Cybernetics: Towards a Unified Theory?

Heikki Hyötyniemi

Cybernetics Group
Helsinki University of Technology, Control Engineering Laboratory
P.O. Box 5500, FIN-02015 HUT, Finland
heikki.hyotyniemi@hut.fi
http://www.control.hut.fi/hyotyniemi

Abstract. Cybernetics, or the study of "Control and Communication in the Animal and the Machine" was one of the cornerstones of systems science. However, it seems that no concrete methodologies exist to make use of the intuitively appealing ideas. — This paper proposes one approach that makes it possible to benefit from the intuitions. The proposed generic model structure is based on local interactions, resulting in globally optimized behavior, and it can be claimed that the underlying ideas are applicable to a wide variety of complex systems, including neuronal, ecological, and economical ones.

1 "Project 42"

The idea of *cybernetics* was presented by Norbert Wiener over half a century ago [21]. In many natural systems it is local interactions that result in emergent global behaviors, like self-stabilization and self-organization. The ability of adaptation, and surviving in changing environments was seen as a prototype of what *intelligent* behavior is. Indeed, cybernetics was one of the cornerstones of the early Artificial Intelligence research. And today, again, it seems that AI field more or less is identified with "Agent Intelligence", or "Ambient Intelligence".

During the decades, the ideals of holism have flourished under different names. For example, the *theory of complex systems* has also been searching for panacea for explaining all systems, big and small, in the same coherent framework. The intuitive feel of similarity among different systems has resulted in hopes of finding theories for "Life, Universe, and Everything" (for example, see [13]). There have been no breakthroughs, and as compared to the wonderful promises, the results have been marginal. Because of the disappointments, the concept of *ironic science* has been coined [9].

The power of theories should be in their capability of providing us with new intuitions that could not be anticipated right from beginning. Are there such intuitions to be reached?

The claim here is that different branches have now matured enough so that new intuitions really are available. Specially, control theory and mathematical modeling have developed considerably since the mid 1900's. Understanding of the dynamics in closed feedback loops has deepened. What is more, understanding of

specific cybernetic systems (neural networks, for example) has dramatically progressed. Still, it seems that the other key to deeper understanding of cybernetic systems also dates back to the late 1940's — namely, the work of the physician Donald O. Hebb is employed here.

Rather than starting from all-embracing hypotheses, the discussion below concentrates on a concrete example. Conceptual tools from *system theory* are applied to reach a *systemic view,* and based on that, assumptions about the *emergent functionality* can be made. It turns out that these *neocybernetic* discussions can be applied to a wide variety of different fields.

2 Starting Bottom-Up ...

When studying complex systems, concrete examples often help to see the bigger picture, making it possible to draw analogies. For this purpose, a specific example of complex systems is studied, namely, a *grid of neurons.*

2.1 About system theory

When searching for models for complex systems, there always exist various ways to proceed. And depending on the selected route, one can either get a long way, or the analyses are inevitably doomed in deadlock. How to select among candidate approaches? In this paper, *system theory* is applied as a guiding principle.

System theory is the conceptual framework for capturing general behavioral principles common to different kinds of systems [2]. In the system theoretic setting, the problems are seen in a wider perspective. Rather than studying single signals, for example, the system of signals is studied as a whole. It may turn out that from the interactions between signals some unanticipated behaviour pops up.

It is essential to select the boundaries of the system being studied appropriately. Correct level of abstraction needs to be selected, and all dependencies within the system need to be captured. For example, in [15] the problem with simple causal models are studied: The basic problem with cause/effect structures is that changes in effect do not mean that the causes have changed, meaning that simple statistical correlations-based analyses become void. The answer to this problem is appropriate selection of system boundaries; *closures* of causal dependencies are only studied here.

One is not trying to study all mathematically possible systems, only physically plausible, or *interesting* ones: In a "smart" — cybernetic — system, there are no unidirectional causality diagrams but there always exist feedbacks, so that all dependencies become tangled. All interacting variables need to be included within the system, whereas the truly independent variables are interpreted as inputs coming from outside. Different kinds of analyses and theoretical tools are needed here as compared to traditional modeling of causalities. Assuming that appropriate feedback mechanisms somehow have been implemented, no matter

where these feedbacks originate from. This way, natural and technical systems alike can be studied.

In a cybernetic system causes and effects are bidirectionally connected (indeed, there exists a multitude of influence mechanisms), and statistical tools can be applied. However, the dimensions of the systems become high as all variables need to be simultaneously represented. And because of the feedback loops, static studies do not suffice, but dynamics has to be taken into account. It is these two things — high dimensionality and dynamicity — that make it possible to reach new system-level functionalities. When seen in the traditional perspective, this high dimensionality and dynamical nature of the systems seems to make analysis difficult. But if appropriate tools are applied, things seem to clear up — there exist special mathematical frameworks for mastering such high-dimensional spaces of parallel dynamic signals: *Linear algebra* and *matrix calculus,* together with *multivariate statistics* and *theory of dynamic systems.*

When trying to be too ambitious the applicability of the system theoretic ideas necessarily becomes vague. To reach something concrete, an additional assumptions is made here:

The structures that are studied are *essentially linear.*

The only justification for this assumption is based on intuition: We are not yet facing the End of Science — for example, the neuronal system, after all, just *has to be* analyzable[1]. Only for linear structures the *scalability* of the models can be reached, and properties of the whole system can be attacked using reductionistic approaches — the only ones there exist today (later, this linearity constraint is relaxed to make the studies better match reality). And, even though one is here studying "non-equilibrium state" systems, to reach concrete results, another basic system theoretic intuition that has to be obeyed is:

The structures that are studied are *essentially stable.*

There are theoretical and philosophical motivations for this stability assumption. First, from the theoretical point of view, powerful analyses are only available for stable systems. The second motivation is more intuitive: Interesting functionalities take time to emerge, and in an unstable environment, there is no possibility for such fragile functionalities to ever become noticeable. As the systems become larger, there is more and more need for explicit emphasis for the stability issue: It is enough that just one of the n dynamic modes in the system is unstable, and the whole system explodes.

The loss of expressional power due to the linearity assumption, and the loss of intuitive appeal due to the stability assumption, are compensated in a system theoretic way: Rather than studying individual neurons, the whole grid of neurons is simultaneously taken into account, and the feedback loops result in dynamic behaviors. The interactions between neurons result in emergent functionalities, as will be seen later.

[1] Just as Gaia preserves the Earth, Pallas Athene preserves scientific progress!

2.2 Modeling a Hebbian neuron

The research on artificial neural networks has departed from the original objectives — today, ANN's are seen only as computing devices, forgetting about the origins in artificial intelligence, when the operation of the brain was being studied and explained. Simultaneously, the models have become increasingly complex, so that efficient analysis methods do no more exist. For example, starting from the linear, intuitively appealing *Hebbian neuron* model, highly complex structures have been developed (for example, see [7], [4]).

In what follows, the goal is to explicitly stick to basic neurophysiological observation, the *Hebbian learning principle.* This neuronal behavior was observed by the physician Donald O. Hebb some half a century ago [8], and it can be formulated as follows:

> **Hebbian law.** Synaptic connection between the neuron and an incoming signal becomes stronger if the signal and the current neuron activity correlate with each other.

As is known, simple Hebbian learning is unstable: Following the basic idea, the synaptic connections become stronger and stronger without limit. In practical Hebbian algorithms, this instability is eliminated by introducing an additional nonlinearity (*Oja's rule,* see [14]). Unfortunately, this nonlinearity makes the analysis of the overall system very difficult.

However, stability can also be reached using linear techniques by applying *negative feedback.*

To study the neuronal system behavior in systemic terms, let the vector $x(t)$ denote the neuron activities at time t, so that vector element x_i represents the activity of the neuron index i. Further, let the vector of input signals be $u(t)$. Assume that the number of neurons, n, is lower than the number of input channels, m. The synaptic strength between neuron i and input signal j, or r_{ij}, is now assumed to change as

$$\frac{d\,r_{ij}}{d\,t}(t) = \rho \bar{x}_i(t) u_j(t) - \frac{1}{\tau} r_{ij}(t). \tag{1}$$

The first term contains the Hebbian learning factor (unscaled correlation between neuronal activity and input signal), whereas the latter term represents the negative feedback from the neuron activity. In what follows, $\bar{x}$ denotes the steady-state value of x after initial transients. Parameter ρ determines the synaptic dynamics, together with the parameter τ that is the time constant determining the rate of decay. Correspondingly, all connections from inputs to neurons can be expressed in matrix form as

$$\frac{d\,R}{d\,t}(t) = \rho \bar{x}(t) u^T(t) - \frac{1}{\tau} R(t), \tag{2}$$

where u^T denotes vector transpose. If it is assumed that the system is (weakly) stationary, that is, the (second order) statistical properties of the data do not

change over time, and if τ is large, one can solve for the steady-state value, so that (using the expectation operator in a somewhat sloppy way)

$$\bar{R} = \rho\tau\, \mathrm{E}\{\bar{x}u^T\}, \tag{3}$$

That is, the matrix of synaptic weights R becomes the (scaled) *cross-correlation matrix.*

2.3 Dynamics in a neuron grid

Above, individual neurons (indeed, individual synapses) were studied — no interesting functionalities can be seen yet. Next proceed to the *grid level adaptation* among a set of Hebbian neurons.

When studying interconnections among individual neurons, stability issues become crucial again. And, again, stabilization of this process can be implemented in linear terms by applying negative feedback.

Assume that each neuron is connected to each input, but, additionally, each of the neurons is also connected to all other neurons. Using the matrix notation, the dynamics in a linear neuron grid can be modeled as

$$\frac{d\,x}{d\,t}(t) = -A\,x(t) + B\,u(t). \tag{4}$$

Matrix A captures the synaptic weights between neurons; because of its construction (as explained below) A is *positive definite,* and negative feedback is reached by explicitly adding the minus sign. When looking at the Hebbian principle above, it is evident that one can decompose (3) for different signals as

$$A = \rho\tau \mathrm{E}\{\bar{x}\bar{x}^T\}, \tag{5}$$

and

$$B = \rho\tau \mathrm{E}\{\bar{x}u^T\}. \tag{6}$$

Despite the simplicity of the formulation, this kind of all-embracing framework has not been studied before. Note that even though the synaptic effects are presented in such a compact form, the interactions and adaptation operations are still completely local, matrix element A_{ij}, for example, representing the synaptic weight from neuron j to neuron i. Unbiased correlation matrix estimates can be updated on-line as follows:

$$\frac{d\,\hat{\mathrm{E}}\{\bar{x}\bar{x}^T\}}{d\,t}(t) = -\lambda\,\hat{\mathrm{E}}\{\bar{x}\bar{x}^T\}(t) + \lambda\,\bar{x}(t)\bar{x}^T(t) \tag{7}$$

and

$$\frac{d\,\hat{\mathrm{E}}\{\bar{x}u^T\}}{d\,t}(t) = -\lambda\,\hat{\mathrm{E}}\{\bar{x}u^T\}(t) + \lambda\,\bar{x}(t)u^T(t), \tag{8}$$

where λ determines the adaptation rate. In what follows, it is assumed that the grid dynamics, as determined by (4), is much faster than the rate of change in

the external input; further, the adaptation of correlation matrices, as determined by (7) and (8), is assumed to be still much slower. Because of the properties of correlation matrices, all eigenvalues of $-A$ are non-positive, and the model (4) is stable.

