AS-74.4192 Elementary Cybernetics

Lecture 4:

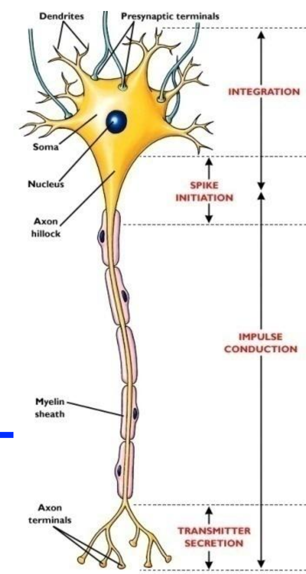
Neocybernetic Basic Models

HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Automation and Systems Technology
Cybernetics Group

- Starting point when modeling real complex systems:
  - Observation: *Bottom-up* approaches (studying the mechanisms alone) is futile
  - Another observation: *Top-down* approaches alone are similarly hopeless – there is no *grounding*
- Mission: Both views have to be combined
  - One needs *vision from top*
  - One needs *substance from bottom*
- Try to apply the ideas to a prototypical example: Modeling of *neural networks* – the best understood complex system (?)
- Remember that combining the two views is a big challenge: *Computationalism* (numeric) and *traditional AI* (symbolic) seem to be incompatible; low-level functions and high-level (emergent) functionalities are very different

1. What is the “deep structure” of the emergent patterns?

2. How to capture the attractors of changing behaviors?



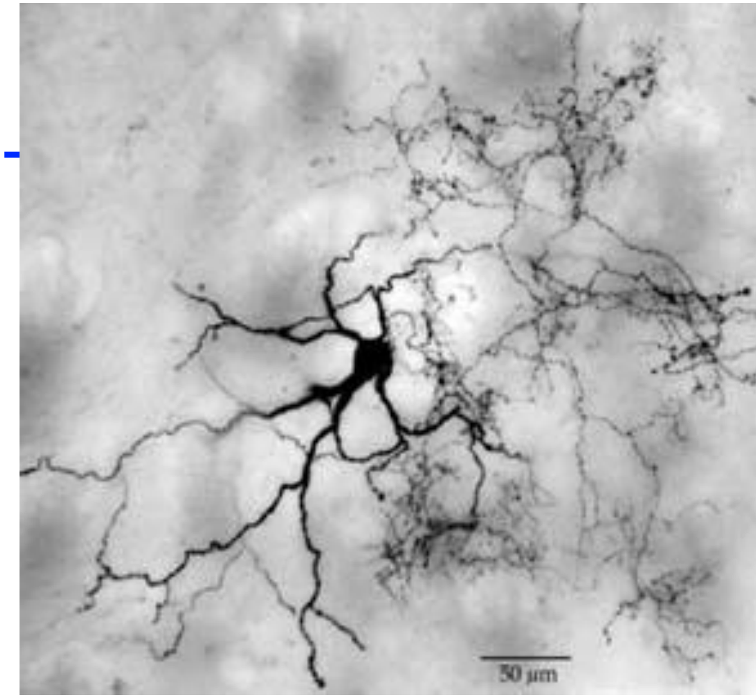
Neocybernetic starting points – summary

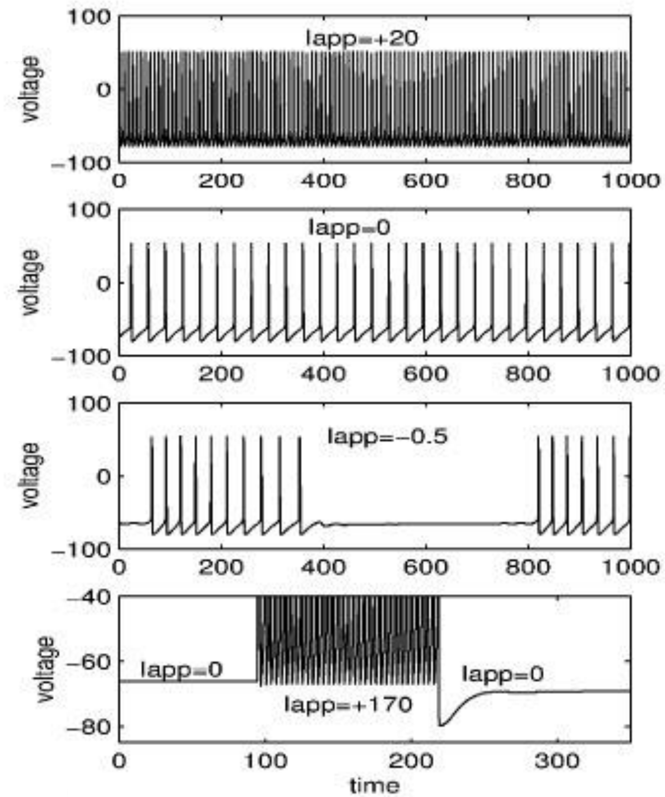
- The details (along the time axis) are abstracted away, holistic view from the above is applied
- There exist local actions only, there are no structures of centralized control
- It is assumed that the underlying interactions and feedbacks are consistent, maintaining the system integrity
- This means that one can assume *stationarity* and *dynamic balance* in the system in varying environmental conditions
- An additional assumption: Linearity is pursued as long as it is reasonable

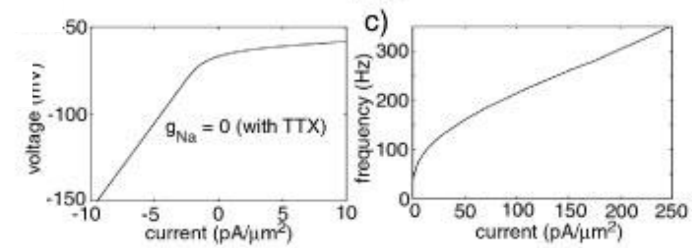
Sounds simple – are there any new intuitions available?

Strong guiding principles for modeling

Modeling a neuron







- Neural (chemical) signals are pulse coded, asynchronous, ... extremely complicated
- Simplification: Only the relevant information is represented – the *activation levels*

Abstraction level #1

- Triggering of neuronal pulses is stochastic
- Assume that in stationary environmental conditions the average number of pulses in some time interval remains constant

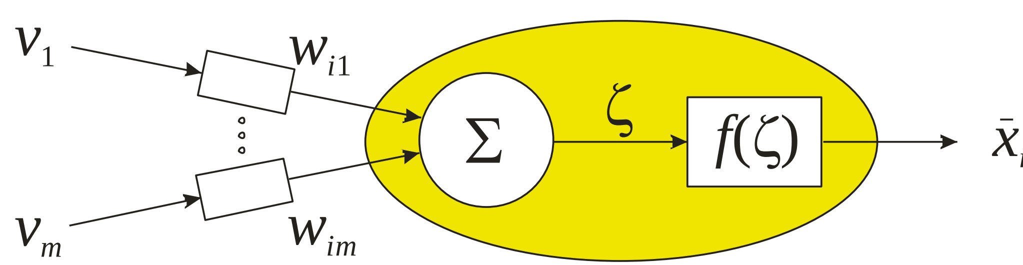
$v = \mathrm{E}\{\text{activity}\}$

- Only study statistical phenomena: Abstract the time axis away, only model *average activity* (cf. weak emergence!)
- *Perceptron*: Linear summation of input signals v_j + activation function:

$$\bar{x}_i = f\left(W_i^T v\right)$$

and linear version

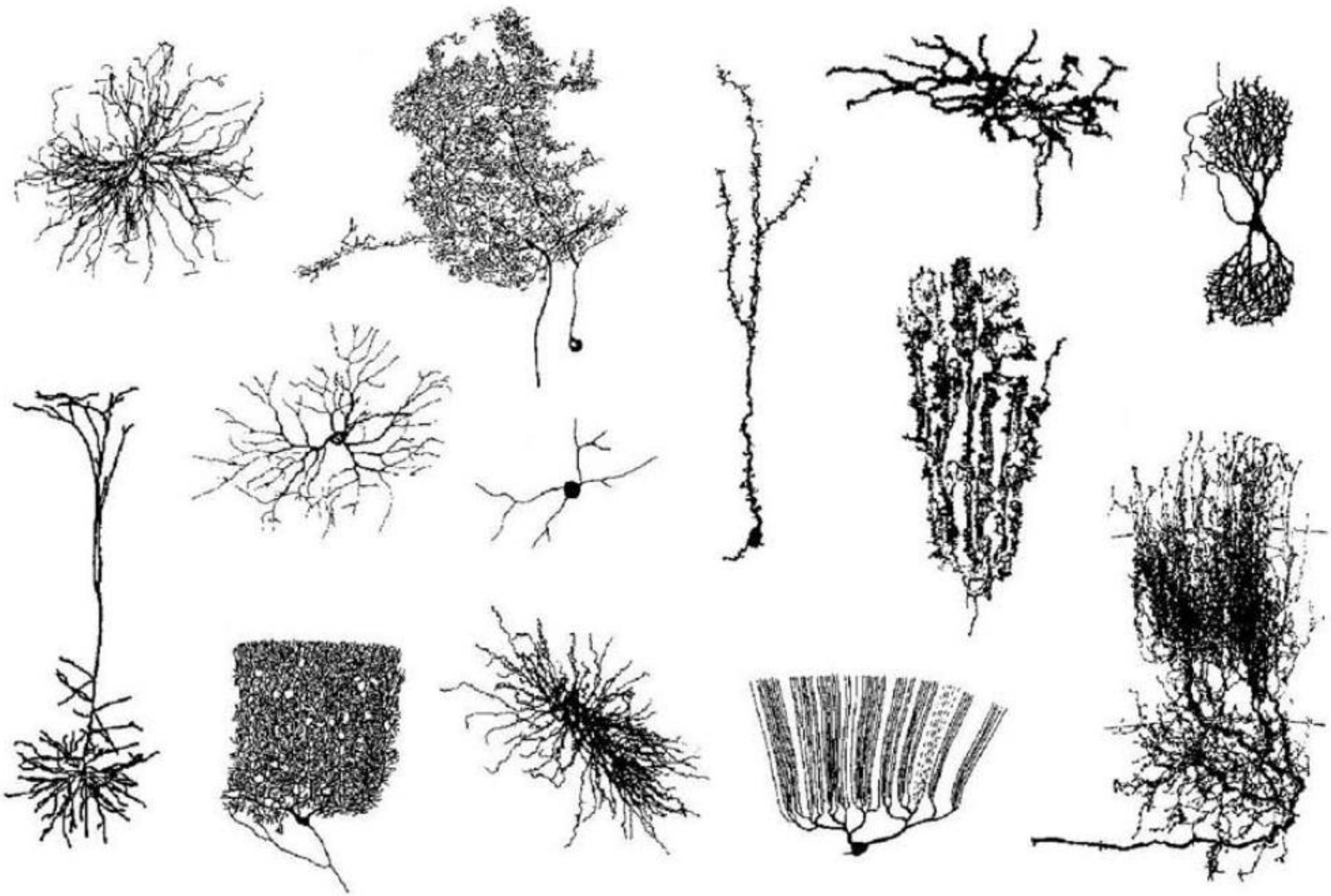
$$\bar{x}_i = W_i^T v = \sum_{j=1}^{m} w_{ij}^T v_j$$



For the whole linear grid $\bar{x} = W^T v$

HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Automation and Systems Technology
Cybernetics Group

Still, remember the reality ...



- The emergence idea is exploited here – deterministic activity variables are employed to describe behaviors
- How to exploit the “first-level” neuron abstraction, how to reach the *neuron grid* level of abstraction?
- Neural networks research studies this – opposite ends:

1. Feedforward perceptron networks
   - Non-intuitive: Black-box model, unanalyzable
   - Mathematically strong: Smooth functions can be approximated to arbitrary accuracy
2. Kohonen’s self-organizing maps (SOM)
   - Intuitive: Easily interpretable by humans (visual pattern recognition capability exploited)
   - Less mathematical: A mapping from m dimensional real-valued vectors to n integers

Now, again, trust deep structures more than surface patterns!

