An extended Hebbian model of unsupervised learning

March 17, 2006

PREPRINT - PLEASE, DO NOT COPY OR USE WITHOUT THE AUTHOR'S CONSENT

Abstract

The goal of learning is to produce such a configuration of the synaptic weights in a network that roughly encodes the properties of the object to be learned. Unsupervised learning relies on two factors: the input data and the network internal processing, with no external report on the state of synapses. Although it initially appeared that some systems have no internal objective monitor, a more careful analysis showed that even they acquire some self organization through an objective internal function. consider, for instance, a network that adjust the weights according to Oja's rule, which is a normalized version of Hebb's postulate of learning. The system will try to extract from the set of input data whatever it seems to find there on a systematic base.

In the present paper, we discuss how this classical result changes in a more realistic context that includes noise. The modified linear network tends to extract as an answer the principal component of an input correlation matrix that includes possible errors due to information leaks. The probability of such leaks is expressed in average by a parameter intrinsic to the network, called the *quality* of the system. We analyze the dependence between the accuracy of the answer and the quality value, for different degrees of input correlations and for various network sizes. This behavior, studied analytically as well as through Matlab simulations, is surprisingly similar to the Eigen error catastrophe model (see [E]).

1 Introduction to the model and previous work

We consider the simple neuron model, linear in the sense that the model output is a linear combination of its inputs. The neuron receives a set of n input signals

$x_1, x_2, ..., x_n$ from a collection of neurons, through corresponding connections with synaptic weights $w_1, ..., w_n$ respectively. The resulting output is defined as the sum of inputs weighted by the strengths:

$$\eta = \sum_{i=1}^{n} x_i w_i$$

We consider the input column vector $x = (x_1 ... x_n)^t$ to be a randomly drawn from a probability distribution $\mathcal{P}(x),\ x \in \mathbb{R}^n$.

In accordance to Hebb's postulate of learning, a synaptic weight w_i will strengthen when x_i and η coincide in sign. Specifically:

$$w_i(s+1) = w_i(s) + \gamma\eta(s)x_i(s)$$

where γ is a time independent learning parameter that describes the learning capability of the network and the argument s represents the dependence on time, or the input draw.

This learning rule in its basic form could clearly be leading to unlimited growth of the synaptic weight w_i, which is unacceptable physically. So a normalization was incorporated, the effect being to introduce some "competition" among the synapses of a neuron over limited resources, condition which will be essential for stabilization and hence learning.

In 1982, Oja simply normalized the weight vector w with respect to the Euclidean metric on $\mathbb{R}^n$:

$$w_i(s+1) = \frac{w_i(s) + \gamma\eta(s)x_i(s)}{\parallel w(s) + \gamma\eta(s)x(s) \parallel}$$

then expanded in Taylor series for γ. Justifiably ignoring the $\mathrm{O}(\gamma^2)$ term, for γ sufficiently small, the result was:

$$w(s+1) = w(s) + \gamma\eta(s)[x(s) - \eta(s)w(s)]$$

The equation can be rewritten as:

$$w(s+1) = w(s) + \gamma\left[x(s)x^t(s)w(s) - \left(w^t(s)x(s)x^t(s)w(s)\right)w(s)\right]$$

where x^t denotes transposition of vectors.

A convergence analysis of a such stochastic, time varying difference equation is in general nontrivial. Define the correlation matrix C of the distribution $\mathcal{P}(x)$ to be the expected value of the matrix $x^s(x^s)^t$. Clearly C is symmetric and semipositive definite. The work has been done under the following additional assumptions:

- the learning process is slow enough for w to be treated as stationary;
- C has distinct eigenvalues;
- $x(s)$ and $w(s)$ are statistically independent.

Under these conditions, we can take conditional expectation over $\mathcal{P}(x)$ and rewrite the learning rule as:

$$E(w(s+1)/w(s)) = w(s) + \gamma\left[Cw(s) - \left(w^t(s)Cw(s)\right)w(s)\right]$$

The conclusion was that, in case $w(s)$ converges as $s \to \infty$, the limit is one of the two normalized eigenvectors corresponding to the maximal eigenvalue of C. (We call the direction of this vector the principal component of the matrix C.)