One can elaborate on the structure in (4). Because of the minus sign, it seems that the basic Hebbian law is inverted if the signal belongs to a feedback structure. Indeed, one can define

> **Anti-Hebbian law.** Synaptic connection between two neurons becomes *weaker* if the neuronal activities correlate with each other (or, actually, *opposite* activation becomes stronger).

The Hebbian laws can be interpreted so that the effects from the prior level are *excitatory,* but lateral connections between the same level neurons are *inhibitory.* As will be seen later, the Hebbian principle makes it possible to learn the data, whereas the role of anti-Hebbian learning is to implement some kind of *competitive learning* resulting in organization of structures. Whereas the Hebbian learning has a long history, anti-Hebbian learning is newer, perhaps the most notable early studies being carried out in [5]. However, all such models have been highly nonlinear, and they have been developed for one neuron at a time.

2.4 Principal subspace analysis

It is also assumed that the system is in stationary state; further, assume that the change of rate in u is low enough, $\frac{d\,u}{d\,t}(t) \approx 0$, so that, perhaps excluding the initial transient, x in the process (4) always follows the input very fast. If this holds, the derivative in (4) vanishes, and one has for the steady-state

$$\bar{x}(t) = A^{-1} B\, u(t), \tag{9}$$

assuming invertibility of A (this always holds if the variables in x are not linearly dependent). Note that the effects of the parameters ρ and τ vanish in this operation. For future reference, define this linear mapping from u to $\bar{x}$ as

$$\phi^T = A^{-1} B = \mathrm{E}\{\bar{x}\bar{x}^T\}^{-1} \mathrm{E}\{\bar{x} u^T\}. \tag{10}$$

Next, study the covariance of the neuronal state, as calculated from (9), taking the expectation of the sidewise outer products:

$$\mathrm{E}\{\bar{x}\bar{x}^T\} = \mathrm{E}\{\bar{x}\bar{x}^T\}^{-1} \mathrm{E}\{\bar{x} u^T\} \mathrm{E}\{u u^T\} \mathrm{E}^T\{\bar{x} u^T\} \mathrm{E}\{\bar{x}\bar{x}^T\}^{-1}. \tag{11}$$

One can multiply the expression from left and from right by the covariance matrix $\mathrm{E}\{\bar{x}\bar{x}^T\}$, so that

$$\mathrm{E}\{\bar{x}\bar{x}^T\}^3 = \mathrm{E}\{\bar{x} u^T\} \mathrm{E}\{u u^T\} \mathrm{E}^T\{\bar{x} u^T\}. \tag{12}$$

There seldom pop up third powers in linear algebra! Remembering the definition of ϕ, this becomes

$$\left(\phi^T \mathrm{E}\{u u^T\} \phi\right)^3 = \phi^T \left(\mathrm{E}\{u u^T\}\right)^3 \phi. \tag{13}$$

As shown in [11], the stable fixed point for ϕ is such that it spans the *principal subspace* of the input data, that is, columns of ϕ are linearly independent combinations of the n most significant eigenvectors of the input correlation (covariance) matrix $\mathrm{E}\{uu^T\}$. Further, the variability in x equals the total variance along the n most significant principal component directions in u (for information on the central role of principal components and principal component analysis (PCA) in modern multivariate data analysis, see [1]).

Because $n < m$, not all variation in $u(t)$ can be explained by $x(t)$. However, the coding based on PCA is efficient, because the latent variables are mutually uncorrelated and do not disturb each other. What is more, it can easily be shown that the largest principal components explain the variations in the inputs in the *best possible* way. Indeed, this global optimality can be elaborated on closer.

2.5 Optimality of representations

As shown in [11], the operation of the neuron grid can be seen in a more abstract perspective. It turns out that (4) determines an algorithm (steepest descent approach in continuous time) for finding the minimum for the quadratic criterion

$$J(x) = \frac{1}{2} (u - \phi x)^T \mathrm{E}\{uu^T\} (u - \phi x) , \tag{14}$$

where ϕ is the stationary solution to (10). It is evident that x's that minimize the above criterion try to *explain* the input data u as accurately as possible. Because of the weighting matrix $\mathrm{E}\{uu^T\}$ it is also clear that the above criterion emphasizes the directions in the data having the highest variation — and, indeed, this property can be detected in simulations (see [11]). The peculiar weighting makes the solutions differ from the traditional maximum likelihood data fitting approaches.

The cost criterion gives us yet higher-level view of what happens in the neuronal process: Rather than having to follow the actual iteration, one can directly concentrate on the final pattern that would finally emerge out from the iteration.

In (general) systems theory two views of looking at a system are distinguished: The *process view* and the *pattern view* [18]. In this perspective, the original approach of seeing the behavior of a system as a dynamic process, or as an iteration, is an example of the process view. The opposite perspective, or trying to see beneath the complicated iterations, trying to perceive the emergent patterns, may open up new horizons. It is not the complicated iterations that determine the essence of complex systems, as claimed by Stephen Wolfram [22] — it is the final pattern that emerges from that iteration!

It should be remembered that one is just constructing models, not claiming that there should exist some fundamental correspondence between the model and the reality. However, as compared to traditional engineering disciplines, system theory is just a step nearer to philosophy and metaphysics: It not only tries to explain behaviors, answering the *how* questions, but it also tries to answer the *why* questions, searching for general principles governing the system behavior.

Here, in this study we started with an *intentional* assumption concerning the neuronal behavior: It was assumed that neurons *try to* maximize correlation or match between some quantities, and now it turns out that simultaneously they try to optimize something else. Whereas the original behavior was local, the optimizing behavior turns out to be an emergent *global* property.

2.6 Relaxing the assumptions

The mathematical approach above is not truly credible: Why should the neurons exactly obey some mathematical formula? Indeed, it turns out that it only needs to be assumed that the formulas (4), (7) and (8) reveal the *tendency,* not the *rate,* that is, the numerical parameters are of no major relevance what comes to global properties of the system. For example, assuming that the actual neuronal dynamics can better be expressed as

$$\frac{d\,x}{d\,t}(t) = -\alpha N\,A\,x(t) + \beta N\,B\,u(t), \tag{15}$$

where α and β are arbitrary scalars and N is an invertible matrix, the structure of the problem remains intact: Even though the correlation matrices may become scaled because of the numerical values of α and β, their structure is not changed. Eigenvalues (or activity variances) become scaled, but the eigenvectors remain the same. The presented mathematics also applies if one just assumes that all signals are treated equally. If $\alpha \equiv \beta$, the above results can directly be applied, no changes taking place whatsoever.

The matrix N, on the other hand, can be introduced to facilitate representing some kind of *topology* among the neurons, so that, for example, *diffusion* (or spread of activation) can be modeled. Note that in this linear case N has no effect on the steady state.

One phenomenon that later turns out to be important when extending the model is that if there exist various neurons having the same behavioral pattern, the results do not change, if the corresponding x_i's are combined into a single variable that contains the cumulative activity of individuals. This fact results from the linearity of the model. One can also freely select the appropriate level of abstraction, switching the emphasis from individuals to populations. From the point of view of extending the model to more complicated environments where the boundaries between subsystems cannot necessarily be exactly determined (see later), this linear scalability property is crucial.

Further, it turns out that if one defines the adaptation of the correlation matrix estimates as

$$\frac{d\,\hat{\mathrm{E}}\{\bar{x}\bar{x}^T\}}{d\,t}(t) = -\Lambda\,\hat{\mathrm{E}}\{\bar{x}\bar{x}^T\}(t) + \Lambda\,x(t)x^T(t) \tag{16}$$

and

$$\frac{d\,\hat{\mathrm{E}}\{\bar{x}u^T\}}{d\,t}(t) = -\Lambda\,\hat{\mathrm{E}}\{\bar{x}u^T\}(t) + \Lambda\,x(t)u^T(t), \tag{17}$$

where Λ is no more scalar but an invertible matrix, the steady-state solutions for A and B still remain intact (and the Λ's in (16) and (17) need not even be the same). This means that, for example, if Λ is diagonal, individual forgetting factors of the neurons being collected on the diagonal, different neurons can have differing adaptation dynamics without affecting the overall behavior at all. This property can become relevant if the local actors are not identical (see later). And for a *nonlinear* system where there can exist various local extrema the local dynamics can play a major role.

Often, specially when the system is locally linearized, the model is not linear but *affine,* that is, there is additionally a constant term in the model:

$$\frac{d\,x}{d\,t}(t) = A\,x(t) + B\,u + c. \tag{18}$$

The original linear model formulation still applies if one introduces the new (non-zero-mean) state variable $x' = x + A^{-1}c$.

2.7 Unification of levels

A single synapse is a lower-level complex system, where the interplay of underlying chemical processes determine its behavior on the macroscopic level. Above, the models were derived essentially in the same way for a single synapse and for the whole neuron grid: Correlation-oriented structures were constructed, and stability was provided by applying negative feedback. Could the same model structure be applied in both cases? It is evident that some extensions are needed to capture both systems in the same framework; the goal here is to search for the *simplest* of such extensions.

