Cognitive Neuroscience II
Lecture 7

Prof. Dr. Andreas Wendemuth

Lehrstuhl Kognitive Systeme

Institut für Elektronik, Signalverarbeitung und Kommunikationstechnik

Fakultät für Elektrotechnik und Informationstechnik
Otto-von-Guericke Universität Magdeburg

http://iesk.et.uni-magdeburg.de/ko/

Resumé of previous lecture 6

- Hebbian-type rules are biologically plausible and motivated
- Ocular dominance is a prominent example which can be modelled with Hebb rules

8 Plasticity and Learning

- 02. May Hebb Rules, PCA

Hebb Rules

- Donald Hebb (1949): If input from neuron A contributes to firing of neuron B, the synaptic strength / weight w from A to B should be strengthened.
- Basic (linear) Hebb rule for one pattern:

$$\tau_w \frac{d\mathbf{w}}{dt} = F(v\mathbf{u}) = v\mathbf{u}$$

Recall firing rate equation

- Fir. Rate eq.: $$\tau_r \frac{dv}{dt} = -v + F(\mathbf{w} * \mathbf{u})$$
- Linear version: $$\tau_r \frac{dv}{dt} = -v + \mathbf{w} * \mathbf{u}$$
  has strong deficiencies (unlimited growth, 2^{nd} order statistics) but for the moment is easier to handle.
- Hebb learning is much slower than firing dynamics, hence $\tau_w >> \tau_r$ and the firing dynamics can be assumed in equilibrium for Hebb learning, i.e. $v = \mathbf{w} * \mathbf{u}$

Hebb Rule for equilibrium firing

- Obtain $\tau_w \frac{d\mathbf{w}}{dt} = v\mathbf{u} = (\mathbf{u}\mathbf{u}^T) * \mathbf{w} = \mathbf{Q} * \mathbf{w}$
- $(\mathbf{u}\mathbf{u}^T)$ is an *outer product*, i.e. forms the *input correlation matrix* **Q** with components $Q_{ij} = (\mathbf{u}\mathbf{u}^T)_{ij} = u_i u_j$
- If we have an ensemble of p input patterns, these can be presented one after the other (sequential learning), or, almost equivalently, –as a thought model – in parallel, which leads to averaging $< (.) >= \frac{1}{p} \sum_{\mu=1}^{p} (.)$ with

$$\tau_w \frac{d\mathbf{w}}{dt} =< v^\mu \mathbf{u}^\mu >=< \mathbf{Q}^\mu > * \mathbf{w} \quad \text{or} \quad \tau_w \frac{dw_k}{dt} =< v^\mu u_k^\mu >= \sum_{i=1}^{N} < Q_{ki}^\mu > w_i$$

Unlimited growth of |**w**|

- „Multiplying“ the Hebb rule $\tau_w \frac{d\mathbf{w}}{dt} = v\mathbf{u}$ with **w:**

$$\tau_w \frac{d\,|\mathbf{w}|^2}{dt} = 2\tau_w \frac{d\mathbf{w}}{dt} * \mathbf{w} = (Hebb) = 2v\mathbf{w} * \mathbf{u} = (Fir.Rate) = 2v^2 \geq 0$$

  i.e. the length (norm) of **w** will increase in every learning step, sequential or parallel, (other than in trivial cases v=0) . Since $v = \mathbf{w} * \mathbf{u} \propto |w|$, these increases will add up unlimitedly.

- This is a consequence of the linearization of the activation function F. If F saturates, growth is limited.

The Covariance Rule

- The basic Hebb rule can be interpreted as modelling the difference in activity against a base level. In this case, the mean <u>=0.
- If <u>≠0, we subtract it as a *presynaptic threshold*, arriving at $\tau_w \frac{d\mathbf{w}}{dt} = v(\mathbf{u} - <\mathbf{u}>) = ((\mathbf{u} - <\mathbf{u}>)\mathbf{u}^T) * \mathbf{w}$
- Since $\mathbf{C} = <(\mathbf{u} - <\mathbf{u}>)(\mathbf{u} - <\mathbf{u}>)^T> = <(\mathbf{u} - <\mathbf{u}>)\mathbf{u}^T>$ is the input *covariance* matrix, we get for <u> ≠ 0 the <u>covariance rule</u> $\tau_w \frac{d\mathbf{w}}{dt} = \mathbf{C} * \mathbf{w}$