Hebbian learning

- Artificial neural networks are mainly seen as computational tools only
- To capture the functional essence of neuronal systems, one has to elaborate on the domain area
- The Hebbian learning rule (by physician Donald O. Hebb) also dates back to mid-1900's:*

  "If the neuron activity correlates with the input signal, the corresponding synaptic weight increases"

- Are there some *goals* for neurons included here?! Is there something teleological taking place?
- Bold assumptions make it possible to reach powerful models

* *"When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."*

Traditional approach

- Assume: Perceptron activity $\bar{x}_i$ is a linear function of the input signal v_j, where the vector w_{ij} contains the synaptic weight:

$$\bar{x}_{ij} = w_{ij} v_j \qquad \text{with} \qquad \bar{x}_i = \sum_{j=1}^{m} \bar{x}_{ij}$$

- Hebbian law applied in adaptation: Correlation between input and neuronal activity expressed as $\bar{x}_i v_j$, so that

$$\frac{dw_{ij}}{dt} = \gamma \cdot \bar{x}_i v_j = \gamma \cdot w_{ij} v_j^2$$

assuming here, for simplicity, that $m = 1$.

- This learning law is unstable – the synaptic weight grows infinitely, and so does $\bar{x}_i$!

Enhancements

- Stabilization by the *Oja's rule* (by Erkki Oja):

$$\frac{dw_{ij}}{dt} = \gamma \cdot w_{ij} v_j^2 - \gamma \cdot w_{ij} \bar{x}_i^2$$

Compare to the logistic formulation of limited growth!

- Motivation: Keeps the weight vector bounded ($|W_i| = 1$), and average signal size $\mathrm{E}\{|\bar{x}_i|\} = 1$
- Extracts the *first principal component* of the data
- Extension: Generalized Hebbian Algorithm (GHA): Structural tailoring makes it possible to deflate pc's one at a time
- However, the new formula is nonlinear: Analysis of neuron grids containing such elements is difficult, and extending them is equally difficult – What to do instead?

Level of synapses

- The neocybernetic guidelines are: Search for *balance* and *linearity*
- Note: Nonlinearity was not included in the original Hebbian law – it was only introduced for pragmatic reasons

Are there other ways to reach stability – in linear terms?

- Yes – one can apply *negative feedback*:

$$\frac{dw_{ij}}{dt} = \gamma_i \cdot \bar{x}_i v_j - \frac{1}{\tau_i} w_{ij}$$ or in matrix form $$\frac{dW}{dt} = \gamma \cdot \bar{x} v^T - \tau^{-1} W$$

The steady-state is

$$\bar{W} = \gamma\tau \cdot \mathrm{E}\left\{\bar{x} v^T\right\} = \Gamma \cdot \mathrm{E}\left\{\bar{x} v^T\right\}$$

Synaptic weights become coded in a correlation matrix!

Level of neuron grids

- Just the same principles can be applied when studying the neuron grid level – *balance* and *linearity*
- Distinguish sources:

$$\bar{W} = \left(A \mid B \right) \quad \text{and} \quad v = \begin{pmatrix} -x \\ \hline u \end{pmatrix}$$

so that $A = \Gamma \cdot \mathrm{E}\left\{ \bar{x}\bar{x}^T \right\}$ and $B = \Gamma \cdot \mathrm{E}\left\{ \bar{x} u^T \right\}$

- To implement negative feedback, one needs to apply the *anti-Hebbian* action between otherwise Hebbian neurons:

$$\frac{dx}{dt} = -Ax + Bu$$

so that the steady state becomes

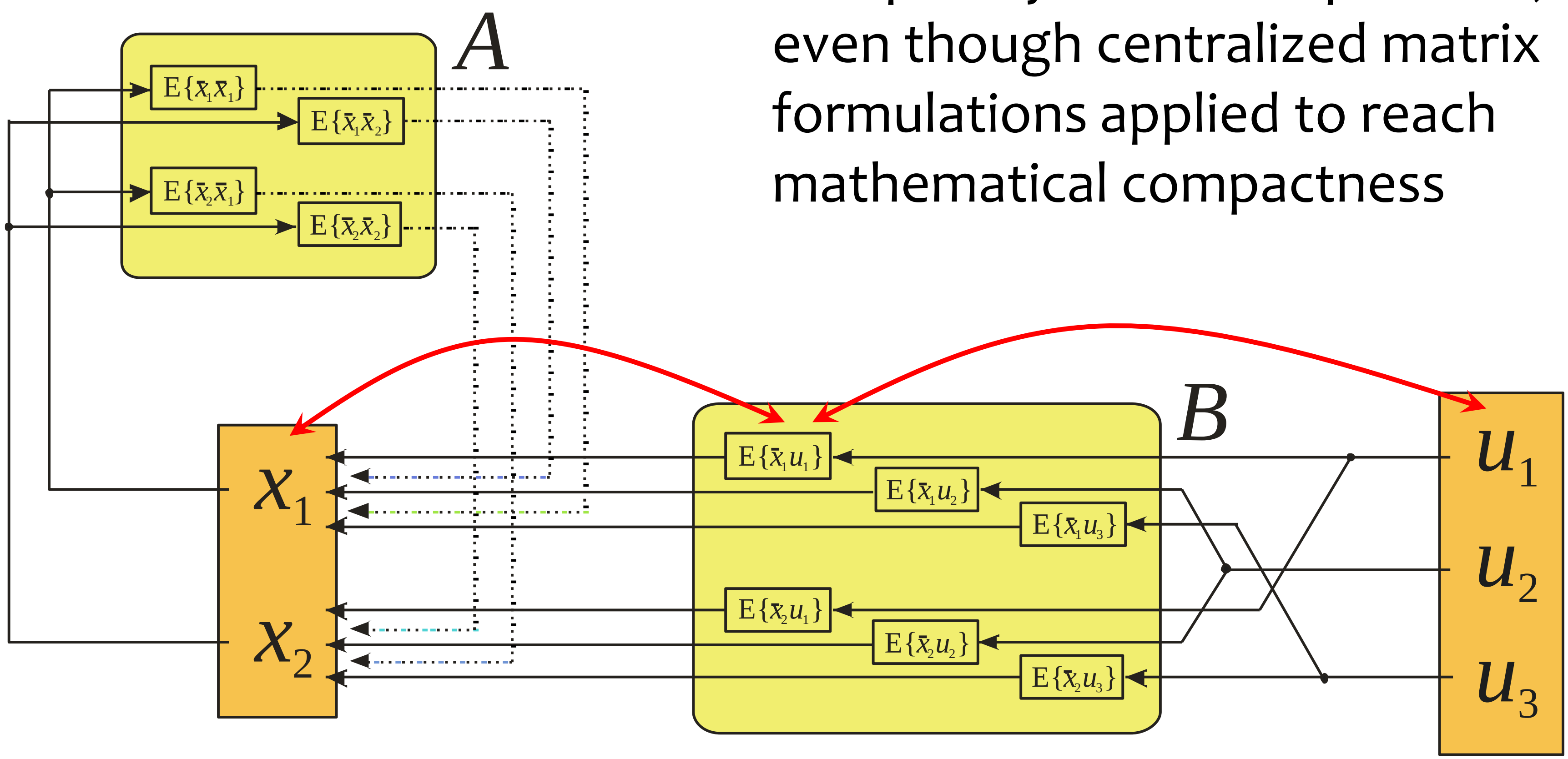
$$\bar{x} = A^{-1} B\, u = \mathrm{E}\left\{ \bar{x}\bar{x}^T \right\}^{-1} \mathrm{E}\left\{ \bar{x} u^T \right\} u = \phi^T u$$

Model is stable! Eigenvalues of *A* always real and non-negative

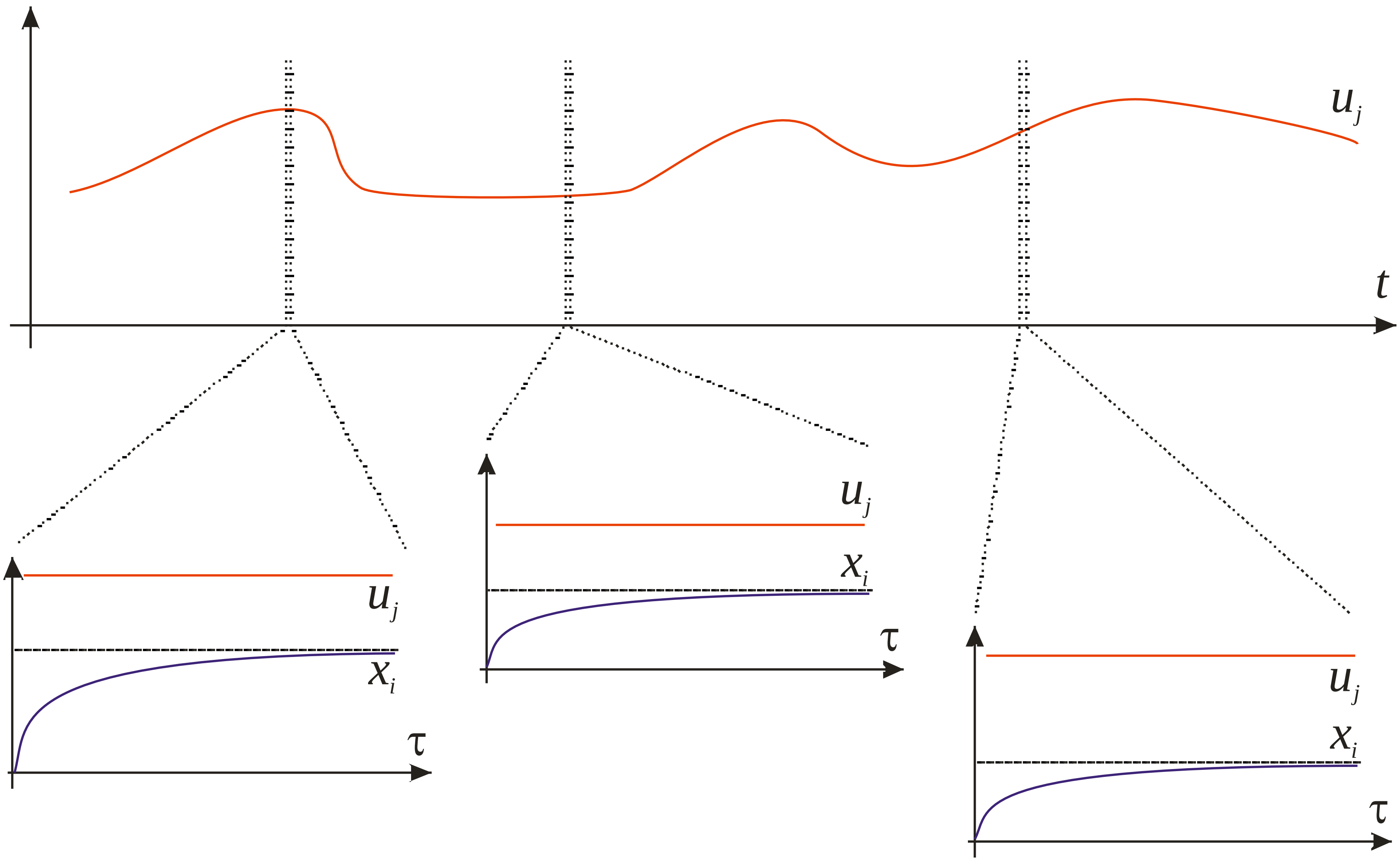
Hebbian/anti-Hebbian system

$$\dot{x} = -Ax + Bu$$

- Explicit feedback structures
- Completely localized operation, even though centralized matrix formulations applied to reach mathematical compactness



Different time scales – after all, *u* also varies



Towards abstraction level #2

- Cybernetic model = statistical model of balances $\bar{x}(u)$
- Assume dynamics of u is essentially slower than that of x and study the covariance properties:

$$\mathrm{E}\left\{\bar{x}\bar{x}^T\right\} = \mathrm{E}\left\{\bar{x}\bar{x}^T\right\}^{-1} \mathrm{E}\left\{\bar{x}u^T\right\} \mathrm{E}\left\{uu^T\right\} \mathrm{E}\left\{\bar{x}u^T\right\}^T \mathrm{E}\left\{\bar{x}\bar{x}^T\right\}^{-1}$$

or

Operator E connects the levels

$$\mathrm{E}\left\{\bar{x}\bar{x}^T\right\}^3 = \mathrm{E}\left\{\bar{x}u^T\right\} \mathrm{E}\left\{uu^T\right\} \mathrm{E}\left\{\bar{x}u^T\right\}^T$$

or

$$\left(\phi^T \mathrm{E}\left\{uu^T\right\} \phi\right)^3 = \phi^T \mathrm{E}\left\{uu^T\right\}^3 \phi \qquad n < m$$

- Balance on the statistical level = *second-order balance*

HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Automation and Systems Technology
Cybernetics Group

Third level of time scales!