Let us *assume* that the synapse operates qualitatively in the same manner as the whole neuron grid does. In this case the input and output are scalars[2], and slightly extending the Hebbian/anti-Hebbian basic formula one can write for the output of a single synapse as

$$\frac{d\,\xi_{ij}}{d\,t}(t) = -\lambda_i \mathrm{E}\{g(\bar{\xi}_{ij})^2\}\,\xi_{ij}(t) + \lambda_i \mathrm{E}\{g(\bar{\xi}_{ij})u_j\}\,u_j(t), \tag{19}$$

so that in steady state

$$\xi_{ij}(t) = \frac{1}{\mathrm{E}\{g(\bar{\xi}_{ij})^2\}} \mathrm{E}\{g(\bar{\xi}_{ij})u_j\}\,u_j(t). \tag{20}$$

In the similar way, also the synapses between two neurons can be expressed analogously (below, variables ζ rather than ξ are used in those cases). The function g offers extended functionality in the model. Assume that function g is defined as

$$g(\xi_{ij}(t)) = \xi_{ij}(t) + \epsilon(t), \tag{21}$$

[2] Note that this scalar nature of the synapse is only a simplification; the synaptic process is complex, involving many variables. Again, the observed synaptic behavior is an emergent phenomenon

where $\epsilon(t)$ is some unknown signal. The main difference in (19) as compared to earlier discussions is the function g: It defines the interaction mechanism between cybernetic subsystems. From the viewpoint of the single synapse, the role of the function g is to introduce noise in the system; however, as seen from outside, however, the noise consists of the contributions of the *other* synapses, that is, the final neuron activation can be written as $x_i(t) = g(\xi_{ij}(t))$, or actually

$$x_i(t) = -\zeta_{i1}(t) - \ldots - \zeta_{in}(t) + \xi_{i1}(t) + \ldots + \xi_{im}(t), \tag{22}$$

where the signs are determined by whether the effects are inhibitory or excitatory. This means that when (22) is differentiated, using (20) one can write

$$\frac{d\,x_i}{d\,t}(t) = \frac{1}{\mathrm{E}\{\bar{x}_i^2\}} \left(-\sum_{j=1}^{n} \mathrm{E}\{\bar{x}_i \bar{x}_j\} x_j(t) + \sum_{j=1}^{m} \mathrm{E}\{\bar{x}_i u_j\} u_j(t) \right). \tag{23}$$

When collected into a matrix form, the set of expressions capturing the whole neuron grid can be written as

$$\frac{d\,x}{d\,t}(t) = -\mathit{\Lambda}\mathrm{E}\{\bar{x}\bar{x}^T\}\, x(t) + \mathit{\Lambda}\mathrm{E}\{\bar{x}u^T\}\, u(t), \tag{24}$$

where $\mathit{\Lambda}$ is a diagonal matrix containing the elementwise adaptation rates on the diagonal; in the above case, this matrix is defined using the *variance matrix* as

$$\mathit{\Lambda} = \mathrm{V}\{\bar{x}\bar{x}^T\}^{-1} = \begin{pmatrix} \mathrm{E}\{\bar{x}_1^2\} & & 0 \\ & \ddots & \\ 0 & & \mathrm{E}\{\bar{x}_n^2\} \end{pmatrix}^{-1}. \tag{25}$$

Comparing (24) to the behavior of the Hebbian/anti-Hebbian neuron grid, where there were only the correlation matrices without the matrix $\mathit{\Lambda}$, one can see that this new formulation is an extension of it: If defined as $\mathit{\Lambda} = \mathrm{V}\{\bar{x}\bar{x}^T\}^{-1}$, the neuron indices reveal that the parameters σ and τ in (3) are now neuron-specific, so that $\rho_i\tau_i = 1/\mathrm{E}\{\bar{x}_i^2\}$. From the physiological point of view, this extensions is motivated: It has been recognized that high variation level exhausts the neuron, making it less sensitive to inputs; this phenomenon can also be modeled in the above form. It turns out that if one in (15) selects $\alpha N = \beta N = \mathrm{V}\{\bar{x}\bar{x}^T\}^{-1}$, as defined in (25), the steady state is identical in any case. However, the new more sophisticated approach gives information also on the transient dynamics: The adaptation rates are given in the matrix $\mathit{\Lambda}$. As compared to the earlier, extremely simple model that was proposed for neuron grids (4), it seems that the model (24) can be accepted as a more sophisticated version of it.

The main result here is, however, that the extended model applies to various domains. Indeed, comparing formulas (19) and (24) it seems that the same model structure captures the behaviors of two different systems, the synapse-level processes as well as processes on the neuron grid level. On the neuron grid level, there are no neighboring systems, so that the function g is an identity mapping. Or is it?

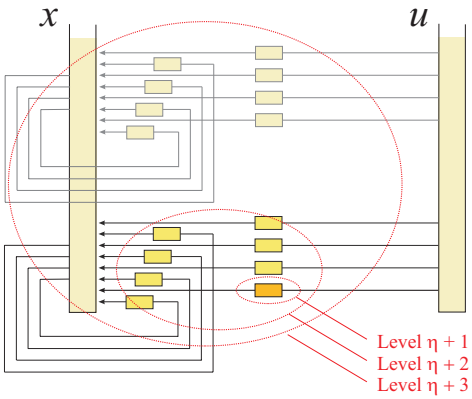

Fig. 1. Schematic illustration of levels in a neural system

When wondering whether the presented neuron model could still be extended, so that there are various coexisting subsystems, one can generalize the above studies, interpreting neuron grids as mere synapses above. Indeed, if one has vector-form sub-blocks in (25), so that $\mathrm{E}_i\{\bar{x}\bar{x}^T\}$ denotes the covariance matrix of the i'th block (only a subset of x's being employed), one can write

$$\mathrm{V}\{\bar{x}\bar{x}^T\} = \begin{pmatrix} \mathrm{E}_{\mathrm{first}}\{\bar{x}\bar{x}^T\} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathrm{E}_{\mathrm{last}}\{\bar{x}\bar{x}^T\} \end{pmatrix}. \tag{26}$$

Similarly, the additive formulation of the function g can be motivated also in more general cases. Assume that there are η subsystems, each generating its own vector-form contribution ξ_η to x, and these vectors are summed:

$$\begin{aligned} x &= \xi_{\mathrm{first}} + \cdots + \xi_{\mathrm{last}} \\ &= \mathrm{E}\{\bar{x}\bar{x}^T\}^{-1}\mathrm{E}\{\bar{x}u_{\mathrm{first}}^T\}u_{\mathrm{first}} + \cdots + \mathrm{E}\{\bar{x}\bar{x}^T\}^{-1}\mathrm{E}\{\bar{x}u_{\mathrm{last}}^T\}u_{\mathrm{last}} \\ &= \mathrm{E}\{\bar{x}\bar{x}^T\}^{-1}\left(\mathrm{E}\{\bar{x}u_{\mathrm{first}}^T\}\middle|\cdots\middle|\mathrm{E}\{\bar{x}u_{\mathrm{last}}^T\}\right)\begin{pmatrix} u_{\mathrm{first}} \\ \vdots \\ u_{\mathrm{last}} \end{pmatrix} \end{aligned} \tag{27}$$

This means that the input vectors u_{first} to u_{last} can be stacked on top of each other, and the overall system assumedly carries out principal subspace analysis of this combined vector. Subsystems can also be freely connected; the appropriate subsystem boundaries are dictated by where the structure adaptation feedback comes from, or where is the focus of system optimization, as determined by the selection of the function g. In the discussions that follow, this function g will be ignored: It is assumed that the scope of the cybernetic system has been determined appropriately, so that it is the system output x that is the feedback signal in the system matrices (see Fig. 1).

The derivations above were based on intuition: One would like the have the same structure on all levels. Applying the model (24) this holds: The synapses have the same model as the neurons do (even though the dimensions are very

different). However, the results sound plausible also from other points of view: The weighting factors in Λ (that is, the input to the neuron x_i are multiplied by $1/\mathrm{E}\{\bar{x}_i^2\}$) resemble the *maximum likelihood weighting* of variables when the error variance levels in different variables are not the same.

Extending such mathematical considerations, there is still more intuition available. Note that the model (4) with (5) and (6) has its fixed point in the same point where the following criterion has its minimum:

$$\mathcal{J}(x) = \frac{1}{2} x^T \mathrm{E}\{\bar{x}\bar{x}^T\} x - x^T \mathrm{E}\{\bar{x}u^T\} u. \tag{28}$$

This must equal to the original formulation (14) if constant terms are omitted, defining the same minimum. The gradient of this criterion is

$$\frac{d\,\mathcal{J}}{dx}(x) = \mathrm{E}\{\bar{x}\bar{x}^T\} x - \mathrm{E}\{\bar{x}u^T\} u, \tag{29}$$

giving the gradient descent process in the familiar form

$$\frac{d\,x}{d\,t}(t) = -\Lambda \left(\mathrm{E}\{\bar{x}\bar{x}^T\}\, x(t) - \mathrm{E}\{\bar{x}u^T\}\, u \right). \tag{30}$$

Any positive definite matrix Λ determining the final search direction can be selected; if this matrix is selected according to the *Newton method* that gives for quadratic criterion the best possible search direction, one has

$$\Lambda = \left(\frac{d^2 \mathcal{J}}{dx dx^T}(x) \right)^{-1} = \left(\mathrm{E}\{\bar{x}\bar{x}^T\} \right)^{-1}. \tag{31}$$

Comparing this to (25), one can see that this formulation is more complicated, also the non-diagonal matrix entries being employed. This more efficient definition cannot be implemented locally, but it can have some practical value: If one wants to simulate a neural system, a practical algorithm has the form

$$\frac{d\,x}{d\,t}(t) = -x(t) + \mathrm{E}\{\bar{x}\bar{x}^T\}^{-1} \mathrm{E}\{\bar{x}u^T\}\, u. \tag{32}$$

Model (32) represents a filtered version of the direct static mapping $x = \phi^T u$.

2.8 Towards sparse coding

Above, linearity was taken as one of the basic goals in modeling. However, linear networks truly *are* too restricted to carry out really interesting tasks. For example, successive layers of linear grids, when connected together, can be substituted with a single layer, where the mapping matrix is constructed by multiplying the original mapping matrices together.

It is well known that PCA is well-motivated from the mathematical point of view, but not so well from the physical point of view. How to introduce nonlinearity in the model to reach something new? One should be extremely cautious when selecting the function form, because, from the point of view of

analysis and well-founded theory, nonlinearities open the *Pandora's box.* One would not like to change the basic PCA-type functionality too much.

How to select the nonlinearity in practice, and where to put it, then? Natural data is often non-Gaussian, being composed of *independent* or *sparse* components (see [6], [16]). The cognitive machinery also seems to decompose input into sparse components, that is, some neural units being active and some being inactive. This means that in the current framework only some of the latent variables should be non-zero whereas other ones should have strictly zero activity when explaining an input sample. The simplest way to reach sparsity is by introducing the *"cut" function* $f_{\text{cut}} : \mathcal{R}^n \to \mathcal{R}_+^n$ that can be defined elementwise as

$$f_{\text{cut},i}(x) = \begin{cases} x_i, & \text{if } x_i \geq 0, \text{ and} \\ 0, & \text{otherwise.} \end{cases} \tag{33}$$

In principle, vector elements below zero are simply cut to zero, dividing the state space in two parts: Active (non-zero) and inert (zero). This function form has the following advantages:

- **Physical plausibility.** No matter if the neuronal activity is based fundamentally on pulse frequencies, or chemical concentrations in synapses, the only structural constraint is that such signals *cannot be negative.*
- **Theoretical applicability.** The model is piecewise linear; sparsity is facilitated when negative variables automatically are inactivated, and do not affect the locally linear dynamics.
- **Pragmatic benefits.** There are no adjustable parameters; in larger systems, this considerably simplifies the tuning of the model behaviors.