Ex 1

- a) Show $<(\mathbf{u}-<\mathbf{u}>)(\mathbf{u}-<\mathbf{u}>)^T> = <(\mathbf{u}-<\mathbf{u}>)\mathbf{u}^T>$
- b) Show that the same effect of covariance normalization can be reached by subtracting a *postsynptic threshold*, i.e. show that

$$\tau_w \frac{d\mathbf{w}}{dt} = (v - <v>)\mathbf{u}$$

also leads to

$$\tau_w \frac{d\mathbf{w}}{dt} = \mathbf{C} * \mathbf{w}$$

Solution of Hebbian dynamics

- The Hebb rule (pattern-averaged or not)

$$\tau_w \frac{d\mathbf{w}}{dt} = \mathbf{C} * \mathbf{w}$$

where **Q** is regarded as a special case of **C**, can be solved by eigenvalue decomposition of **C** with eigenvalues λ_j and eigenvectors $\mathbf{e}^j$.

$$\mathbf{w}(t) = \sum_{j=1}^{N} (\mathbf{e}^j * \mathbf{w}^j(t=0))\ \mathbf{e}^j \exp(\frac{\lambda_j}{\tau_w} t)$$

- The $\mathbf{e}^j * \mathbf{w}^j(t=0)$ are the projections of the initial weights on the eigenvectors.

Ex 2:

- Show that the covariance rule

$$\tau_w \frac{d\mathbf{w}}{dt} = \mathbf{C} * \mathbf{w}$$

has the solution

$$\mathbf{w}(t) = \sum_{j=1}^{N} (\mathbf{e}^j * \mathbf{w}^j (t=0)) \ \mathbf{e}^j \exp(\frac{\lambda_j}{\tau_w} t)$$

Long-time development

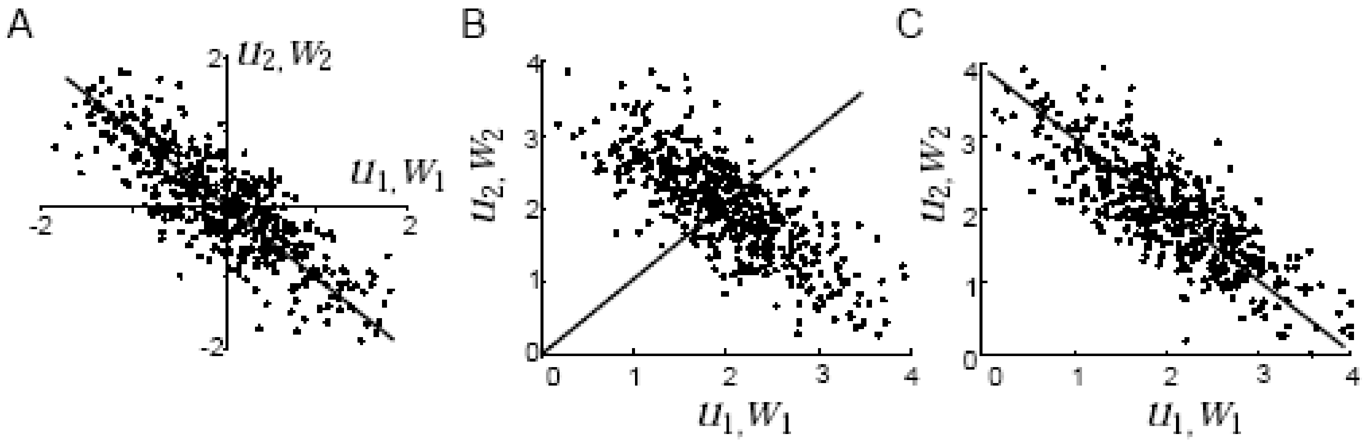
- If the initial weight vector (t=0) has components in all eigenvector directions, the long-time development will be governed by the *largest eigenvalue*, i.e.

$$\mathbf{w}(t) \xrightarrow{t\to\infty} \mathbf{e}^{j=\max} \exp\left(\frac{\lambda_{\max}}{\tau_w} t\right)$$

- The eigenvector with largest eigenvalue is called the *principal* eigenvector.
- Clearly, |**w**| will grow unlimitedly.

