

The Neocortex as a Hebbian Proofreader

Kingsley J. A. Cox, Anca Radulescu and Paul R. Adams

Department of Neurobiology and Behavior, College of Arts and Sciences, SUNY Stony Brook, State University of New York, Stony Brook, NY 11794-5230, USA

Abstract

We propose that the neocortex is a machine for learning high order correlations which avoids error catastrophes induced by relatively unstructured environments using an online Hebbian proofreading canonical microcircuit. We argue that there is inevitable crosstalk between individual weight updates, for example caused by spine-spine calcium spillover. This reflects the incompatible requirements for voltage spread and calcium localization in networks that use voltage to compute and calcium to learn; the ratio of their space constants sets a plasticity error level. We show that for linear neurons which only learn pairwise statistics such errors do not completely prevent selforganisation, although learning gets worse as the network enlarges. But if neurons are nonlinear (essential for learning high-order statistics), a learning error collapse occurs at a network size given by the reciprocal error rate, around 1 thousand. Since the cortex cannot know in advance what environments could trigger a collapse, and the column size is around 1000 neurons, it must use an independent online assessment of current input statistics to gate learning by feedforward networks, a form of Hebbian proofreading. We show that a microcircuit which guarantees catastrophe avoidance and which would therefore allow neocortex to act as a universal learning machine closely resembles the physiology and anatomy of layer 6 cells and connections. Proofreading guarantees catastrophe avoidance at the expense of slowed learning.

1. Introduction

The neocortex has distinctive layers with rather stereotyped connections (such as layer 6 feedback to thalamus), suggesting that it is specialised to perform some basic, rather general, function. Neocortex is particularly good at learning subtle patterns, and it's possible that it is designed to solve some fundamental, universal, problem connected with synaptic learning.

It is likely that the neocortex learns statistically efficient models of world regularities using local, activity-dependent synapse-adjustment rules similar to a Hebb rule. We propose, heretically, that the main problem confronting the neocortex is that Hebbian adjustments are not completely synapse-specific, and that most neocortical circuitry is devoted to preventing catastrophic accumulation of learning errors rather than to traditional learning/information processing. Lack of learning specificity inevitably results from the finite ratio of the voltage and calcium space constants ($\lambda_w/\lambda_c \sim 10^3$), which reflects hard biophysical constraints.

2. Cortical microcircuit enhancing specificity

Our proposed canonical microcircuit mitigating Hebbian inspecificity is shown in Figure 1. If learning were completely synapse-specific a feedforward connection from cell J_j to cell I_i would be exactly Hebbian – its strength $w_{i,j}$ changing as a function of the product of the input and output activities x_j and y_i (or as a result of the nearly coincident firing of these cells, in a spiking network). However, because learning cannot be completely synapse specific, we postulate that some of the weight change that actually occurs is contributed by the conjoint activity of other presynaptic and postsynaptic cells (not shown in Fig 1), and that the conjoint activity at the J-I connection also triggers strength changes at connections other than $w_{i,j}$. We represent the crosstalk between synapse updates as an error matrix \mathbf{T} , which becomes the identity matrix in the zero error case. Our model is thus a generalisation of a standard feedforward perceptron network.