The sparsity introduces new possibilities in the model structure. Experiments are presented in [11]: Hand-written digits are modeled, etc. However, from now on the approach differs here from that presented in [11]: There the selected approach was plagued by theoretical and practical problems. The goal is now to keep the iteration structure as simple as possible, only extending the linear model. For that purpose, define the extended nonlinearity $f : \mathcal{R}^n \to \mathcal{R}_+^{2n}$ as

$$f(x) = \begin{pmatrix} f_{\text{cut}}(x) \\ f_{\text{cut}}(-x) \end{pmatrix}. \tag{34}$$

In fact, this function f represents *feature extraction* — how a system is seen by the outside world. When this nonlinearity is added in the output mapping of the system, so that previous layer of signals is mapped onto the next layer as

$$u_{\text{next}} = f(\bar{x}) = f\left(A^{-1} B\, u_{\text{prev}}\right), \tag{35}$$

it is evident that the nonlinear model is simply an extension of the linear one: Assuming that there exists a linear mapping from the linear system variables x to some other set of variables y, so that $y = c^T x$, the same mapping can be implemented also using the nonlinear variables because there holds $y = \left(c^T \middle| -c^T\right) f(x)$. If the actual underlying mapping is nonlinear, the parameter weightings in the approximated piecewise linear mapping typically become

unsymmetric. In both cases, the parameters can be solved applying linear regression techniques[3].

This expansion of the variable space, or changing the n variables to $2n$, corresponds to the multivariate statistical approach to attacking nonlinear mappings. The structural complexity is changed into dimensional complexity: Hoping that in the higher-dimensional space the problem is monotonous and better linearizable, the efficient tools of linear algebra can take care of the remaining problems of redundancy compression (and, indeed, the PCA nature of the Hebbian/anti-Hebbian network is well suited for this compression task).

Adding the nonlinearity to the output of a single system makes it possible to introduce nontrivial multi-layer systems. Because of the positivity of the subsystem outputs, they cannot naturally be made zero-mean; this means that the variables in the sparse basis typically remain correlated.

2.9 Networks of neuron systems

Intuitively, it is evident that in a neuronal system, there are subsystems at many levels. Successive (or strictly hierarchic) neural layers are not problematic: The latter just uses the outcome of the former ones, the former ones remaining ignorant of the subsequent processing phases. Theoretical problems emerge if the former layers also refer to the latter ones, so that the hierarchy changes into a network.

Indeed, intuition tells us that the structure of a complex system is somehow *fractal,* so that the same kinds of substructures are found in different scales — and, somehow, the neuronal functionalities gradually change to cognitive ones when higher levels are studied. It would be nice if the same modeling principles could be applicable at all levels. Above, the synaptic systems within single neurons, and the neuron grid itself, were seen to constitute a subsystem/supersystem pair having the same model structures; how about higher-level structures? When going from the synaptic level to neuron grid level, the data structures changed from scalars to vectors; in the same way, one could assume that on the higher level, *tensors* are needed. However, in this context, simple methods are preferred, and it is assumed that standard matrices can still be applied also on the higher levels. Whereas the *system semantics* can become infinitely complex, the *system structures* remain the same.

Where does to added expressive power come from when lower-level structures are connected? The answer here is, of course, that the structures are not linear. It is not only through explicit nonlinearities, as in Sec. 2.8, but also through implicit structures that give rise to nonlinearities. In the above cases, the networks between neurons were fully connected, but there can be different sets of variables for subdomains, structural constraints affecting the emergent structures

[3] These methods typically necessitate covariance matrix invertibility; now, on average, half of the variables are zero for any given input, and some variables can *always* remain zero, meaning that the covariance matrix becomes singular. These cases need to be taken care of, for example, always adding a small constant ϵ in the diagonal elements of the covariance/correlation matrix

(see [23]). Typically, in a neural system, there are rather sparsely connected networks coupling individual densely connected subsystems, and it is reasonable to retain such functional decomposition also in the model. However, each subsystem is highly iterative; how to achieve *pragmatic,* practical models?

In multi-level systems, the inner time scales are shorter, the inner dynamics being faster. When seen from the higher-level system, the lower-level system looks like a static mapping, the inner-level dynamics being abstracted away. The inertia in the dynamic nonlinear system is unavoidable, and when there are many layers, time scales become difficult to master; what is more, convergence can take a considerable amount of time. However, in systems with such a simple underlying linear structure, different kinds of enhancements are possible: Higher-level phenomena can be modeled using lower-level tools, collapsing the structures, and the whole structure of different time-scale dynamics vanishes to a singularity, as explained below.

When there are various such sub-blocks with different sets of input and output variables, how to represent the overall system in a compact form? And, what is more acute, when neural sub-blocks are connected together, there again emerges the stability prolem. It seems that the most practical way to accomplish this is to collect all variables in all x's in a single vector χ, stacking them on top of each other, and collect all completely independent inputs in another vector μ; and, again, the stability problem can be solved applying negative feedback. The model between vectors μ and χ can be expressed in the form

$$\frac{d\,\chi}{d\,t}(t) = -\mathbf{A}\, f(\chi(t)) + \mathbf{B}\,\mu, \tag{36}$$

starting from $\chi(0) = \mathbf{0}$. Matrices $\mathbf{A}$ and $\mathbf{B}$ reflect the network structure, being sparse if the network is not fully connected. Matrix $\mathbf{A}$ is an $n \times 2n$ matrix and naturally uninvertible; dynamic process is thus needed to determine the steady state $\bar{\chi}$ — but just one dynamic process is involved instead of many fractal ones. Function f transforming outputs of previous levels to inputs for later (or same level) layers is defined in (34). Whereas in the linear case the data structures are determined by the principal subspaces that can be solved explicitly, in the extended nonlinear case the correlation matrices cannot be determined without iteration. What is more, the resulting correlation structures depend on the order of introducing new layers: The structure has to be trained gradually step by step, in a "constructivistic" manner. Because of the (different kinds of) nonlinearities, the dim of χ can become higher than that of μ.

The simple structure in the model (36) is deceptive. There is no hierarchy fixed in the structure, the structure being homogeneous and flat, but within the correlation matrices the dependencies can be explicitly maintained. If one wants to retain the known *a priori* structure, assuming it is known, the elements in the correlation matrix $\mathbf{A}$ corresponding to no connection can be explicitly zeroed. Indeed, this issue needs to be elaborated on: It is also the hierarchy in the network that is dictated by the correlation learning strategies.

Remember that it was the Hebbian learning rule that resulted in positive feedforward and emergence of structures, and anti-Hebbian rule that resulted

in negative feedback and stabilization, and self-organization of structures. This intuition can be extended: The connections from lower layer to a higher one follow the Hebbian principle, and connections among the same layer units, or from higher layer to lower ones follow the anti-Hebbian principle. This means that depending of the ordering among neurons, adaptation is either positive or negative; this can be formulated in the correlation matrix estimate calculation as

$$\frac{d\,\hat{\mathrm{F}}\{\bar{\chi}f^T(\bar{\chi})\}}{d\,t}(t) = -\lambda\,\hat{\mathrm{F}}\{\bar{\chi}f^T(\bar{\chi})\}(t) + \lambda\,K \odot \bar{\chi}f^T(\bar{\chi}). \tag{37}$$

Symbol "F" is used instead of "E" to emphasize the structure modifications. Above, operator $\odot$ denotes elementwise multiplication, that is, the matrix K is a mask determining whether there is connection, and if so, in which direction the adaptation takes place. Matrix element K_{ij} is zero if neuron j is not connected to neuron i, otherwise the element is either 1 or -1, depending on whether learning is Hebbian or anti-Hebbian (in the connections from the "minus block", or the lower part of $f(\chi)$, the reasoning applies in the same way). Matrix K can be utilized also in another masking role: Densely connected subsystems (of the form studied above) carrying out principal subspace analysis do not necessarily converge; it is better to force different variables select their roles, so that unique PCA representation is created. This is reached (as explained in [11]) by making the connections explicitly unsymmetric, that is, zeroing the elements above (or below) the diagonal in the corresponding correlation matrix block using the matrix K. This means that instead of being full of ones, the corresponding subblock (on the diagonal of K) is triangular.

Combinations of active variables determine the *state* of the neural system, facilitating "on-the-fly" structures, making it perhaps possible to integrate quantitative (numeric) and qualitative (symbolic) representations in the same model framework (for cognitive concequences, see [23]).

The above discussions were still rather concrete, being grounded on the neuronal realm. In retrospect, it seems that there was need of balance (dynamic equilibrium) in synapses, and there is need of balance in the neuron grids in different levels: Applying these starting points, far-reaching hypotheses could be made. This suggests that it is *balances* that are the key issue in complex cybernetic systems also in more general terms. This starting point offers a compact structural framework for further studies. In what follows, very brave generalizations of these intuitions are presented. The question being elaborated on in what follows is the following: What is the structure of a *cybernetic basic block* when characterizing complex systems?

3 ... Continuing Top-Down ...

Above, the behavior of a neuron grid was studied from the *reductionistic* point of view. It turned out that *self-stabilization* was reached when the neurons were connected to each other and *feedback* was applied. Further, it was recognized

that if structural constraints are introduced, or if there is nonlinearity in the system, some kind of *order* emerges in the system. Is this all there is, or are there still more intuitively appealing results to be discovered? Can the results be generalized? Indeed, in what follows, a more *holistic* approach is applied. Concrete examples are presented in [23].

3.1 Cybernetic models

When looking at the obtained higher-level principles concerning neuronal behavior, it seems that the details about the system have been abstracted away. Rather than speaking of pulse trains, etc., one concentrates on the information processing level, speaking of correlations; and still higher levels of abstraction are reached, when optimality issues are addressed. One is using concepts that have nothing with actual neurons to do any more. Just as in the case of neurons, in other areas of complex systems, analogues pop up when higher level of abstraction is selected, when details are ignored, and appropriate concepts are employed.