Principal component analysis (1)

- The eigenvectors of a covariance matrix (!) select the directions of an approximative Gaussian multinomial distribution. Large eigenvalues correspond to large variances. Example: Gaussian data:



Eigenvalues of Covariance Matrix

- Eigenvalue conditions: $\lambda \mathbf{v} = \mathbf{C}\mathbf{v} = \sum_\mu \mathbf{u}^\mu \mathbf{u}^{\mu T} \mathbf{v}$
- Multiply from left with $\mathbf{v}^\mathrm{T}$:

$$\lambda = \mathbf{v}^\mathbf{T} \lambda \mathbf{v} = \mathbf{v}^\mathbf{T} \mathbf{C}\mathbf{v} = \sum_\mu \mathbf{v}^\mathbf{T} \mathbf{u}^\mu \mathbf{u}^{\mu T} \mathbf{v} = \sum_\mu (\mathbf{u}^{\mu T} \mathbf{v})^2 \geq 0$$

- The last sum is called a „perfect square“
- Hence the eigenvalues of a real-valued covariance matrix are not negative.

Principal component analysis (2)

- Note that if the distribution of patters is non-Gaussian, a best Gaussian fit to the data is assumed implicitely by PCA.
- Non-Gaussian distributions have central correlation moments of higher order, $< (\mathbf{u} - < \mathbf{u} >)^n > \neq 0$ for some n = 3,4,....
- These are not modelled by PCA. Neural models with nonlinear activation function model those so-called *higher order statistics*. (higher than 2)

Example: ocular dominance (1)

- Consider a single layer 4 cell which receives input from 2 LGN afferents, associated with the 2 eyes (R,L), with activities u. Both eyes are statistically equivalent.
- Cov.: $\mathbf{Q} = \langle \mathbf{u}\mathbf{u}^T \rangle = \begin{pmatrix} \langle u_R u_R \rangle & \langle u_R u_L \rangle \\ \langle u_L u_R \rangle & \langle u_L u_L \rangle \end{pmatrix} = \begin{pmatrix} q_S & q_D \\ q_D & q_S \end{pmatrix}$

  where „S“=Same and „D“=Different
- PCA: $e^1 = (1,1)\ ;\ \lambda_1 = q_S + q_D \qquad e^2 = (1,-1)\ ;\ \lambda_2 = q_S - q_D$

Ocular dominance (2)

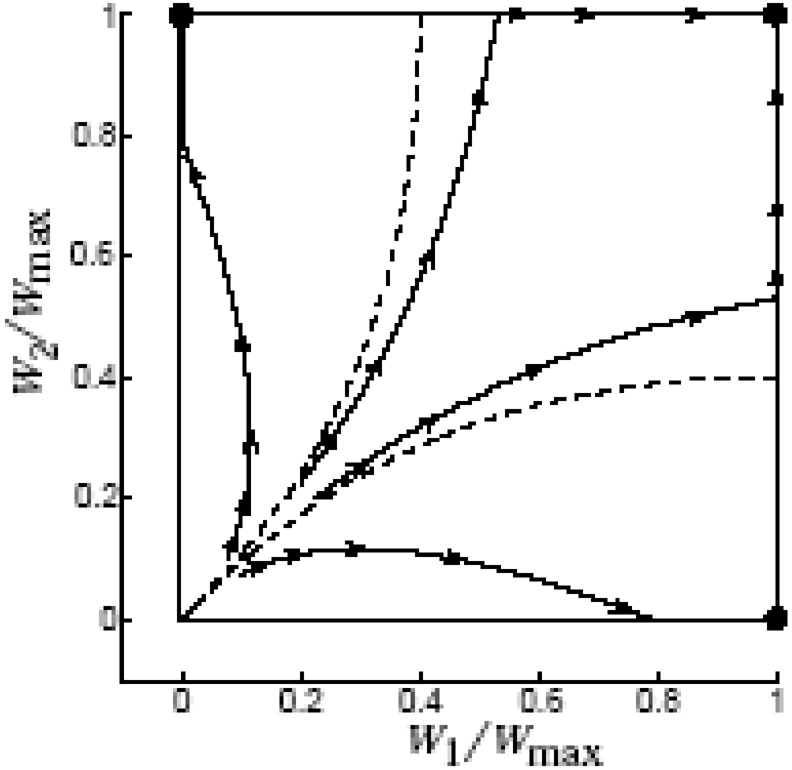
- If correlation between eyes is positive, $q_D > 0$. Then the principal eigenvector is $e^1 = (1,1)\ ;\lambda_1 = q_S + q_D$, representing the combined weight vector $w_R + w_L$.
- After some Hebbian Learning time, the weights will be proportional to $w_R + w_L$, whereas the other eigenvector is suppressed, i.e. $w_R - w_L \rightarrow 0$.
- This means that both eyes contribute equal innervation, so no ocular dominance occurs.
- Hebb has failed ?????