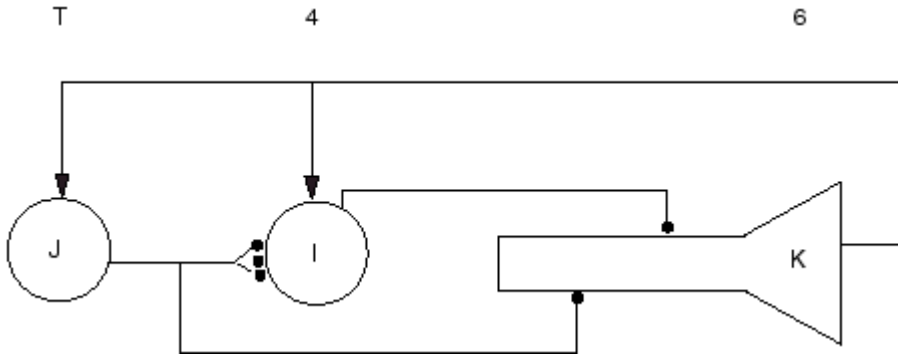


Figure 1 The Hebbian proofreading circuit

The update errors, arising because the Hebbian adjustment of w_{ij} is not completely synapse specific, can only be decreased in 2 ways; better isolating the Hebbian synapses comprising w_{ij} (for example, by reducing diffusional intersynapse coupling), or by using a second *independent* measurement of the product $x_j y_i$ (or, equivalently, of the coincident firing of J_j and I_i) to “gate” the plasticity of the synapses comprising w_{ij} . The first method requires intersynapse distances and/or spine neck resistances to be increased, which prevents efficient collection of voltages, and may necessitate recourse to the second method, shown in Fig 1, which we call the thalamocortical algorithm. A third type of neuron K acts as a coincidence detector of spikes in J and I (or, in rate networks, computes the product $x_j y_i$). The firing of neuron K_{ij} enables the plasticity of the connection w_{ij} (arrow connections), so the accuracy of the 2 independent coincidence measurements (by the Hebbian synapses comprising w_{ij} and by the K_{ij} cell) are multiplied together to determine the update of w_{ij} . Because it is biologically impossible to wire the plasticity gating signal (the firing of K) directly to the synapses comprising w_{ij} , we suggest that the gating signal be led both to the presynaptic neuron J_j and the postsynaptic neuron I_i , and that only the conjunction of plasticity gating signals in J and I enables the plasticity of w_{ij} . We call this arrangement “outer product plasticity gating”. In general, while J_j and I_i may both also receive plasticity gating signals from other K cells (computing correlations across other J - I connections), if the K activity is sparse the outer product rule will uniquely identify the correct connections. The microcircuit shown in Fig 1 acts as a “Hebbian proofreader”, in the same way that the proofreading exonuclease activity of DNA polymerase increases the accuracy of polynucleotide replication from $\sim 10^{-4}$ to $\sim 10^{-8}$.

3. Hebbian error in neural networks

Three questions arise from this proposal. First, is there any evidence for Hebbian inspecificity or for the intersynapse calcium diffusion that underlies it? Second, can such inspecificity completely prevent network learning? Third, is there evidence that the physiology and anatomy of the neocortex correspond to scheme 1?

There is experimental evidence for minor Hebbian inspecificity and calcium diffusion, despite some claims to the contrary [3,4,5,6]. Thalamocortical synapses are widely spaced and spine necks widen *pari passu* with synapse strengthening 5, completely consistent with the view that synapse isolation is a major biological goal. Necks cannot be narrowed or dendrites lengthened without performance impairment. We argue that the error rate ϵ_2 is on the order $a^h \lambda_c / l$; a is a spine head/shaft Ca attenuation factor, h is twice the effective Hill coefficient for calcium-triggering of strengthening and l is the mean dendritic length. In addition to these “type 2 errors” spine head calcium spontaneous fluctuations will cause “type 1 errors” of size $\epsilon_1 \sim 1/\alpha!$ where α is the factor by which Ca must increase to trigger LTP.

We studied the effect of Hebbian inspecificity in 2 classical feedforward mini-networks: a linear single neuron Principle Component analyser and an Independent Component analyser [1,2], each with n inputs. The erroneous PC analyzer converges to the leading eigenvector of TC (C is the input covariance matrix),

which provides a less efficient representation than an error-free network, which converges to the first principle component of the input distribution. However, the representation, which is sensitive only to pairwise statistics, remains somewhat useful for values of $n < 1/\epsilon_2$, and there is no true error catastrophe (complete loss of all stored information) at finite n . We compare this model with data on the specificity of connections onto lateral geniculate relay cells and cerebellar Purkinje cells.

It is likely that pairwise statistics have already been removed from thalamic input to neocortex, which must therefore learn higher order statistics, using suitable response nonlinearities. We propose that such networks, which must learn potentially astronomically large numbers of higher order correlations, are subject to an error catastrophe if the network size exceeds $\sim 1/\epsilon$. This catastrophe not only prevents future learning but wipes out all previous learning. We studied the ICA model, using an inspecific learning rule (see figure 3). In this model the nonlinearity is chosen to match the cdf of the input statistics. We separate out the nonlinearity and the erroneous learning dynamics by representing the single neuron IC model as a virtual 3 layer network which has p^n “hidden” neurons each of which gets input from the n input neurons via a fixed (implastic) weight vector digitized to p levels. By making p sufficiently large these hidden weight vectors can be made to match the possible weight vectors of the original single neuron IC model. They can be regarded as “candidate ICs”, and a correct IC can be selected in the second half of the 3 layer virtual network using a Hebb rule and a linear output neuron (Fig 2), which functions as a high dimensional PCA network.

