STUDIES ON EMERGENCE AND COGNITION
PART 1: LOW-LEVEL FUNCTIONS

Heikki Hyötyniemi
Helsinki University of Technology
Control Engineering Laboratory
P.O. Box 5400, FIN-02015 HUT, Finland

It has been demonstrated how iteration of functions with very simple structure can result in every imaginable outcome. However, from the point of view of implementing some cognitive functions having real relevance this result has only academic interest. This paper discusses the challenges one is facing when trying to match the theory of complex systems against the realm of cognitive systems. This first part concentrates on the basic principles, and high-level functionalities are studied in [8].

1 INTRODUCTION

It should not be a surprise that the field of complex systems is itself a complex mixture of ideas and intuitions. There are different views of what kind of characteristics distinguish a complex system from "simple" ones. Now it is assumed that the key phenomenon taking place in a complex system is *emergence:* When simple functions are iterated massively, some kind of new order emerges. In this context, *cognitive systems* are concentrated on — in this special field, the emergent phenomenon is *intelligence* (see Fig. 1). It seems to offer new, fresh possibilities for attacking the mysteries of mind when intelligence (and other cognitive functions, even the taboo of *consciousness* itself?) are seen as emergent holistic phenomena — more concrete, more reductionist definitions can directly be implemented (and, indeed, they have been implemented, resulting in different kinds of "shallow AI" applications).

The approaches to studying complexity have traditionally been rather heuristic, being based on simulations and intuitively appealing patterns. But how to control these patterns, how to make something interesting emerge out from an iteration? Even though it has been claimed that mathematics has no role in the study of complex systems, it seems that mathematics is still the *best available language for discussing emergence.* This claim will be concentrated on in this paper, and in the follow-up paper [8].

2 WOLFRAM'S WORLD

The algorithmic forms, complicated, unforeseen patterns generated by the iterated functions can resemble natural forms. Is this a coincidence — or does Nature itself fundamentally function in this way? The history of artificial intelligence is full of fluctuations between hope and despair, and so is the history of complex systems research. How about combining these two paradigms? The turbulent vawes will form a rip-tide. A major splash

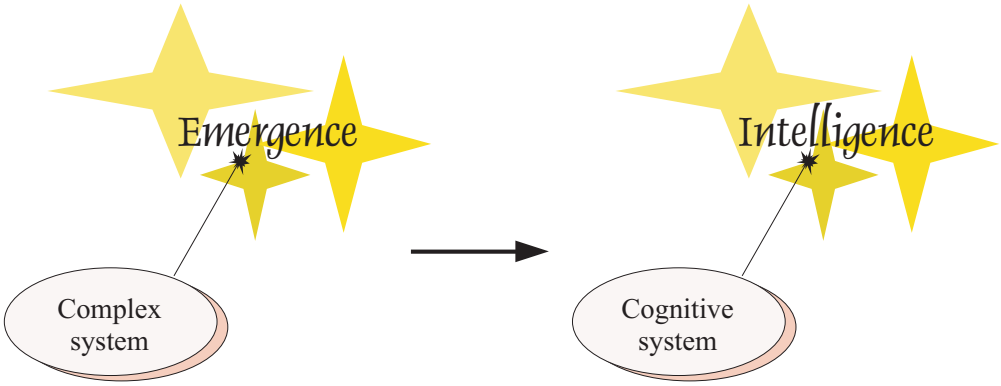

Figure 1: Emergent behavior is characteristic to complex systems in general ... and intelligent behavior is characteristic to cognitive systems in special

in the concoction is now produced by Stephen Wolfram and his book "A New Kind of Science" [11]. The Wolframian idea is that all natural processes are best explained in terms of *cellular automata*. He says that this scheme is appropriate in all kinds of systems, and at all levels of systems. The problem here is that this approach is incompatible with old theories – old mathematics, physics, biology, etc. (even ethics!) should be *forgotten* altogether. This is, indeed, what Stephen Wolfram claims: Everything must be started from the beginning! There are different reasons to criticize Wolfram's claims.

- First, from the point of view of scientific work in general, it seems that Wofram has entered the "fiddler's paradise": It is easy to put up new theories; the hard task is to put the theories fit with what there already exists — how to integrate the new ideas with old theories? After all, the evolution of theories has always been based on cumulation of ideas, how could this special field be so fundamentally different?

- From the point of view of complex systems research in special, there are also some issues that may turn out to become problems. It is as with the "Turing test" in artificial intelligence: This criterion states that something is intelligent if it is capable of behaving intelligently. This *shallow view* of AI has plagued this field ever since. Similarly, the "Wolfram test" for emergence and order is very behavioristic, not taking into account the internal underlying phenomena: Something is interesting if it *looks interesting*. This kind of "shallow view of complexity" may result in the complex systems research being stuck in admiration of superficial patterns.