The goal here is to find a concrete definition for what a *cybernetic system* is. A neuronal system is intuitively a very characteristic example of such a system, and the above intuitions are employed here: Neural system is, after all, one of the best understood cybernetic systems, and its subsystems (synaptic level and grid level) are easily quantifiable. What kind of general lessons have been learned?

First, there is the powerful basic structure — dynamic state-space models — that will be employed. The second objective is simplicity, constraining the effects of nonlinearities to minimum. In what follows, the most important assumption concerning cybernetic systems is the emphasis on *equilibria:* At each level in the neuronal system it is the steady state after transients that was of relevance, and this stabilization at various levels seems to be the key point in a cybernetic system.

Now we are ready to define the "cybernetic standard systems". There are three levels of such models[4]:

1. **First-order cybernetic systems** can be represented in the form

$$\frac{d\,x}{d\,t}(t) = -\Lambda A\,x(t) + \Lambda B\,u, \tag{38}$$

with the system output being defined in the steady state as $f(\bar{x})$. Because it is this fixed state that is mainly of interest, the model could easily be formulated also in discrete time. It is assumed that the matrix A is stable, and the internal dynamics of the system is much faster than the changes in input u. This model describes a simple balance model of dynamic equilibria; only the most concrete adaptation, or reacting to environmental phenomena, takes place in such a system. In a way, any stable state-feedback control system is cybernetic in this sense.

[4] The formulations can be somewhat modified without affecting the basic functionalities, as explained in Sec. 2.6

2. **Second-order cybernetic systems**[5] are first-order systems where

$$A = \mathrm{E}\{\bar{x}\bar{x}^T\} \qquad \text{and} \qquad B = \mathrm{E}\{\bar{x}u^T\}. \tag{39}$$

This structure means that there is not only the evident balance, but also the *second-order balance* among the statistical properties of the system. The system spans the principal subspace of the input data. As compared to control engineering systems, this model corresponds to an adaptive control system with special goals of adaptation algorithms. **Higher-order cybernetic systems** have the same form as the second-order system, but the correlation matrices can be deformed to reflect the structure among subsystems (elements corresponding to missing connections are zeroed).

3. **Enhanced/optimized cybernetic systems** are second-order systems where

$$\Lambda = \mathrm{V}\{\bar{x}\bar{x}^T\}^{-1} \qquad \text{or} \qquad \Lambda = \mathrm{E}\{\bar{x}\bar{x}^T\}^{-1}, \tag{40}$$

respectively, where the variance matrix $\mathrm{V}\{\bar{x}\bar{x}^T\}$ only contains the diagonal of the correlation matrix. Seen from outside, the steady-state behaviors of such systems do not differ from those of the second-order systems, but the transient behaviors are more streamlined and better applicable to practical implementations.

Indeed, a truly cybernetic system can be characterized in terms of balances: Or, what is more, the goal of a truly cybernetic system is a *balance of balances,* or a *higher-order balance* with the environment; or, just as well, it is a *higher-order match* with the environment, as measured in terms of correlation structures.

A truly cybernetic system is balanced not only on one level. Qualitatively, when going from a lower-level balance to studying the higher-level balances, nothing new takes place. It is just the time scales, etc., that are different: When the previous-level phenomena are studied statistically in a wider perspective, searching for balance at that higher level, new emergent properties pop up.

The function f in the models is either identity mapping (strictly linear system, where f can be ignored altogether), or the cut function with augmentation (different kinds of function forms could also be proposed). When model is developed for a locally linearized behaviors and small deviations from the nominal state or the linearization center, the linear model is appropriate, whereas when constructing global models, the nonlinearity is needed. Note that when a model is linearized, the resulting model is generally *affine;* it is here assumed that the variables are not zero-mean, and there are excessive variables, so that the constant terms need not be explicitly included in the model.

[5] Note that this has not very much to do with studies of "second-order cybernetics" (by Heinz von Foerster), where the higher level denotes the *observer system,* also being a cybernetic system. Such studies where separate systems are explicitly mixed together easily result in "cyber-semiotics" and other esoteric studies; now there is no need to introduce Heisenbergian-style uncertainties ("observer disturbs the system"), or there is no need for infinite regress; the studies can be kept mathematically concrete. On the other hand, this mathematical basis results in a new platform that facilitates studies on the Foersterian second-order cybernetics, too (see [23])

In the higher-order systems, the scaling of the variables is of crucial relevance. Due to the role of the correlation matrices, the models are by no means unique: Because the underlying model constructs are the (sparse coded) principal components, variation levels of variables in u (or, for uncentered variables, the average deviation from zero) dictates how visible those variables will be in the results. Typically, the absolute values of the input variables should be directly applied, if the inputs are equally valuable as resources (see later); however, if there is no *a priori* knowledge about the relevances of very different types of variables, the best initial guess is to normalize them, or scale the variables to have identical variation range. Another important issue is the selection of variables, of course: Different sets of variables result in different models. Specially, omission of some relevant variables may make the model misleading, however cybernetic it may be.

In technical terms, it is perceptron-like basic blocks that could be duplicated to implement complicated systems for modeling of signals. It is an open question how such correlations-based adaptation of weight factors could be implemented on silicon; in software, however, such modules can readily be implemented, and such a framework is being developed (in Matlab/Simulink environment).

3.2 "Humble systems"

The system level studied in the previous section is an emergent level, not visible in the actual lower level. As experienced at the local level, what is it like to be a member of a cybernetic system? After all, it is these individual agents that autonomously implement the emergent functionalities of the system: *How can they do it, how do they know what they are expected to do?*

One has to generalize intuitions gained when studying the Hebbian neurons. First, it needs to be recognized that the network structure is just the framework to facilitate communication among the neurons: When all neurons are connected to each other, the network paradigm loses its intuitive appeal and explanation power. The all-or-nothing style of thinking about connections in a network is misleading, because the phenomena are not qualitative but quantitative. A more appropriate way is to forget about the physical connections and study the system in an information theoretic way; the grid of neurons changes into a *population of neurons.* This view can be generalized to more abstract environments.

From now on, the data structures can be given new interpretations: The vector x represents any *population,* vector elements being *activities* of the individuals, whatever is the realization of these activities, and u is the vector of *resources,* whatever these resources are. Elements of x and u are real-valued scalar signals, even though the signal carriers can be nerves, chemicals, or ants (see later). It is assumed that relevant phenomena (activities) can be quantified in a scalar form, and, similarly, interactions among the actors can be expressed as scalar functions of the activations. The elements of x can be activities of individual actors, or, in other cases, they can be the activities of a group of identical actors. It is assumed that the actions of actors are local, behavioral decisions being carried out independently by individual actors.

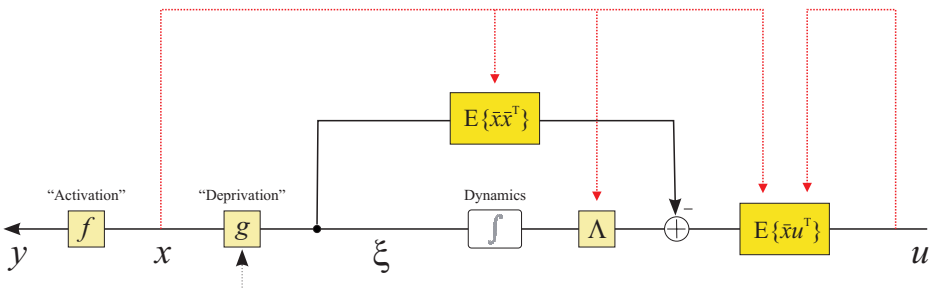

Fig. 2. General structure of a cybernetic (sub)system

Often, the cybernetic agent can be characterized as shown schematically in Fig. 2. According to traditional terminology concerning neural systems, the function f can be called *activation function*, mapping from internal state $\bar{x}$ into the system output $f(\bar{x})$, as determined by the physical constraints. In the same spirit, g can be called *deprivation function* mapping the agent's internal activity $\bar{\xi}$ into $\bar{x}$, representing the actual contribution of the individual in the overall system, being constrained by other subsystems. This function offers a mechanism for integrating separate subsystems.

The (sub)system tries to affect the environment by modifying $\bar{x}$, but it does not necessarily succeed; in this sense, the system has two inputs, u and $\bar{x}$, and the role of the cybernetic agent is to get along there in between. Variables in u are such that nothing can be done about them, whereas variables in $\bar{x}$ are such that something *perhaps* can be done (as restricted by the function g). Only if no compromises are needed (g vanishing), the optimal case characterized by the eigenvectors, etc., is reached already on the lower level. The topmost system with no constraints can reach the optimum, not the subsystems; one could speak of *exocybernetic* systems, when the goal is given from outside, and of *endocybernetic* systems, when everything is self-contained.

If one assumes *homogeneity* of the actors, that is, if one assumes that the actors in the system are equal, the effects of individual actors can be directly summed; similarly, (appropriately scaled) resources are assumed to be additive. This means that the signals in the system input can be assumed to be linear. The linearity assumption means that the connection weights among variables can be characterized as *profiles* in a vector form. In concrete terms, competition is harsher if the actor profiles are similar. In a (higher-order) cybernetic system, the key property is that these profiles are adjusted in a consistent way. The compromising, or balancing between demands is characteristic to a cybernetic system; it is the way to reach coordination among subsystems without central control. In concrete terms, The behaviors of individual actors, or "cybernetic agents" are governed by the following principles:

– **Strive after resources:** Better availability of resources results in increasing activity (excitation).

– **Strive against competitors:** Higher number of competitors means less available resources resulting in decreasing activity (inhibition).

In a word, it seems that a cybernetic agent is simply *sensible.* There is egoism — you try to prosper — but most of all, there is realism: When you are too weak and when you are alone, you have to retreat (truly, the agents are alone to start with — note that cooperation is an emergent higher-level phenomenon). The agent tries to do something; if it does not manage (remember function g) it adapts to the reality. In this sense, cybernetic systems are "humble systems", as contrasted with the man-made *arrogant* ones that apply excessive power to carry out actions. The opportunism can also be interpreted in other ways: Cybernetic systems conforming and adapting to their environments, whatever the properties of that environment happen to be, could be called "chameleon systems". A philosophically motivated way to characterize the essence of such systems that always strive towards something or against something (more or less hopelessly) would perhaps be to speak of "Schopenhauer systems".

Adaptation to the environment means that the connection weights are updated in a "locally consistent" way. Local consistency here means that behavior is locally adapted towards the direction where there seems to be resources easiest available and where there is less competition, in a more or less consistent manner. It does not matter very much what are the details of the adaptation process; seen in the global scale, the resulting steady state is assumedly always the same (at least in the linear case). One can ignore the dynamics of the complex stochastic process, and directly study the situation where the system inevitably ends in, assuming stationarity of the environment and long enough time span. The details of the process still determine how the different "roles" in the environment are distributed among actors, so that, as seen from the local rather than the global perspective, the dynamics makes a difference.