The purpose of this paper is to prove that, in some more realistic conditions, the linear network governed by the self-organizing rule we have described still tends to extract the principal component of a (this time) ”modified” correlation matrix of the input patterns.

What we have in mind is a modification based on the possible imperfections of the mechanical-chemical hardware that supports learning. Namely: the request to modify the i-th component w_i of the weight vector could ”leak” from the neighboring synaptic contacts. Therefore, each w_i will receive direct information from x_i in some proportion Q and information from its neighbors in proportion $1 - Q$. In this context, we will need to carefully define what we mean by “neighbors”.

We will again try to analyze the convergence of the weight vector w, where the new stochastic learning equation can be obtained in the same way as Oja's rule, under the same assumption of stationary process.

$$w_i(s+1) = w_i(s) + \gamma\eta(s)[Tx(s)]_i \quad \Rightarrow$$

$$E(w(s+1)/w(s)) = w(s) + \gamma\left[TCw(s) - (w^t(s)Cw(s))\, w(s)\right]$$

$T \in \mathcal{M}_n(\mathbb{R})$ is an error matrix (or noise matrix) that could be assumed to be symmetric and with positive entries. T will depend on the value of Q, hence will reflect directly the quality of the information transfer in the network. We also make the obvious requirement that T is the identity in case of no error (i.e. $Q = 1$).

In section 4 we will present a few particular examples of such network geometries which we considered plausible.

2 A dynamical analysis

Let us first analyze the trivial case when the error matrix T is the identity matrix. In other words, we look at the classical deterministic Oja model, with zero error.

For a fixed size $n \in \mathbb{N},\ n \geq 2$, we want to know if our linear network with learning constant $\gamma > 0$ is able to learn a probability distribution $\mathcal{P}$ with correlation matrix C. That is, we want to see if relevant information on the input distribution $\mathcal{P}$ is reflected into the values of the synaptic-strength vector w, provided the process of reading from $\mathcal{P}$ and adjusting w is allowed to run long enough.

Namely, our objective is to research if a vector $w \in \mathbb{R}^n$ could stabilize under iterations of the function:

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

$$f(w) = w + \gamma[Cw - (w^tCw)w]$$

We expect the discussion to involve the properties of the distribution to be learned (through properties of the symmetric, positive definite matrix C), as well as the magnitude of γ.

In even more precise terms, we are looking to find out if f has any hyperbolic attracting fixed points (see [**?**]).

The condition for w to be fixed by f is:

$$f(w) = w + \gamma[Cw - (w^tCw)w] = w \ \Leftrightarrow \ Cw = (w^tCw)w$$

An equivalent set of conditions is:

$$\begin{cases} Cw = \lambda_w w \\ \lambda_w = w^tCw \end{cases} \Leftrightarrow \begin{cases} Cw = \lambda_w w \\ \lambda_w = \lambda_w w^t w \end{cases}$$

In case C is invertible (i.e. all its eigenvalues are nonzero), these conditions translate as: "w is an eigenvector of C with unit Euclidean norm".

A such vector w is a hyperbolic attractor for f if all eigenvalues of the differential matrix Df_w are less than one in absolute value. We calculate Df_w, for a fixed vector w:

Lemma 2.1. $Df_w = I + \gamma\,[C - 2w(Cw)^t - (w^tCw)I]$

Proof. Call $g(w) = (w^tCw)w$, so $f(w) = w + \gamma(Cw - g(w))$

$$g_i(w) = (w^tCw)w_i$$

If $i \neq j$:

$$\frac{\partial g_i}{\partial w_j}(w) = \frac{\partial}{\partial w_j}(\sum_{k,l} C_{kl}w_kw_l)w_i = 2(\sum_k C_{kj}w_k)w_i \ = \ 2(Cw)_jw_i$$

