Journal of Theoretical Biology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Hebbian errors in learning: An analysis using the Oja model

Anca Rădulescu [a,b,*], Kingsley Cox [b,c], Paul Adams [b,c]

[a] University of Colorado, UCB 526 Boulder, CO 80309-0526, USA
[b] Stony Brook University, Stony Brook, NY 11794, USA
[c] Kalypso Institute, Stony Brook, USA

ARTICLE INFO

Article history:
Received 26 March 2008
Received in revised form
27 November 2008
Accepted 23 January 2009

Keywords:
Synaptic plasticity
Information transfer
Neural computation
Learning algorithm
Dynamical system

ABSTRACT

Background: Recent work on long term potentiation in brain slices shows that Hebb's rule is not completely synapse-specific, probably due to intersynapse diffusion of calcium or other factors. We previously suggested that such errors in Hebbian learning might be analogous to mutations in evolution.

Methods and findings: We examine this proposal quantitatively, extending the classical Oja unsupervised model of learning by a single linear neuron to include Hebbian inspecificity. We introduce an error matrix **E**, which expresses possible crosstalk between updating at different connections. When there is no inspecificity, this gives the classical result of convergence to the first principal component of the input distribution (PC1). We show the modified algorithm converges to the leading eigenvector of the matrix **EC**, where **C** is the input covariance matrix. In the most biologically plausible case when there are no intrinsically privileged connections, **E** has diagonal elements Q and off-diagonal elements $(1-Q)/(n-1)$, where Q, the quality, is expected to decrease with the number of inputs n and with a synaptic parameter b that reflects synapse density, calcium diffusion, etc. We study the dependence of the learning accuracy on b, n and the amount of input activity or correlation (analytically and computationally). We find that accuracy increases (learning becomes gradually less useful) with increases in b, particularly for intermediate (i.e., biologically realistic) correlation strength, although some useful learning always occurs up to the trivial limit $Q = 1/n$.

Conclusions and significance: We discuss the relation of our results to Hebbian unsupervised learning in the brain. When the mechanism lacks specificity, the network fails to learn the expected, and typically most useful, result, especially when the input correlation is weak. Hebbian crosstalk would reflect the very high density of synapses along dendrites, and inevitably degrades learning.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Various brain structures, such as the neocortex, are believed to use unsupervised synaptic learning to form neural representations that capture and exploit statistical regularities of an animal's world. Most neural models of unsupervised learning use some form of Hebb rule to update synaptic connections. Typically, this rule is implemented by updating a connection according to the product of the input and output firing rates. Other forms of the update rule are sometimes used, but they are still typically local and activity dependent, and often Hebbian in the sense that they depend on both input and output activity. Biological networks may also use spike-timing dependent rules, but these are also Hebbian in the sense that they depend on the relative timing of pre- and postsynaptic spiking. The key element in Hebbian learning is that the update should depend on the extent to which the input appears to "take part in" firing the output (Hebb, 1949; we added "appears" to emphasize that a single neural connection knows nothing about actual causation, and merely responds to a statistical coincidence of pre- and postsynaptic spikes).

One of us has previously proposed, in this journal (Adams, 1998), that such Hebbian errors might be analogous to genetic mutations. Specifically, we are interested in the possibility that if the Hebb rule is not completely local (in the sense there might be some, possibly very weak, dependence of the local update on activity at other connections) unsupervised learning might fail catastrophically, not only preventing new learning, but wiping out previous learning. We have proposed that the basic task of the neocortex is to avoid such hypothetical learning catastrophes (Adams and Cox, 2002a, b, 2006; Cox and Adams, 2000). In this paper we modify a classical model of unsupervised learning, the Oja single neuron principal component analyzer (Oja, 1982), to include Hebbian inaccuracy. By "inspecificity", or "inaccuracy", we mean that part of the local update calculated using a Hebb rule (for example, proportional to the product of input and output

* Corresponding author at: University of Colorado, UCB 526 Boulder, CO 80309-0526, USA. Tel.: +1 303 786 1616; fax: +1 303 492 4066.
E-mail address: radulesc@colorado.edu (A. Rădulescu).

0022-5193/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.01.036

firing rates) is assigned to connections other than the one at which the product was calculated. We also refer to this postulated nonlocality as "leakage", "crosstalk" or simply "error". Some other papers on this topic have appeared (Adams and Cox, 2002a; Botelho and Jamison, 2002, 2004), and we will try to clarify the relationship between these various studies. We conclude that while the modified Oja model, and perhaps others that are only sensitive to second-order statistics, do not show a true error catastrophe at finite network size, their behaviour gives important clues to understanding the difficulties that brains might encounter in learning higher-order statistics.