3.3 Elegance through minimalism

It is the same low-level strategies that have to be shared by all of the actors — otherwise the emergent functionalities do not pop up. In all environments, there always exist more than one strategies that can be employed — why should one assume that it is the above scheme that is followed by all the agents in a complex system? And, specially, what is the justification for generalizing the assumed principles over all very varying cybernetic systems?

Indeed, it needs to be assumed that there have existed various strategies in sub-cybernetic systems. The subsystems with different strategies have competed for survival, in the Darwinian sense, and the fittest has become dominant. The question should also be posed as follows: What kind of evolutionary advantage does the presented scheme have? The answer here is very simple — a higher-order cybernetic system structure as presented above is *optimal.* This global-level optimality that is fundamentally based on the observations in Sec. 2.5 can be paraphrased in different ways, because in different environments different kinds of semantical entities are relevant. In general, the presented system structure as

a whole exhausts the resources available in the environment in an optimal way. The system as a whole operates applying minimum amount of effort. Simultaneously, robustness against variations becomes optimized (see [23]): Cybernetically "less mature" systems are (on average) less prepared to environmental surprises, whereas in an optimized system the outside disturbances are compensated as efficiently as possible by the system. This could also be expressed in information theoretic terms, if information is measured in terms of variations from the expected values.

The optimality pursuit applies to subsystems as well to the overall system — in linear system the problems can be decomposed and analyzed separately. An intuitively holistic system *can* be distributed and studied in parts. Cybernetically constructed system of cybernetic subsystems is itself cybernetic. The successive layers of cybernetic subsystems have the same form, so that a fractal hierarchy of similar abstract structures is spanned. Each of the subsystems is optimal, minimizing the efforts for achieving the goals; this fractal optimality pursuit is the key to the marvellous elegance of Nature.

A crucial point in the mathematical derivations above was the assumption of system linearity, at least on the lowest level. However, it is well known that real systems are fundamentally nonlinear, and observing the very different domains where there are cybernetic systems, this problem becomes acute. However, here it is assumed that *Nature applies the imperfect machinery it has available in such a way that linear behavior is approximated.* It is claimed that in *truly* cybernetic, or "interesting systems" with many interconnected cybernetic sublevels, linearity *can* be assumed. Again, the motivation is based on evolutionary considerations[6]. As seen from the Nature's point of view, trying to optimize complex systems, linearity has an evolutionary advantage. If the structures were nonlinear, different optimization strategies would be needed in all levels of the systems; the Universe is simply not old enough. If linearity applies, solutions scale up, and the same strategies can be applied in all levels.

If the above assumptions truly are motivated, powerful predictions about the system behaviors also beyond the neuronal realm can be made. The system carries out *principal components oriented sparse coding* of the resource variations, offering views for understanding large-scale systems as seen from *above.* The methodology extracts the underlying structures beneath the observations, no matter what the application domain is: The cybernetic framework is the same for a wide scale of systems. This is, indeed, near to the promises that have been hypothetized in complex systems research community.

4 ... Putting Inside-Out

In what follows, the above cybernetic intuitions are followed to the extreme, extrapolating the ideas beyond their nominal validity area. Because of the prob-

[6] Another point that needs to be noted that feedback systems virtually *linearize* smooth nonlinearities

lems when trying to quantify phenomena in very abstract domain fields, only rather loose interpretations can be made.

4.1 Economies and ecologies

There has been some activity for applying system theoretical principles in the analysis of economical environments; however, the models derived using the ideas of System Dynamics (see [20]) are typically rather heuristic, the large number of model parameters being determined more or less intuitively. There never exists enough data to uniquely determine the high number of parameters. A more consistent approach to modeling of business dynamics would be invaluable.

For example, assume that there exists a selection of industrial enterprises within some branch, or *niche.* Such a system is a prototypical example of the above type: Together the companies fulfill the demands of the market, while competing against each other. The activity of a company can be measured in terms of its annual turnover. What is more, such an economic system not only strives towards balance, they also try to reach the higher-order balance:

1. The companies explicitly optimize their production based on the market demands and competition[7].
2. The similarity of companies can be measured in terms of similarity in their production profiles.

This means that the "match of matches" interpretation is applicable now. However, the similarities in production can be difficult to quantify; for the purposes of determining the model parameters, a still simpler idea can here be applied:

> The similarities among companies can be measured in terms of *co-activity;* if the companies have similar activity histories, having been operating in the same environments and experiencing the same boundary conditions, their profiles have assumedly become similar (assuming consistency of actions).

This motivates why the correlations in activities are used as a measure of similarity — correlations are emergent key figures, reflecting the underlying realm and abstracting details away. Remember that because one can extend the model, as presented in (15), the correlations are needed only for determining the trend directions in activities, and numeric values are irrelevant from the systemic point of view. From the point of view of an individual company, the numeric values are relevant, of course: It is well known that the difference between stable and

[7] Actually, companies just try to maximize *profit.* The hypothesis here is that (assuming that x_i's have the unit of *money*) the Hebbian adaptation rules (7) and (8) accomplish this peek-a-boo optimization, and the vector x reveals how the available money Σu_j is redistributed among competitors. — On the other hand, the presented model "proves" the heuristic claim that free market economy where the companies behave in a selfish way is also optimal for a customer (indeed, free market economy is *more cybernetic* than the Soviet-style centrally controlled economy)

unstable behavior is in the parameter values. Indeed, closer studies of adaptation factors λ_i may cast some light on the heuristic notion of *edge of chaos:* Fast reactions (high λ_i) may give a company the competitive advantage, but too fast adaptations result in a catastrophe, when the activity starts following some random fluctuations in the markets. If the company profile drops out from the principal subspace describing the market, it is a catastrophe for the company, but not for the system: The system feels no pity for individuals, the more conservative companies soon fill in the gap. What is a good rate of adaptation — this is dependent of the *signal-to-noise ratio* in the market (a rule of thumb: Acquired correlation structures should be forgotten at the same rate as the acquired information becomes obsolete). This kind of studies of adaptation rates may make it possible to answer some paradoxes: Why is it so that in some systems of interacting agents (like gases, for example), energy is wasted to increase entropy, whereas in truly cybernetic systems, energy is utilized to *decrease* the entropy level.

Following the cybernetic intuitions, a company can, for example, differentiate its activity *dynamically* rather than statically in its environment. The company can perhaps optimize its behavior in its environment, not by trial-and-error, but once-and-for-all. In any case, different scenarios can be easily simulated. Changing ones behavior also changes the market, changing the boundary conditions, so that successive optimizations are needed.

Just in the same way, in ecological systems, plants in the same area share the same resources. Indeed, in the case of an ecosystem, one is modeling *populations of populations,* so that the activity of an individual species is characterized by the number of individuals. Together all these plants exhaust the available resources (soil, light, water, etc.), and, similarly, correlations in their temporal/lateral distribution reveal their similarities. The herbivores, on the other hand, share the resources provided by the previous level, the plants, and carnivores share the production of this intermediate level. In this sense, a hierarchy of models is formed; the trophic levels constitute a succession of (more or less mixed) cybernetic systems.

In Figs. 3 and 4, a single population of bacteria is simulated, x denoting biomass, assuming that there is a step change in substrate u from 0.01 to 1 (dotted line). The one-species "enhanced" model with no nonlinearities can be written as

$$\begin{cases} \frac{dx}{dt}(t) = -x(t) + \frac{b(t)}{a(t)}\, u(t) \\ \frac{da}{dt}(t) = -\lambda a(t) + \lambda x^2(t) \\ \frac{db}{dt}(t) = -\lambda b(t) + \lambda x(t) u(t). \end{cases} \tag{41}$$

In Fig. 3, the behavior of the "cybernetically optimized" population is shown. This does not truly look familiar, and it is evident that the standard model is not directly applicable: In a completely cybernetic system there is no inertia included in the structures, whereas in practice the growth rate is limited, a population can multiply only following the exponential growth curve. Additional dynamics emerges from the fact that the number of offspring is related to the number of parents.

Dynamics can be modified, for example, by ignoring the extra weighting, so that the results of Fig. 4 are obtained if one modifies the first part in (41):

$$\frac{d\,x}{d\,t}(t) = -a(t)x(t) + b(t)u(t). \tag{42}$$

There exist a plenty of different kinds of growth models (Monod, Lotka–Volterra, etc.); typically their behavior is exponential in the beginning, stabilizing when the constraints of the environment are met. These principles are obeyed also by the latter model: Looking at the steady state, there is the positive factor of the form $u^2\,x$ (assuming that λ is large, and u is constant), resulting in exponential growth, and the constraint is essentially of the form $-x^3$ (compare to logistic model, where the limiting factor is only quadratic)[8]. As compared to the standard growth models, there is an interesting difference — it seems that now there always is an *overshoot* before stabilization.

Another major difference in the new model as compared to old ones is that there is only one tunable parameter, the forgetting factor λ now having the interpretation as growth factor. If there were various species, it would be natural to modify the original formulations (7) and (8) so that each matrix row would have its individual adaptation rate, scalar λ being substituted with diagonal matrix Λ, as shown in (16) and (17). In such a case, matrix A could temporarily become non-symmetric during adaptation (steady-state solution still remaining intact).

It needs to be recognized that adaptation in economical and ecological domains is not as smooth as in the case of neuronal systems: In economical systems, modifications of business strategies are based on more or less random distinct decisions, and in ecological systems, adaptation is based on genetic changes, being an equally discontinuous and spurious process. However, in the long run, such stochastic optimization still wanders towards optimum — and, assuming that the system is cybernetic, this final state is essentially unique. Unfortunately, if there is too much "intelligence" in the system — that is, if there is some master mind trying to globally optimize the company behavior ("let's concentrate on our main business!") — the cybernetic balance in the system can be disturbed, and the analyses become void altogether. The harsh reality sooner or later assumedly eliminates such non-optimal anomalies, but this can take a long time.

Of course, the basic model framework can be extended if there exists some *a priori* information available. For example, modeling the age distributions in a population, letting each age class have a separate variable of its own, makes it possible to explicitly model transitions between the classes. The presented model structure only helps in capturing the large numbers of interactions among and within classes in a compact framework.

