

## An extended Hebbian model of unsupervised learning

Recent work in Long Term Potentiation in brain slices shows that Hebb's rule is not completely synapse-specific, possibly due to intersynapse Calcium diffusion.

We extend the classical Oja unsupervised learning model of a single linear neuron to include Hebbian infidelity, by introducing an error matrix which expresses the cross-talk between the Hebbian updating at different connections.

The current linear model is sensitive only to pairwise statistics.

We expect that a more powerful nonlinear model that takes into consideration higher correlations would show an error catastrophe under biologically plausible conditions.

## Linear neuron model:

- Input signals on a linear layer of  $n$  neurons, randomly drawn from a probability distribution  $\mathcal{P}(\mathbf{x} = (x_1 \dots x_n)^t)$
- A unique output neuron:  $h = \mathbf{S} \mathbf{x}_k \mathbf{w}_k$
- Connections with plastic synaptic weights:  $\mathbf{w} = (w_1 \dots w_n)^t$  which update following Hebb's rule of learning:

$$\mathbf{w}_k(s+1) = \mathbf{w}_k(s) + \mathbf{g} h(s) \mathbf{x}_k(s)$$

Normalize, expand in Taylor series w.r.t  $\mathbf{g}$  and ignore the  $O(\mathbf{g}^2)$  terms for  $\mathbf{g}$  small. Assume the process is stationary (slow) and that  $\mathbf{x}(s)$  and  $\mathbf{w}(s)$  are statistically independent:

$$\mathbf{w}(s+1) = \mathbf{w}(s) + \mathbf{g} [\mathbf{C}\mathbf{w} - (\mathbf{w}^t \mathbf{C} \mathbf{w}) \mathbf{w}]$$

$\mathbf{C}$  = covariance matrix of the input distribution:  $\mathbf{C} = \langle \mathbf{x}^t \mathbf{x} \rangle$

## Equivalent dynamical system to be studied:

We study the stability of the synaptic vector  $\mathbf{w}$  under iterations of the function:

$$\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$\mathbf{f}(\mathbf{w}) = \mathbf{w} + \mathbf{g} [\mathbf{C}\mathbf{w} - (\mathbf{w}^t \mathbf{C} \mathbf{w}) \mathbf{w}]$$

The result should depend on the distribution to be learnt (through the symmetric, positive definite matrix  $\mathbf{C}$ ) and on the size of  $\mathbf{g}$ .

$\mathbf{w}$  fixed point for  $\mathbf{f} \Leftrightarrow \mathbf{w}$  is a unit eigenvector of  $\mathbf{C}$

$\mathbf{w}$  is a hyperbolic attractor for  $\mathbf{f} \Leftrightarrow \mathbf{Df}_{\mathbf{w}}$  has only stable eigendirections (i.e. with eigenvalues  $|| \lambda | < 1$ )

$$Df_w = I + g [C - 2w(Cw)^t - (w^t C w)I]$$

Complete  $w$  to an orthonormal eigenbasis  $\mathcal{B}$  of  $C$ .

Then  $\mathcal{B}$  is also an eigenbasis for  $Df_w$ :

$$Df_w v = (1 - g [l_w - l_v])v$$

$$Df_w w = (1 - 2l_w)w$$

$w$  is a hyperbolic fixed point of  $f$  if and only if:

$$|1 - g (l_w - l_v)| < 1$$

$$|1 - 2l_w| < 1$$

**Conclusion:** The fixed vector  $w$  is asymptotically stable if

$$l_w > l_v, \text{ for all } v \neq w \text{ and } g < l_w^{-1}$$

To generalize for a system whose information transmission is imperfect, we encode the errors in a matrix  $\mathbf{T} \in \mathcal{M}_n(\mathbb{R})$ , positive, symmetric,  $\mathbf{T} = \mathbf{I}$  for zero error:

$$\mathbf{f}^T(\mathbf{w}) = \mathbf{w} + \mathbf{g} [\mathbf{T}\mathbf{C}\mathbf{w} - (\mathbf{w}^T\mathbf{C}\mathbf{w})\mathbf{w}]$$

If the matrix  $\mathbf{T}\mathbf{C}$  has a positive maximal eigenvalue  $\lambda$  with multiplicity one, then its corresponding eigenvector  $\mathbf{w}$  normalized by  $\mathbf{w}^T\mathbf{C}\mathbf{w} = 1$  is the only asymptotically stable fixed vector of  $\mathbf{f}^T$ , provided that  $\mathbf{g}$  is small.

In the absence of error, the network learnt the principal eigenvector of the correlation matrix  $\mathbf{C}$ . A reasonable measure of the output error for a given error matrix  $\mathbf{T}$  is:

$$\cos(\mathbf{q}) = \langle \mathbf{w}^C, \mathbf{w}^{TC} \rangle \cdot \|\mathbf{w}^C\|^{-1} \cdot \|\mathbf{w}^{TC}\|^{-1}$$

- Analytic results in the particular case of uncorrelated inputs and neurons equivalently exposed to error
- Simulations for more general settings.

$$C = \begin{pmatrix} \lambda & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$T = \begin{pmatrix} Q & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & Q & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & Q & \varepsilon \\ \varepsilon & \varepsilon & \varepsilon & Q \end{pmatrix}$$

$$\lambda > 1, \quad Q = q^n, \quad \varepsilon = \frac{1 - q^n}{n - 1}$$