Solution

- Expression fulfilled for $\phi = \theta_n D$, where θ_n is a matrix of n of the covariance matrix eigenvectors, and D is orthogonal

- Loose motivation: This is because left-hand side is then

$$\left(\phi^T \mathrm{E}\left\{uu^T\right\}\phi\right)^3 = \left(D^T \theta_n^T \mathrm{E}\left\{uu^T\right\}\theta_n D\right)^3 = \left(D^T \Lambda_n D\right)^3 = D^T \Lambda_n^3 D$$

and right-hand side is

$$\phi^T \mathrm{E}\left\{uu^T\right\}^3 \phi = D^T \theta_n^T \mathrm{E}\left\{uu^T\right\}^3 \theta_n D = D^T \Lambda_n^3 D$$

- Stable solution when θ_n contains the *most significant* data covariance matrix eigenvectors

Accurate proof: *Adaptation converges to PS*

Study how the mapping matrices become adapted; assume that this process is divided in k steps:

$$\bar{x}(0) = A^{-1}(0)B(0)\,u$$

$$\bar{x}(1) = A^{-1}(1)B(1)\,u = A^{-1}(1)\,\mathrm{E}\left\{\bar{x}(0)u^T\right\}u = A^{-1}(1)\,\mathrm{E}\left\{A^{-1}(0)B(0)\,uu^T\right\}u = A^{-1}(1)A^{-1}(0)B(0)\,\mathrm{E}\left\{uu^T\right\}u$$

$$\bar{x}(2) = A^{-1}(2)B(2)\,u = A^{-1}(2)\,\mathrm{E}\left\{\bar{x}(1)u^T\right\}u = A^{-1}(2)\,\mathrm{E}\left\{A^{-1}(1)A^{-1}(0)B(0)\,\mathrm{E}\left\{uu^T\right\}uu^T\right\}u = A^{-1}(2)A^{-1}(1)A^{-1}(0)B(0)\,\mathrm{E}\left\{uu^T\right\}^2 u$$

$$\vdots$$

$$\bar{x}(k) = A^{-1}(k)B(k)\,u = \quad\cdots\quad = \prod_{i=0}^{k} A^{-1}(k-i)\;B(0)\,\mathrm{E}\left\{uu^T\right\}^k u$$

The first part of this expression does not affect the subspace being spanned; so only $B(k)$ is now of interest. Write it applying the basis axes spanned by the principal components of data, so that

$$\mathrm{E}\left\{uu^T\right\} = \Theta\Lambda\Theta^T$$

and

$$B(0) = D\Theta^T.$$

Now one has

$$B(k) = B(0)\,\mathrm{E}\left\{uu^T\right\}^k = D\Theta^T\;\Theta\Lambda^k\Theta^T = D\,\Lambda^k\Theta^T,$$

meaning that in the mapping matrix the relevance of the principal component direction j is weighted by λ^j. Because the variables x_i are linearly independent, it is the n most significant of those directions that only will remain visible in the mapped data after adaptation. – This completes the proof.

Principal subspace analysis

- Any subset of input data principal components can be selected for ϕ
- The subspace spanned by the n most significant principal components gives a stable solution
- Conclusion:

  Competitive learning (combined Hebbian and anti-Hebbian learning) without any structural constraints results in self-regulation (balance) and self-organization (in terms of principal subspace).

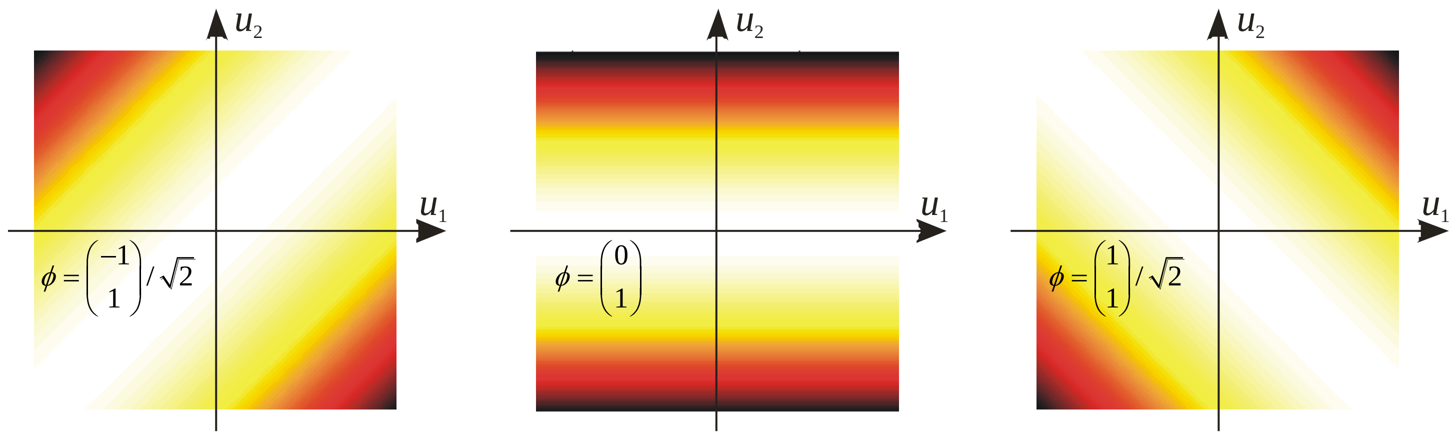
Emergent patterns

- The process (convergence of x) can be substituted with the final pattern: Details are lost, but the essence remains (?)
- The pattern is characterized in terms of a *cost criterion*

$$J(x,u) = \frac{1}{2} x^T \mathrm{E}\left\{ \bar{x}\bar{x}^T \right\} x - x^T \mathrm{E}\left\{ \bar{x} u^T \right\} u$$

Process itself = Gradient descent minimization for the criterion!

- Models of local minima (m = 2, n = 1):



Mathematics vs. reality

1. Correlations vs. covariances
   - The matrices being studied are *correlation matrices* rather than *covariance matrices* (as is normally the case in PCA)
   - This means that now data u is not assumed to be zero-mean, there is no need for preprocessing; in practice, the variables are always non-negative
   - From physical point of view, this is beneficial: Note that the actual signal carriers (chemical concentrations / pulse frequencies) cannot be negative
2. Principal subspace vs. principal components
   - When applying the linear structure, the actual principal components are not distinguished, only the subspace spanned by them
   - This means that the variables can again all be non-negative, so that the signals x again can be physically plausible
   - Indeed: If all u_j are non-negative, and initially all x_i have non-negative values, the x_i's will always be non-negative (so that one has a *positive system*)

Unification of layers

- Again, study a synapse; it is not a static mapping but a dynamic system (much faster than the grid dynamics)
- Assume this (trivial) system is also cybernetic:

$$\frac{dx_{ij}}{dt} = -\gamma \mathrm{E}\left\{x_i^2\right\} x_{ij} + \gamma \mathrm{E}\left\{x_i v_j\right\} v_j$$

or in balance

$$\bar{x}_{ij} = \frac{\mathrm{E}\left\{x_i v_j\right\}}{\mathrm{E}\left\{x_i^2\right\}} v_j = w_{ij} v_j$$

Extra term!

- Comparing this to the grid model, one can see that the two layers are qualitatively identical if one selects

$$\boxed{\Gamma = \mathrm{Var}\left\{xx^T\right\}^{-1}}$$

HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Automation and Systems Technology
Cybernetics Group

Technical experiments

- In technical terms, one can define the algorithm

$$\bar{x} = \hat{\mathrm{E}}\left\{\bar{x}\bar{x}^T\right\}^{-1} \hat{\mathrm{E}}\left\{\bar{x}u^T\right\} u$$

where

$$\frac{d\hat{\mathrm{E}}\left\{\bar{x}\bar{x}^T\right\}}{dt} = -\lambda\, \hat{\mathrm{E}}\left\{\bar{x}\bar{x}^T\right\} + \lambda\, \bar{x}\bar{x}^T$$

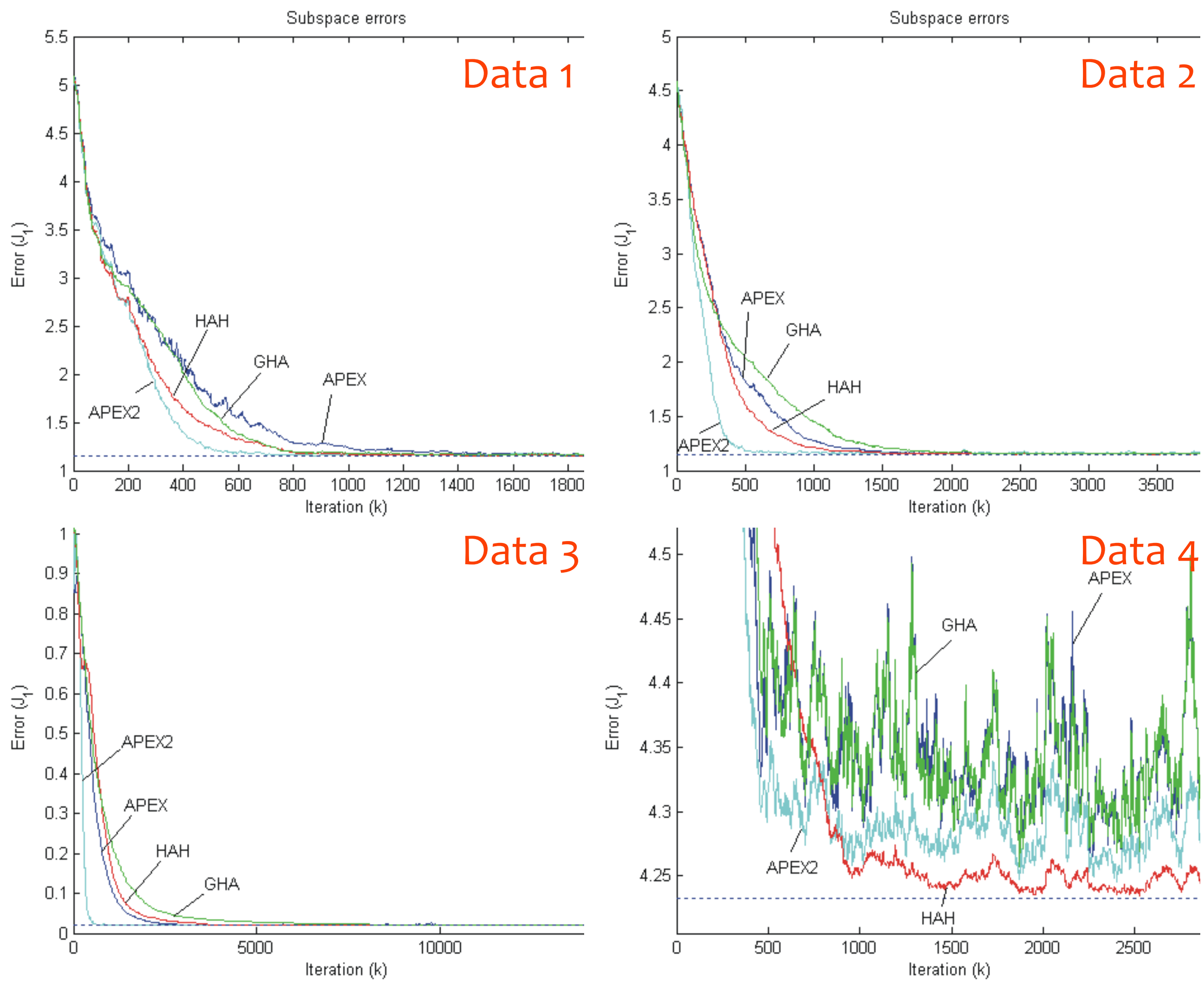
$$\frac{d\hat{\mathrm{E}}\left\{\bar{x}u^T\right\}}{dt} = -\lambda\, \hat{\mathrm{E}}\left\{\bar{x}u^T\right\} + \lambda\, \bar{x}u^T$$

