Paper submitted Dec. 18, 2008, to **Neural Networks**

Hebbian-Style Feature Extraction

—

From Neural Systems to Neocybernetics

Abstract

This paper proposes a new systemic view to interpret and exploit Hebbian learning in linear neurons. Rather than introducing any explicit control mechanisms, a stabilizing negative feedback is implemented through *exhaustion of the incoming signals.* Even though the resulting models are essentially linear, they are far from trivial; it turns out that the model spans the principal subspace of the input data. What is more, the basis axes are rotated to implement *sparse coding.* The linearity of the structures facilitates concrete analyses and a deeper view into emergent structures. As it can be claimed that the proposed neuron structure is the *simplest possible,* wider horizons seem to open up.

Key words: Hebbian neuron, principal components, factor analysis, sparse coding; systems, control and feedback, cybernetics.

1 Introduction

Perhaps the most fundamental principles of neuronal functioning were found some fifty years ago by Donald O. Hebb [9]. Similarly, the ideas of cybernetic systems were found almost at the same time by Norbert Wiener [22]. Both of these innovations give clues about approaches for understanding complex interactions in the brain; still, it seems that this far these ideas have not truly been put together.

When the Hebbian principles were functionalized, it soon turned out that the learning rule as such is unstable. This shortcoming was circumvented by introducing different kinds of structural or functional nonlinearities; interesting behaviors in such neural networks were observed (for example, see [18], [21], [5], and [4]). However, when exploiting nonlinear models, there is an explosion of the space of possibilities, and one soon ends in a "fiddler's paradise". One gets lost in the details — but, still, the *beauty lies in the details.*

One has to select the correct details. The systems theoretical modeling understanding reveals that one has to characterize the *real non-idealities applying the transfer of real signals* rather than the artificial nonlinearities applied in the transfer of idealized information. The cybernetic models reveal that stability can be reached also through *linear negative feedback*; such feedback loop is created when *exploitation of signals exhausts them.* No phenomena in nature are based on pure information flows but there always exists the material flow. Introduction of nonlinearities, on the other hand, would limit the models to non-scalable "toy worlds".

Such intuitions give strict guidelines when pondering in which direction to search for the neuronal model. Counterintuitively, detailed analyses make it possible to reach high-level Platonian "ideals" beyond the non-ideal reality. It turns out that the general model structures are applicable also in domains beyond neural networks.

2 Another look at Hebbian learning

When looking at complex systems one observes that there is too much data and too many possible projections (explanations) of that data. To limit the search, one needs strong undrelying theories.

The operation of a natural neuron is extremely complex. However, there seem to exist some general properies characterizing their functioning — as seen from above, the neurons almost seem to aspire towards something, in a teleological spirit; indeed, the Hebbian learning principle [9] is here paraphrased as follows:

> When the neuronal activity of a neuron and the activity of an incoming signal correlate positively, the synaptic strength between them increases.

Assume that there are n neurons with activities $\bar{x}_i$, where $1 \leq i \leq n$, and there are m input signals $\bar{u}_j$, where $1 \leq j \leq m$. Then the effect of an adapted Hebbian

synapse between input j and neuron j can assumedly be characterized in linear terms as

$$\bar{x}_{ij} = \mathrm{E}\left\{\bar{x}_i \bar{u}_j\right\} \bar{u}_j. \tag{1}$$

Here $\mathrm{E}\left\{\bar{x}_i \bar{u}_j\right\}$ is the long-term (unnormalized) correlation between $\bar{x}_i$ and $\bar{u}_j$, determining the coupling between the input signal and the neuron according to the Hebbian principle. The signals are assumed to be appropriately sampled. The activity of of the linear Hebbian perceptron can then be expressed as

$$\bar{x}_i = q_i \sum_{j=1}^{m} \bar{x}_{ij}, \tag{2}$$

where the additional activation parameter $q_i > 0$ is here employed; for reasons to become clear later, this is called the *coupling factor*. The operation of the whole grid can be compactly represented in a matrix form

$$\bar{x} = Q\, \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} \bar{u}, \tag{3}$$

where $\bar{x}$ is an $n \times 1$ vector containing all variables $\bar{x}_i$, the $m \times 1$ vector $\bar{u}$ contains the inputs $\bar{u}_j$, and the matrix Q contains the parameters q_i on its diagonal. The synaptic strengths are now captured by the covariance matrix $\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}$; strictly speaking one should speak of an *inner product matrix,* as the variables are assumed to be neither mean-centered nor normalized. Regardless of this matrix formulation, *all operations in the system are strictly local,* and adaptations in a synapse can take place without any information about what happens in other synapses. This strict locality of operations applies throughout this discussion.

Hebbian neurons have been studied for a long time, but it turns out that there still exist fresh approaches. Key point here is to recognize the dynamic nature of the processes: signals and variables represent *dynamic equilibria.* This dynamics is studied in the following section; here the *fractal hierarchy of balances* is elaborated on, just assuming that those balances can be reached.

2.1 Emergence in neuron grids

Dynamic neuron models have been studied directly; however, in such models complexity becomes overwhelming [20]. One needs appropriate simplifications;

the "average neuronal activity" applied above is already an abstraction — but more abstractions are needed.

The key phenomenon in complex systems is *emergence*: something qualitatively new comes up when some kind of *infinity* or *singularity* is reached and individual lower-level phenomena are abstracted away. Typically, the concept of emergence is rather heuristic, but here *weak emergence* is defined compactly as

$$\zeta = \mathrm{E}\left\{f\left(\xi(t)\right)\right\} = \lim_{t\to\infty}\left\{\frac{1}{t}\int_{-t}^{0} f(\xi(\tau))\,d\tau\right\}. \tag{4}$$

Here, ζ is the emergent variable (vector) emerging from individual observations of variables ξ at the lower level; f is some function ($t = 0$ here representing current time). In practice, expectations reduce to averages over some finite data.

This definition of emergence makes it possible to connect successive levels of abstraction in a system, combining two time scales. Because

$$\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = \lim_{t\to\infty}\left\{\frac{1}{t}\int_{-t}^{0} \bar{x}(\tau)\bar{x}^{\mathrm{T}}(\tau)\,d\tau\right\}, \tag{5}$$

Hebbian learning is an example of weak emergence, exploiting a coupling between time axes. The transition between the formal levels will be exploited below. The fast dynamics of signals is combined with slow dynamics of the synaptic adaptation processes. The model based on covariances is defined on the higher level of statistical hierarchy; to make the correlation matrix exist, one has to assume stationarity of signals and variables. Balances are of primary importance here. Indeed, one has to study the "second-level dynamic balances" of statistical properties; this will be explained in the next section.

When concentrating on vectors x, the variables in u look static, being very slow, and when concentrating on u, the variables in x look static, being very fast. This means that when looking the system from outside, one can always apply static analysis, even though the convergence of the state is an asymptotic phenomenon. In Fig. 1, the hierarchy of time scales is illustrated schematically. The long-term behavior of u is assumed to be stochastic but stationary.

When looking at the system in a functional perspective, it can be said that the emergent functionality is *memory,* storing information of past associations among signals. Such memory constitutes a *filter* that defines how the system sees the surrounding world. Indeed, as will turn out, the memory becomes a *model* of the world.