- From the point of view of research methodology, the all too common "big hammer" syndrome makes everything look like a nail ... designing general-purpose tools should not be an end in itself — it is still nature one wants to explain. The approaches and tools have to be selected so that they match the phenomenon being studied, and the level of abstraction should be appropriate, hiding irrelevant details[1]. The general-purpose tools often result in clumsier models than more

[1] After all, Wolfram has a point here — it would be a nice bonus if a single framework could span the system properties on different levels. And, when studying cognitive processes, it is information processing all the way from bottom to top ... indeed, also in this paper, some kind of "unified theory" is searched for where functionalities on different mental levels could be collapsed into the same general framework

special-purpose modeling methods do. Another point is that Wolfram happily ignores deeper expertise in special fields to make his theories match the application; this approach is easy to understand because devil lives in details — but without details the discussions remain on the level of hand-vawing.

- Also from the point of view of relevance there are also problems: After all, it was noticed already for a long time ago that all material consists of elementary particles, all living things consist of cells. Neither from the point of view of cellular automata, the results are not so revolutionary or new. So what is really new here? Is there some new understanding or are there new conceptual tools? It is possible to reduce all phenomena to the elementary level, but it is not a reasonable level of abstraction when explaining complex things. Rather than escaping back to reductionism, *emergent phenomena should be attacked "from above".*

Still, there is potential in the idea of emergence — some things cannot naturally be defined in any other framework. For example, it seems that the best way to define such holistic phenomena as *intelligence* is through emergence: More concrete definitions have already been implemented as software, but the essence of intelligence still seems to escape.

3 MODELING VIEW

3.1 About models

When modeling cognition, one first has to admit that *models are always false.* Model is an abstraction where some details are more or less consciously ignored. All details cannot be captured, but if the model reveals something essential, giving insight into the phenomenon, then it is a *good model* — that is all one can hope for.

Still, let us not be too humble: *Assuming* that the brain itself is just an information processing device, trying to analyze the surrounding world, it is implicitly solving the same modeling problem as we are trying to solve explicitly — and finding the principles of mental processing, one could construct mechanical devices carrying out the essentially identical tasks of information processing (see [4]). Even though the physical realm cannot be captured in the models, phenomena that are interesting from the cognitive point of view can directly be attacked on the more abstract level of information processing and representation. It has been ironically commented that "simulation of a hurricane does not make you wet" — meaning that something essential is lost if one tries to mimic cognitive phenomena in an artificial substrate outside the brain. But one is not studying hurricanes now: *Simulation of information processing is still information processing!*

Good model is a compromise between different objectives: To serve the user of the model, *analyzability* and *understandability* are essential; to serve the system to be modeled, sufficient *expressional power* is necessary, and to serve applications, *practical applicability* has to be considered. These viewpoints will briefly be studied below.

3.2 Analyzability

Mathematics is the only available analysis tool now. And mathematical analyzability actually means *linearity.* In a multivariate environment, a linear function can be expressed in the form

$$y = A \cdot x, \tag{1}$$

where x and y are n_x and n_y dimensional vectors, respectively, and A is a compatible real-valued matrix. The mathematical benefits of this starting point are illustrated in Sec. 6; from the physiological point of view, the above linear formulation can be seen as an approximation of the operation of a real grid of neurons. Applying this function, the (unnormalized) correlations between the "signals" (vector x) and "synaptic weights" (rows of A) are calculated; as shown by Donald O. Hebb, this kind of correlations are the underlying essence beneath the neuronal behavior and adaptation. Hypothetically, it can even be assumed that nature tries to be linear, but the devices and mechanisms that are available are hopelessly nonideal and nonlinear[2]. Or, it can be claimed that the world, as we see it, simply *must be* (piecewise) linear[3].

Properties of (finite-dimensional) linear mappings are exactly known and possible behaviors of dynamic linear systems are well understood. The nice feature about linear structures is that they are scalable, and qualitatively the same phenomena take place also in high dimensions. On the other hand, this is also a drawback: New functionalities cannot emerge if linear structures are combined:

$$\begin{aligned} y &= A_1 \cdot \quad \cdots \quad \cdot A_n \cdot x \\ &= (A_1 \cdots A_n) \cdot x \\ &= A \cdot x, \end{aligned} \tag{2}$$

that is, no matter how many linear structures are combined and how many variables there are in the hidden layers, the functional complexity can still be expressed as a single matrix. To reach the capacity of something unexpected emerging, some kind of nonlinearity is necessary.

3.3 Expressional power

Here, when choosing the approach to extend the linear structure, one is facing major challenges: What kind of nonlinearity would be powerful enough, still "gentle", so that not all benefits offered by the linearity would be lost — and, perhaps, what kind of nonlinearity could also be motivated from the physiological point of view?

It turns out that *positivity* constraint is a nice compromise. In a positive system, variable values never become negative. When looking at a neural system, the positivity constraint seems to be well justified:

- If the neural activity is studied on the level of chemistry, so that different kinds of chemical concentrations explain the neural phenomena, it can be noticed that *concentrations never become negative.*
- If the neural activity is studied on the level of signal processing, so that pulse coded signals transfer the information, it can be noticed that *pulse frequencies never become negative.*

[2]If the brain *is* an analog computer rather than digital, the experiences from analog electronics can perhaps be applied: The nonideal, nonlinear amplifiers (transistors) are routinely used to implement linear functions; it is all dependent of which part of the characteristic curve is utilized

[3]If the brain essentially is a *tabula rasa* to start with, it is the properties of the surrounding world that have to be revealed by the brain to construct appropriate representations. This means that the brain can only recognize analyzable phenomena! No matter what the "real world" is like, our subjective world, perhaps being only a tiny part of it all, has to be composed of linear constructs

Figure 2: Remember the power of *Positive Thinking!* Positive systems theory just *might* help in understanding the underlying functions beyond thinking processes

How about higher cognitive levels? It can be claimed that also the "negative thoughts", even though they may be sinister, still having positive activity ... and remember the intuitive appeal of positivity (see Fig. 2)!