[8] In the simulation it is assumed that adaptation rate λ in the population is rather fast, and the covariance structure is not in steady state; if this were not the case, if there were no adaptation in the population of bacteria, so that the (co)variances a and b were kept practically constant, the curve forms would be qualitatively very different, the biomass approaching the final value exponentially as a first-order linear process

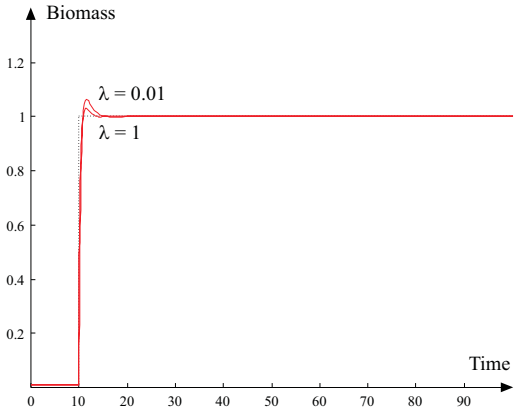

Fig. 3. Step changes in substrate level: "Completely cybernetic" system

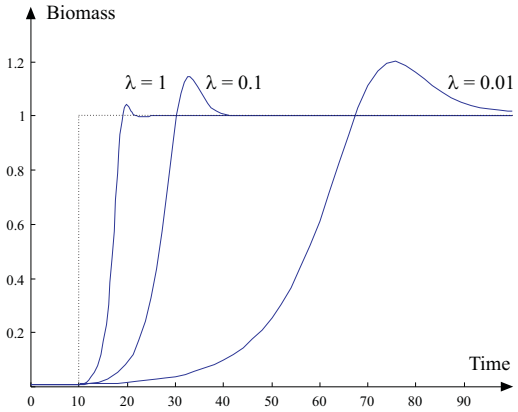

Fig. 4. Step changes in substrate level: Growth rate adjusted

4.2 Further domains

The above examples were rather straightforward, because the optimality assumption was more or less natural, and the role of correlations could be motivated. However, there are different kinds of cybernetic systems where the basic model structure still is applicable, even though the optimality cannot necessarily be assumed (and predictions about matrices A and B cannot be based on observed u alone).

The more abstract the domain field is, the more qualitative are the results of cybernetic analyses. In narrower fields, quantitative and more concrete analyses can be carried out. This applies also to *synthesis* of systems: Concrete engineering applications of the methodology are presented in [23].

Let us briefly study some examples of different kinds of populations. The underlying realms and the physical laws are very different, but on the more abstract level, general principles seem to exist (as has always been assumed in the field of complexity research). Quite concretely, the society of individuals extends the capacity of an individual, and it is a very tempting idea that the

model structure remains the same on all levels — it is only the signal carriers that are changed.

- **Genetic system** consists of a population of genes within a nucleus; the genes compete for the *transcription factors,* excitation of a single gene being a weighted sum of excitatory and inhibitory effects. The Dawkins' "selfish genes" are not competing only at the population level, but also within each individual, and within each cell, where the feedbacks are much more imminent and concrete. On the other hand, the lower-level metabolic processes can also be studied in the same balance framework (see [12]).
- **Ant colony** is based on distributed operation of individual ants; the control of ants is carried out by spatially distributed *pheromone levels.* Individual ants are more like mere signal carriers, determining a more or less stochastic spatial distribution, whereas the society itself is the living organism in this perspective. The local activation, or the ant frequency, is proportional to the pheromone gradient along that path, and it is the pattern of such activations where the ant colony functions are manifested: Different tasks and paths compete for activation, or ants.
- **Project team** is an example of another kind of society that is explicitly directed towards some goal; in that sense, analysis of such societies is simpler, assuming that the goals can be appropriately quantified. The seemingly optimal "assembly line style" of division of labour — everybody specializing on a single task — is not robust; the workers' ability profiles should be cybernetically matched against the dimensions in the "task space". Larger human societies (associations, companies, etc.) can also be modeled using the same model structures.
- **Democratic society** is a straightforward instantiation of (implicit) cybernetic ideas: Political parties compete for popularity, trying to differentiate between their profiles, and the resources, or given votes, are distributed among them accordingly. This offers yet another perspective to explaining why democracy is so successful despite all its shortcomings (inefficiency, etc.) — it can be claimed that the democratic system is the best possible approximation of a cybernetic one, assumedly being robust and reacting relatively fast to changes in the environment.
- **Memetic system** abstracts the above view: Ideas share the common infosphere, similar ones (ideologies, religions, etc.) competing with each other in human minds, having varying profiles what comes to their capabilities of explaining observations. Specially in science, where one tries to be objective, so that "goodness" of explanations can be explicitly quantified, shifts between theories and paradigms can be abrupt when time is right. The self-correcting nature of science is that of a cybernetic system. And within a single language, concepts (words) compete with each other, each having its own profile of connotations ... There is an infinite number of cybernetic subdomains within the memetic world. Whereas *free will* is not necessarily only a myth, there are very stringent laws ("statistical ethics") that have to be followed for an individual to prosper.

It needs to be noted that the proposed cybernetic model formulation is *information theoretic,* being applicable only in populations where there locally is *complete information available.* For example, the distances between actors, or the actual locations of the actors, was not assumed to affect their mutual interactions. In this sense, "populations" of elementery particles, or atoms and molecules, cannot (directly) be studied in the above framework, even though it is again interactions and feedbacks that determine the individual behaviors also in such systems.

4.3 New tools and intuitions

In [22] it is claimed that totally New Science is needed to attack the complex systems. On the contrary, it may be that age-old tools only are needed — they just need to be applied in new ways, and new interpretations are needed.

How to attack *emergence* that by definition is beyond the capacities of reductionistic science? How to reach a qualitatively new level? Here, one can proceed one step at a time. Whereas signal realizations are actual, statistical parameters are a step towards more abstract system characterizations. Another step is taken when these statistical quantities are studied statistically. If there is an *infinite* number of such levels in this continuum, something qualitatively new can pop up. Essentially, in (11), statistical expectation operators are applied recursively, taking "expectations of expectations" indefinitely. To emphasize the shift in thinking, the expectation operator E could be substituted with the "emergence operator" $\mathcal{E}$.

Multivariate statistics, when applied appropriately, is an efficient way to study cybernetic systems. For example, in *foraging theory* [19] rotated principal components have been applied for modeling forage profiles in ecosystems. Now these profiles are the columns of ϕ. Rather than being only analysis tools the statistical phenomena can be thought to convey some essence of the system. For example, the number of species in trophic levels can be estimated by studying the variability structure in the previous level.

Indeed, multivariate statistics offers not only practical but also *conceptual* tools for attacking cybernetic systems. For example, assume that an economical system is being studied; because of the turmoils in economy, time series data from the past is not comparable to current data (the markets expand or contract, old companies get bankrupt and new ones are founded), and there is no possibility of finding statistically credible models. However, utilizing statistical intuitions, assuming that the system is fundamentally *ergodic,* that is, statistical properties over time and over individual realizations are equal, one can collect simultaneous data from different markets (from different countries, say) and apply modeling to that spatial data instead of temporal data. Whether or not an economic system truly is an "ergosystem" is another interesting issue[9].

[9] For example, it has been claimed that as there is continuous evolution in an ecosystem, no stationarity assumptions are justified; however, for some reason, it seems that there are long periods of balance between the rapid developments

The central role of simple (co)variances in the new framework also suggests natural extensions: The mapping between u and $\bar{x}$ above was assumed to be static (x having reached its steady-state value after transient in u) — however, also in identification of dynamic systems, different kinds of statistical correlation structures play a central role. If one defines a *dynamic agent* so that it not only operates on instantaneous signal values but also has *memory,* its history and past activities being taken into account, one can define a formula for *power spectra* in frequency domain:

$$\bar{X}(\omega) = |H_{xu}(\omega)|^2 \, U(\omega). \tag{43}$$

This means that one can define formally similar-looking expressions for truly "cybernetic dynamics" as shown in (4): Matrices $\mathcal{A}$ and $\mathcal{B}$ corresponding to A and B are now matrices of auto spectrum and cross spectrum functions, respectively (introducing yet another level of statistical quantities in the analyses). This means that the changes in the dynamic behavior (as revealed by the spectra) become modeled. Updating the matrices has to be carried out in a very different way, of course, but because of the algebraic nature of the Laplace transformed signals, involved convolution integrals are avoided, and, essentially, simple frequency-wise calculations are enough to determine changes in $H(\omega)$. Note that positivity constraint (cut nonlinearity) is again appropriate when studying power spectra. For example, the spectrograms studied in voice analysis could benefit from such analyses.

The presented approach — determining both x and ϕ, or the state and the structure together when only the input data sequences u are given — is one form of *blind source separation.* Finding structure beneath the observations is a huge (philosophical) challenge, and some assumptions are needed to make it possible. In *Independent Component Analysis (ICA)* [10] this additional assumption is that the original underlying signals are maximally non-Gaussian; now, on the other hand, the assumption is that the underlying system is *cybernetically balanced.* Indeed, one could speak of *Cybernetic Component Analysis (CCA).* Similarly, in traditional control engineering different kinds of *canonical representations* play a central role; corresponding to *balanced realizations,* one could define a "cybernetic realization" between input u and output y, where the state vector is selected as presented above.

4.4 Technical applications

Today, there is great need for analysis and design methods for different kinds of networks and distributed agent systems. This is clear in environments like Internet; but also in social systems, for example, one would need tools for analysing networks where individual actors have differing ability profiles. Agent systems consist of software architectures with no systemic theories. The resulting control schemes are centrally controlled rather than truly distributed. Now, on the other hand, quantitative analysis and synthesis tools may be available. In truly distributed systems new functionalities are reached that cannot be foreseen when

centralized approaches are applied. To illustrate this, different kinds of simulations have been carried out (for closer information, see [23]).

In the first example a distributed sensor system was studied. It has been proposed that the state of distributed parameter models that are governed by a partial differential equation models could efficiently be captured by more or less randomly distributed sensors. To make the promises come true, clever sensor fusion techniques are needed. In the example, the temperature in a heated rod was to be measured. Because of the continuity of the temperature function, spatially neighboring temperature readings u_i are correlated, and this fact can be utilized to filter the measurements, having the latent variables x_j in (9). Regression from x_j back to filtered u_i is implemented as explained in [11]; the filtering scheme that was applied was linear. If the sensors are fully connected, the network carries out principal component filtering, projecting the measurements onto the latent basis determined by the most significant principal components, utilizing the correlations among measurements, and from there back to temperature readings. In a sense, this behavior is trivial, being the same as when using a centralized architecture; more interesting functionalities emerge if the network of sensors is *not* fully connected. It turns out that when only the nearest neighbors are connected, the latent variables become localized in an interesting way, and, even though information is incomplete, the resulting estimates seem to be more accurate. Such emergent properties cannot be implemented in a centralized manner; or, rather, in the distributed system there emerge properties that cannot be foreseen in the centralized framework. The distributed framework seems to give new substance to the agent paradigm.

