —1] (see dashed lines). This fact may help us to partially explain some properties of *CMN* to be discussed later on.

Some interesting things can be observed by considering the V-transformation between z-layer and η-layer. Each η_i specifies by the vector $v_i = [v_{i0}, v_{i1}, \ldots v_{ik}]$ a k-dimensional hyperplane in k-dimensional space, totally n in number ($1 \leq i \leq n$). Now you can count the number of possibilities, where the vector $z = [z_0, z_1, \ldots z_k]$, which came out of lower part of the net, is placed w. r. t. the convex subspaces dissected by the hyperplanes in the following sense. Relying on a theorem from combinatorial geometry (as proved e. g. in Edelsbrunner's monograph) there are (in the general case)

$$\sum_{i=0}^{d} \binom{n}{k-i} \binom{k-i}{d-i} \tag{14}$$

d-dimensional convex subspaces arising by the dissection of n k-dimensional hyperplanes. This formula may sometimes help to estimate the number of hidden layer neurons (one of the most difficult problems in the design of proper *CMN* realizing the wanted input/output mapping). In the simplest (classifier) task when we wish that outputs of all output neurons should tend to 0 or 1, all z's should be as far from the separating hyperplanes as possible, so that η_i's have the largest possible value [remember the note about the distances in section 2] and are thus far away from the 0 of nonlinear function S (tending to $\mp$ infinity); it follows that in such cases we can consider only full dimensional (i. e. k-dimensional) convex subspaces and set $d = k$. The formula (14) then simplifies to

$$\sum_{i=0}^{k} \binom{n}{i} \tag{15}$$

which shows how many different 0/1 (approximately) output vector responses we can maximally expect. For a net *—2—4 we get the number 11. See *Fig. 17* to see which separating lines actually arise and to see where the z points were placed (in a 2-dimensional plane) during some stage of the adaptation process (when we keep in mind the restriction to 11 successful possibilities). It is seen that to transfer over the net 4—2—4 all of the 16 vertices of a 4-dimensional cube is an impossible task. If you try to realize it, some of the 16 inputs necessarily lead to the same output. Moreover, which 11 out-of 16 will be successfully transmitted over the net, also depends on the orientation of the hyperplanes.

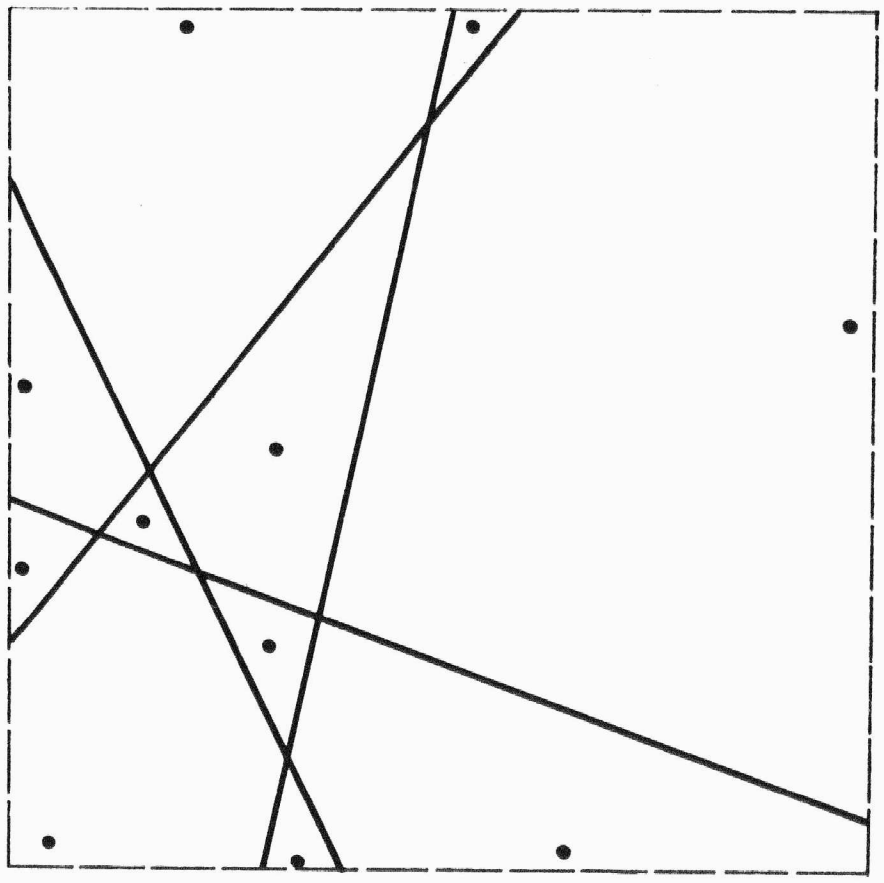


Fig. 17 The 11 aut—of 16.