In this context[4], the positivity constraint is implemented as the "generalized cut" function:

$$f_{\text{cut},i}(x) = \begin{cases} x_i, & \text{if } x_i > 0, \text{ and} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Function f_{cut} is vector valued; each element i, where $i = 1, \ldots, \dim(x)$, is calculated independently, simply zeroing the entry if it is negative. If the variables in x remain positive, this function is transparent, and the nonlinear nature does not at all pop up. When studying the information representations, this kind of nonlinearity gives a practical way to implement *sparse coding.*

3.4 Pragmatic issues

It seems that the "Turing power" of universality is a rather common capability among nonlinear function structures — and, indeed, also the above positivity restriction results in such omnipotent power: Any algorithm can be implemented in that framework (for example, see [7]). Even though this sounds like a strength, it is a weakness: it makes the systems unanalyzable. And when modeling phenomena that defy explicit definitions, one cannot utilize the power anyway.

Another point is that even though something is *possible* it is not necessarily cognitively *plausible.* Implementations of some functionality using such general-purpose framework tend to become high-dimensional: When algorithms are implemented in the framework of [3], the structural complexity of the algorithms is not avoided, it is only transformed into *dimensional complexity.* The number of the free variables needed to represent the algorithms in this kind of matrix form is typically high. There are many variables that do not have any straightforward interpretation in terms of any observable quantities. From the point of view of learning from data, for example — and assuming that there do not exist (too many) hardwired, preprogrammed constructs in the brain, this kind of observation-orientedness is crucial — these "latent variables" result in extreme complexities. If the input and output were known, the parameters within a fixed structural framework could be more or less easily optimized; on the other hand, if there exist various

[4]Because of their physical relevance, *positive systems* have been studied a lot — however, the studies are often limited to positive *linear* systems, where special constraints to system parameters are imposed to keep the variables of a strictly linear system always positive. It is easy to understand why this kind of limitations are used — as was shown in [3], even a "simple" nonlinearity like the presented "cut" function results in unanalyzable behaviors

layers with either the input or output (or both) being unknown, iteratively searching for the appropriate parameter values soon becomes an untractable problem.

In this paper, and specially in [8], more "optimized" representations are presented, so that cognitively relevant phenomena can be implemented in a maximally compressed form with *no extra variables,* so that the inner representation has (in somewhat loose terms) *maximum expressional power but minimum complexity.* Each data structure can be explicitly interpreted in terms that are natural in the domain area, so that no latent variables remain. It turns out that this kind of chrystallization increases the transparency, and perhaps also the cognitive plausibility of the models. For example, one of the main objections against connectionist approaches can be attacked: When the underlying variables are conceptual enough, it turns out that all data structures need not be learned solely based on the observation data; the learning can also be based on explicit rules or examples, and declarative knowledge can be integrated in the same basically data-oriented model structure.

4 FRAMEWORK FOR COMPLEX SYSTEMS

4.1 Synthesis

The ideas of linearity and positivity are now integrated: The function form that will be discussed from now on is

$$f_{\text{cut}}(Ax). \tag{4}$$

In what follows, it will be assumed that *this is the only operation that can be carried out by the available machinery,* and all functionalities somehow *emerge* from this. To make this possible, this kind of simple functions have to be chained:

$$x(k+1) = f_{\text{cut}}(Ax(k)), \tag{5}$$

starting from some initial vector $x(0)$ and continuing *ad infinitum,* hoping that complicated behaviors emerge when the iteration goes on.

Essentially, (5) is a linear mapping of the vector $x(k)$, the nonlinearity only nullifying negative values. Regardless of the seemingly simple system structure, it is extremely difficult to say what is the faith of the iteration without actually running the process. When the mapping is iterated indefinitely, the process (hopefully) finally converges to some fixed point $\bar{x}$ so that there holds $\bar{x} = f_{\text{cut}}(A\bar{x})$. This vector $\bar{x}$ this vector is the emergent pattern.

4.2 Note on the formalism

In what follows it turns out that the iterations are often easiest to present in the form

$$x(k+1) = f_{\text{cut}}(Ax(k) + a), \tag{6}$$

where a is a constant vector dimensionally compatible with x. The constant term in the above formulation introduces no additional expressional power; it only helps to restructure the simple starting point. If one defines another vector x_{new} of dimension $2n$, obeying the following dynamics

$$\begin{aligned} x_{\text{new}}(k+1) &= f_{\text{cut}}\left(\left(\begin{array}{c|c} A & \operatorname{diag}\operatorname{sign}(a) \\ \hline \mathbf{0} & I \end{array}\right) \cdot x_{\text{new}}(k)\right) \\ &= f_{\text{cut}}(A_{\text{new}} x_{\text{new}}(k)) \end{aligned} \tag{7}$$

with

$$x_{\text{new}}(0) = \begin{pmatrix} x(0) \\ \hline |a| \end{pmatrix}, \tag{8}$$

exactly the same behavior can be observed as in (6) — the constants in a are just stored in the augmented state vector.