Typically, technical implementation tasks can be formulated as optimization problems. The key question is how to formulate the optimality criteria appropriately. Assuming that the proposed structure truly is the essence of cybernetic systems, this approach gives a way to determine the optimality in a consistent way. This approach was experimented in a power plant simulation: There are m consumers and n producers, and the demand has to be balanced. Here, the original model framework had to be modified: the sum of x_i's, or the power production, has to equal the sum of u_j's, or the power consumption, at any time, and an additional optimality criterion was added to emphasize the energy production costs at each plant. When the system was simulated, consumers having more or less redundant consumption behaviors, the converged "production profiles" for the plants, or the columns of ϕ, revealed how the plants should react to behaviors of individual consumers. There were various local minima in the cost criterion, and the resulting distribution structures were dependent of the initial configuration. Again, the experiences were interesting.

It has been claimed that natural systems are more robust than the man-made ones. For example, a single fault can result in a domino effect in an energy supply system, but single collapses of some species do not escalate into ecocatastrophes. But what is this "natural robustness" in the first place? The power plant case simulations offer some intuitions:

– Globally optimized control always runs the power plants so that only one of them is active, others being in their extreme values (assuming affine cost increase between minimum and maximum, that is, for each active plant i there is additional cost $k_i x_i + c_i$ for some constants k_i and c_i). The novel scheme, on the other hand, seems to avoid extreme values. It is evident that as more plants are active, there is more buffer against sudden changes in consumption.
– Explicitly distributed systems with some fixed profiles (for example, a few plants taking care of a set of consumers) is vulnerable to domino effects: If one of the plants is out, the sudden excessive load can collapse the other ones too. Now, the profiles are based on (sparse) principal components, meaning that the profiles are (almost) orthogonal. Collapse of one plant does not excessively strain any of the other plants; rather, the additional load is distributed evenly among the reserve plants.
– Also, because the profiles are based on (sparse) principal components, the plants are insensitive against random noise (compare to principal component analysis). The plants only react to real underlying changes in consumption, probably resulting in smoother production.

From the technical point of view, the problem with the cybernetic models is that their causality structures are — by definition — deeply tangled. As compared to traditional input/output models, it is difficult to construct controllers, for example, based on such models. However, it should be remembered that the traditional models are only a pragmatic simplification of reality: When the dependencies truly are cyclic, simplified unidirectional models often result in false interpretations and questionable control strategies. What is more, causality structures cannot be automatically induced from data (remember Hume) — but "pancausal" cybernetic structures *can.*

The cybernetic models with no clear-cut causal flows cannot perhaps be used for changing the system behavior, but they can efficiently be applied for prediction, and for gaining intuition of the system behavior.

4.5 About the cognitive system

This discussion started from neurons. These neurons offer a cybernetic medium for another cybernetic system, the cognitive machinery.

The cognitive machinery is based on the neural one, and — at least, what comes to some specific mental pattern matching functions — if the neuronal level is available, it is only a scaling matter to reach some cognitively relevant functionalities. And, it truly seems that cognitive issues can also be addressed to some extent by the presented methodology. Such cognitive concepts are, for example, *short term memory* or *working memory* (vector x) and *long-term memory* or *chunks* (columns of ϕ). The field of cognitive science is far from mature; some intuitions are presented, for example, in [17] and [3].

As an example, *chess configurations* have been trained in such a cybernetic model. The configurations were coded as 768 dimensional sparse binary vectors,

and these vectors were used as inputs u one at a time. When the data structures had converged, it turned out that some columns in ϕ represented some kinds of centers ("categories") in the chess configuration space, and the other ones were used to fine-tune the patterns around the category centers ("features" or "attributes"). The non-zero entries in x (maximum number of active entries being limited by the "short-term memory capacity") revealed how the patterns were reconstructed using the data structures ("mental representations"). The learning dynamics and the recall performance of the model resembled those of human test subjects (see [23]).

Assuming that the essence of mental machinery has been captured (at least to some level), interesting perspectives open up. If a computer is given a set of data, and the appropriate modeling approaches are applied, the resulting data structures should have something in common with those mental representations that a human would construct if given (only) the same data and enough time. Even though objective reality beyond the data cannot be seen, the possibility of interpreting between the beliefs of a computer and a human makes it possible to reach "intersubjective" world views. This opens up new horizons what comes to applying the today's number crunching capacity for knowledge mining in complex environments. For example, in industrial processes smart preprocessing of measurement data can perhaps be implemented.

In any case, a search engine has already been implemented where the above views are implemented: This tool searches contextual similarities among textual documents, and models them within the cybernetic model. Vector u contains the document "fingerprint", that is, the histogram of words within the document, and, after convergence, the columns of ϕ represent the "generalized keywords" assumedly being relevant for characterizing the documents and distinguishing between them. The resulting model looks like a higher-level table of contents into the body of documents; but there may be also more interesting conclusions to be made (see [23]).

5 "Panta Rhei"

Can the above hypotheses be proven? Certainly not. But no theories can be proven.

A more appropriate way to test the claims is to study whether some specific cybernetic system really behaves as was assumed. This way, one can gain understanding to what extent the model structures are applicable, and what are their limitations. Another, more philosophical approach is to study whether the hypotheses are intuitively reasonable — do the abstractions capture the essence of what cybernetics actually is? If it turns out that the new models explain phenomena more economically than the old ones, or if new phenomena can be studied that earlier remained in darkness altogether, they are useful abstractions. Even if incorrect, the new ideas can give new intuitions when focusing and redirecting the further research efforts.

For example, have you ever wondered *biodiversity,* how the variability in ecosystems seems to flourish after all these millions of years. Why the exponential decay does not continue to extinction, why the dominating species has not specialized in being the best in all respects, wiping the losers away? The same question applies to business enterprises. Similarly, why the best ideas (memes) never seem to reach the final victory, or why the best genes are still accompanied by the more inferior ones? Why is this volatile state so non-volatile?

The presented cybernetic model demonstrates that in a dynamical environment it is not some kind of blind explosion that takes place — feedback structures stabilize the system. However, this "stability" is far from being some placid *status quo.* Indeed, it is well known that representations based on PCA explicitly try to *maximize* the variability that is observed in the input. The statistical *maximum likelihood* seems to be replaced by *maximum livelihood* in nature. The new approaches to looking at existing systems offer a novel top-down view for analysis of complex systems: Just study the environmental conditions and their variability, and draw conclusions concerning the distribution of individual actors.

It is astonishing how far-sighted the ancient Greeks were. For example, according to Heraclitus "wisdom is knowing how all things are steered by all things". Another observation by Heraclitus is the following:

Everything changes, everything remains the same.

Paradoxes, like the one above, help to see the contradictions in the everyday style of thinking. Just as the Epimenides Paradox (liar's paradox), when appropriately formalized, resulted in the famous *Gödel's incompleteness theorem,* it may be that the Heraclitus paradox helps us to see the fundamental questions underlying the cybernetic systems: The world is one unified whole which is constant yet contains perpetual change.

Whereas 42 is known to be the answer to "Life, Universe, and Everything" (as claimed by Douglas Adams), the key point is to find the correct *questions.* The Greeks have pondered the eternal questions — but we have the language the present these questions. The correct tools for formalizing ideas are offered by mathematics. Even though everything may remain the same, the memetic evolution is not a cycle; perhaps the above discussions help us to take yet one step towards *Logos,* the center of the spiral.

References

1. A. Basilevsky: *Statistical Factor Analysis and Related Methods.* John Wiley & Sons, New York, NY (1994).
2. L. von Bertalanffy: *General System Theory: Foundations, Development, Applications.* George Braziller, New York, NY (1969, revised edition).
3. W.G. Chase and H.A. Simon, "The minds eye in chess". In W. Chase (ed.), *Visual information processing* (Academic Press, New York, 1973).
4. K.I. Diamantaras and S.Y. Kung: *Principal Component Neural Networks: Theory and Applications.* Wiley, New York, NY (1996).

5. P. Földiák: Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics,* **64**, pp. 165–170 (1990).
6. P. Földiák: Sparse coding in the primate cortex. In M.A. Arbib (ed.): *The Handbook of Brain Theory and Neural Networks.* MIT Press, Cambridge, MA (2002, second edition).
7. S. Haykin: *Neural Networks — A Comprehensive Foundation.* Prentice–Hall, Upper Saddle River, NJ (1999).
8. D.O. Hebb: *The Organization of Behavior: A Neuropsychological Theory.* John Wiley & Sons, New York, NY (1949).
9. J. Horgan: *The End of Science: Facing the Limits of Knowledge in the Twilight of the Scientific Age.* Helix Books, New York, NY (1997).
10. A. Hyvärinen, J. Karhunen, and E. Oja: *Independent Component Analysis.* John Wiley & Sons, New York, NY (2001).
11. H. Hyötyniemi: Hebbian and Anti-Hebbian Learning: System Theoretic Approach. Submitted to *Neural Networks* (2004).
12. H. Hyötyniemi: Processes of Life — Towards a Unified Model? Submitted to the Finnish Artificial Intelligence Conference (STeP'04), Vantaa, Finland (September 2004).
13. S.A. Kauffman: *At Home at the Universe.* Oxford University Press, New York, NY (1995).
14. E. Oja: Principal components, minor components, and linear neural networks. *Neural Networks,* **5**, pp. 927–935 (1992).
15. J. Pearl: *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge, MA (2000).
16. B.A. Olshausen and D.J. Field: Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research,* Vol. 37, pp. 3311–3325 (1997).
17. E. Rosch: Principles of Categorization. In E. Rosch and B.B. Lloyd (eds.): *Cognition and Categorization.* Erlbaum, Hillsdale, NJ (1978).
18. H.A. Simon: *Sciences of the Artificial.* MIT Press, Cambridge, MA (1996, third edition).
19. D.W. Stephens and J.R. Krebs: *Foraging theory.* Princeton University Press (1986).
20. J.D. Sterman: *Business Dynamics — Systems Thinking and Modeling for a Complex World.* Irwin McGraw-Hill, Boston, MA (2003).
21. N. Wiener: *Cybernetics: Or Control and Communication in the Animal and the Machine.* Wiley, New York, NY (1948).
22. S. Wolfram: *A New Kind of Science.* Wolfram Media, Champaign, IL (2002).
23. *Additional material will be available in public domain in near future at* `http://www.control.hut.fi/cybernetics`.