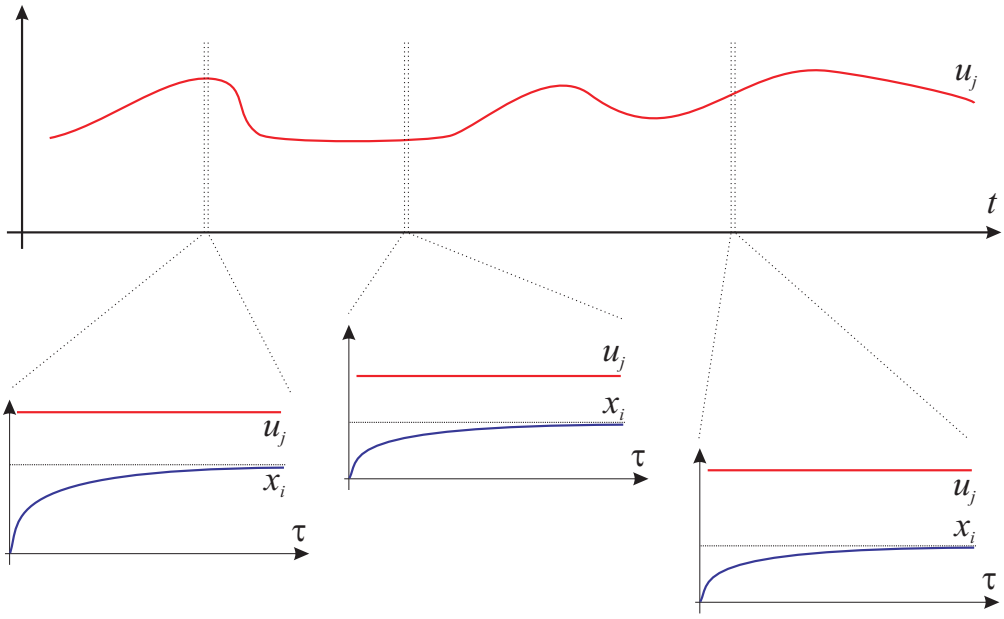

Figure 1: Illustration of two time scales. It is assumed that the dynamics of u (on the t scale) is much slower than that of x (τ scale)

From now on, assume that there is *excitation* in the data, so that there are at least n linearly independent directions in the data space spanned by the input signals. Only then the matrix $\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}$ can be invertible.

2.2 Interplay among emergent levels

To connect the level of signals and their emergent counterparts, one has to concentrate on their statistical properties. Now, one can find many expressions governing signal covariances. When multiplying (3) from the right by $\bar{x}^{\mathrm{T}}$ and taking expectation, one has the following expression:

$$\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q\,\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}. \tag{6}$$

The transpose of this gives yet another expression

$$\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}Q. \tag{7}$$

From these, it is evident that there must hold[1]

$$Q\,\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} Q, \tag{8}$$

so that also

$$f\left(Q\right)\, g\left(\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right) = g\left(\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)\, f\left(Q\right), \tag{9}$$

where f and g are any functions that can be defined in terms of matrix power series. This commutativity property means that many mathematical manipulations of the matrix data structures become very much like scalar algebra in later analyses.

Further, assuming invertibility of $\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}$, and noting (9), from (6) or (7) one has

$$I_n = Q^{1/2}\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}Q^{1/2}. \tag{10}$$

When defining

$$\theta = \mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}Q^{1/2}, \tag{11}$$

one has

$$I_n = \theta^{\mathrm{T}}\,\theta. \tag{12}$$

The columns in θ are also orthonormal. What else can one say by applying strictly formal analyses?

By multiplying (3) from the right this time by $\bar{u}^{\mathrm{T}}$ and taking expectation, one has

$$\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} = Q\,\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}. \tag{13}$$

[1]For example, $\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = \alpha Q^{-1}$ fulfills this for any scalar α; or, if $Q = \beta I_n$ for some scalar β, $\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}$ can be arbitrary. These classes of solutions help to understand the results later.

Because $\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}$ is symmetric, also Q must be symmetric, so that $Q = Q^{\mathrm{T}}$; in what follows, it is assumed that $q_i \neq q_\iota$ for all $i \neq \iota$, but the diagonality assumption of Q can in some cases be relaxed

Substituting this in (7), there holds

$$\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q\,\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}Q. \tag{14}$$

Assuming invertibility of $\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}$, and noting (9), this can be changed to read

$$Q^{-1} = Q^{1/2}\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2}Q^{1/2}, \tag{15}$$

so that

$$Q^{-1} = \theta^{\mathrm{T}}\,\mathrm{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}\theta. \tag{16}$$

This means that if ever the basic assumption (3) is fulfilled, the statistical properties of the input $\bar{u}$ are fixed. But how can the system dictate the properties of its environment? This question is the key towards extending the analyses towards general cybernetics (see later).

2.3 Relation to *principal subspace*

To understand the properties of the Hebbian neuron grids, the structure of input data needs to be studied closer. For stationary data u, one can always write the *eigenvalue decomposition* for the covariance matrix $\mathrm{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}$ as (for example, see [1])

$$\mathrm{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\} = \bar{\Theta}\,\bar{\Lambda}\,\bar{\Theta}^{-1}, \tag{17}$$

where the $m \times m$ matrix $\bar{\Theta}$ contains the *eigenvectors* of the covariance matrix, and the diagonal matrix $\bar{\Lambda}$ contains the corresponding *eigenvalues* on its diagonal. Because of the structure of the covariance matrix, all of its eigenvalues are real and non-negative, and they can be ordered in the order of descending significance, revealing the proportion of variation that is distributed in that eigenvector direction. Because of the symmetricity of the covariance matrix, all eigenvectors are normal to each other, so that when they are normalized, there holds $\bar{\Theta}^{T}\bar{\Theta} = I_m$, or $\bar{\Theta}^{-1} = \bar{\Theta}^{\mathrm{T}}$.

When data is projected onto the basis determined by the covariance matrix eigenvectors, so that $\bar{z} = \bar{\Theta}^{\mathrm{T}}\,\bar{x}$, the new latent variables $\bar{z}$ are known as *principal components.*

Now, together from (12) and (16) one can see that the columns in θ are orthonormal, and if Q is diagonal, they diagonalize the data covariance. If $\bar{\Theta}_{[n]}$ is used to denote a matrix with only n of the m columns in $\bar{\Theta}$ being selected, one can write

$$\theta = \bar{\Theta}_{[n]}\, D, \tag{18}$$

where D is some orthogonal $n \times n$ mapping matrix, $D^{\mathrm{T}} = D^{-1}$. This means that the columns of θ span a subspace of n eigenvectors of $\mathrm{E}\left\{\bar{u}\bar{u}^{\mathrm{T}}\right\}$, the eigenvalues being given by the diagonal entries in Q^{-1}. It needs to be noted here that the selected subspace directions are not necessarily the most significant ones, as measured in terms of the corresponding eigenvalues.

From (8) it becomes evident that if the diagonal entries in Q are distinct, $q_i \neq q_j$ for $i \neq j$, the matrix $\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}$ also has to be diagonal; this means that it is not only the subspace but the columns in θ are the actual principal component directions themselves. This means that the "behavioral modes" become separated from each other if they are coupled to the environment in different degrees, variation levels in variables $\bar{x}_i$ being determined by the coupling factors q_i. On the other hand, if one has $Q = q\, I_n$, all eigenvalues are *equalized,* $\bar{\lambda}_j$ equalling $1/q$, but then the principal component directions do not become separated.

The above analyses apply if such a mapping matrix exists as proposed in (3). How to avoid the excessive growth of $\bar{x}_i$ and the resulting instability of adaptation? How to supply the "intelligence" to assure the balance on the "edge between order and chaos"?

3 Facing the reality and exploiting it

It is well known that the basic Hebbian learning strategy is *unstable.* If the synapses adapt according to observed correlations, the correlating signals grow ever larger without limit. The traditional approach to fix this problem is to apply some nonlinearity: for example, the *Oja's rule* normalizes the synapse vector after each step [16].