Recent experimental work has shown that long term potentiation (LTP), a biological manifestation of the Hebb rule, is indeed not completely synapse specific (Engert and Bonhoeffer, 1997; Schuman and Madison, 1994; Bi, 2002). For example, Engert and Bonhoeffer have shown that LTP induced at a local set of connections on a CA1 pyramidal cell "spills over" to induce LTP at a nearby set of inactive connections on the same cell. In earlier work using less refined methods, it had been concluded that LTP was synapse-specific (Andersen et al., 1977; Levy and Steward, 1979). Even in the Engert–Bonhoeffer experiments (Engert and Bonhoeffer, 1997), it is likely that, because the "pairing" method used to induce LTP was rather crude, the inspecificity was far greater than would ever be actually seen in an awake brain. More recent work has shown that at least one type of Hebbian inspecificity, induced by theta burst stimulation of retinotectal connections, reflects dendritic spread of calcium (Tao et al., 2001). Even more recent LTP experiments at single synapses (Tao et al., 2001) have shown that, while LTP is only expressed at the synapse where it is induced (Matsuzaki et al., 2004), the threshold for LTP induced at neighboring synapses is reduced (Harvey and Svoboda, 2007). Thus, some degree of Hebbian inspecificity is probably inevitable, and its effects on learning need to be evaluated.

2. Overview

Here we briefly review the classical Oja model (Oja, 1982; Oja and Karhunen, 1985; Diamantaras and Kung, 1996; Hertz et al., 1991) and define terms as background to the new analysis. The model network consists of a single output neuron receiving n signals $x_1, x_2, \ldots, x_n$ from a set of n input neurons via connections of corresponding strengths $\omega_1, \ldots, \omega_n$ (see Fig. 1). We assume throughout that the inputs have zero-mean (see Discussion).

The resulting output y is defined as the weighted sum of the inputs:

$$y = \sum_{i=1}^{n} x_i \omega_i \tag{2.1}$$

The input column vector $\mathbf{x} = (x_1 \ldots x_n)^T$ is randomly drawn from a probability distribution $\mathscr{P}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^n$ (where T denotes transposition of vectors).

In accordance to Hebb's postulate of learning, a synaptic weight ω_i will strengthen proportionally with the product of x_i and y:

$$\omega_i(t+1) = \omega_i(t) + \gamma y(t) x_i(t) \tag{2.2}$$

Here γ is a time independent learning rate and the argument t represents the dependence on time (or on the input draw). The relation between this formulation and neural processes such as LTP is considered in the Discussion.

Oja (1982) modified this by normalizing the weight vector $\boldsymbol{\omega}$ with respect to the Euclidean metric on $\mathbb{R}^n$:

$$\omega_i(t+1) = \frac{\omega_i(t) + \gamma y(t) x_i(t)}{\|\boldsymbol{\omega}(t) + \gamma y(t)\mathbf{x}(t)\|} \tag{2.3}$$

Expanding in Taylor series with respect to γ, making $\|\boldsymbol{\omega}\| = 1$ and ignoring the $\mathscr{O}(\gamma^2)$ term for γ sufficiently small, the result is

$$\boldsymbol{\omega}(t+1) = \boldsymbol{\omega}(t) + \gamma y(t)[\mathbf{x}(t) - y(t)\boldsymbol{\omega}(t)] \tag{2.4}$$

Henceforth, we omit the variable t whenever there is no ambiguity. The equation can be then rewritten as:

$$\boldsymbol{\omega}(t+1) = \boldsymbol{\omega} + \gamma[\mathbf{x}\mathbf{x}^T\boldsymbol{\omega} - (\boldsymbol{\omega}^T\mathbf{x}\mathbf{x}^T\boldsymbol{\omega})\boldsymbol{\omega}] \tag{2.5}$$

Consider the covariance matrix of the distribution $\mathscr{P}(\mathbf{x})$, defined by $C = \langle \mathbf{x}\mathbf{x}^T \rangle = \langle \mathbf{x}(t)\mathbf{x}^T(t) \rangle$ (where $\langle\,\rangle$ stands for expectation). Clearly $\mathbf{C}$ is symmetric and semipositive definite. With the following additional assumptions:

- the learning process is slow enough for $\boldsymbol{\omega}$ to be treated as stationary,
- $\mathbf{x}(t)$ and $\boldsymbol{\omega}(t)$ are statistically independent.