This kind of state augmentation makes it easy to include also inputs y from the outside (the observations) in the presented framework (see [4]). Indeed, yet another formulation is possible — in all applications that will be studied, it would also be possible to have the same dynamics by explicitly representing the external input y:

$$x(k+1) = f_{\text{cut}}(Ax(k) + By). \tag{9}$$

Further, this can be extended to a formulation that resembles a *state-space model* commonly applied in systems theory:

$$\begin{array}{rcl} x(k+1,t) & = & f_{\text{cut}}(Ax(k,t) + By(t)) \\ \hat{y}(t) & = & f_{\text{cut}}(C\bar{x}(t)). \end{array} \tag{10}$$

Note that there are some peculiarities, though: First, now the input is called y rather than u, and the output is the estimate for y, or $\hat{y}$. Second, note that y is static, it does not change during iteration, and $\hat{y}$ is valid only after the system has converged to some fixed state $\bar{x}$; for completeness, the input sample index t is also included in the model.

Because of the cut function, the coordinates will always be strictly positive if applying iterations with the cut function. This means that the fixed point solution is searched for only in the first (hyper)quadrant. The behavior of the strictly linear process $x(k+1) = Ax(k) + a$ can be simulated by the nonlinear structure (14)) using the following model (note that either x_i^+ or x_i^- has to be zero at any time):

$$\begin{pmatrix} x^+(k+1) \\ \hline x^-(k+1) \end{pmatrix} = f_{\text{cut}} \left(\left(\begin{array}{c|c} A & -A \\ \hline -A & A \end{array} \right) \cdot \begin{pmatrix} x^+(k) \\ \hline x^-(k) \end{pmatrix} + \begin{pmatrix} a \\ \hline -a \end{pmatrix} \right). \tag{11}$$

The state vector — possibly containing negative values — can then be reconstructed as

$$x(k) = x^+(k) - x^-(k). \tag{12}$$

5 MODELING OF COGNITION

5.1 About semantics

Running the iteration (5), something perhaps pops up. However, it is not some obscure fractals that would now make us happy, but it is some (cognitively) relevant functionalities that should emerge. How could something new and interesting come out automatically from such brainless formal activity? Real intelligence involves behaving in a reasonable way in a new environment without guidance; some *understanding* of the *meaning* of different entities in the environment is necessary to accomplish such task successfully. Understanding of meaning is deeply connected to *semantics* — all these are very difficult and loaded concepts, and some "engineering-like" simplification is necessary.

In formal manipulations, syntax remains syntax if no semantics is somehow involved in the structures being manipulated. It is now assumed that it is *naturalistic* and *contextual*

semantics that will be concentrated on: Connections to measurements from outside world or connections to other processing elements determine the meaning of a signal (see [4]). This makes it possible that something meaningful (when "meaning" is defined in such a narrow sense) can emerge from associative manipulation of correlation structures.

5.2 Observation vs. perception

As presented in [5], perhaps the most fundamental cognitivist concept to be studied is that of a *mental image* or *perception* (these two concepts being daringly identified here[5]). It turns out that many cognitive functionalities can be formulated in this extremely powerful conceptual framework. In what follows, this "inner image" will be denoted as x, and the *observation* inspiring this image is y. On the lowest level, these "observations" are direct sensations coming from the senses/sensors, but on the higher levels, these observations can also be some kind of combinations of lower-level perceptions. These simple data structures, real-valued vectors x and y, are now assumed to stand for the very abstract mental representations, whatever is their contents.

Assume that the above function calculation (5) is all that can be accomplished in the cognitive framework. There are two main tasks that need to be carried out (see Fig. 3):

1. **Reconstruction of the observation.** Assuming that x is the inner image corresponding to an observation y, according to the assumptions in [4], etc., ideally the reconstruction should be calculated simply as $y = Cx$. Here C contains the features, or "numeric chunks" (see [2]) as collected into a matrix, and x contains the weight for each of them, that is, variable x_i indicates how relevant the chunk C_i is when explaining y. If one restricts to positive values, this can be readily approximated in the proposed framework:

$$\hat{y} = f_{\text{cut}}(Cx). \tag{13}$$

2. **Determination of the inner image.** As compared to the previous task, this is much more complicated. In fact, the objective is to somehow invert the mapping from x to y in (13), so that best possible x could be found when y is given. Because of the nonlinearity (and because of the unmatching dimensions) this inversion is by no means trivial. Having only the function form (5) available, the appropriate x best explaining y must be determined by some iteration; starting from some initial observation-dependent vector $x(0) = g(y)$, where $g(\cdot)$ is *some* function of y, and using *some* appropriate matrix A, the following is repeated until the process (hopefully) converges:

$$x(k+1) = f_{\text{cut}}(Ax(k)). \tag{14}$$

If the above iterative scheme works, the inner image $\bar{x}$ where the iteration converges can be regarded as an emergent phenomenon. Note that because of the function form in (14) there always exists the trivial solution $\bar{x} = \mathbf{0}$; to have non-degenerate solutions emerge,