However, following the system theoretical understanding, it is clear that stability can be assured in linear terms, for example, by applying *negative feedback.* Indeed, this idea is applied implicitly in the *subspace learning algorithm* [17] and its derivations, and explicitly in *negative feedback networks* [8]. Applying the current notation, however, the mapping matrix ϕ from $\bar{u}$ to $\bar{x}$, rather than being $\phi =$

$\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}$, it is now

$$\phi' = \mathrm{E}\left\{\bar{x}\left(\bar{u} - \phi'^{\mathrm{T}}\,\bar{x}\right)^{\mathrm{T}}\right\}. \tag{19}$$

The inner expression $-\phi'^{\mathrm{T}}\,\bar{x}$ can be interpreted as negative feedback from $\bar{x}$ back to $\bar{u}$. This formula has a complex structure as there are different signals to be operated on and being used for training: the input signal itself is filtered and the feedback signal for adaptation. A completely local synapse can only use the signal it immediately sees, $\bar{u}_j$ being the locally visible effective input (see later). Indeed, the feedback assumption has not been taken here to the logical conclusion.

In [11], there is explicit feedback $\frac{dx}{d\tau} = -\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}x + \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\bar{u}$, with $\bar{x}$ being the converged x, that also assures global system stability. Here, the locality is implemented as claimed above, and, again, the structure converges to principal subspace. However, such combination of Hebbian and anti-Hebbian like adaptations means that the roles of signals are different, and one needs information about the hierarchy among signal sources; what is more, as n grows, the number of explicit feedbacks grows as n^2, the scalability of the structure becoming poor. Such neurons with global-level view about their neighbors could be called "social" or "intelligent" agents (as opposed to the "selfish" agents studied below).

In what follows, the above shortcomings are fixed. The structural complexity of synaptic filtering and adaptation is changed to *functional complexity.* It turns out that when asymptotic properties of dynamic processes are appropriately employed, the network structure becomes the simplest possible, and, thus — one can claim — such structure is also physically plausible.

3.1 Feedback through environment

The key to solve the problems is to look closer at real systems and non-idealities therein. There are no pure information flows in nature. Applying the electrical engineering analogy, the *fan-out* of the feeding port cannot be infinite. When an input is observed by the system, the input is affected: exploitation of a signal means exhaustion. This exhaustion defines an *implicit* negative feedback through the environment. In this way there is a coupling of information and matter/energy flows in real systems. As it turns out, the *observer effect* means that the environment becomes part of the system, the system reflecting the properties of the environment.

The inputs from environment as modified by the system (closed loop) are characterized by the vector $\bar{u}$, whereas the original undisturbed inputs (open loop)

are hereafter denoted u. How to express the $\bar{u}$ in terms of u?

If $q_i \, \mathrm{E}\{\bar{x}_i \bar{u}_j\}$ is the contribution of an input what comes to the activity of a neuron, it is reasonable to assume that the loading among the inputs is distributed in the same way — that is, the change in input is $\Delta u_j = -\mathrm{E}\left\{\bar{u}_j \bar{x}_i{}^{\mathrm{T}}\right\} q_i \, x_i$, or when all feedback effects are expressed in a matrix form[2]

$$\Delta u = -\mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\} Q \, x. \tag{21}$$

Both $\bar{u}$ and $\bar{x}$ represent dynamic equilibria, being only virtually static signals. The asymptotic values are defined (in a somewhat sloppy way) as

$$\bar{u} = \lim_{\tau \to \infty} \left\{ u - \mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\} Q \, x(\tau) \right\} \tag{22}$$

and

$$\bar{x} = \lim_{\tau \to \infty} \left\{ x(\tau) \right\}. \tag{23}$$

Here it is assumed that one only studies some kind of "local infinities" at the local time scale τ as seen on the time scales relevant to the dynamics of x. When the expressions are combined,

$$\bar{u} = u - \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} Q \bar{x}. \tag{24}$$

When the balance is reached,

$$\bar{x} = Q \, \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} \bar{u} = Q \, \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} u - Q \, \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} \mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\} Q \, \bar{x}, \tag{25}$$

[2]If it cannot be assumed that the feedback effects are directly given by the feedforward effects, the system state variables can be changed, or feedback effects can be scaled without affecting the derivations. If the expression (3) is assumed valid for $\bar{x}$ it is valid also for $\bar{x}' = C\bar{x}$ because

$$C^{-1}\bar{x}' = Q\mathrm{E}\left\{C^{-1}\bar{x}'\bar{u}^T\right\}\bar{u} = QC^{-1}\mathrm{E}\left\{\bar{x}'\bar{u}^T\right\}\bar{u} = C^{-1}Q\mathrm{E}\left\{\bar{x}'\bar{u}^T\right\}\bar{u} \tag{20}$$

or $\bar{x}' = Q\mathrm{E}\left\{\bar{x}'\bar{u}^T\right\}\bar{u}$ for diagonal and invertible C (and Q diagonal). More generally, if it can be assumed that the feedback effects just somehow stabilize the overall system (as the case must be in existing systems), however stability is guaranteed the mathematical structure of (3) always implements the eigensystem

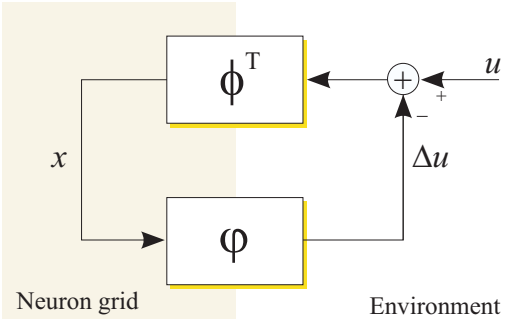

Figure 2: Exhaustion of the environment presented as negative feedback

or, when solved,

$$\bar{x} = \left(I_n + Q\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}Q\right)^{-1} Q\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u. \tag{26}$$

Using (7), one has

$$\bar{x} = \left(I_n + Q\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} Q\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u, \tag{27}$$

and, further,

$$\bar{x} = \left(Q^{-1} + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u. \tag{28}$$

Then $\bar{u} = u - \mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}Q\left(I_n + Q\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} Q\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u$ (see Fig. 2).

3.2 Stability of mapping

When the system affects its environment as studied above, the overall structure becomes an *algebraic loop,* and this dictates many of the system properties. However, in real life there are always delays, and the system becomes dynamic; this internal dynamics is needed to find the solutions fulfilling the static constraints, recursion refining the "eigenforms". The challenge here is that feedbacks are notorious: making system dynamic jeopardizes the system stability, just one unstable mode suffices to ruin the orchestration.

Now, assume that there is a time delay of h when signals traverse through the

system. One can approximate the situation in terms of a discrete-time model

$$\begin{aligned} x((k+1)h) &= Q\,\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\bar{u}(kh) \\ &= Q\,\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u - Q\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}Q\,x(kh). \end{aligned} \tag{29}$$

Reordering, one has

$$\begin{aligned} &\frac{x((k+1)h) - x(kh)}{h} \\ &\quad = \frac{1}{h}\left(Q\,\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}u - \left(I_n + Q\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}Q\right)\,x(kh)\right). \end{aligned} \tag{30}$$

As $h \to 0$, the left-hand side becomes the derivative $\dot{x}$, and the corresponding continuous time system matrix becomes

$$-\frac{1}{h}\left(I_n + Q\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}\mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}Q\right). \tag{31}$$

All eigenvalues of this matrix are always strictly negative; as h goes towards zero, the poles go towards minus infinity. This means that one has stable system regardless of the mapping matrices; the signals remain bounded, and so do the covariances, no matter what are the (bounded) inputs.

Another dynamic phenomenon that also deserves some attention is the practical evaluation of covariances. True covariance is an abstraction, involving expectations that involve infinite data sequences. In real life one can use the following exponentially weighted estimator for covariance:

$$\hat{\mathrm{E}}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}(k+1) = \mu\,\hat{\mathrm{E}}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}(k) + (1-\mu)\;\bar{x}(k)\bar{u}^{\mathrm{T}}(k), \tag{32}$$

where $0 \ll \mu < 1$ is a forgetting factor. No matter how erroneous $\hat{\mathrm{E}}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}(0)$ happens to be, signal-level convergence to a unique value is assured because of linearity; but how about the uniqueness of $\hat{\mathrm{E}}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}(k)$ when $k \to \infty$?