There is another restriction deducible from simple linear algebra. If $k < n$, so that in (11b) you have more

restrictions (equalities) than variables (coordinates of vectors z), the vectors η are constrained to satisfy some linear dependency (and because S is at least formally one-to-one, output vectors y are not free, but constrained in some way as well).

To be consistent with our optimistic assertion from section 2 that three layers are able to implement any (boolean) function, it should be noted that there we were not too much interested in the number of hidden neurons. While dimensionality of input and output spaces (i. e. m and n) is fixed by the environment and the task, the number of hidden neurons has to be established and to keep their number optimal is not easy. Our final considerations show that if k is "too small", we have to be prepared for some unpleasant restrictive constraints on mappings which we would like to realize by *CMNs*.

If a *CMN* has all neurons equipped with the same nonlinear function S, it is called *homogeneous*. But experience shows that sometimes it is desirable to let the different neurons have different transfer functions. The simplest and most useful way to construct such an *inhomogenous* net is to introduce a parameter λ which governs the shape of S. He formula (7) then changes to

$$S_\lambda(\eta) = 1/(1 + \exp(-\lambda\eta)) \qquad (16)$$

In this case λ gives the slope of the sigmoid at point 0 and is sometimes called the *gain* of S; in *Fig. 13* it is $\lambda = 1$. If we choose $\lambda = 0$, we get just a straight line (constant 0.5), while $\lambda \to$ infinity derives from the S_λ the hard nonlinearity shown before. Every neuron can have its own λ and the mapping φ realized by the net (cf. *Fig. 18*) then depends not only on the weights (including thresholds), but also on all the λ's. Denoting their vector λ, we can then indicate this dependency by writing $\varphi_{w\lambda}$. As we will note in the next section, the individual gains can be also adapted automatically, sometimes bringing a better or faster solution to the adaptation process.

In 1957 Soviet mathematician Kolmogoroff proved a theorem that a many-valued continuous function

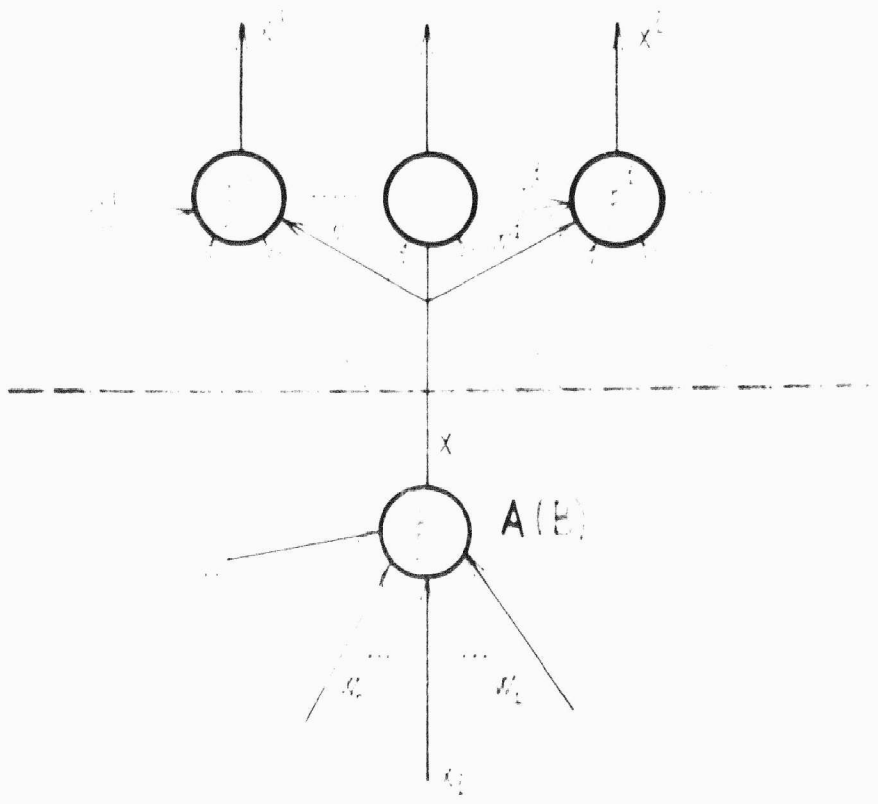


Fig. 18

can be expressed by suitable sums of single-valued functions, answering thus after more than half century one of Hilbert's famous problems. As noticed by Hecht-Nielsen and others, this theorem established in fact the possibility of realizing any continuous function φ by a three layered *NN*. Because, however, the theorem was not constructive enough to solve the problem, how to do it, it took some time before efficient adaptive algorithms for *CMNs* were presented; nevertheless the theorem (and its various improvements) encouraged the search for better understanding of the whole problem. one result of such a search is described in the next section.