[5]Indeed, every cognitive concept used here should be preceded by the word "artificial" — that is, one is studying some kind of *artificial cognition*. Sticking to the strictly philosophical/cognitive concepts would limit the freedom that is now necessary

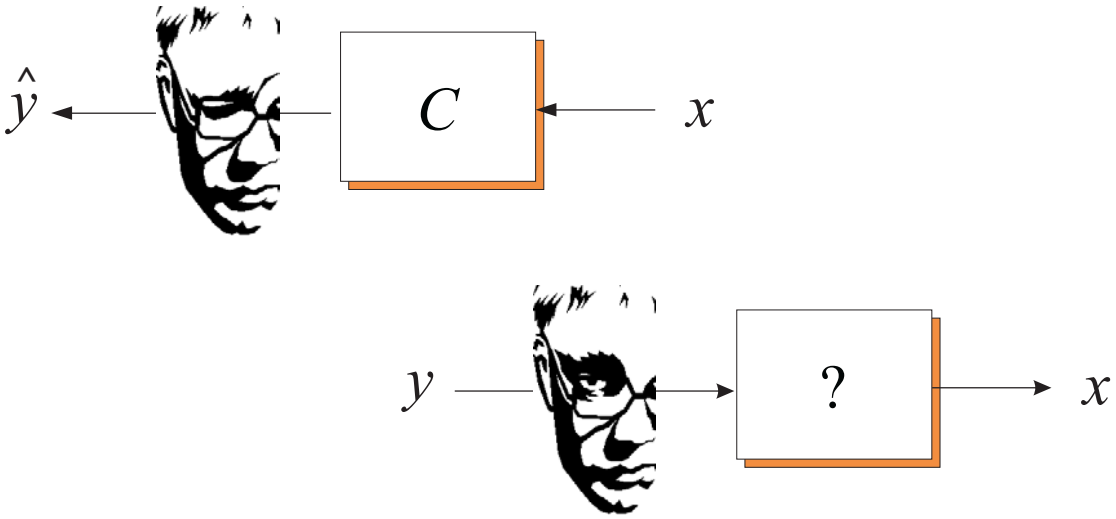

Figure 3: According to hypotheses, it is assumed that given the "inner image", reconstruction of an estimate of the "outer image" is simple; the difficult task is *modeling* of the environment, abstracting and compressing the incoming information into the mental image

the construction of A and $x(0)$ has to be studied closely — and, indeed, this is the main objective in what follows.

Comparing the proposed inner image determination structure to that presented in [4], it can be seen that some evolution has taken place. In the earlier formulation, only a fixed number of substructures could be active at any time; now all non-zero correlations are active. This makes the matching process simpler and better parallelizable — but, on the other hand, the cognitivist short-term memory constraints cannot be put into practice so efficiently. Another difference is, of course, the nonlinearity f_{cut}.

5.3 Example: Forward chaining

First, study a simple AI application in the proposed framework — implementation of a *forward-chaining rule system.* To realize the logical reasoning machinery in the numerical form (14), let us assume that value 1 means "true" and 0 means "false". Then the negation of a proposition P, denoted ¬P, can be implemented as

$$x_{\neg \mathrm{P}} = f_{\text{cut}}(1 - x_{\mathrm{P}}), \tag{15}$$

where x_{P} denotes the logical value of the proposition P. Further, conjunction of propositions, P_1 AND ... AND P_n, can be implemented as

$$x_{\mathrm{P}_1 \wedge \ldots \wedge \mathrm{P}_n} = f_{\text{cut}}(x_{\mathrm{P}_1} + \cdots + x_{\mathrm{P}_n} - n + 1). \tag{16}$$

Using *de Morgan rules,* all logical functions can be constructed from these. Assume that the knowledge has been coded in the rule form as IF P_1 AND ... AND P_n THEN P_{new}, meaning that the new proposition holds only if a set of other propositions all hold; this kind of rules can be implemented as

$$x_{\mathrm{P}_{\text{new}}} = f_{\text{cut}}(x_{\mathrm{P}_1} + \cdots + x_{\mathrm{P}_n} - n + 1). \tag{17}$$

For example, if there is just one rule IF P_1 AND P_2 THEN P_3, and the truth values of P_1 and P_2 are assumed to be known, the corresponding "reasoning system" becomes

$$x(k+1) = f_{\mathrm{cut}}\left(\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \cdot x(k) + \left(\begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} + \begin{pmatrix} x_{\mathrm{P}_1} \\ x_{\mathrm{P}_2} \\ 0 \end{pmatrix}\right)\right) \tag{18}$$

with

$$x = \begin{pmatrix} x_{\mathrm{P}_1} \\ x_{\mathrm{P}_2} \\ x_{\mathrm{P}_3} \end{pmatrix}. \tag{19}$$

The initial state $x(0)$ can be arbitrary. The solution may become more complicated, for example, if there are various disjunctive rules for some P_i; it may be necessary to include additional variables in the model to limit the maximum values to 1. However, it turns out that all logic constructs that are needed to implement the rule system can be written in the form (14). Collecting all x_{P_i}, $x_{\neg\mathrm{P}_i}$ (only those that are needed) in the vector x, A and a can be constructed so that iteration carries out forward reasoning. Vector a contains the observed "world state", and A contains the inference rules as shown above.