The matrix $\hat{\mathrm{E}}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}$ in open loop is *not* unambiguously determined by the input data u alone. This is easy to see, for example, by multiplying both sides of (3) by some scalar α: the expression still holds but now for new variables $\bar{x}' = \alpha\bar{x}$ and correlation matrix $\mathrm{E}\left\{\bar{x}'\bar{u}^{\mathrm{T}}\right\} = \mathrm{E}\left\{\alpha\bar{x}\bar{u}^{\mathrm{T}}\right\}$. However, when the loop is closed, and when the coupling is strong enough, $\bar{x}$ is uniquely determined by the input, as seen in (28), and so is $\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}$.

3.3 Modification of the environment

In formulas (28), etc., there is a discrepancy: the input is u but the covariances are given in terms of $\bar{u}$. This can be resolved by manipulating the expression:

$$\begin{aligned} \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} &= \mathrm{E}\left\{\bar{x}\left(u - \mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\}Q\bar{x}\right)^{\mathrm{T}}\right\} \\ &= \mathrm{E}\left\{\bar{x}u^{\mathrm{T}}\right\} - \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}Q\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}. \end{aligned} \tag{33}$$

Solving this for $\mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\}$, one has

$$\begin{aligned} \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} &= \left(I_n + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}Q\right)^{-1}\mathrm{E}\left\{\bar{x}u^{\mathrm{T}}\right\} \\ &= \left(Q^{-1} + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1}Q^{-1}\mathrm{E}\left\{\bar{x}u^{\mathrm{T}}\right\}. \end{aligned} \tag{34}$$

Combining (28) and (34):

$$\bar{x} = \underbrace{\left(Q^{-1} + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-2}Q^{-1}}_{M_1}\underbrace{\mathrm{E}\left\{\bar{x}u^{\mathrm{T}}\right\}}_{M_2}u. \tag{35}$$

Using this expression, one can study the connection between the undisturbed u and $\bar{x}$. If the statistical properties of the input data u are assumed to remain intact, one has

3.3.1 Theorem.

If data is is rich enough (non-zero variation dimensions in data $d \geq n$), and if each mode remains cybernetic *(see later), after convergence the neuronal mapping from u to $\bar{x}$ spans the principal subspace of data variation in u, corresponding to the n most significant eigenvector directions of the covariance matrix $\mathrm{E}\left\{uu^T\right\}$.*

3.3.2 Proof.

Rather than studying the adaptation process as a continuous process, the time axis is here assumed to be divided in long enough subparts; these subparts are indexed below using numbers in parentheses. The expectations, when calculated as sample averages within each interval, are already assumed to be accurate

enough. If one starts from some arbitrary mapping matrices $M_1^{(0)}$ and $M_2^{(0)}$, the step-by-step covariance adaptation proceeds as

$$\begin{aligned}
\bar{x}^{(0)} &= M_1^{(0)} M_2^{(0)}\, u \\
\bar{x}^{(1)} &= M_1^{(1)} \mathrm{E}\left\{\bar{x}^{(0)} u^{\mathrm{T}}\right\} u = M_1^{(1)} \mathrm{E}\left\{M_1^{(0)} M_2^{(0)} u u^{\mathrm{T}}\right\} u \\
&= M_1^{(1)} M_1^{(0)} M_2^{(0)} \mathrm{E}\left\{u u^{\mathrm{T}}\right\} u \\
\bar{x}^{(2)} &= M_1^{(2)} \mathrm{E}\left\{\bar{x}^{(1)} u^{\mathrm{T}}\right\} u = M_1^{(2)} \mathrm{E}\left\{M_1^{(1)} M_1^{(0)} M_2^{(0)} \mathrm{E}\left\{u u^{\mathrm{T}}\right\} u u^{\mathrm{T}}\right\} u \\
&= M_1^{(2)} M_1^{(1)} M_1^{(0)} M_2^{(0)} \mathrm{E}\left\{u u^{\mathrm{T}}\right\}^2 u \\
&\vdots \\
\bar{x}^{(k)} &= M_1^{(k)} M_2^{(k)}\, u = \left(\prod_{i=0}^{k} M_1^{(k-i)}\right) M_2^{(0)} \mathrm{E}\left\{u u^{\mathrm{T}}\right\}^k u.
\end{aligned} \tag{36}$$

The former part $M_1^{(k)} = \prod_{i=0}^{k} M_1^{(k-i)}$ is a scaling matrix of dimension $n \times n$ and it does not affect the subspace being spanned by the mapping. On the other hand, $M_2^{(k)}$ deserves more attention. Assume that the eigenvalue decomposition of the data covariance is written as[3]

$$\mathrm{E}\left\{u u^{\mathrm{T}}\right\} = \Theta\, \Lambda\, \Theta, \tag{37}$$

and one expresses the mapping matrix $M_2^{(0)}$ using the basis determined by the columns in Θ, so that

$$M_2^{(0)} = D\, \Theta^{\mathrm{T}}, \tag{38}$$

with D being some invertible matrix, the resulting mapping matrix $M_2^{(k)}$ becomes

$$M_2^{(k)} = M_2^{(0)} \mathrm{E}\left\{u u^{\mathrm{T}}\right\}^k = D\Theta^{\mathrm{T}}\, \Theta \Lambda^k \Theta^{\mathrm{T}} = D \Lambda^k \Theta^{\mathrm{T}}. \tag{39}$$

This means that in the mapping matrix the relevance of the principal component direction j is weighted by λ_j^k. At each iteration, the eigenvectors become better aligned with the most significant eigenvectors. Because the variables $\bar{x}_i$ are linearly independent, it is the n most significant covariance matrix eigenvectors that determine the mapping after adaptation. These define the same subspace

[3]For simplicity, assume that the eigenvalues of the data covariance matrix are distinct, so that $\lambda_i > \lambda_j$ for $i < j$, $1 \leq i, j \leq m$

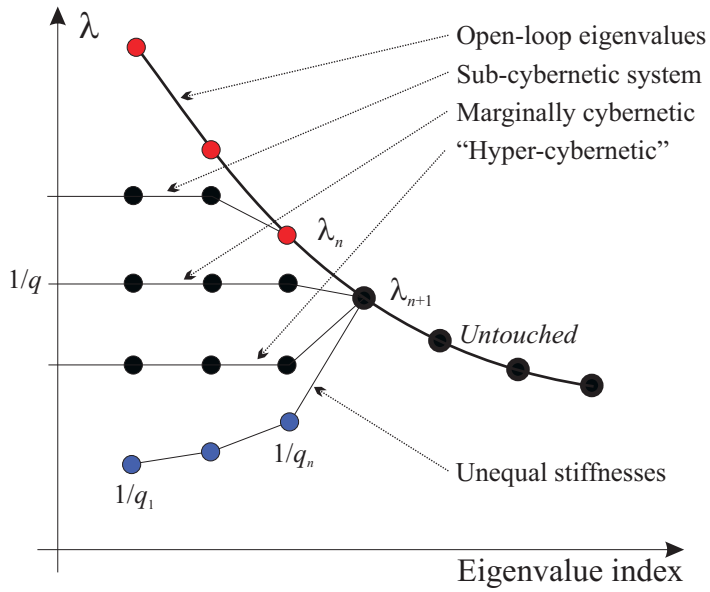

Figure 3: The effect of the neuronal system being coupled to external data. The feedback "sucks out" variation in the most significant data directions — but *only if the coupling manages to make the neurons cybernetic* (see Sec. 4)

as in the case of $\bar{x}$ vs. $\bar{u}$, but the eigenvalues differ. □

Indeed, again it is the principal component directions, but more can be said about the mapping from u to $\bar{x}$ than what one could say about the mapping from $\bar{u}$ to $\bar{x}$. It is the most significant variation directions in u that the system concentrates on, but because of the feedback, they are *not necessarily the most significant variation directions any more in* $\bar{u}$. This situation is visualized in Fig. 3. And this result concerning the principal subspace is not all — one can also say something about the actual basis vectors spanning that subspace.