Initialization: Matrix A has to be always positive definite

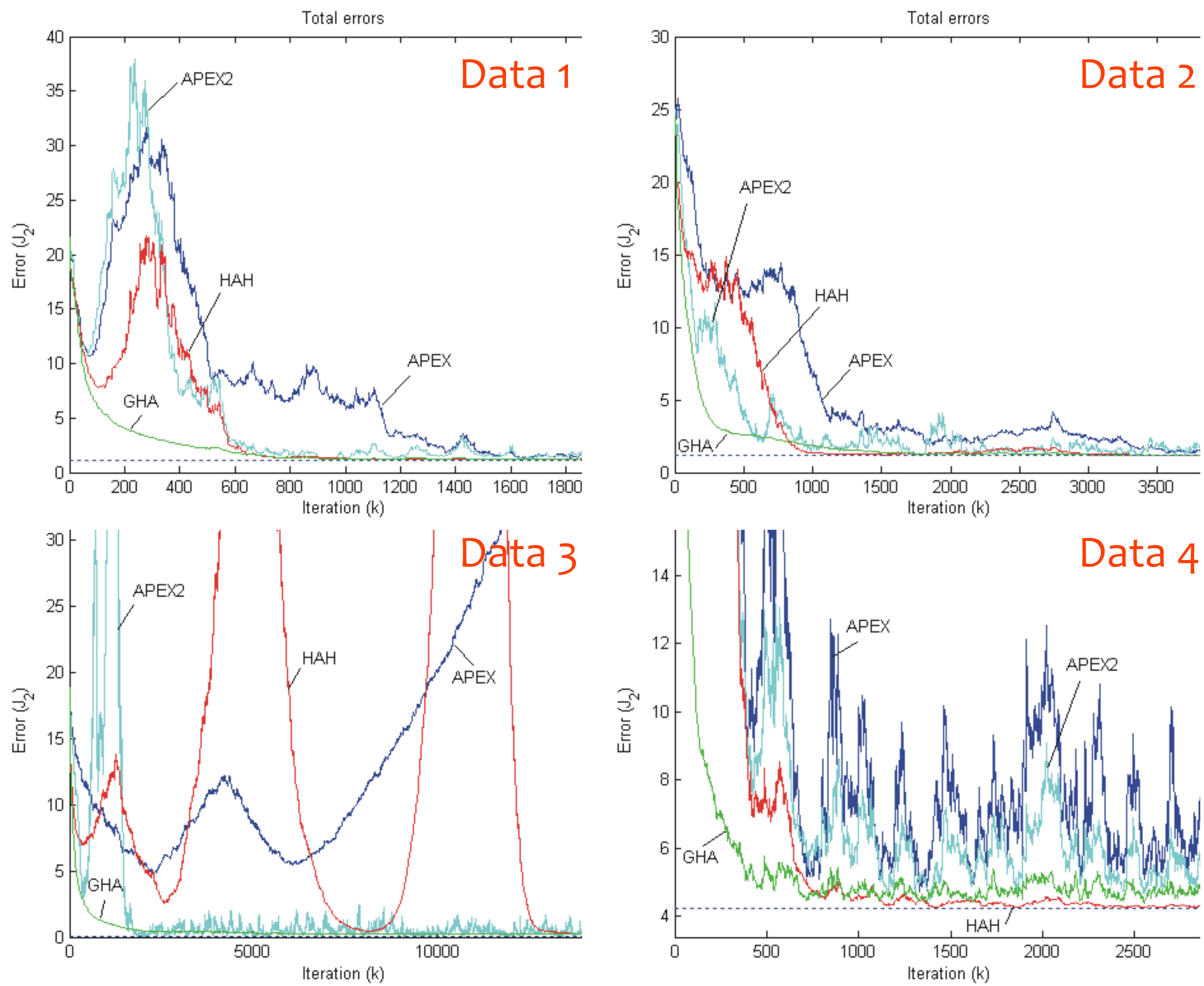
- In practice, easiest implemented in discrete-time
- *Matrix inversion lemma* can be applied
- Matrices can be *masked* to implement hierarchies, etc.

- Principal subspaces found OK for different data sets

HAH = Hebbian/Anti-Hebbian algorithm



- PC’s cannot always be told apart



Summary: Neocybernetic models

- **First-order cybernetic system**: For any stable A, assume that there holds

$$\frac{dx}{dt} = -Ax + Bu \qquad \text{with} \qquad \bar{x} = A^{-1}B\,u = \phi^T u$$

- **Second-order cybernetic system**: Additionally, assume that the matrices are

$$A = \Gamma \cdot \mathrm{E}\left\{\overline{xx}^T\right\} \qquad \text{and} \qquad B = \Gamma \cdot \mathrm{E}\left\{\bar{x}u^T\right\}$$

- **Higher-order (optimized) cybernetic (parameterless!) system**: Additionally, assume that

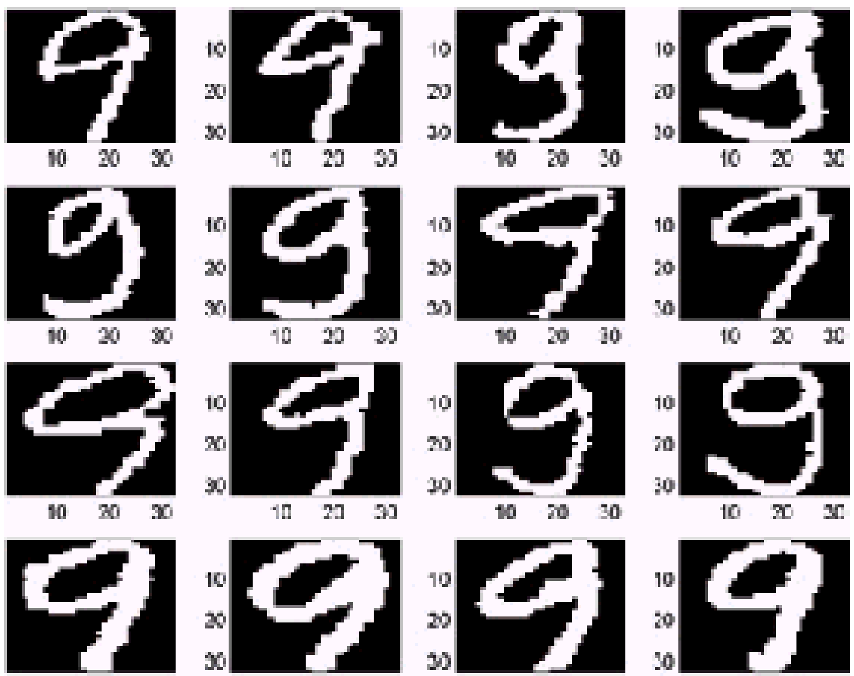
$$\Gamma = \mathrm{Var}\left\{\overline{xx}^T\right\}^{-1} \qquad \text{or} \qquad \Gamma = \mathrm{E}\left\{\overline{xx}^T\right\}^{-1} \longleftarrow$$

Newton algorithm: Second-order convergence

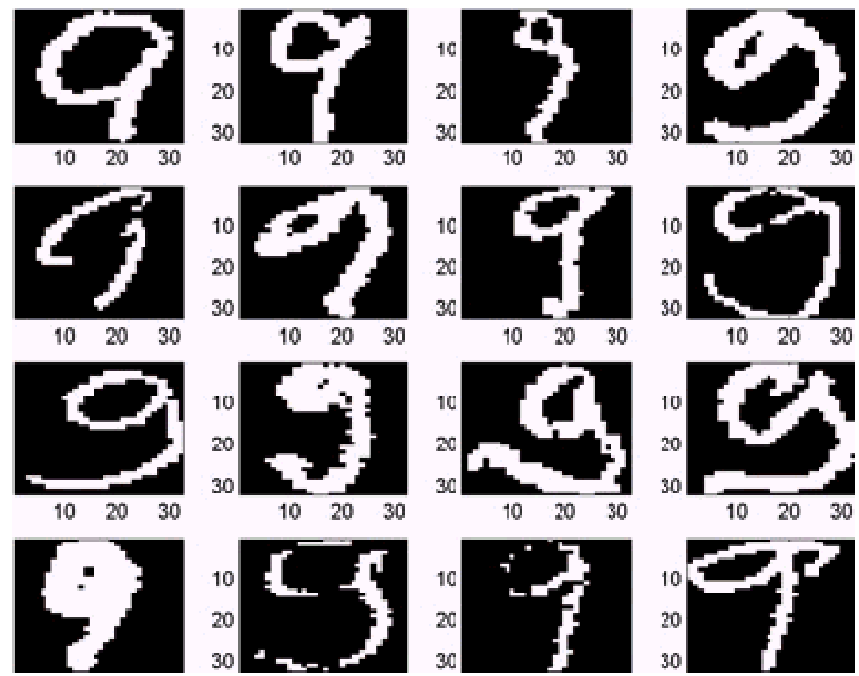
Example: Hand-written digits

- There were a large body of 32×32 pixel images, representing digits from 0 to 9, over 8000 samples

Examples of typical "9"



Examples of less typical "9"



Algorithm for Hebbian/anti-Hebbian learning ...

```
LOOP – iterate for data in the kxm matrix U

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

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

  % PCA rather than PSA through structural constraints
  Exx = tril(ones(n,n)).*Exx;

END

% Recursive algorithm can be boosted with matrix inversion lemma
```