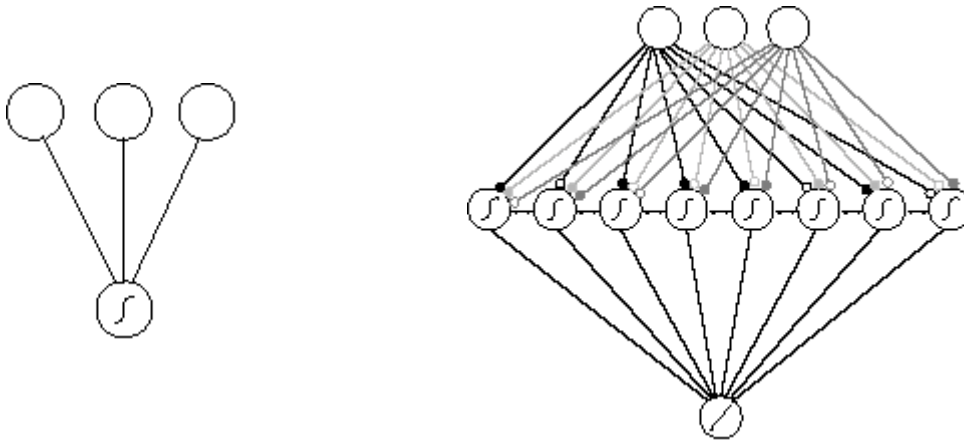


Figure 2 Real (left) and virtual (right) ICA networks

If the hidden network is “winner-take-all” (recurrent inhibition silences all the hidden neurons except the one receiving the largest input), then in the absence of update error, the output neuron will become wired exclusively to the hidden neuron that has the highest variance (averaged over the input ensemble). When a candidate weight vector is a true IC its corresponding hidden neuron will have a maximally uniform output distribution 1, and the distribution of its activity will be single-peaked, so it will also have maximum variance, and will therefore be “picked out” by the Hebb rule acting on the linear output neuron.

In the presence of error the network will converge to the leading eigenvector of \mathbf{MF} , where \mathbf{M} is a symmetric p^n dimensional “mutation” matrix and \mathbf{F} is a diagonal “fitness” matrix whose elements are the variances of the p^n hidden neurons. For “Type 1” errors the elements of \mathbf{M} are $(1-\epsilon_1)^{n-d} \epsilon_1^d$ where $d \leq n$ (d is a Hamming distance between error classes). For plausible instances of \mathbf{F} (which depends on input statistics and the choice of nonlinearity) this leads to an error catastrophe at $n \sim 1/\epsilon_1$. However, because ϵ_1 can be made very small (for example, by enlarging the spine head), assembly of networks as large as the human neocortex is possible. This is fortunate, because the proofreading mechanism we outline above cannot prevent type 1 errors.

Type 2 errors are even more interesting. They arise from coupled fluctuations in spine calcium induced by coincident activity. The result is an error matrix in which odd transitions are “forbidden”. The general matrix element (for binary connections, $p = 2$) for the even transitions appears to be $\epsilon_2^{n/2-d} (\epsilon_2)^d$. This

shows an error catastrophe at a larger network size $n \sim 2/\epsilon_2$ but because $\epsilon_2 \gg \epsilon_1$ we conclude that without Hebbian proofreading large cortical networks cannot selforganise.