One can take conditional expectation over $\mathscr{P}(\mathbf{x})$ and rewrite the learning rule as:

$$\langle \boldsymbol{\omega}(t+1) | \boldsymbol{\omega}(t) \rangle = \mathbf{w} + \gamma[\mathbf{C}\mathbf{w} - (\mathbf{w}^T\mathbf{C}\mathbf{w})] \tag{2.6}$$

Oja concluded that, if $\boldsymbol{\omega}(t)$ converges as $t \to \infty$, the limit is expected to be one of the two opposite normalized eigenvectors corresponding to the maximal eigenvalue of $\mathbf{C}$ (i.e., the "principal component" of the matrix $\mathbf{C}$—Oja, 1982; Oja and Karhunen, 1985). If the input elements are Gaussian, the output of the Oja neuron provides the statistically optimal representation of the current input vector.

Three main types of synaptic learning error could occur. First ("Type 0"), one could add uncorrelated noise to the input elements, for example reflecting fluctuations in transmitter release, which would simply add to the variances along the diagonal of $\mathbf{C}$. Second ("Type 1"), the updates themselves may be imprecise (for example because of spine head calcium fluctuations), with these fluctuations occurring independently at different synapses. Third ("Type 2"), the updates could be inaccurate, in the sense of depending on updates occurring at other synapses (e.g., because of intersynapse calcium diffusion). This is the case we study here. To introduce such inspecificity into the learning equation, we assume that, on average, only a fraction Q of the intended update reaches the appropriate connection, the remaining fraction $1 - Q$ being distributed amongst the other connections according to a defined and biologically plausible rule. The actual update at a given connection thus includes contributions from erroneous or inaccurate updates from other connections. The erroneous updating process is formally described by a (possibly time-dependent, but input-independent) error matrix $\mathscr{E} = \mathscr{E}(t)$, whose elements reflect at each time step t the fractional contribution that the activity across weight ω_i makes to the update of ω_j.

If one wanted to ensure that the weight vector retains the same norm as the error-free rule, we would introduce $\mathscr{E}$ into Eq. (2.2), obtaining as a first order approximation for the single weight one pattern rule[1]:

$$\omega_i(t+1) = \omega_i + \gamma y([\mathscr{E}\mathbf{x}]_i - \boldsymbol{\omega}^T\mathscr{E}\mathbf{x}\omega_i) \tag{2.7}$$

However, while the averaged form of this rule has the same fixed points as the rule we actually used (see Eq. (2.8)), we could not prove their stability. Furthermore, the rule would imply biologically that the normalizing component "knew" the pattern-to-pattern form of $\mathscr{E}$, which is highly implausible. More generally we could assume that the Hebbian and normalizing steps would have different error matrices, reflecting their different physical

[1] Here the notation $[\mathscr{E}\mathbf{x}]_i$ stands for the i-th component of the vector $[\mathscr{E}\mathbf{x}]$.

Please cite this article as: Rădulescu, A., et al., Hebbian errors in learning: An analysis using the Oja model. J. Theor. Biol. (2009), doi:10.1016/j.jtbi.2009.01.036