… resulting in Principal Components

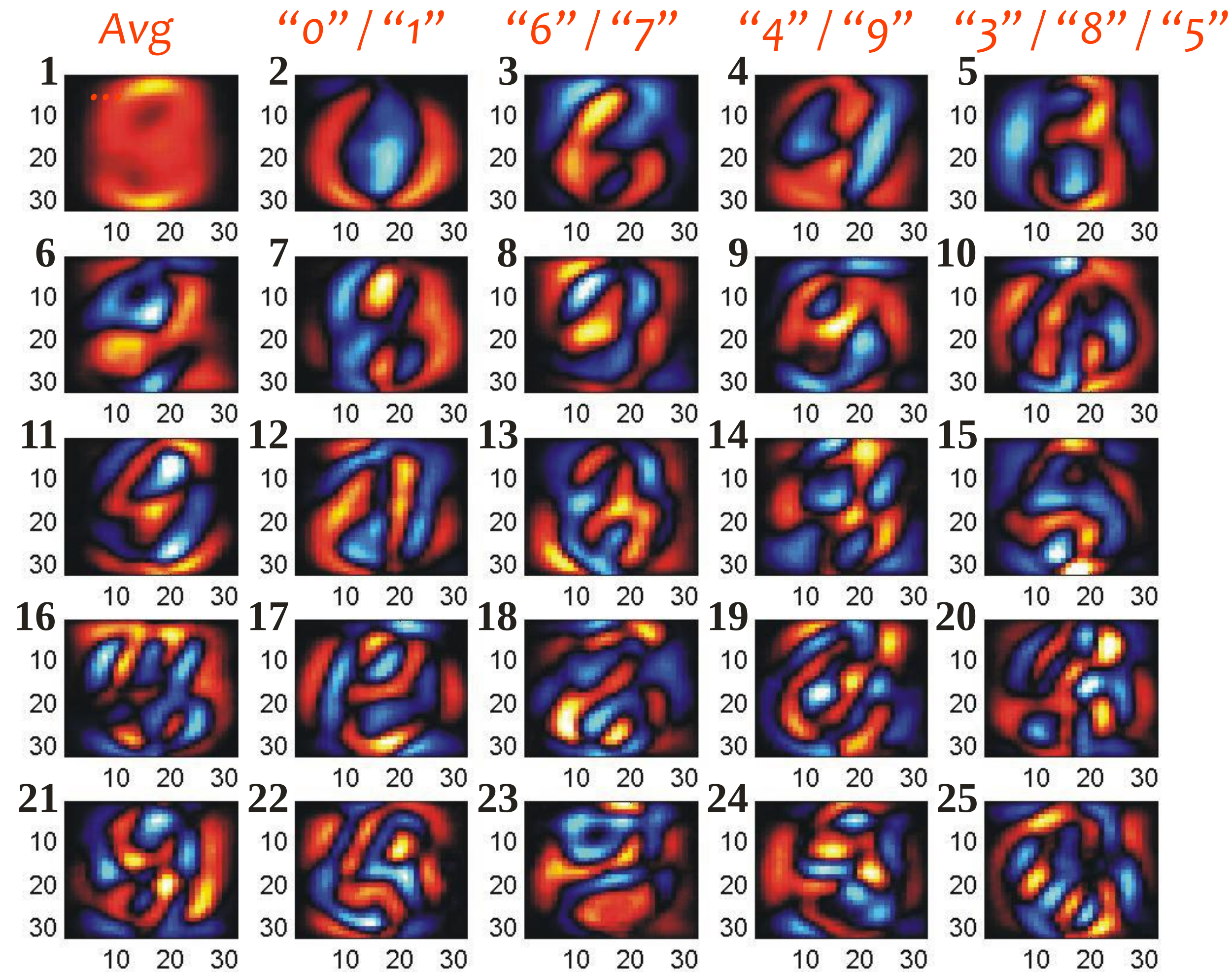
- Parameters:

  m = 1024

  n = 25

  $\lambda = 0.5$

- Rather fast convergence, starting from θ_1 on top left
- Compression of data takes place



Summary of "Clever Agents"

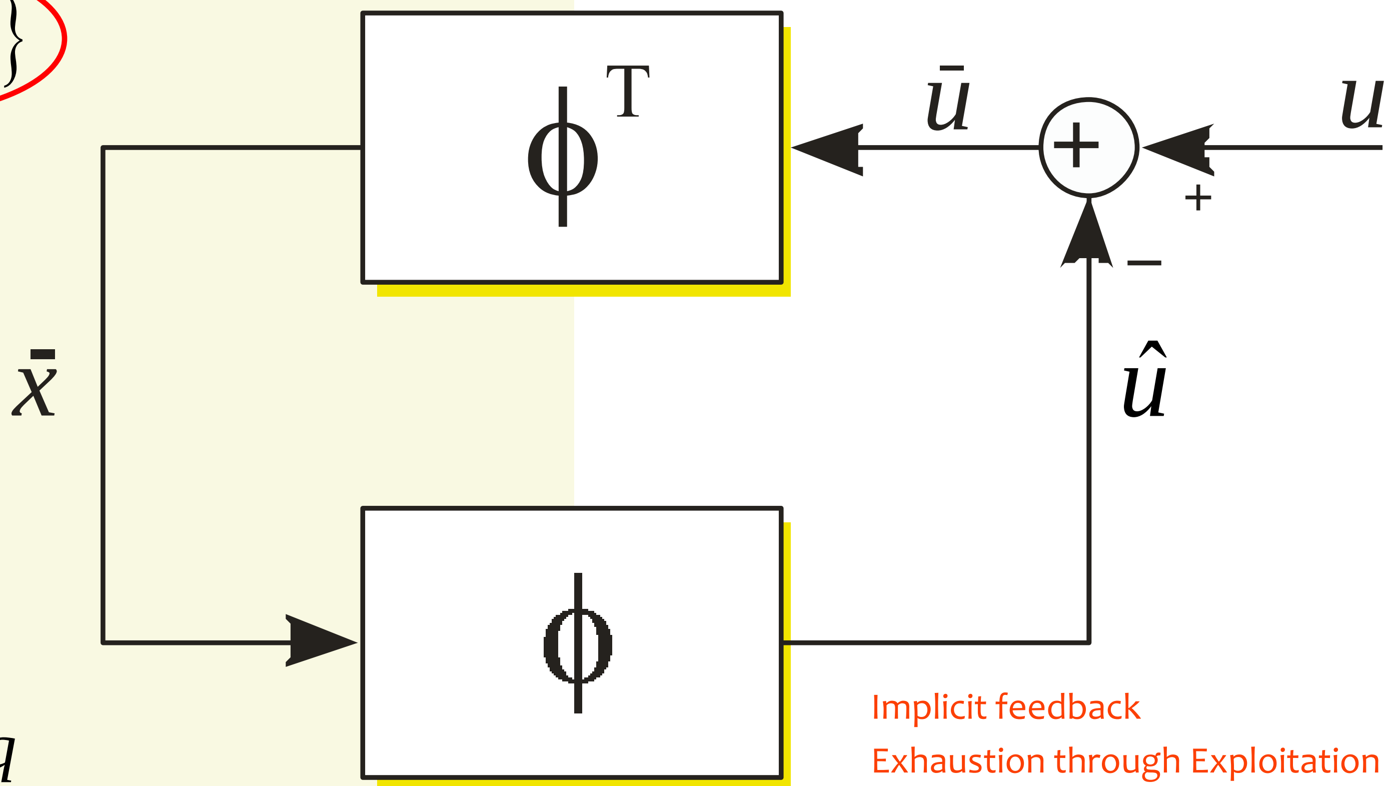
- Emergence in terms of self-regulation (stability) and self-organization (principal subspace analysis) reached
- This is reached applying physiologically plausible operations and model is linear – scalable beyond toy domains
- Learning is local – but not *completely* local: Need "communication" among neurons (anti-Hebbian structures)
- Roles of signals different: How to motivate the inversion in adaptation direction (anti-Hebbian learning)?

- Solution next: Apply non-idealities – in an unorthodox way!
- There exist no unidirectional causal flows in real life systems
- Feedback: Exploiting a signal exhausts that signal

"Stupid Agents"

$$\phi^{\mathrm{T}} = q\,\mathrm{E}\left\{\overline{x}\overline{u}^T\right\}$$

"stiffness",
coupling factor



$$\phi = \mathrm{E}\left\{\overline{u}\overline{x}^T\right\} q$$

Implicit feedback
Exhaustion through Exploitation

HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Automation and Systems Technology
Cybernetics Group

"Generalized Boltzmann Machine"?

Simply “go for resources”

- Again: Balancing is reached by feedback, but now not explicitly but implicitly through the environment

$$\begin{cases} \bar{x} = \phi^T \bar{u} \\ \bar{u} = u - \phi \bar{x} \end{cases}$$

- Also environment finds its balance as time evolves and the iteration converges (see next slide)
- Only exploiting locally visible quantities, implement evolutionary adaptation as

$$\phi^T = q \, \mathrm{E}\left\{ \bar{x} \bar{u}^T \right\}$$

How to characterize this “environmental balance”?

Self-regulation assured ...

- Study how activity in the loop behaves for short time steps:

$$x\big((k+1)h\big) = \mathrm{QE}\left\{\overline{x}\overline{u}^T\right\} \overline{u}(kh)$$

Q can be matrix, too

$$= \mathrm{QE}\left\{\overline{x}\overline{u}^T\right\} u - \mathrm{QE}\left\{\overline{x}\overline{u}^T\right\} \mathrm{E}\left\{\overline{x}\overline{u}^T\right\}^T Q\, x(kh)$$

- One can write an approximation for the derivative as $h \to 0$

$$\frac{x\big((k+1)h\big) - x(kh)}{h} = -\frac{1}{h}\left(I + \mathrm{QE}\left\{\overline{x}\overline{u}^T\right\} \mathrm{E}\left\{\overline{x}\overline{u}^T\right\}^T Q\right) x(kh) + \cdots$$

- The system matrix has (regardless of the covariance matrix) always negative eigenvalues – thus the signals remain bounded, and so do the signals.

... How about self-organization?

- Because $\bar{x} = q\mathrm{E}\left\{\overline{xu}^T\right\}\bar{u}$, one can write two covariances:

$$\mathrm{E}\left\{\overline{xu}^T\right\} = q\mathrm{E}\left\{\overline{xu}^T\right\}\mathrm{E}\left\{\overline{uu}^T\right\}$$

and

$$\mathrm{E}\left\{\overline{xx}^T\right\} = q^2\,\mathrm{E}\left\{\overline{xu}^T\right\}\mathrm{E}\left\{\overline{uu}^T\right\}\mathrm{E}\left\{\overline{xu}^T\right\}^T = q\,\mathrm{E}\left\{\overline{xu}^T\right\}\mathrm{E}\left\{\overline{xu}^T\right\}^T$$

so that

$$\begin{cases} I_n = \underbrace{\sqrt{q}\,\mathrm{E}\left\{\overline{xx}^T\right\}^{-1/2}\mathrm{E}\left\{\overline{xu}^T\right\}}_{D^T\theta^T}\;\underbrace{\mathrm{E}\left\{\overline{xu}^T\right\}^T\mathrm{E}\left\{\overline{xx}^T\right\}^{-1/2}\sqrt{q}}_{\theta D} = \theta^T\theta \\ \dfrac{1}{q} I_n = \underbrace{\sqrt{q}\,\mathrm{E}\left\{\overline{xx}^T\right\}^{-1/2}\mathrm{E}\left\{\overline{xu}^T\right\}}_{D^T\theta^T}\mathrm{E}\left\{\overline{uu}^T\right\}\underbrace{\mathrm{E}\left\{\overline{xu}^T\right\}^T\mathrm{E}\left\{\overline{xx}^T\right\}^{-1/2}\sqrt{q}}_{\theta D} \end{cases}$$

Forget the trivial solution where x_i is identically zero

- Similarly, if $\bar{x} = Q\mathrm{E}\left\{\bar{x}\bar{u}^T\right\}\bar{u}$ for some (diagonal) matrix Q:

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

and

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

Note: this has to be symmetric, so that

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

Stronger formulation is reached:

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

For non-identical q_i, this has to become diagonal also

Equalization of environmental variances

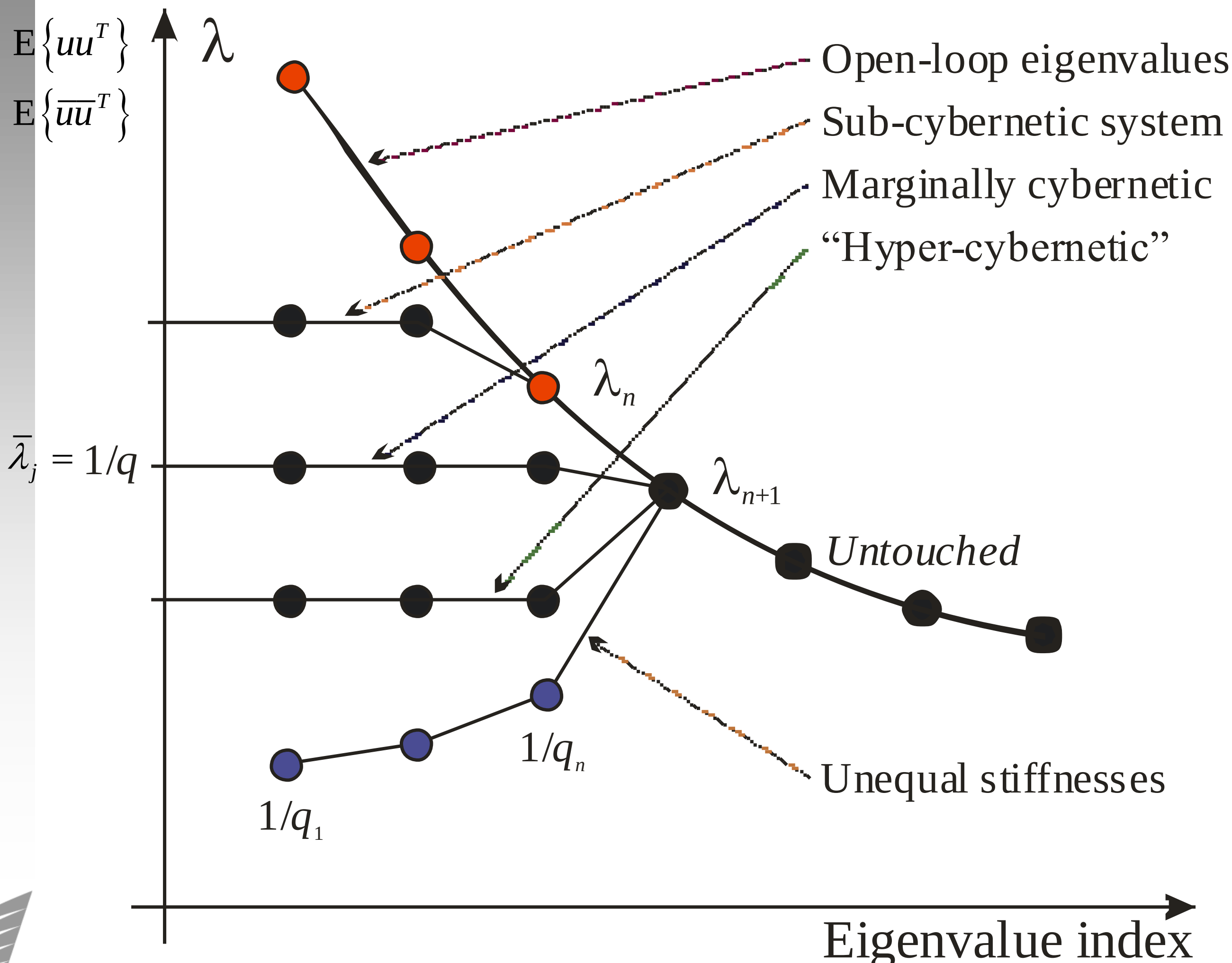
- Because $\theta^T \theta = I_n$ and $\theta^T \mathrm{E}\left\{\overline{u}\overline{u}^T\right\} \theta = Q^{-1}$, θ consists of the n (most significant) eigenvectors of $\mathrm{E}\left\{\overline{u}\overline{u}^T\right\}$, and $\mathrm{E}\left\{uu^T\right\}$
- If $n = m$, the variation structure becomes trivial:

$$\mathrm{E}\left\{\overline{u}\overline{u}^T\right\} = \frac{1}{q} I_m \qquad \text{or} \qquad \mathrm{E}\left\{\overline{u}\overline{u}^T\right\} = Q^{-1}$$

- Visible data variation becomes *whitened* by the feedback
- Relation to ICA : Assume that this whitened data is further processed by neurons (FOBI) – but this has to be nonlinear!

- On the other hand, if q_i are different, the modes become separated in the PCA style (rather than PSA)

Still try to avoid nonlinearity!



DEMO

`equalvar.m`

“Emergence of Intelligence”

- Solving for the balance signals, one can write

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

- This can be rewritten in a form where there is $\mathrm{E}\left\{ \overline{xu}^T \right\}$ as the feedforward matrix, and $\mathrm{E}\left\{ \overline{xx}^T \right\} + Q^{-1}$ as the feedback matrix
  Of course, u has changed to $\bar{u}$ as it is the only signal visible
- It turns out that if the coupling increases, so that $q_i \to \infty$, this structure equals the prior one with explicit feedback

- This means that in evolution (to be studied later closer) a selfish agent finally becomes a clever agent
- Without preprogramming, an agent becomes “context-aware”

- Because

$$\mathrm{E}\left\{\overline{xx}^T\right\} = Q\mathrm{E}\left\{\overline{xu}^T\right\}\mathrm{E}\left\{\overline{xu}^T\right\}^T = \mathrm{E}\left\{\overline{xu}^T\right\}\mathrm{E}\left\{\overline{xu}^T\right\}^T Q$$

there must hold

$$\mathrm{E}\left\{\overline{xx}^T\right\}Q = Q\,\mathrm{E}\left\{\overline{xx}^T\right\}$$

and, further, $f\left(\mathrm{E}\left\{\overline{xx}^T\right\}\right) g(Q) = g(Q)\, f\left(\mathrm{E}\left\{\overline{xx}^T\right\}\right)$ for any functions *f* and *g*

- What does this mean for covariance for different *Q* ?!
- The commutativity makes manipulations with the expressions simpler, for example...

Variance inheritance

- Further – study the relationship between $\bar{x}$ and original u:

$$\bar{x} = \left(I_n + q^2 \ \mathrm{E}\left\{ \bar{x}\bar{u}^T \right\} \mathrm{E}\left\{ \bar{x}\bar{u}^T \right\}^T \right)^{-1} q\mathrm{E}\left\{ \bar{x}\bar{u}^T \right\} u$$

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

Multiply from the right by transpose, and take expectations:

$$\left(I_n + q \ \mathrm{E}\left\{ \bar{x}\bar{x}^T \right\} \right) \mathrm{E}\left\{ \bar{x}\bar{x}^T \right\} \left(I_n + q \ \mathrm{E}\left\{ \bar{x}\bar{x}^T \right\} \right)$$

$$= \mathrm{E}\left\{ \bar{x}\bar{x}^T \right\}^{1/2} \left(I_n + q \ \mathrm{E}\left\{ \bar{x}\bar{x}^T \right\} \right)^2 \mathrm{E}\left\{ \bar{x}\bar{x}^T \right\}^{1/2}$$

$$= q^2 \mathrm{E}\left\{ \bar{x}\bar{u}^T \right\} \mathrm{E}\left\{ uu^T \right\} \mathrm{E}\left\{ \bar{x}\bar{u}^T \right\}^T$$

$$\left(I_n + q\,\mathrm{E}\left\{\overline{xx}^T\right\}\right)^2$$

$$= q\,\underbrace{\sqrt{q}\,\mathrm{E}\left\{\overline{xx}^T\right\}^{-1/2}\,\mathrm{E}\left\{\overline{xu}^T\right\}}_{D^T\theta^T}\,\mathrm{E}\left\{uu^T\right\}\underbrace{\mathrm{E}\left\{\overline{xu}^T\right\}^T\,\mathrm{E}\left\{\overline{xx}^T\right\}^{-1/2}\sqrt{q}}_{\theta D}$$

Solving for the latent covariance:

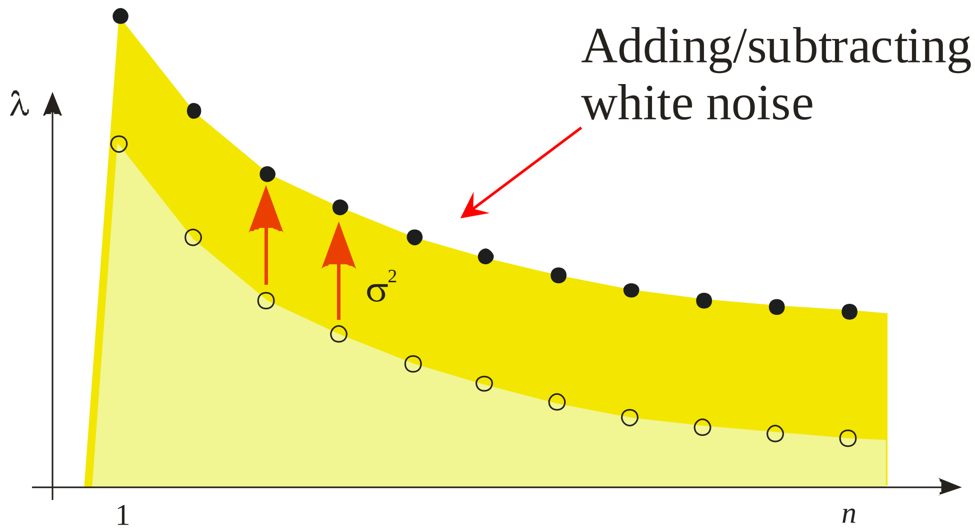
$$\mathrm{E}\left\{\overline{xx}^T\right\} = \frac{1}{q}\left(q\,D^T\theta^T\,\mathrm{E}\left\{uu^T\right\}\theta D\right)^{1/2} - \frac{1}{q}\,I_n$$

This means that the external and internal eigenvalues (variances) are related as follows:

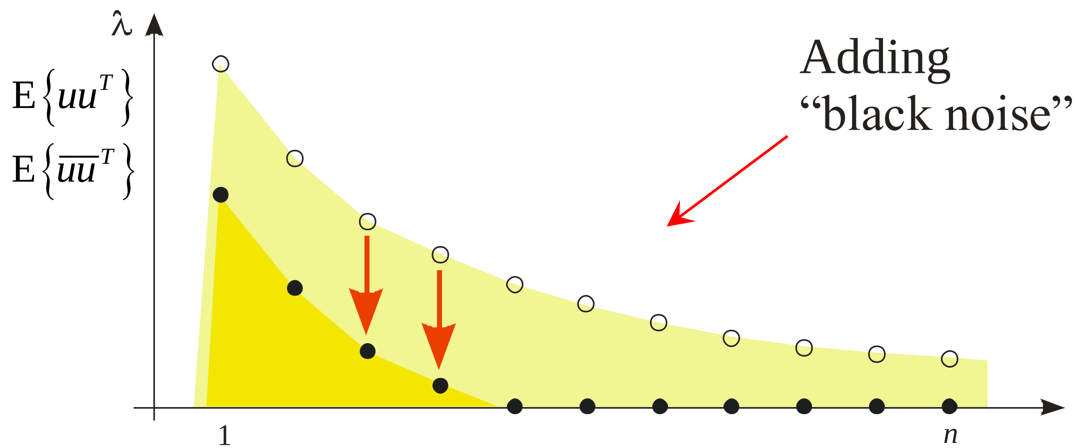
$$\lambda_i = \frac{\sqrt{q_i\lambda_j}-1}{q_i}$$ – for pairs, there must hold $q_i\lambda_j > 1$

HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Automation and Systems Technology
Cybernetics Group

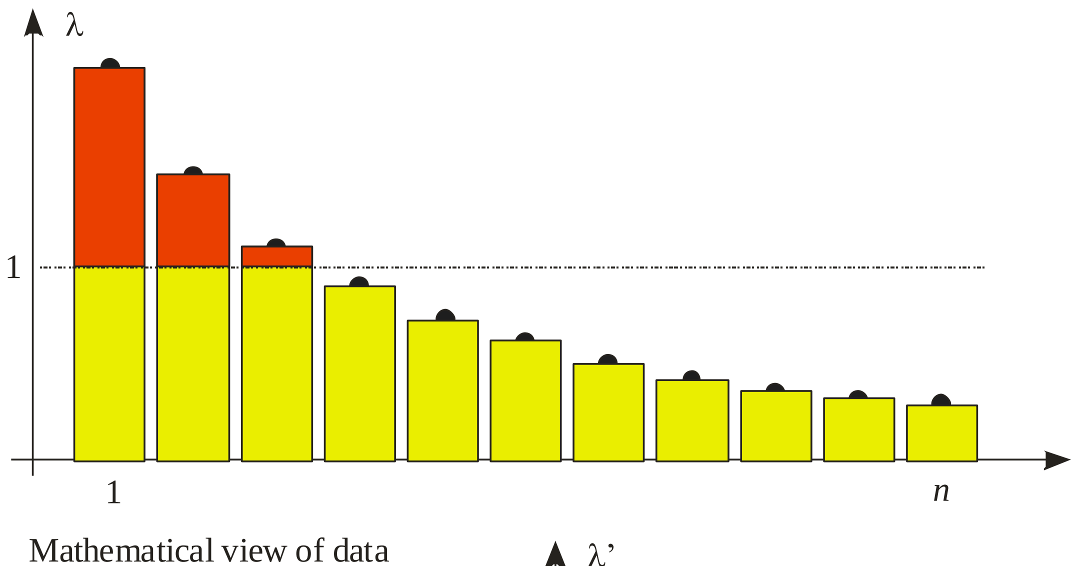
Effect of feedback = add "black noise"