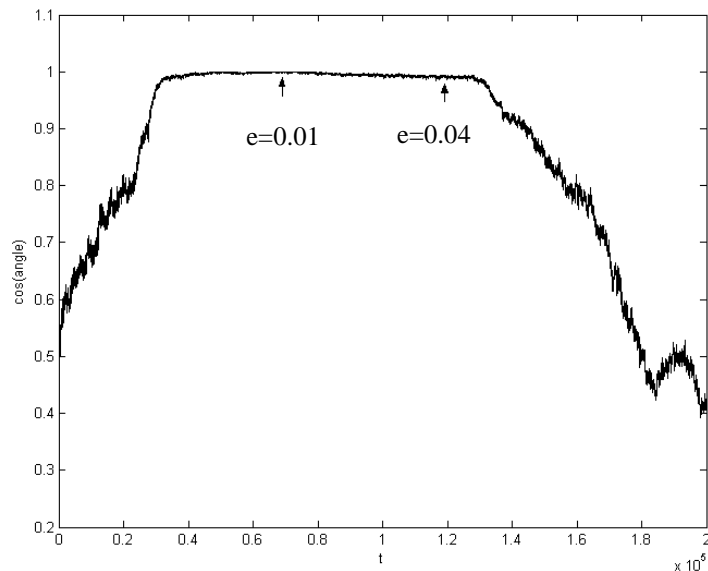


Figure 3. MATLAB ICA computer simulation. 5 Laplacian distributed source signals were mixed by a mixing matrix M generating 5 ‘mix’ input neurons. 5 nonlinear output neurons (sigmoid function) recover the original sources using an anti-Hebbian learning rule (Bell & Sejnowski [1]). The plot shows the angle of one of the rows of the weight matrix against one of the rows of the inverse of M (I.e. an unmixing vector for one of the sources). The x-axis is time in epochs.

Figure 3 shows a simulation of an ICA network in which there seems to be a critical error threshold beyond which the system shows apparently chaotic behaviour. Initially there was no update error and the network converges almost exactly to an IC. When error e of 0.01 is introduced at 70,000 epochs the performance is slightly degraded. Error of 0.04 at 120,000 epochs results in a complete breakdown in performance. At this error level the network was quite unable to extract any of the independent components.

4. Discussion

Both brains and species “learn” the structure of the environment, the former (“Nurture”) by activity-dependent synapse adjustment and the latter (“Nature”) by selective replication of polynucleotides. The key feature that allows successful Darwinian evolution is the accurate amplification of particular specific base sequences, out of the astronomical numbers of possible sequences, ultimately as a result of the elementary Crick-Watson base-pairing mechanism. Similarly, neural learning must require the selective amplification of particular combinations of synaptic weights, ultimately as a result of the Hebbian adjustment of particular connections. We propose that the crucial factor that limits the usefulness of these elementary mechanisms in building complex information-rich structures (nucleic acids or weight vectors) is the accuracy of the elementary process itself (polynucleotide replication or adjustment of synaptic weights), and that vast improvements can be achieved by proofreading.

At the most basic level, learning involves the adjustment of the weights supplying individual neurons by local, activity-dependent mechanisms, and we propose that the neocortex is essentially a device which allows individual neurons to find a particular, uniquely appropriate, set of synaptic input weights from the enormous potential weight space. In this view learning, at the single neuron level, is tantamount to locating a particular weight configuration, or ‘sequence’, just as Darwinian evolution locates particular base configurations. If the input-output relation of the neuron is linear, then the simple Hebb rule allows

only orthogonal weight vectors to grow, and the fastest growing vector corresponds to the leading eigenvector of the correlation matrix: this is NOT a general sequence-finding device, since only a few of the possible sequences can grow. However, if the neuron input-output relation is nonlinear, it becomes possible for all possible sequences to grow, not just orthogonal ones. Support for this view comes from a statistical-mechanical analysis of Hebbian learning in nonlinear neurons by Prugel-Bennett and Shapiro [7]. Looking at the virtual network sketched in Fig 2, the rate of growth of each of the possible sequences is represented by the variance of one of the “hidden” virtual neurons. From this viewpoint, the development of the neocortical microcircuitry would be analogous to the switch from the RNA world to the DNA/protein world, allowing vastly more complex structures (chemical or neural) to be built using elementary steps.