In a rule system with no cyclic structures, matrix A will be triangular with zero diagonal. After a finite number of steps (the forward rule matrix determining a so called dead-beat system), no more changes take place in the state vector x. It needs to be noted that the "straight-forward chaining" can be extended in this framework: Allowing non-binary weights, and fuzzy logical truth values, the resolution results can also be recirculated in the system. This means that A no more needs to be triangular with zero diagonal; finding a stable result becomes an infinite process that hopefully converges.

6 WHY NOT FORGET MATHEMATICS?

6.1 Connections to linear algebra

It seems that some people are specially relieved about Stephen Wolfram's declaration that "old mathematics is dead". It is like in the field of control theory — the new methods that promise that "no understanding is necessary" have become popular even though the old methods would often be unbeatable. There is perhaps a need to briefly discuss the issue why abandoning mathematics should be avoided.

The first reason to stick to mathematics is that it simply cannot be avoided: Mathematics is formalized logic, and if something is expressed exactly, it *is* mathematics. Mathematics is a language — actually, it is the natural language of Nature — and it seems to offer good abstractions and ready-to-use concepts for applications.

Mathematics is also a tool for getting *insight*. A glimpse of what this means will be provided in this section. For this purpose, let us study the linear version of (10):

$$x(k+1,t) = Ax(k,t) + By(t). \tag{20}$$

One does not need to iterate the function (20) to know how it will behave. First, the *eigenvalues* of A that can readily be calculated, dictate the stability properties of the system: If all eigenvalues are inside a unit circle in the complex plain, the system will be stable, no matter what the input y is. It turns out that what is actually solved by

the iteration — if it converges — is just another way of implementing a matrix inversion problem:

$$\bar{x}(t) = (I - A)^{-1} B y(t). \tag{21}$$

Further, note that if $y = \mathbf{0}$, the direction of x in the observation space will generally turn towards the direction of the most significant eigenvector (if the eigenvalue is inside the unit circle, the length will exponentially decay towards zero, though).

Mathematics can also offer analogies and (more or less) well motivated hypotheses can be based on such intuitions. Assume that the parameters (synaptic weights) in the data structures have been determined according to the *Hebbian law,* so that the parameters connecting signals are determined by the long-term correlations between the corresponding (converged) signals. This means that

$$A = \begin{pmatrix} \mathrm{E}\{\bar{x}_1(t)\bar{x}_1(t)\} & \cdots & \mathrm{E}\{\bar{x}_1(t)\bar{x}_{n_x}(t)\} \\ \vdots & \ddots & \vdots \\ \mathrm{E}\{\bar{x}_{n_x}\bar{x}_1(t)\} & \cdots & \mathrm{E}\{\bar{x}_{n_x}\bar{x}_{n_x}(t)\} \end{pmatrix}, \tag{22}$$

or $A = \mathrm{E}\{\bar{x}(t)\bar{x}^T(t)\}$, and, correspondingly, $B = \mathrm{E}\{\bar{x}(t)y^T(t)\}$. If these are substituted in (21), one has

$$\bar{x}(t) = \Big(I - \mathrm{E}\{\bar{x}(t)\bar{x}^T(t)\}\Big)^{-1} \mathrm{E}\{\bar{x}(t)y^T(t)\} \cdot y(t). \tag{23}$$

Now, study the structure of the general *multilinear regression model* (for example, see [6]) solving for the least-squares estimate for $\bar{x}(t)$ when $y(t)$ is known:

$$\hat{\bar{x}}(t) = \mathrm{E}\{\bar{x}(t)y^T(t)\} \Big(\mathrm{E}\{y(t)y^T(t)\}\Big)^{-1} \cdot y(t). \tag{24}$$

Note that as y typically has huge dimension, the covariance matrix would be non-invertible in practice. However, this strange analogy between (24) and (23) persuades us to make a slight modification in the model structure:

$$z(k+1,t) = z(k,t) - M \cdot \mathrm{E}\{\bar{z}(t)\bar{z}^T(t)\} \cdot z(k,t) + M \cdot \mathrm{E}\{\bar{z}(t)y^T(t)\} \cdot y(t), \tag{25}$$

where z has been introduced instead of x, and M is an arbitrary invertible matrix. This means that there now holds

$$\bar{z}(t) = \Big(\mathrm{E}\{\bar{z}(t)\bar{z}^T(t)\}\Big)^{-1} \mathrm{E}\{\bar{z}(t)y^T(t)\} \cdot y(t). \tag{26}$$

Comparing this to (24), one can see a remarkable difference: The inverted matrix operates in the (low-dimensional) space of $\bar{x}$ rather than in the (high-dimensional) space of y. In least-squares regression, the role of this operation is to compensate for the covariance structure in y before projecting the data onto $\bar{x}$; now, on the other hand, the "postprocessing" makes the elements of $\bar{z}$ less correlated — and, indeed, looking closer at (25), it is evident that it implements combined Hebbian – Anti-Hebbian learning: Strongly correlating elements in $\bar{z}$ try to inhibit each other. It is a well-known fact that Anti-Hebbian learning (at least when augmented with some appropriate nonlinearity) tries to implement *sparse coding* (see [1]).