4 Towards sparse coding

Despite the analyses above, there are *two* classes of solutions to (3). In addition to the case that was discussed in Sec. 3, the trivial solution $\bar{x} \equiv 0$ for all inputs, or $\bar{x}_i \equiv 0$ for a subset of them, also satisfies the assumed constraint, the mapping matrix becoming $\mathrm{E}\{\bar{x}_i \bar{u}_j\} \equiv 0$. To understand the faith of a neuron i, whether it fades away or stays "alive", depends on the corresponding coupling factor q_i.

As studied briefly in Sec. 5, one is facing deeper questions here — these properties are characteristic not only to neuron systems, but to all systems that are *cybernetic,* being tightly coupled to their environments, and competing for resources with their neighbors. This bold claim can be motivated when one understands all facets of the "feedback Hebbian" network.

4.1 Fine structure of the neuronal mapping

From (28) one can write yet another expression for the covariance:

$$\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = \left(Q^{-1} + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} \mathrm{E}\left\{uu^{\mathrm{T}}\right\} \mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\} \left(Q^{-1} + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1}.$$

Eliminate the matrix inverses by multiplication, so that

$$\begin{aligned}\left(Q^{-1} + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right) \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} \left(Q^{-1} + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right) \\ = \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} \mathrm{E}\left\{uu^{\mathrm{T}}\right\} \mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\},\end{aligned} \tag{40}$$

and observe the commutativity of the matrices:

$$\begin{aligned}&\left(Q^{-1} + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^2 \\ &\quad = Q^{-1/2}\, Q^{1/2} \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2} \mathrm{E}\left\{\bar{x}\bar{u}^{\mathrm{T}}\right\} \mathrm{E}\left\{uu^{\mathrm{T}}\right\} \mathrm{E}\left\{\bar{u}\bar{x}^{\mathrm{T}}\right\} \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}^{-1/2} Q^{1/2}\, Q^{-1/2} \\ &\quad = Q^{-1/2}\, \theta^{\mathrm{T}}\, \mathrm{E}\left\{uu^{\mathrm{T}}\right\} \theta\, Q^{-1/2},\end{aligned}$$

and, further, because of the diagonalizing properties of θ,

$$Q^{-1} + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\} = Q^{-1/4}\, \theta^{\mathrm{T}}\, \mathrm{E}\left\{uu^{\mathrm{T}}\right\}^{1/2} \theta\, Q^{-1/4}, \tag{41}$$

or

$$\mathrm{E}\left\{\bar{x}'\bar{x}'^{\mathrm{T}}\right\} + Q^{-1/2} = \theta^{\mathrm{T}}\, \mathrm{E}\left\{uu^{\mathrm{T}}\right\}^{1/2} \theta, \tag{42}$$

where

$$x' = Q^{1/4}\, x. \tag{43}$$

This is almost like a similarity transform between $\mathrm{E}\left\{uu^{\mathrm{T}}\right\}$ and $\mathrm{E}\left\{\bar{x}'\bar{x}'^{\mathrm{T}}\right\}$ again, but there are some essential differences: First, there is the square root, meaning that the eigenvalues change in the process, or variation is lost in the closed loop; second, there is the additional term $Q^{-1/2}$ in the formula. This is the key to understand the new functionalities of the cybernetic mapping.

Because the eigenvalues of $\mathrm{E}\left\{\bar{x}'\bar{x}'^{\mathrm{T}}\right\}$ always must be non-negative, meaning that variances in each direction must have real values, one can see that the non-trivial solutions are only possible if the variation level in input data is high enough, so that the additional factor $Q^{-1/2}$ becomes fully compensated. If the eigenvalue λ_j in u data has become coupled with variable x'_i, one can write

$$\mathrm{E}\left\{\bar{x}_i'^2\right\} = \sqrt{\lambda_j} - \frac{1}{\sqrt{q_i}}. \tag{44}$$

There must hold

$$\lambda_j > \frac{1}{q_i}, \tag{45}$$

or, on the other hand, if one wants to make the mode i visible, one has to apply more power in the observation according to

$$q_i > \frac{1}{\lambda_j}. \tag{46}$$

Thus, it has to be assumed that there is some local mechanism assuring that the neuron i remains "alive" by increasing the value of q_i if the activity in the neuron i seems to be vanishing altogether. This assures that that the studies above are relevant; this also assures that the matrix $\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}$ remains invertible. From the plausibility point of view, such additional local activity control is not as disturbing as it seems, because also natural neurons turn out to implement similar activity controls [3].

4.2 Rotations of basis vectors

When looking the properties of (44) closer, it turns out that the system carries out *active noise suppression.* The variation in the n most significant data directions becomes attenuated, and could speak of "black noise" being added to signals (see Fig. 4).

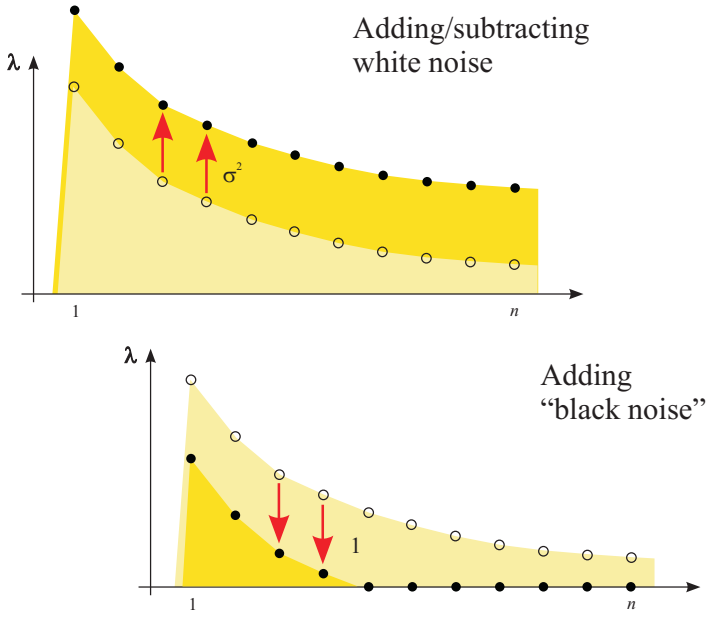

Figure 4: Consequences of adding "black noise" are opposite to white noise: The variation decreases in all directions — but only if it is possible

But this is not the only way to interpret the effects of the cybernetic feedback. In the spirit of PCA, the adaptation tries to maximize the amount of variation that is inherited by the system variables $\bar{x}_i$. Now, when there is the threshold, the system tries to maximize the overall variation *above that threshold.* Whereas orthonormal axis rotations do not alter the total variation above zero level, they can alter the *visible variations* in the cybernetic loop. As the total variation level cannot be changed, one should select the basis vectors so that *whenever the variable differs from zero, it does that in a spectacular way,* getting high above the threshold (see Fig. 5).

This all is near the ideas of *factor analysis,* where, too, one tries to apply basis rotations so that the latent variables would become better separated. In factor analysis one applies criteria like *varimax,* etc., where the goal is similar: to maximize the variations in factor scores. Such selection of variables has turned out to often result in physically meaningful models, enhancing the underlying features in data.