The proofreading circuitry in Fig 1 corresponds remarkably well to the actual connections of layer 6 neurons. For example, in visual cortex layer 6 neurons receive weak input from both relay cells and layer 4 simple cells, and they feedback to the same structures using special neuromodulatory “drumstick” synapses. We propose that these layer 6 cells are “doubly simple” because they inherit their simplicity from the conjunction of their simple inputs. We suggest that the presynaptic enabling signal is a shift of relay cell firing from tonic to burst mode; the postsynaptic enabling signal would be mGluR activation, which would also switch off the recurrent interaction of spiny stellate cells. The shift simultaneously binarises the input. Similarly, layer 6 complex cells would be doubly complex, and they would feed back to their conjoint inputs, as is observed. However, we suggest that instead of merely lowering the effective overall error rate (at the cost of slowed learning), layer 6 cells evaluate (via recurrent connections) the relative strengths of correlations across both current and incipient connections. This strategy can keep the learning rate as high as possible, while guaranteeing that unfavorable statistics (low regularity) will not wipe out previously learned information.

5. Conclusion

We are interested in the possibility that the neocortex has a unique and rather stereotyped structure because it is specialized to solve some rather basic problem associated with learning. One obvious aspect of this structure, which has not hitherto been adequately explained, is that almost all input to cortex arrives via thalamic relay cells, which in turn receive rather specialized feedback influence from layer 6 of cortex. The prototype for cortical learning is the Oja rule, which allows a neuron to “discover” aspects of the overall statistical structure of its inputs, using an elementary, synapse-specific, activity-dependent adjustment process. The neuron finds the structure because it corresponds to the fastest-growing weight vector, and it is important that the “learning” is progressive, involving interaction between the incoming patterns and the evolving weight vector. This means that errors in the accuracy of the adjustment process can compound over time. However, in the case of the linear Oja neuron, since only orthogonal weight vectors can grow, all that happens is that as Hebbian error increases, the advantage of the first principal component over the other principle components is gradually eroded, without any catastrophic failure. We think that when the neuron is nonlinear, each of the possible weight vectors (depending on the resolution of the weight adjustment process itself) can grow, and because the number of possible weight vectors grows exponentially with the number of inputs, an error catastrophe can ensue if the Hebbian process is not sufficiently accurate. An obvious (and perhaps unique) way to make the Hebbian process more accurate is to combine 2 independent measurements of neural “coincidence”, i.e. “proofreading”. This leads to the novel, and as yet untested, idea that particular cortical neurons act as coincidence-detectors that gate neural plasticity. In support of this idea, we examined a simple example of a nonlinear neuron learning the “independent components” of input ensembles. The key idea here is that in order to find ICs, potentially all orders of statistics (and not just second-order, as in the Oja model) must be exploited, and the number of higher-order statistics grows exponentially with input dimension. In order to find a neural filter for the “right” set of higher-order statistics (implicit in the idea of Independent Components), all possible weight vectors must be available to the neuron, and locating the correct vector (a particular “sequence”) requires an extremely accurate Hebb rule. (Of course, the accuracy required depends on how “obvious” the statistical structure of the input ensemble – the “environment” – is). We observe that an accurate rule leads to successful learning of the independent components, but an inaccurate rule leads to

chaotic behaviour. We are not proposing that the neocortex performs independent component analysis, merely that its neurons are confronted with a similar task: locating unique weight configurations out of zillions of possibilities, and that, just as Manfred Eigen has shown for molecular evolution, the final achievable complexity is set by the accuracy of the elementary steps. We think this offers a completely new picture of neocortical logic, even though there are many gaps in our analysis.

References

- [1] A. Bell, T. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, *Neural Comput.* 7, (1995)1129-311.
- [2] E. Oja, A simplified neuron model as a principal component analyzer, *J. Math. Biol.*15 (1982) 267-273.
- [3] F. Engert, T. Bonhoeffer, Synapse specificity of long-term potentiation breaks down at short distances, *Nature* 388 (1997) 279-284.
- [4] B.S. Sabatini, T. Oertner, K. Svoboda, The life-cycle of Ca²⁺ ions in dendritic spines, *Neuron* 33 (2002) 439-452.
- [5] M. Matsuzaki, N. Honkura, G.C.R. Ellis-Davies, H. Kasai, "Structural basis of functional synaptic plasticity in single dendritic spines", *Nature* 429 (2004) 761-766.
- [6] Noguchi et. al., Spine-Neck Geometry Determines NMDA Receptor Ca²⁺ Signaling in Dendrites, *Neuron* 46 (2005) 609-622.
- [7] A Prugel-Bennett, J L. Shapiro, Statistical mechanics of unsupervised Hebbian learning, *J. Phys. A: Math. Gen.* 26 (1993) 2343-2369.