Looking closer at (25), one can see that there is some freedom still available: Matrix M can be selected, for example, as follows:

$$M = \begin{pmatrix} \frac{1}{\mathrm{E}\{\bar{x}_1(t)\bar{x}_1(t)\}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\mathrm{E}\{\bar{x}_{n_x}\bar{x}_{n_x}(t)\}} \end{pmatrix}. \tag{27}$$

This selection means that the iteration (25) always remains stable; what is more, the eigenvalues of the process all are located in the origin, so that the iteration will reach its final value after n_x steps. To make the process more cautious, M can still be multiplied by some small constant μ:

$$z(k+1) = z(k) - \mu \cdot \bar{\mathrm{E}}\{\bar{z}\bar{z}^T\} \cdot z(k) + \mu \cdot \bar{\mathrm{E}}\{\bar{z}y^T\} \cdot y, \tag{28}$$

where $\bar{\mathrm{E}}\{\bar{z}\bar{z}^T\}$ and $\bar{\mathrm{E}}\{\bar{z}y^T\}$ denote the variance-compensated covariance matrices. In stationary state where no more adaptation takes place the (uncentered) covariance of $\bar{z}$ is

$$\mathrm{E}\{\bar{z}\bar{z}^T\} = \left(\mathrm{E}\{\bar{z}\bar{z}^T\}\right)^{-1} \mathrm{E}\{\bar{z}y^T\} \cdot \mathrm{E}\{yy^T\} \cdot \mathrm{E}^T\{\bar{z}y^T\} \left(\mathrm{E}\{\bar{z}\bar{z}^T\}\right)^{-1}. \tag{29}$$

This can be written as

$$\mathrm{E}^3\{\bar{z}\bar{z}^T\} = \mathrm{E}\{\bar{z}y^T\}\mathrm{E}\{yy^T\}\mathrm{E}^T\{\bar{z}y^T\}. \tag{30}$$

If one assumes that there holds $\bar{x} = \theta^T \cdot y$ for some $n_y \times n_x$ matrix θ, one has

$$\left(\theta^T \cdot \mathrm{E}\{yy^T\} \cdot \theta\right)^3 = \theta^T \cdot \left(\mathrm{E}\{yy^T\}\right)^3 \cdot \theta. \tag{31}$$

This is a strange result. It is evident that any subset of eigenvectors of $\mathrm{E}\{yy^T\}$ as collected in the matrix θ satisfy (31); it turns out that the cumbersome GHA-like sequential extraction of covariance matrix eigenvectors (or principal components) is not necessary in the Hebbian framework, but a parallel, linear, and neurally perhaps more plausible process suffices!

When the mapping model θ^T has been found, the least-squares estimate for the reconstructed y can be seen from (26) to be

$$\hat{y}(t) = \theta \cdot \bar{z}(t). \tag{32}$$

If there is the nonlinearity f_{cut} included in the model, the matrix no more consists of principal components; the hypothesis here is that, rather, they are *sparse components:* Only those elements of $\bar{z}_i(t)$ are active that correspond to features θ_i that are existent in the $y(t)$ input sample (similar "positive feature" decomposition assumption is explicitly implemented by the `HUTCH` model; see [9]).

6.2 Shift from novice to expert

An interesting problem in cognitive science is that of *shift from novice to expert* [10]. It can be assumed that a novice follows rules — in the forward-chaining manner — whereas an expert applies pattern recognition using a pool of domain-oriented features. The outlook of the reasoning processes is very different in these two cases; how can the qualitative leap be explained and overcome? Actually, this dilemma between different types

of reasoning is a painstaking paradox in cognitive science — but, as studied below, the mathematical model can perhaps give some intuition and conceptual tools for attacking the problem.

First let us study the structure of forward-chaining (assumedly being the basic mechanism underlying novice reasoning). Loosely speaking, the rule-form representations ("x_1 causes x_2") are crisp, typically asymmetric, not reciprocal; associative, expert-like, feature-based representations ("x_1 correlates with x_2 AND x_2 correlates with x_1") are less crisp, and typically more symmetric, facilitating fast and consistent excitation of features related to an observation. How to implement a change from one to the other?