Ex 3

- Derive the ocular dominance behaviour with Hebbian learning in the simple presented model.

Ocular dominance with saturation (3)

- With the (biologically plausible) saturation of weights $0<w<w_{max}$, the outcome of Hebbian learning depends on the initial overlaps e*w and the products λt:
- If „few“ time has elapsed and saturation is already reached, the outcome is rather determined by the initial overlaps than by the largest eigenvalue [here = (1,-1)]:



The Oja rule (1982)

- The Oja rule affects weight normalization by only requiring information local to the synapses, but ”**w** (multiplicative normalization):

$$\tau_w \frac{d\mathbf{w}}{dt} = v\mathbf{u} - \alpha v^2 \mathbf{w}$$

- The weights grow as:

$$\tau_w \frac{d\,|\mathbf{w}|^2}{dt} = 2\tau_w \frac{d\mathbf{w}}{dt} * \mathbf{w} = (Oja) = 2v\mathbf{w} * \mathbf{u} - 2\alpha v^2\,|\mathbf{w}|^2$$

$$= (Fir.rate) = 2v^2(1 - \alpha\,|\mathbf{w}|^2)$$

so finally weights are normalized $|\mathbf{w}|^2 = 1/\alpha$

Oja Rule (2)

- Expressing the Oja Rule fully in terms of **w**:

$$\tau_w \frac{d\mathbf{w}}{dt} = \left(\mathbf{C} - \alpha(\mathbf{w}^T * \mathbf{C} * \mathbf{w})\mathbf{I}\right) * \mathbf{w}$$

- This is highly nonlinear in **w**. Writing **w** in C-eigenvector coordinates gives for component k:

$$\tau_w \frac{dw_k}{dt} = \left(\lambda_k - \alpha \sum_j \lambda_j w_j^2\right) w_k$$

- Since the sum term is the same for all components k, the maximum in () will be at component k with maximum λ. So the Oja rule selects the principal component of **C** as well.

Oja rule (3)

- After some time, all components other than w_1 have been suppressed, and we obtain

$$\tau_w \frac{dw_1}{dt} = \left(1 - \alpha w_1^2\right)\lambda_1 w_1$$

- Even though the factor λ_1 would increase w_1, this increase is brought to a halt by the factor $\left(1 - \alpha w_1^2\right)$ which will not allow for a further increase after $|\mathbf{w}|^2 = 1/\alpha$,which we saw already.

Subtractive Normalization (1)

- A biologically non-plausible (non-local) way of suppressing the principal eigenvector $\mathbf{e}^1$ is to force the solution to be orthogonal to it:

$$\tau_w \frac{d\mathbf{w}}{dt} = v\mathbf{u} - v(\mathbf{e}^1 * \mathbf{u})\mathbf{e}^1 \qquad \text{[subtr. normalization]}$$

- The orthogonality is strictly enforced:

$$\tau_w \frac{d(\mathbf{e}^1\mathbf{w})}{dt} = v\mathbf{e}^1\mathbf{u} - v(\mathbf{e}^1 * \mathbf{u})(\mathbf{e}^1\mathbf{e}^1) = 0$$

Subtractive Normalization (2)

- Writing **w** and **u** in **C**-eigenvector coordinates gives for component k≠1: $\tau_w \frac{dw_k}{dt} = \lambda_k w_k$,

  hence standard Hebbian behaviour. Of course an initial component of **w** in these directions is required.
- For the component k=1 we show the behaviour as follows:

Subtractive Normalization (3)

- Writing all vectors **w,u** with a component in $\mathbf{e}^1$ –direction and a component (‘) orthogonal to $\mathbf{e}^1$ gives:

$$\tau_w \frac{d\left[(\mathbf{e}^1 * \mathbf{w})\mathbf{e}^1 + \mathbf{w}'\right]}{dt} = v[(\mathbf{e}^1 * \mathbf{u})\mathbf{e}^1 + \mathbf{u}'] - v(\mathbf{e}^1 * \mathbf{u})\mathbf{e}^1 = v\mathbf{u}'$$

- Hence $\tau_w \frac{d\left[(\mathbf{e}^1 * \mathbf{w})\mathbf{e}^1\right]}{dt} = 0$,i.e. the component of the initial weight vector in $\mathbf{e}^1$ –direction is never changed.