To conclude our discussion of complete multilayered nets, enormous variability of the device should be emphasized. This consists not only in the huge number of parameters involved (one will extend their set when we shall discuss their adaptation algorithm), but mainly in the fact that in represents a complex mapping with several stages of linear and nonlinear (although continuous!) functions interleaved.

(Continuation)

Literature Survey

Grajski K. A., Merzenich M. M.: Hebb-Type Dynamics is Sufficient to Account for the Inverse Magnification Rule in Cortical Somatotopy
Neural Computation, Vol. 2, 1990 No. 1, pp. 71—84

Abstract: The inverse magnification rule in cortical somatotopy is the experimentally derived inverse relationship between cortical magnification (area of somatotopic map representing a unit area of skin surface) and receptive field size (area of restricted skin surface driving a cortical neuron). We show by computer simulation of a simple, multilayer model that Hebb-type synaptic modification subject to competitive constraints is sufficient to account for the inverse magnification rule.

Maas van der H. L. J., Verschure P. F. M. J., Molenaar P. C. M.: A Note on Chaotic Behavior in Simple Neural Networks
Neural Networks, Vol. 3, 1990, No. 1, pp. 119—122

Key words: autoassociator; backpropagation; Hebbian learning; periodic windows; bifurcation.

Abstract: Local dynamics in a neural network are described by a two-dimensional (backpropagation or Hebbian) map of network activation and coupling strenght. Adiabatic reduction leads to a nonlinear one-dimensional map of coupling strenght, suggesting the presence of a period-doubling route to chaos. It is shown that smooth variation of one of the parameters of the original map-learning rate gives rise to period-doubling bifurcations of total coupling strenght.

Instructions to authors

1. Manuscript
Two copies of the manuscript should be submitted to the Editor-in-Chief.

2. Copyright
Original papers (not published or not simultaneously submitted to another journal) will be reviewed. Copyright for published papers will be vested in the publisher.

3. Language
Manuscripts must be submitted in English

4. Text
Text (articles, notes, questions or replies) double space on one side of the sheet only, with a margin of at least 5 cm, (2") on the left. Any sheet must contain part or all of one article only. Good office duplication copies are acceptable. Titles of chapters and paragraphs should appear clearly distinguished from the text.
Complete text records on 5 1/4" floppy discs is prefered.

5. Equations
Mathematical equations inserted in the text must be clearly formulated in such a manner that there can be no possible doubt about meaning of the symbols employed.

6. Figures
The figures, if any, must be clearly numbered and their position in the text marked. They will be drawn in Indian ink on white paper or tracing paper, bearing in mind that they will be reduced to a width of either 7,5 or 15 (3 or 6") for printing. After scaling down, the normal lines ought to have a minimum thickness of 0,1 mm and maximum of 0,3 mm while lines for which emphasis is wanted can reach a maximum thickness of 0,5 mm. Labelling of the figures must be easy legible after reduction. It will be as far as possible placed across the width of the diagram from left to right. The height of the characters after scaling down must not be less than 1mm. Photographs for insertion in the text will be well defined and printed on glossy white paper, and will be scaled down for printing to a width of 7,5 to 15 cm (3 to 6"). All markings on photographs are covered by the same recommendations as for figures. It is recommended that authors of communications accompany each figure or photograph with a descriptive title giving sufficient information on the content of the picture.

7. Tables
Tables of characteristics or values inserted in the text or appended to the article must be prepared in a clear manner, preferably as Camera Ready text. Should a table need several pages these must be kept together by sticking or other appropriate means in such a way as to emphasize the unity of the table.

8. Summaries
A summary of 10 to 20 typed lines written by the author in the English will precede and introduce each article.

9. Required information
Provide title, authors, affiliation, data of dispatch and a 100 to 250 word abstract on a separate sheet. Provide a separate sheet with exact mailing address for correspondence

10. Reference
References must be listed alphabetically by the surname of the first author. List author(s) (with surname first), title, journal name, volume, year, pages for journal references, and author(s), title, city, publisher, and year for the book references. Examples for article and book respectively:

Dawes, R. M. and Corrigan, B.: Linear models in decision making, Psychological Bulletin, Vol. **81,** 1974, 95-106.