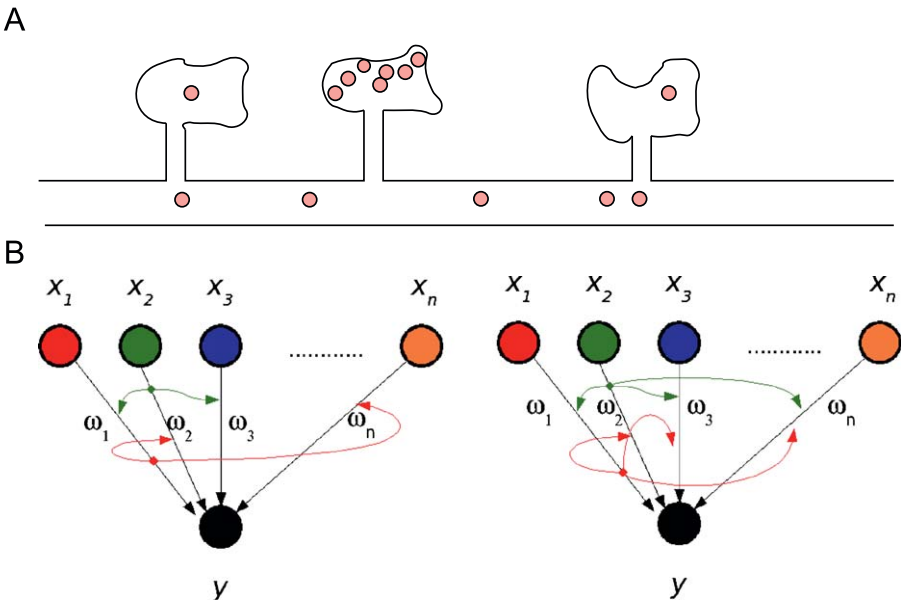

Fig. 1. (A) Three spines on a short dendritic segment are shown. If Hebbian adjustment occurs at the middle synapse, a factor (red dots), such as calcium, diffuses to nearby synapses and affects Hebbian adjustment there. (B) Input neurons (activities x_i) converge on an output neuron (output y) via weights ω_i. Coincident activity at the synapses comprising a weight (e.g. ω_1 or ω_2) leads to modification of that weight and of other weights. The left diagram shows the case where only the immediate neighboring connections (each made up of one synapse) are affected. The right diagram shows the case where all connections are equal neighbors (either because each has many synapses dispersed randomly over the dendrite, or because synapses move around; see Figure S1). The curved red arrows from ω_1 to ω_n shows that periodic boundary conditions are assumed (i.e., ω_1 affects ω_2 and ω_n equally). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

implementation. We assumed for simplicity that the normalizing step is error free, so the rule becomes:

$$\omega_i(t+1) = \omega_i + \gamma y([\mathscr{E}\mathbf{x}]_i - y\omega_i) \tag{2.8}$$

Taking conditional expectation of both sides and rewriting the equation in matrix form leads to (as a "meanfield" approximation, assuming that the errors are independent from the inputs and the weights)

$$\langle \boldsymbol{\omega}(t+1)|\boldsymbol{\omega}(t)\rangle = \mathbf{w} + \gamma[\mathbf{ECw} - (\mathbf{w}^T\mathbf{Cw})\mathbf{w}] \tag{2.9}$$

where we defined $\mathbf{w} = \langle \boldsymbol{\omega} \rangle$ and $\mathbf{E} = \langle \mathscr{E} \rangle$. $\mathbf{E}$ is then a symmetric circulant matrix; in the zero error case of $Q = 1$, $\mathbf{E}$ would become the identity matrix.

3. Methods

As a prelude to analyzing the dynamics of inspecific learning, we revisit the Oja original model with zero error, and the methods used to establish its asymptotic behavior (Botelho and Jamison, 2004; Oja, 1982; Oja and Karhunen, 1985; Hertz et al., 1991). In other words: for a size $n \in \mathbb{N}$, $n \geqslant 2$, we want to know whether or not a vector $\mathbf{w} \in \mathbb{R}^n$ stabilizes under iterations of the function:

$$f\colon \mathbb{R}^n \to \mathbb{R}^n, \quad f(\mathbf{w}) = \mathbf{w} + \gamma[\mathbf{Cw} - (\mathbf{w}^T\mathbf{Cw})\mathbf{w}] \tag{3.1}$$

A nonzero vector $\mathbf{w} \in \mathbb{R}^n$ is a fixed point for f if

$$f(\mathbf{w}) = \mathbf{w} + \gamma[\mathbf{Cw} - (\mathbf{w}^T\mathbf{Cw})\mathbf{w}] = \mathbf{w} \;\Leftrightarrow\; \mathbf{Cw} = (\mathbf{w}^T\mathbf{Cw})\mathbf{w} \tag{3.2}$$

An equivalent set of conditions is:

$$\begin{cases} \mathbf{Cw} = \lambda_{\mathbf{w}}\mathbf{w} \\ \lambda_{\mathbf{w}} = \mathbf{w}^T\mathbf{Cw} \end{cases} \Leftrightarrow \begin{cases} \mathbf{Cw} = \lambda_{\mathbf{w}}\mathbf{w} \\ \lambda_{\mathbf{w}} = \lambda_{\mathbf{w}}\mathbf{w}^T\mathbf{w} \end{cases} \tag{3.3}$$

These conditions translate as: "$\mathbf{w}$ is an eigenvector of $\mathbf{C}$". In case $\mathbf{C}$ is invertible (i.e. all its eigenvalues are nonzero), $\mathbf{w}$ is a unit eigenvector of $\mathbf{C}$ with the Euclidean norm.

Consider an orthonormal basis $\mathscr{B}$ of eigenvectors of $\mathbf{C}$ (with respect to the Euclidean norm on $\mathbb{R}^n$). An eigenvector $\mathbf{w} \in \mathscr{B}$ with eigenvalue $\lambda_{\mathbf{w}}$ is a hyperbolic attractor for f if all eigenvalues of the $n \times n$ Jacobian matrix $(Df_{\mathbf{w}})_{ij} = ((\partial f_i/\partial w_j)(\mathbf{w}))$ are less than one in absolute value.