Multiple Subtractive Normalization (4)

- Subtraction of the k largest eigenvectors can be enforced by setting

$$\tau_w \frac{d\mathbf{w}}{dt} = v\mathbf{u} - v\sum_{j=1}^{k} (\mathbf{e}^k * \mathbf{u})\mathbf{e}^k$$

- This can be used to have several neurons be sensitive to the largest, 2nd largest, ... , kth largest eigenvector.

Subtractive Normalization combined with Oja rule (1)

- Combining both rules gives:

$$\tau_w \frac{d\mathbf{w}}{dt} = v\mathbf{u} - v(\mathbf{e}^1 * \mathbf{u})\mathbf{e}^1 - \alpha v^2 \mathbf{w}$$

- Again, we write this rule in **C**-eigenvector coordinates. k=1 gives: $\tau_w \frac{dw_1}{dt} = -\alpha v^2 w_1$

  i.e. the first component will decay exponentially.

- This is better than the former subtractive normalization where the initial value $w_1(t=0)$ remained.

Subtractive Normalization combined with Oja rule (2)

- Generally:

$$\tau_w \frac{d\mathbf{w}}{dt} = (\mathbf{C} - \alpha v^2 \mathbf{I})\mathbf{w} - v(\mathbf{e}^1 * \mathbf{u})\mathbf{e}^1$$

- In **C**-eigenvector components, k ≠ 1 :

$$\tau_w \frac{dw_k}{dt} = \left(\lambda_k - \alpha \sum_j \lambda_j w_j^2 \right) w_k$$

  has Oja characteristics.

- Summary: exponential decay in first ev, selection of 2nd largest ev, weight normalization.

Ocular dominance with subtractive normalization + Oja(1)

- Recalling $\mathbf{Q} = <\mathbf{u}\mathbf{u}^T> = \begin{pmatrix} <u_R u_R> & <u_R u_L> \\ <u_L u_R> & <u_L u_L> \end{pmatrix} = \begin{pmatrix} q_S & q_D \\ q_D & q_S \end{pmatrix}$
- We use subtractive normalization + Oja with

$$\tau_w \frac{d\mathbf{w}}{dt} = v\mathbf{u} - 0.5v(u_R + u_L)\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \alpha v^2 \mathbf{w}$$

- The time-discrete version is:

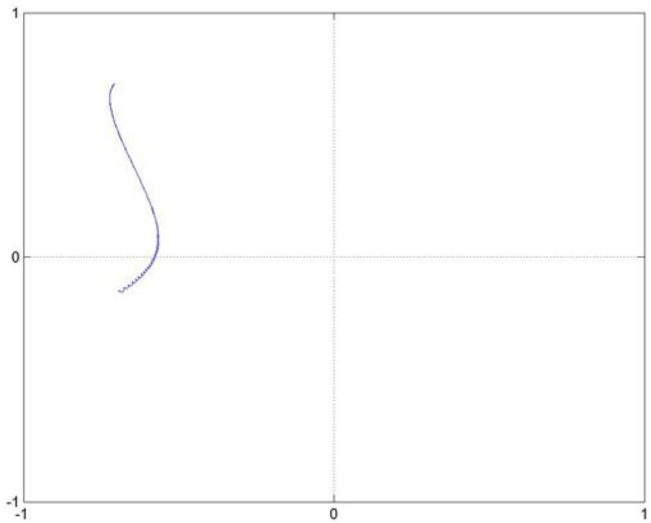
$$\Delta\mathbf{w} = \frac{\Delta t}{\tau_w}\left[v\mathbf{u} - 0.5v(u_R + u_L)\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \alpha v^2 \mathbf{w} \right] \quad \text{and} \quad v = \mathbf{u} * \mathbf{w}$$

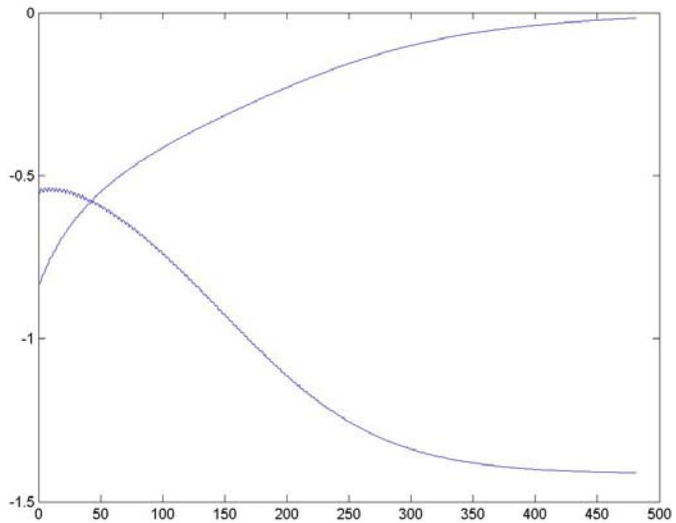
Ocular dominance with subtractive normalization + Oja (2)