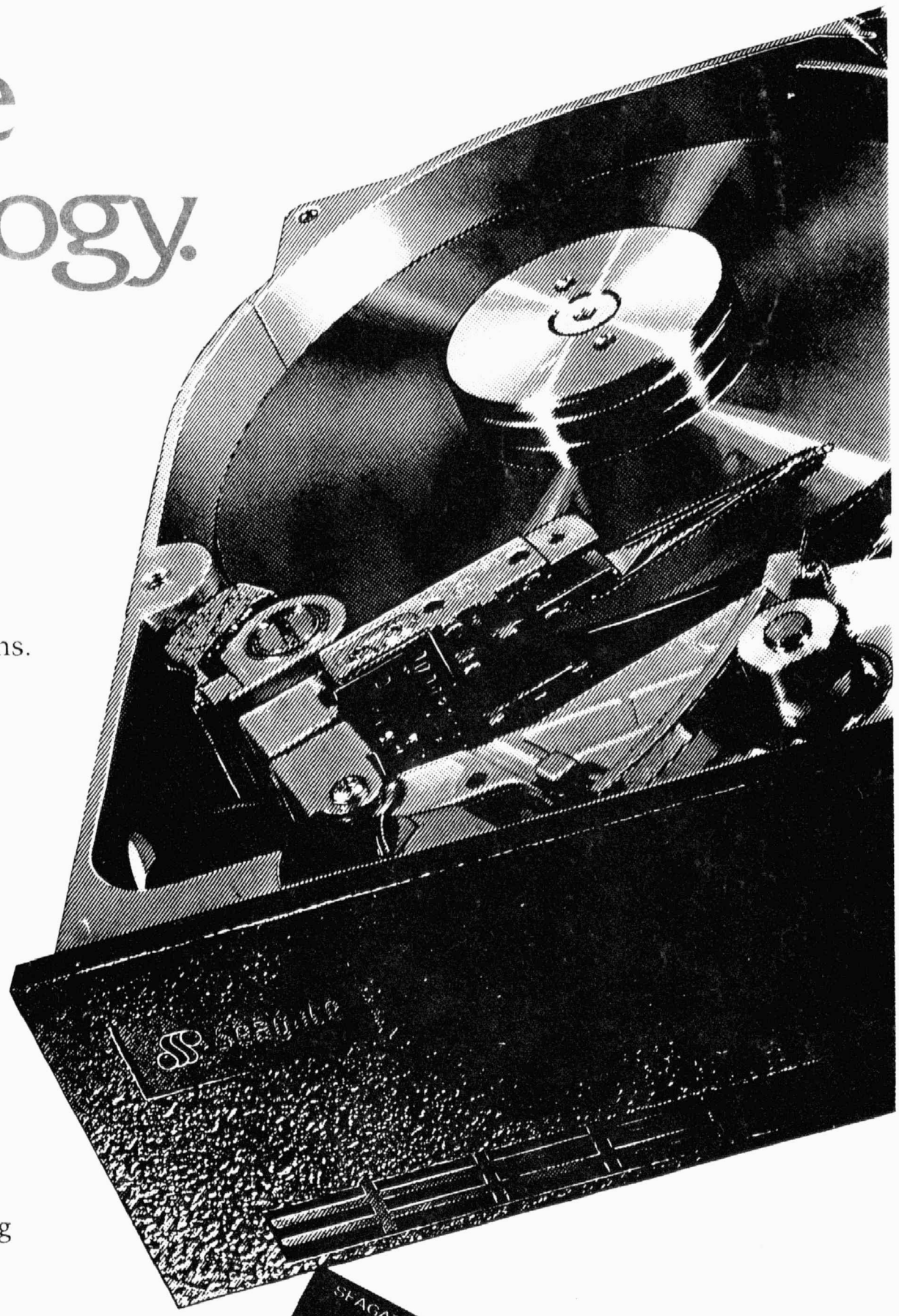
Brown, R. G.: Statistical Forecasting for Inventory Control, New York: McGraw-Hill, 1959.

All references should be indicated in the manuscript by the author's surname followed by the year of publication (e.g., Brown, 1959).

11. Reprints
Each author will receive 25 free reprints of his article.

PD 3818

This is Seagate Technology.



Seagate's line of hard disc drives is packed with high technology. And every one is built to the highest quality and reliability standards in the industry.

And now, Seagate drives are available locally for all your Personal Computer applications.

Only Seagate can offer you full technical support, and a one-year warranty . through our authorised representatives in your country.

Complete technical and interface details are included in the Seagate product brochures, which are free of charge to professional PC buyers and users. Simply use the coupon below to request your copies.

You'll soon see why Seagate has become the world's leading independent manufacturer of disc drives.

Seagate

Seagate Technology Europe
Seagate House, Fieldhouse Lane, Globe Park, Marlow SL7 1LW Great Britain.
Tel: 0628 890366 Fax: 0628 890660 Telex: 846218 SEAGAT G

To: Seagate Technology Europe,
Seagate House, Fieldhouse Lane,
Globe Park, Marlow SL7 1LW Great Britain.
Please send me technical details of Seagate disc drives
Name _____
Job Title _____
Organisation _____
Address _____
_____
Country _____
Type of business _____
Number of employees _____ Number of PCs _____
☐ I use a PC ☐ I authorise the purchase of PCs
☐ I am a technical support manager