We calculate the Jacobian matrix $Df_{\mathbf{w}}$, for a fixed vector $\mathbf{w} \in \mathscr{B}$:

Lemma 3.1. $Df_{\mathbf{w}} = \mathbf{I} + \gamma[\mathbf{C} - 2\mathbf{w}(\mathbf{Cw})^T - (\mathbf{w}^T\mathbf{Cw})\mathbf{I}]$

Proof. Call $g(\mathbf{w}) = (\mathbf{w}^T\mathbf{Cw})\mathbf{w}$, so $f(\mathbf{w}) = \mathbf{w} + \gamma(\mathbf{Cw} - g(\mathbf{w}))$

$g_i(\mathbf{w}) = (\mathbf{w}^T\mathbf{Cw})w_i$

If $i \neq j$:

$$\frac{\partial g_i}{\partial w_j}(\mathbf{w}) = \frac{\partial}{\partial w_j}\left(\sum_{k,l} C_{kl}w_k w_l\right)w_i = 2\left(\sum_k C_{kj}w_k\right)w_i = 2[\mathbf{Cw}]_j w_i$$

If $i = j$:

$$\frac{\partial g_i}{\partial w_i}(\mathbf{w}) = \frac{\partial}{\partial w_i}\left(\sum_{k,l} C_{kl}w_k w_l\right)w_i + \sum_{k,l} C_{kl}w_k w_l = 2\left(\sum_k C_{ki}w_k\right)w_i + \mathbf{w}^T\mathbf{Cw} = 2[\mathbf{Cw}]_i w_i + \mathbf{w}^T\mathbf{Cw}$$

So

$$Dg_{\mathbf{w}} = 2\mathbf{w}(\mathbf{Cw})^T + (\mathbf{w}^T\mathbf{Cw})\mathbf{I} \qquad \square \tag{3.4}$$

Pick any $\mathbf{v} \in \mathscr{B}, \mathbf{v} \neq \mathbf{w}$. Call $\lambda_{\mathbf{w}}$ and $\lambda_{\mathbf{v}}$ their corresponding eigenvalues.

$$\begin{aligned} Df_{\mathbf{w}}(\mathbf{v}) &= v + \gamma[\mathbf{Cv} - 2\mathbf{w}(\mathbf{Cw})^T\mathbf{v} - (\mathbf{w}^T\mathbf{Cw})\mathbf{v}] \\ &= \mathbf{v} + \gamma[\mathbf{Cv} - 2\mathbf{w}\mathbf{w}^T\mathbf{Cv} - (\mathbf{w}^T\mathbf{Cw})\mathbf{w}] \\ &= v + \gamma[\lambda_{\mathbf{v}}\mathbf{v} - 2\mathbf{w}\mathbf{w}^T\lambda_{\mathbf{v}}\mathbf{v} - \lambda_{\mathbf{w}}\mathbf{w}] = (1 - \gamma[\lambda_{\mathbf{w}} - \lambda_{\mathbf{v}}])\mathbf{v} \end{aligned} \tag{3.5}$$

$$\begin{aligned} Df_{\mathbf{w}}(\mathbf{w}) &= \mathbf{w} + \gamma[\mathbf{Cw} - 2\mathbf{w}(\mathbf{Cw})^T\mathbf{w} - (\mathbf{w}^T\mathbf{Cw})] \\ &= \mathbf{w} + \gamma[\lambda_{\mathbf{w}}\mathbf{w} - 2\mathbf{w}\mathbf{w}^T\lambda_{\mathbf{w}}\mathbf{w} - \lambda_{\mathbf{w}}\mathbf{w}] \\ &= \mathbf{w} + \gamma[-2\lambda_{\mathbf{w}}\|\mathbf{w}\|\mathbf{w}] = [1 - 2\gamma\lambda_{\mathbf{w}}]\mathbf{w} \end{aligned} \tag{3.6}$$

So $\mathscr{B}$ is also a basis of eigenvectors for $Df_{\mathbf{w}}$. The corresponding eigenvalues are $1 - 2\gamma\lambda_{\mathbf{w}}$ (for the eigenvector $\mathbf{w}$) and $1 - \gamma(\lambda_{\mathbf{w}} - \lambda_{\mathbf{v}})$ (for any other eigenvector $\mathbf{v} \in \mathscr{B}$, $\mathbf{v} \neq \mathbf{w}$). Therefore, a set of

Please cite this article as: Rădulescu, A., et al., Hebbian errors in learning: An analysis using the Oja model. J. Theor. Biol. (2009), doi:10.1016/j.jtbi.2009.01.036