4.3 "Symmetry breaking"

The structure of the cybernetic neural network is simple and symmetric — indeed, it seems to be *too* symmetric. As in cosmic environments, say, it is only "symmetry breaking" that makes it possible for the structure and differentiation

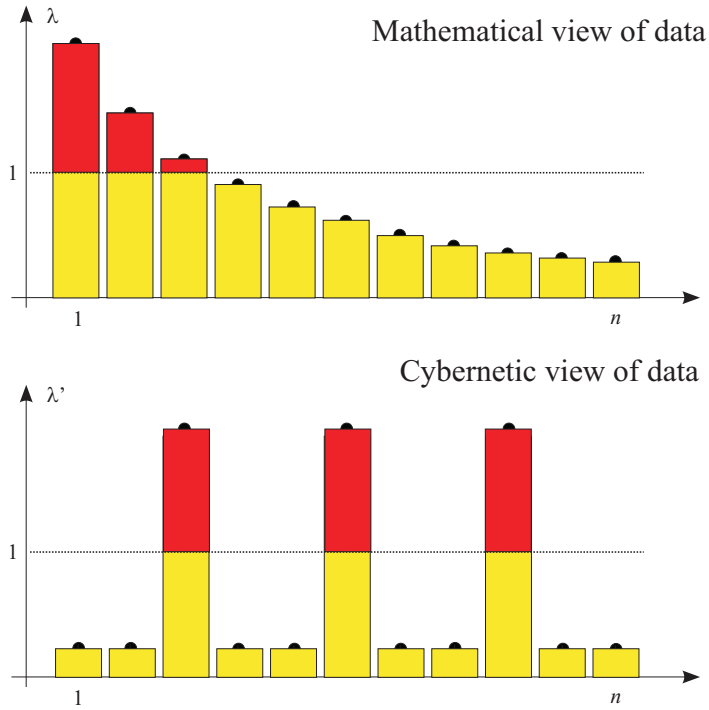

Figure 5: How black noise results in sparsity pursuit: Area above the threshold becomes maximized in the process of variation pursuit

to pop up in a system. And here, too: it seems that in practice a finishing step is needed when applying the above approach. Rather that defining the Hebbian neuron as in (2), introduce a nonlinearity in the model:

$$\bar{x}_i = f_i \left(q_i \sum_{j=1}^{m} \bar{x}_{ij} \right) . \tag{47}$$

Note that as the nonlinearity is applied in such a late phase, the basic functionality of the system can still be interpreted in terms of linear theory.

One can employ one's intuition of natural nonidealities to determine the outlook of the nonlinearity. The "neuronal activity" is only an abstraction; in reality, it is manifested either as a *pulse frequency* or as *concentration of transmitting chemicals.* In either case, the signals can never become negative. In the neuron model, this can be taken into account by manipulating the activity values as follows:

$$f_i(\xi) = \begin{cases} \xi, & \text{if } \xi > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{48}$$

Such nonlinearity cuts out all negative values, resulting in *sparse coding* where only a subset of the latent variables are non-zero at a time. Sparse coding behavior has been observed also in real neuron systems (for example, see [19]).

Artificial sparse coding networks with explicit recurrent structures have been constructed before (see [6], [7]), and it seems that sparse coding is a rather general functionality in dynamic feedback structures. The current approach to sparse coding has some special benefits: For example, as studied in Sec. 4.6, the system tries to minimize the error between the input and the system-defined estimate; as the matching is optimized, the variable values are effectively pushed from zero into the positive (hyper)quadrant. This neural structure is also specially suited for data where the features are assumed to be either non-existent, or the features have positive weights in patterns.

It turns out that such nonlinear extension essentially enhances the convergence properties of the implemented sparse coding algorithms.

4.4 Application example

The presented mathematical operations can be written in a form of a simple algorithm as shown in Fig. 6. In the Matlab form code, `U` is a $k \times m$ matrix of input vectors $u^{\mathrm{T}}(k)$, and `Xbar` is $k \times n$ matrix of neuronal activities. The matrices representing the covariances are denoted `Exx` and `Exu`. In addition to `Q`, there are additional parameters for affecting the adaptation: `lambda` is the forgetting factor, and `P` is the controller gain determining the adjustment rate of the coupling matrix Q, trying to keep variances at a constant level `Vref`. The model matrices are initialized to random values, and the algorithm is iterated for the data until convergence is reached. To make the algorithm parallelizable and to avoid low-level iteration, the nonlinearity is implemented in a sloppy way here, leaving it out fom the loop.

As an example, a case of coding hand-written digits is represented. As data material, there were 8000 samples of digits written in a 32×32 grid of binary (black vs. white) intensity values [15]. These intensity vectors were used as data `U`, with $k = 8000$ and $m = 1024$. The results for $n = 25$ are shown in Fig. 7.

4.5 Related neural algorithms

In addition to the subspace algorithms and the sparse coding algorithms, as discussed above, there are also other neural approaches that need to be mentioned.

```
while ITERATE

  % Balance of latent variables
  Xbar = U * (inv(inv(Q)+Exx)*Exu)';

  % Enhance model convergence by nonlinearity
  if nonlin
    Xbar = Xbar.*(Xbar>0);
  end

  % Balance of the environmental signals
  Ubar = U - Xbar*Q*Exu;

  % Model adaptation
  Exu = lambda*Exu + (1-lambda)*Xbar'*Ubar/k;
  Exx = lambda*Exx + (1-lambda)*Xbar'*Xbar/k;

  % Maintaining system excitation
  Qmult = diag(exp(P*(Vref-diag(Exx))));
  Q = Qmult * Q * Qmult;

end
```

Figure 6: **Algorithm**. Feedback Hebbian feature extraction

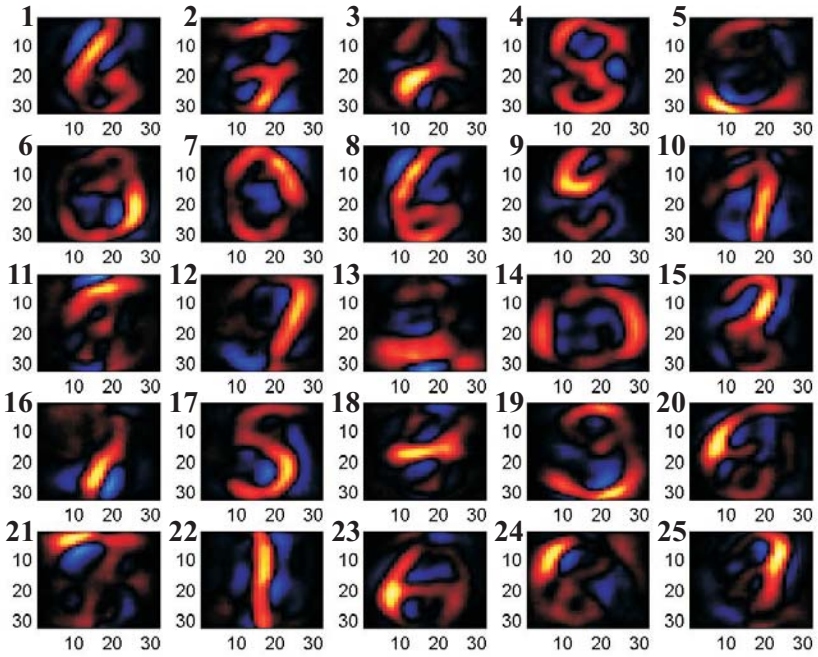

Figure 7: The 25 sparse components extracted from the handwritten digits. It seems that different kinds of "strokes" become manifested

The *Boltzmann machine* [10] is similarly based on energy functions, as the cybernetic neural network is, and the convergence towards the "pattern" is a dynamic process. And, similarly, the Boltzmann machine can be trained applying strictly local learning. However, the training is a complicated process with separate positive and negative training samples. Whereas the Boltzmann machine learns a single energy landscape, the input missing there, now the energy landscape is determined by the input data. There is only a finite number of predetermined training samples, the Boltzmann machine carrying out pattern matching among them, whereas now the pattern classicication is more continuous. Indeed, the cybernetic network tries to construct a model over the whole relevant input data space.