- Choose different **w**(t=0) and let the sequence of (u_R, u_L) be mean-free ($< u > = 0$) to avoid the need for covariances:
- u_R =1,2,1,-1,-2,-1 and cyclic repetition
  u_L =2,1,-1,-2,-1,1 and cyclic repetition
- This gives a reasonable $q_S = 2$ and $q_D = 1$, and $< u_R > = < u_L > = 0$.
- Let $\frac{\Delta t}{\tau_w} = 1/100$

Ocular dominance: Initial conditions 1:

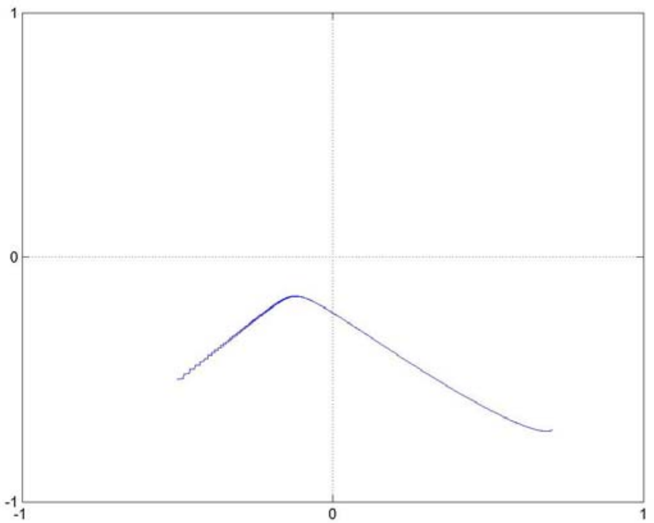
- Let α=1 which should lead to |**w**|=1, i.e. with suppression of first ev., to $w_1 = -w_2 = \pm 0.7$
- w-plane (left) $w_1 - w_2$ / $w_1 + w_2$ (right)

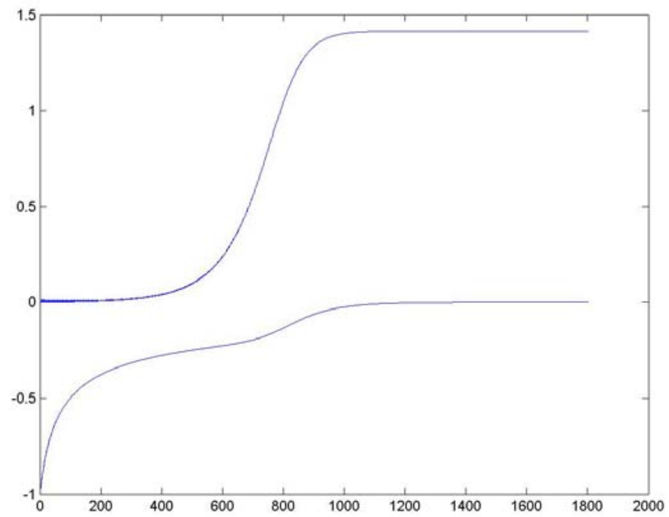




Ocular dominance: Initial conditions 2:

- Initial condition = first ev. : since (u_R, u_L) have fluctu-ations, this leads, after a long time, to 2nd largest ev.
- w-plane (left) w_1-w_2 / w_1 +w_2 (right)





Ocular dominance: Initial conditions n:

- Live demo !!!

Ex 4

- Play with the combined Subtr. Normalization/ Oja rule in provided matlab programme!
- Examine the role of:
  - Initial w-values
  - Initial w-normalization b
  - Final w-normalization a
  - Removing Subtr. Normalization and/or Oja terms
  - Changing the time factor

Resumé: Hebb Rules and PCA

- PCA forces multinomial Gaussian distribution on data, i.e. is sensitive only to 2nd order statistics.
- Basic Hebb Rule selects for **w** the principal eigenvector of the data‘s correlation matrix, **w** grows unlimitedly.
- Subtractive normaliz. suppresses the principal ev.(s)
- Oja‘s rule normalizes **w**.
- Combined rule still works locally and biologically plausible, if prior knowledge exists about desired **w**-behaviour suppression.
- Simple one-cell ocular dominance model can be realized.