Note that the covariance matrices discussed above contain the connections between constructs; at least in special cases this covariance structrure can be interpreted as determining a numeric, continuous-valued *semantic net* among concepts. The covariance matrix can be learned little by little starting from the "rule matrix", gradually adapting the matrices towards those data structures that are dictated by the Hebbian – Anti-Hebbian learning. After this transfer phase is over, the matrix A is completely symmetric — or "two-way". The matrix elements are not crisp but *fuzzy* as compared to the original rule matrix, and there are typically no zero entries. This means that the *spread of activation* takes place fast among the constructs, and because of the number of additional connections the robustness is also enhanced. Note that the explicit, rule-based forward-chaining phase is necessary to originally initialize the pool of $\bar{x}$ vectors before further adaptation can take place; it is the nearest local minimum in the "feature space" where the process converges[6].

Technically speaking, to implement the transfer from declarative to "tacit" data structures in the presented framework, one needs to recognize that the model

$$x(k+1) = A_{\text{decl}} \cdot x(k) + B_{\text{decl}} \cdot y \tag{33}$$

can be transformed into the form

$$x(k+1) = (I - \eta \cdot (I - A_{\text{decl}})) \cdot x(k) + \eta \cdot B_{\text{decl}} \cdot y \tag{34}$$

without affecting the final outcome; the process just becomes slower if η is small. What is more important, is that the new formulation is directly compatible with (28) — indeed, one can combine the two:

$$\begin{aligned} x(k+1) = \Big(I - \eta \cdot (I - A_{\text{decl}}) - (1-\eta) \cdot \bar{\text{E}}\{\bar{x}\bar{x}^T\}\Big) \cdot x(k) \\ + \Big(\eta \cdot B_{\text{decl}} + (1-\eta) \cdot \bar{\text{E}}\{\bar{x}y^T\}\Big) \cdot y. \end{aligned} \tag{35}$$

When enough data has been observed, so that the covariances can be reliably estimated, the shift from novice to expert can be simulated by letting η decay from 1 (purely declarative model) towards 0 ("tacit" representation). This transition process has to be gradual: Changing system matrices change the state vectors, and *vice versa.* The nonlinearity f_{cut} can assumedly be included in the model without pathological effects.

In the forward inference implementation, calculation and reasoning is seen as a process, starting from some initial values and ending up in some final state where no more rules are active. Another interpretation of the expert knowledge, on the other hand, is that the expert data structures appropriately span substructures in the high-dimensional variable

[6]Because of the variance normalization, the first principal components, or the features grabbing most attention, do not dominate exclusively; any of the covariance eigenvectors can pop up in the process

space, determining the "space of expertise". The goal is to determine a location in that space where the given constraints for variables are fulfilled (or where the observation data can best be matched). The originally dynamic process of firing the declarative rules, reasoning thus consisting of a number of sequential, causal steps, has changed into a static problem — all that is needed is to find a point in an unchanging space. Reasoning becomes a search process in that space, and the iterative processes are needed to implement the search. It can be claimed that the different AI formalisms are just *different ways of spanning the high-dimensional space of expertise.*

This kind of mathematically oriented issues and approaches are concentrated on in the latter part of this paper: The main observation there is that *many cognitive behaviors can be formulated as optimization problems.* This can be seen as a key to reaching "holistic level mathematics".

REFERENCES

[1] Földiák, P.: Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics,* Vol. 64, 1990, pp. 165–170.

[2] Hyötyniemi, H. and Saariluoma, P.: Chess — Beyond the Rules. In Timo Honkela (ed.): *Games, Computers and People (Pelit, tietokone ja ihminen),* Finnish Artificial Intelligence Society, Helsinki, Finland, 1999, pp. 100-112.

[3] Hyötyniemi, H.: On Unsolvability of Nonlinear System Stability. In *Proceedings of the European Control Conference (ECC'97),* Brussels, Belgium, July 1-5, 1997 (CD-ROM format).

[4] Hyötyniemi, H.: On Mental Images and 'Computational Semantics'. In *Proceedings of the 8th Finnish Artificial Intelligence Conference STeP'98* (eds. Koikkalainen, P. and Puuronen, S.), Finnish Artificial Intelligence Society, Helsinki, Finland, 1998, pp. 199–208.

[5] Hyötyniemi, H.: *Mental Imagery: Unified Framework for Associative Representations.* Helsinki University of Technology, Control Engineering Laboratory, Report 111, August 1998.

[6] Hyötyniemi, H.: *Multivariate Regression — Techniques and Tools.* Helsinki University of Technology, Control Engineering Laboratory, Report 125, 2001.

[7] Hyötyniemi, H.: *Complex Systems — Searching for Gold.* Arpakannus 2/2002, special issue on Complex Systems, pp. 29–34.

[8] Hyötyniemi, H.: *Studies on Emergence and Cognition — Part 2: High-Level Functionalities.* Finnish Artificial Intelligence Conference (STeP'02), December 16–17, 2002, Oulu, Finland.

[9] Hyötyniemi, H.: `HUTCH` *Model in Information Structuring.* Finnish Artificial Intelligence Conference (STeP'02), December 16–17, 2002, Oulu, Finland.

[10] Kellogg, R.T.: *Cognitive Psychology.* SAGE Publications, London, 1995.

[11] Wolfram, S.: *A New Kind of Science.* Wolfram Media, Champaign, Illinois, 2002.