There are connections to more exotic neural approaches, too. For example, Kohonen SOM's, self-organizing maps [14], can be given a new, more general formulation in the presented framework. Note that in the above approach the matrix Q determines the connection strengths in the system; letting Q be non-diagonal one can simulate "neighborhood effect", so that nodes close to each other assumedly become somehow related. Symmetricity and positive definiteness claims are typically fulfilled also by such neighborhood matrices. In a "Hebbian feedback SOM"

there are various differences as compared to the traditional SOM:

- Now there are new interpretations and new cost criteria that can be employed to understand and exploit self-organization.
- Now there is no globality in the algorithm as the selection of the winner is ignored — there are "many winners" every time.
- Now the representation allows various dimensions, there is no need to project the data onto a low-dimensional local manifold.
- Now there is a continuous high-dimensional coding of data, rather than having a fixed number of discrete node indices as output.
- Now there is automatic distributed tuning of adaptation parameters, keeping all variables active and boosting convergence.

4.6 Relation to statistical regression

One can interpret the formulas governing the neuronal system in terms of multivariate analysis and regression models. It turns out that there is an interesting connection between *principal subspace analysis* and *linear regression,* on the one hand, and between *factor analysis* and *regularized regression,* on the other. Indeed, in a way they are *inverse operations* on data.

Assume that there is a mapping from some given vector $\bar{x}$ to a variable u_j, interpreted here as output,

$$u_j = {\phi_j}^{\mathrm{T}} \bar{x}, \tag{49}$$

and given some data, one would like to find an estimate for the mapping matrix minimizing the criterion

$$J_j(\phi_j) = \mathrm{E}\left\{\left(u_j - {\phi_j}^{\mathrm{T}}\bar{x}\right)^2\right\} + {\phi_j}^{\mathrm{T}} Q^{-1} \phi_j. \tag{50}$$

Here it is not only the quadratic matching criterion that is employed, but one also tries to keep the model parameters small, thus making the estimates less sensitive to errors, and making the mapping more robust. Assuming that this criterion

has been optimized for a group of variables u_j, where $1 \leq m$, the solution for estimates of u_j has the familiar form

$$\hat{u} = \left(Q^{-1} + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} \mathrm{E}\left\{u\bar{x}^{\mathrm{T}}\right\} \bar{x}. \tag{51}$$

Indeed, in the familiar *ridge regression* one selects $Q = q\, I_n$ for some scalar q; and if $Q \to \infty$ so that no parameter weighting is applied, one has traditional *multilinear regression* formula

$$\hat{u} = \left(\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} \mathrm{E}\left\{u\bar{x}^{\mathrm{T}}\right\} \bar{x}. \tag{52}$$

Now it is evident that these regression formulas are closely related to the derived feature extraction formula

$$\bar{x} = \left(Q^{-1} + \mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} \mathrm{E}\left\{\bar{x}u^{\mathrm{T}}\right\} \bar{u}, \tag{53}$$

and to the principal subspace analysis formula with explicit (anti-Hebbian) feedback (see [11]), respectively:

$$\bar{x} = \left(\mathrm{E}\left\{\bar{x}\bar{x}^{\mathrm{T}}\right\}\right)^{-1} \mathrm{E}\left\{\bar{x}u^{\mathrm{T}}\right\} u. \tag{54}$$

In Fig. 8, the above observations have been exploited to implement a scheme for technical data analysis. Traditionally, if the dimension of the latent basis is not minimal, problems may emerge, but it can be assumed that the feature-based regression tolerates extra variables better as there are no matrix invertibility problems. When modeling complex environments it can be beneficial when the system state can be expanded to include additional information.

If the variables $\bar{y}$ are not recirculated into $\bar{x}$, the latent structure is determined solely based on the properties of the input u, in the spirit of *principal component regression;* on the other hand, however, if there is some kind of "attention control" switching between the alternative inputs u and y, the feedback loop corresponding to the contemporary output remaining open during signal recirculation, the model structure is assumedly a compromise between the properties of the input and those of the output (in the spirit of *partial least squares,* for example).

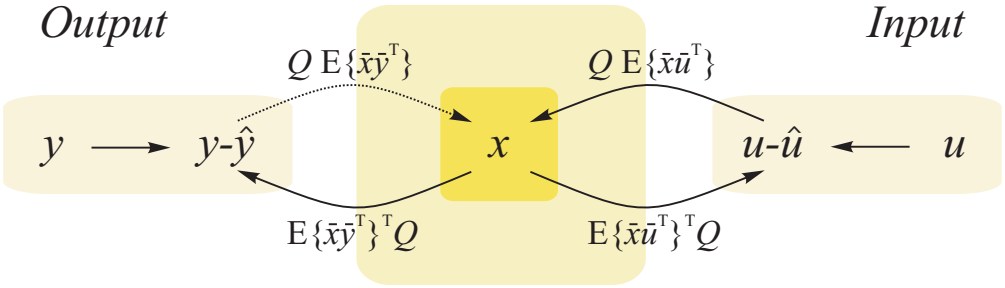

Figure 8: Extension of the framework towards technical data analysis applications

5 Neocybernetics — extension of intuitions

It can be claimed that the above discussions create a framework for studying *complexification* in general terms.

As the observations in (51) to (54) reveal, a system with Hebbian feedback learning constructs the best possible model of the input data (capturing the maximum of available variation) and applies the best possible attenuation of data variation (subtracting the most accurate estimate from the data). Indeed, one could speak of *emergent model-based control* eliminating variation in the environment; or, as seen from the opposite point of view, the system can be said to suck variation, exploiting information in the environment.

The thermodynamic analogue makes it possible to put the above observation in a more general setting. Remember that *entropy* can be defined in terms of probability; the more probable a state is, the higher is its entropy level. In a well-controlled system the state remains in the vicinity of its nominal value, so that its entropy level is high. Complete elimination of variation denotes "heat death". This means that when a system is defined in an appropriate way, cybernetic adaptation struggles towards increasing entropy: the increasing complexity in the emerging control structures is compensated by the loss of information in the controlled data. In a way, a cybernetic system essentially *inverts the arrow of entropy,* so that such systems with increasing structure and order are consistent with other physical systems; perhaps the idea of *maximum entropy production* can even be applied.

The above intuitions make it possible to make brave hypotheses, escaping the neuronal framework.

The input u can be seen as a set of forces pressing the system; the vector $\bar{x}$ reveals how much the system yields in that force field along the corresponding *degree of*

freedom. In mechanical systems it is easy to see that the product of force and resulting deformation reveals the energy stored in the system; correspondingly, for "generalized forces" and "generalized deformations", or action and reaction, one can define their product as "generalized energy" or *emergy.* These definitions make it possible to see the Hebbian learning principle in a yet wider perspective.

Generally, the vector u can be seen as the *resources* available in the environment. For example, neurons compete for activation, and the Hebbian-type learning strategy maximizes the average intake of that resource. In practice, the very natural-sounding rule ("generalized Hebbian principle") for individual agents to follow is to *go for resources.* Successful exploitation of the resources makes it possible for the system (population) to grow and reproduce faster; this means that the Hebbian learning strategy is *evolutionary optimal* and *outperforms others in the struggle among systems,* and it has survived for us to be seen.