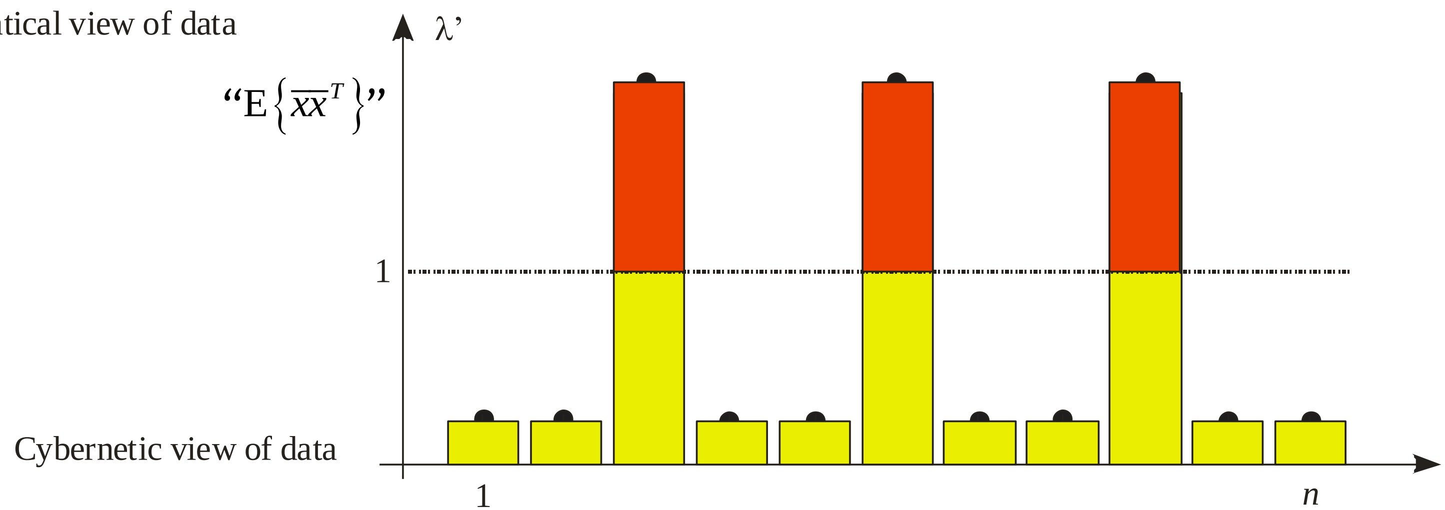
- White noise = Constant increase in all directions
- "Black noise" = Decrease in all directions (if poss.)



Results of orthogonal basis rotations

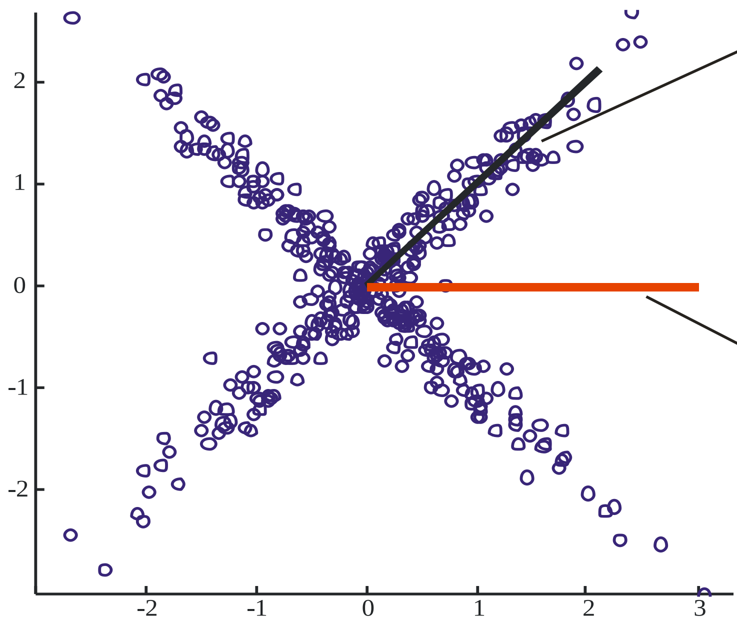


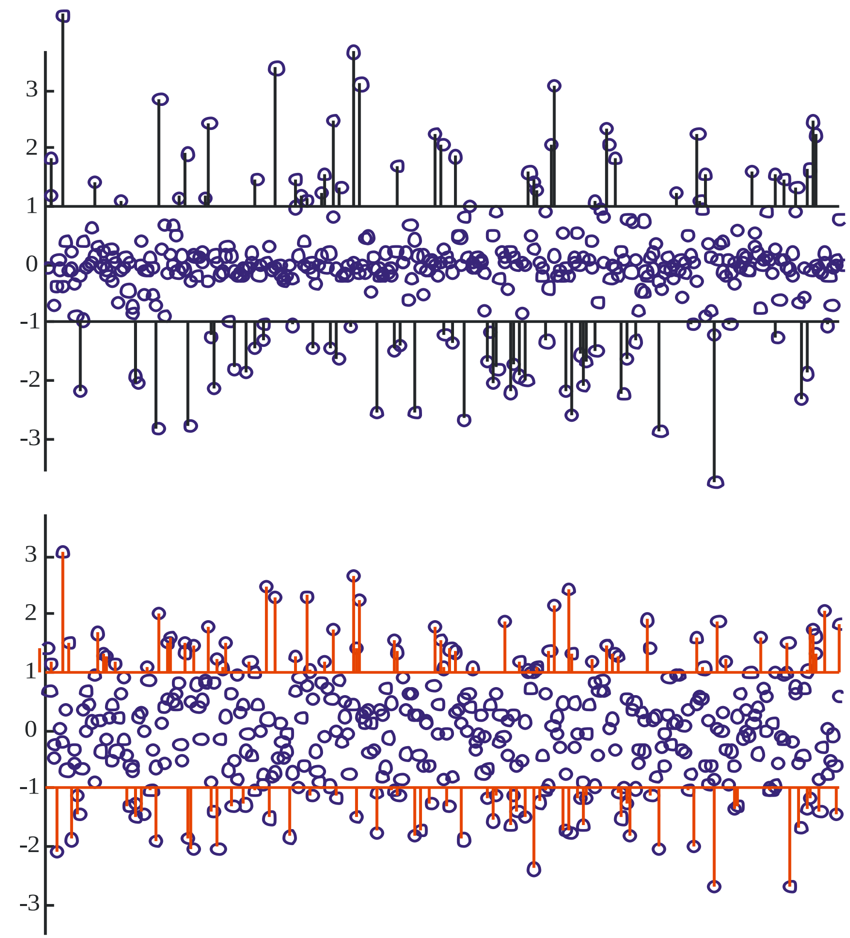
- Total variance above zero level intact regardless of the rotations
- Total variance above 1 changes!



"Factor Analysis"

- Algorithm rotates the PCA subspace axes

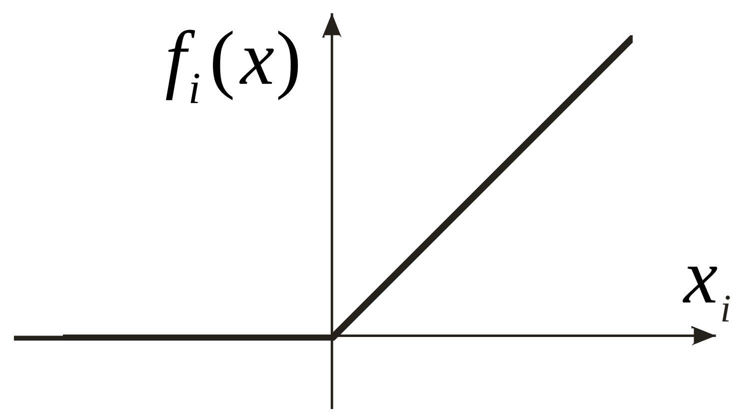




Towards differentiation of features

- A simple example of nonlinear extensions: CUT function
- If variable is positive, let it through; otherwise, filter it out – Well in line with modeling of activity in neuronal systems:
  - Frequencies cannot become negative (interpretation in terms of pulse trains)
  - Concentrations cannot become negative (interpretation in terms of chemicals)
  - Sizes of neuron populations cannot become negative
  - Power / information content cannot become negative, etc.
- Makes modes separated
- Still: *End result almost linear!*

$$f_i(x) = \begin{cases} x_i, & \text{when } x_i > 0 \\ 0, & \text{when } x_i \le 0 \end{cases}$$



Symmetry breaking?!

Algorithm for Hebbian feedback learning...

```
LOOP – iterate for data in kxm matrix U

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

  % Enhance sparsity by cut nonlinearity (or iterate by row)
  Xbar = Xbar.*(Xbar>0);

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

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

  % To maintain system activity qi can be adapted according to diag(Exx)

END
```

… resulting in *Sparse Components!*

- Parameters:

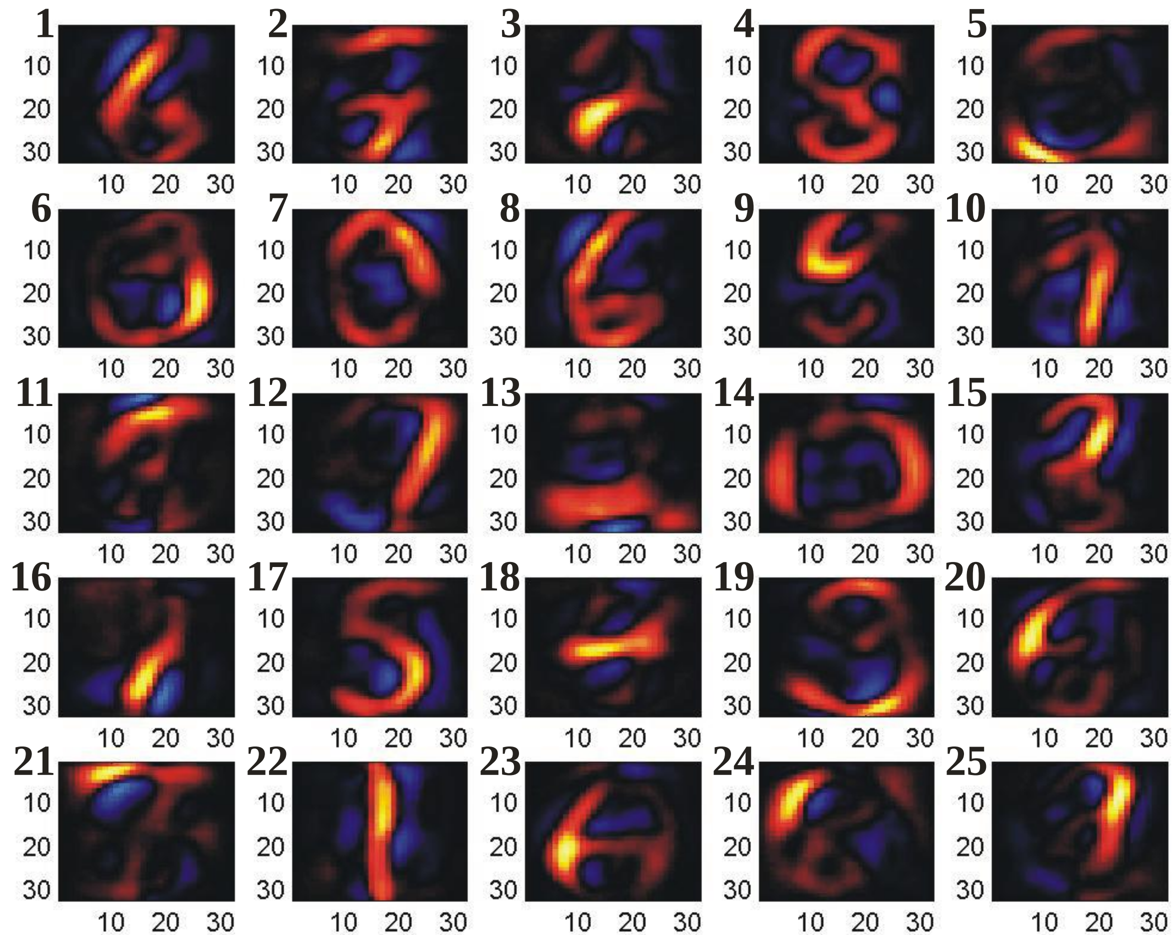
$m = 1024$

$n = 25$

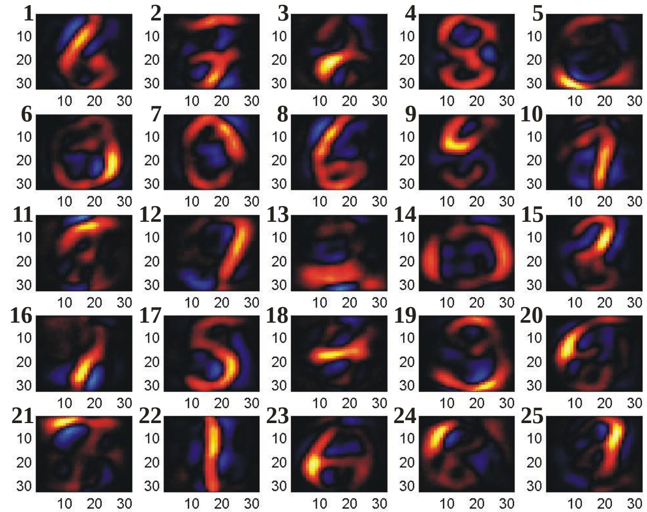
$\lambda = 0.97$

DEMO

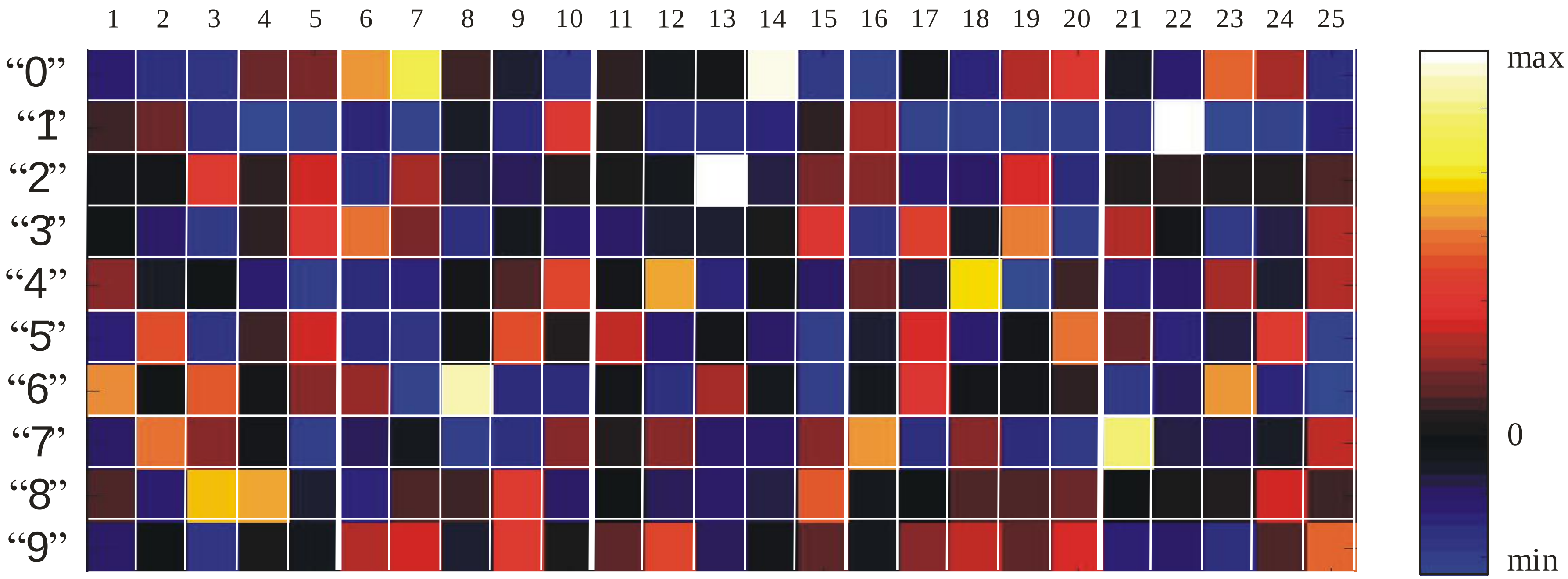
`digitfeat.m`



HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Automation and Systems Technology
Cybernetics Group



- “Work load” becomes distributed
- Correlations between inputs and neuronal activities shown below:



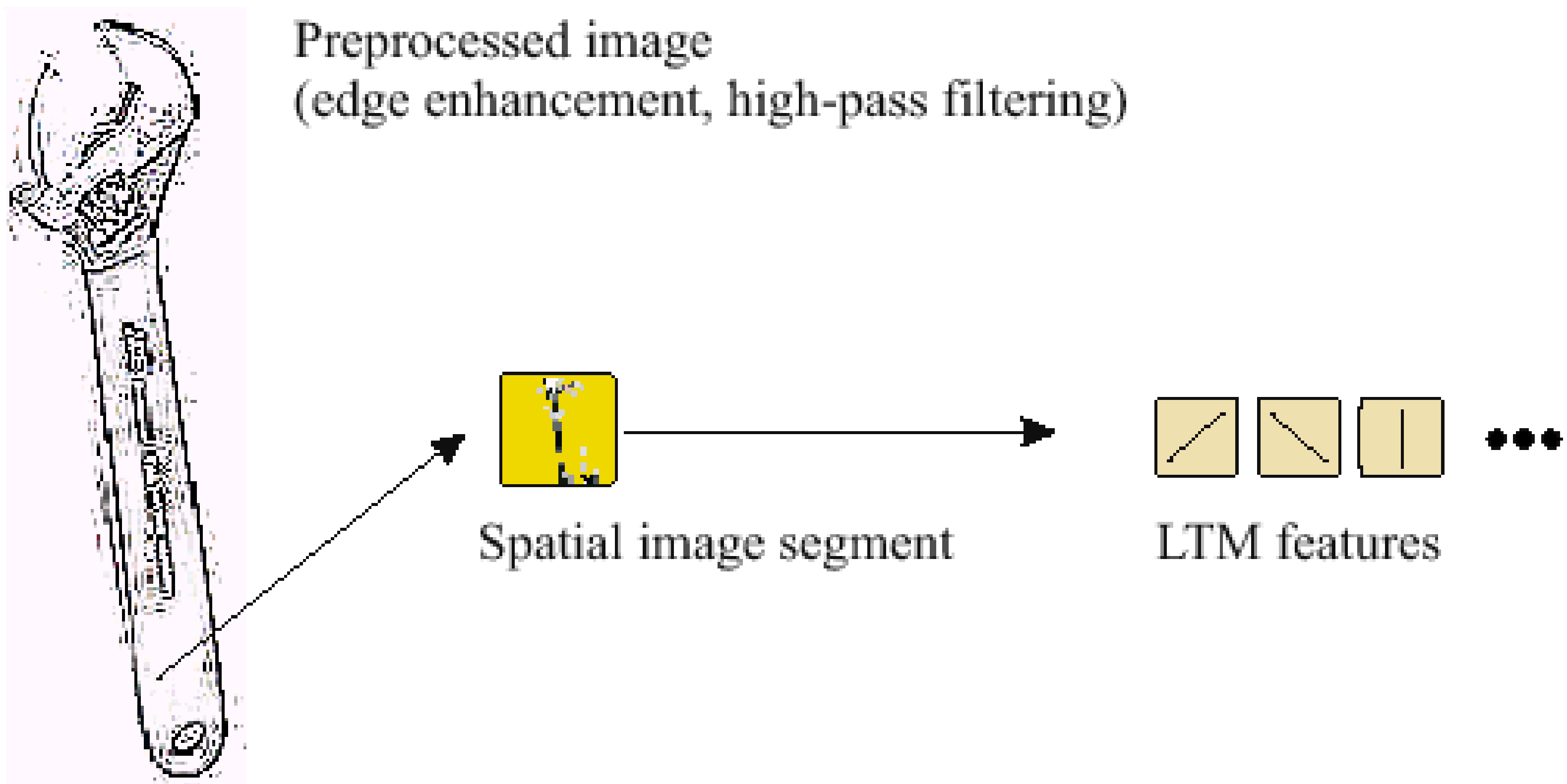
Sparse coding

- Traditional modeling goal – mathematically optimal and simple to implement: *Minimize the overall size of the model*
- Sparsity-oriented modeling goal – often physically relevant, but cumbersome to characterize: *Minimize the number of simultaneously active representations* while there are no acute constraints on the overall model size
- There is a tradeoff: too sparse model becomes inaccurate

- Note the (intuitive) connection to cognitive system:
  There are no acute constraints on the long-term memory (LTM), whereas the size of the working memory / short-term memory is limited to 7 +/– 2

Studied later ...

Visual V1 cortex seems to do this kind of decomposing



NEW

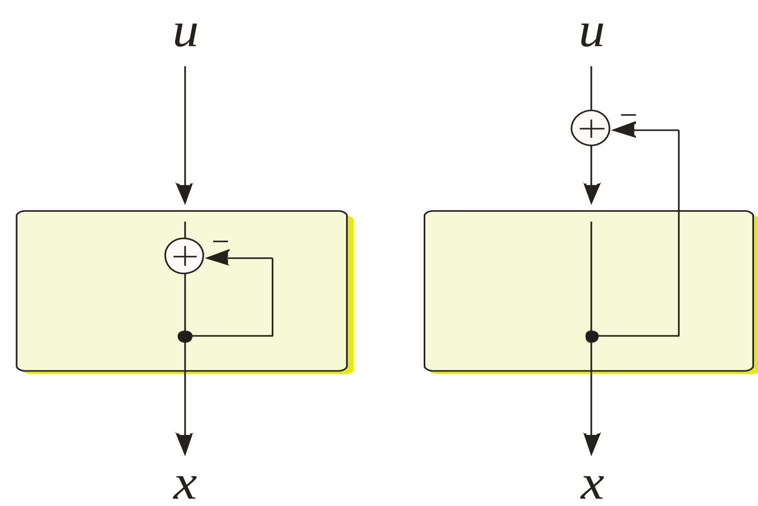
Evolution towards the neocybernetic model?!

- *Hopfield networks* (Hopfield 1982)
  - Energy function and iterative process towards minimum = balance
- *Boltzmann machines* (Hinton, Sejnowski 1983)
  - Simple local learning principle
- *Restricted Boltzmann machines* (Smolensky 1986, Hinton)
  - Feedback only through the environment, real variables
- … Issues still to be solved to reach true acceptance:
  - Stochastic learning to be made more consistent + faster
  - Theoretical work to be done – complex task because of the distributions
  - Explicit restrictions (like "no reciprocity") to be relaxed, …

Conclusion: Clever vs. stupid agents

Both agent types can exist – resulting systems very different

“Intelligent agent”	**“Stupid/selfish agent”**
Hebbian/anti-Hebbian learning	Hebbian environmental learning
$\bar{x} = \mathrm{E}\{\bar{x}\bar{x}^T\}^{-1}\mathrm{E}\{\bar{x}u^T\}\ u$	$\bar{x} = q\mathrm{E}\{\bar{x}\bar{u}^T\}\ \bar{u}$
Principal Component (Subspace) Analysis	Sparse Component (Subspace) Analysis
Complete insulation (*or* elimination) of environment	Stiffness regulation, equalization of environment



Once more

Applying the neocybernetic principles there is *emergence*:

Local model of agent's own behavior (average match among input and state) changes to a global model of the whole system's behavior (sparse coded principal subspace basis).