These observations are elaborated on in the studies of *neocybernetics* [12] — the "difference that makes a difference" (see [2]) is very simply *covariation in data* now[4]. It can be claimed that neocybernetics offers a versatile framework for networked agent systems where there is no centralized control among the actors, making it possible to quantify studies of distributed intelligence. In neocybernetic systems a higher-level structure emerges in the form of principal subspace based sparse components; such systems implement *pattern recognition* among environmental signals. In more abstract systems the vector u can be seen to constitute the vector of *needs* or *functionalities,* and one ends in the field of *(bio)semiotics.* Interpretation of the environmental signs, or the selection and weighting of inputs determines the resulting structures in a more or less unique manner. The dualism between information and matter/energy can be attacked from a fresh point of view (see Fig. 9).

The same ideas of semiosis apply when extending the models from the infosphere to *ideasphere,* or when explaining the basis of *cognitive systems:* the mental model is a cybernetic balance among concepts that represent more or less explicit attractors in the ideasphere. Perhaps the multi-layered *model of models* is the key to the emergent consciousness?

There are a plenty of intuitions readily available to be exploited when trying to reach a "holistic" view of complex real-life systems — here are some examples:

- In an ecosystem, it seems that the magnificent diversity in the natural networks is more robust than the optimized monotony in man-made networks. Now there are some fresh ideas available: biodiversity and the structure of

[4]Information on neocybernetics can be found at
`http://www.control.tkk.fi/research/cybernetics/`

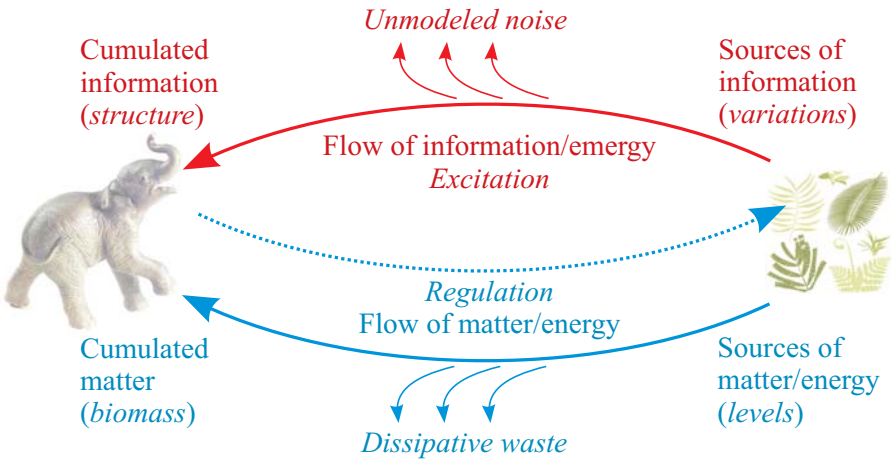

Figure 9: Abstract flows between two trophic levels in a neocybernetic system

"niches" can be explained in terms of principal component structure in the resources. Similarly, such principal component structure makes the system robust against random noise.

- In social systems there also exist trophic layers, and the new intuitions perhaps can even help in fine-tuning political systems. For example, it has been argued what would be the correct way to distribute the seats in the Parliament of the European Union; how should one weigh the population differences in member states? Looking at the formula (44), one could propose a resolution: the number of MEPs should be proportional to the square root of the population.

There are also more fundamental consequences. For example, assuming that the interaction mechanism between a system and its environment can be expressed in terms of distinct variables, the emergy maximization reduces to analysis of covariances, and *linearity of structures has evolutionary advantage* as such structures are optimal in that case — meaning that natural systems try to become linear!

There are philosophical dimensions, too — assuming that nature is essentially trying to model the environment to exhaust its resources, and if natural systems are actually manifestations of such models, one can address even the very principles of scientific work: man-made models are not necessarily only shadows of reality, in the Platonian sense; rather, if one applies the same model structures as nature does, one can capture the *essence* of the systems in one's models ...

and, at the same time, the eternal Einsteinian mystery why natural systems are modellable in the first place need not remain a mystery forever.

6 Conclusion

Neural networks should not be seen as mere data filters; after all, the driving force in articicial neural network research is the magnificent capabilities of the brain. When evaluating architectures and algorithms, one should have a systemic, wider view.

It seems that the whole cognitive machinery is there to model and simulate the surrounding world. This means that there are various sets of constraints that need to be addressed when constructing artificial neuron systems capable of carrying out such modeling:

1. **View from outside**. The *ontological* challenge is that the same basic neuronal structure should be capable of capturing very different kinds of real-world data.
2. **View from inside**. The *epistemological* challenge is that the resulting (emergent) models should be compatible with the qualitatively higher level cognitive system models.

And, in between these extremes there is all the time the engineering-like *modeling view* that is needed to keep the studies plausible. For example, do the models truly scale up beyond toy worlds? As there are no pre-determined structures or controls among neurons, does self-organization and self-regulation emerge from local interactions among the identical low-level entities that do not know the "big picture"?

And, after all, why is the proposed structure *so* good that it has survived in evolution, outperforming all other candidates?

As studied in [12], it seems that the neocybernetic neural network structure is a candidate model that can address all of the above challenges.

References

[1] Basilevsky, A. (1994). **Statistical Factor Analysis and Related Methods.** New York, NY: John Wiley & Sons.

[2] Bateson, G. (1972). *Steps to an Ecology of Mind. Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology.* Chigago, IL: The University of Chicago Press.

[3] Carandini, M. (2000). Visual cortex: Fatigue and adaptation. *Current Biology,* **10**(16), R1–R3.

[4] Cichocki, A. and Amari, S. (2002). **Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications.** New York, NY: Wiley.

[5] Diamantaras, K.I. and Kung, S.Y. (1996). **Principal Component Neural Networks: Theory and Applications.** New York, NY: Wiley.

[6] Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics,* **64**, 165–170.

[7] Falconbridge, M.S., Stamps, R.L., and Badcock, D.R. (2006). A Simpl3e Hebbian/Anti-Hebbian Network Learns the Sparse, Independent Components of Natural Images. *Neural Computation,* **18**, 415–429.

[8] Fyfe, C. (2005). Hebbian Learning and Negative Feedback Networks. Berlin, Germany: Springer–Verlag.

[9] Hebb, D.O. (1949). **Organization of Behavior.**

[10] Hinton, G. E. and Sejnowski, T. J. (1986). "Learning and Relearning in Boltzmann Machines". In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Vol. 1: Foundations, 282–317. Cambridge: MIT Press.

[11] Hyötyniemi, H. (2004). **Hebbian Neuron Grids: System Theoretic Approach.** Helsinki University of Technology, Control Engineering Laboratory, Report 144, September 2004.

[12] Hyötyniemi, H. (2006). **Neocybernetics in Biological Systems.** Helsinki University of Technology, Control Engineering Laboratory, Report 151, August 2006.

[13] Hyvärinen, A., Karhunen, J., and Oja, E. (2001). **Independent Component Analysis.** New York, NY: John Wiley & Sons.

[14] Kohonen, T. (2001). Self-Organizing Maps. *Springer Series in Information Sciences,* Vol. 30. Berlin, Heidelberg, New York: Springer (third edition).

[15] Laaksonen, J. (1997). Subspace Classifiers in Recognition of Handwritten Digits. *Acta Polytechnica Mathematica, Mathematics, Computing and Management in Engineering series,* No. 84, Espoo, Finland.

[16] Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology,* **15**, No. 3, 267-273.

[17] Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems,* **1**, 61–68.

[18] Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks,* **5**, 927–935.

[19] Olshausen, B.A. and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature,* **381**, 607–609.

[20] Rabinovich, M., Varona, P., Selverston, A., and Abarbanel, H. (2006). Dynamical principles in neuroscience. *Review of Modern Physics,* **78**, 1213–1265.

[21] Sanger, T.D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks,* **12**, 459–473.

[22] Wiener, N. (1948). *Cybernetics: Or the Control and Communication in the Animal and the Machine.* Cambridge, MA: MIT Press.