If $i = j$:

$$\frac{\partial g_i}{\partial w_i}(w) = \frac{\partial}{\partial w_i}(\sum_{k,l} C_{kl}w_kw_l)w_i + \sum_{k,l} C_{kl}w_kw_l = 2(\sum_k C_{ki}w_k)w_i +$$

$$+ w^tCw = 2(Cw)_iw_i + w^tCw$$

So:

$$Dg_w = 2w(Cw)^t + (w^tCw)I$$

□

Take now an orthonormal basis $\mathcal{B}$ of eigenvectors of C (with respect to the Euclidean norm $\|\cdot\|$ on $\mathbb{R}^n$).Fix a vector $w \in \mathcal{B}$. Pick any $v \in \mathcal{B}, v \neq w$. Call λ_w and λ_v their corresponding eigenvalues.

$$\begin{aligned} Df_w(v) &= v + \gamma[Cv - 2w(Cw)^tv - (w^tCw)v] = \\ &= v + \gamma[Cv - 2ww^tCv - (w^tCw)w] = \\ &= v + \gamma[\lambda_v v - 2ww^t\lambda_v v - \lambda_w w] = (1 - \gamma[\lambda_w - \lambda_v])v \end{aligned}$$

$$\begin{aligned} Df_w(w) &= w + \gamma[Cw - 2w(Cw)^tw - (w^tCw)] = \\ &= w + \gamma[\lambda_w w - 2ww^t\lambda_w w - \lambda_w w] = \\ &= w + \gamma[-2\lambda_w\|w\|w] = [1 - 2\lambda_w]w \end{aligned}$$

So $\mathcal{B}$ is also a basis of eigenvectors for Df_w.

We want w to be an attracting hyperbolic fixed point (hence an attracting fixed point in the topological sense). The equivalent set of conditions on Df_w is that the eigenvalues all have absolute values strictly less than 1:

$$\mid 1 - \gamma(\lambda_w - \lambda_v) \mid < 1 \ \ , \forall\, v \in \mathcal{B}\,,\ v \neq w$$

$$\mid 1 - 2\gamma\lambda_w \mid < 1$$

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

(1) $\lambda_w > \lambda_v\,,\ \forall\, v \neq w$ (i.e. λ_w is the maximal eigenvalue)

(2) $\gamma < \frac{1}{\lambda_w}$

(3) $\gamma < \frac{2}{\lambda_w - \lambda_v}\,,\ \forall v \neq w$ (weaker than (2))

These conditions are always satisfied provided: (I) C has a maximal eigenvalue of multiplicity one and (II) γ is small enough ($\gamma < \frac{1}{\lambda_w}$).

Conclusion. We have shown that, under the conditions (I) and (II), the network will "learn" the principal component of the correlation matrix C.

The result is certainly not surprising and not new in the literature (see, for example [**?**]). However, this method brings with itself a fairly easy extension that will encompass more general settings.

3 An extension

Our next goal is to generalize this argument for an iteration function that includes errors. The new model introduces an error matrix, $T \in \mathcal{M}_n(\mathbb{R})$ that has positive entries, is symmetric and equal to the identity matrix $I \in \mathcal{M}_n(\mathbb{R})$ in case the error is zero.

$$f^T(w) = w + \gamma[TCw - (w^tCw)w]$$

We will work from now under the assumption that TC has a unique principal eigenvalue.

Note that the symmetric, positive definite matrix $C \in \mathcal{M}_n(\mathbb{R})$ defines a dot product in $\mathbb{R}^n$ as:

$$\langle v, w \rangle_C = v^tCw$$

If v and w are eigenvectors of TC corresponding to the eigenvalues $\lambda_v \neq \lambda_w$, then they are orthogonal with respect to the dot product $\langle , \rangle_C$. Indeed:

$$TCv = \lambda_v v \ \Rightarrow\ \langle w, TCv \rangle_C = \lambda_v \langle w, v \rangle_C$$

$$TCw = \lambda_w w \ \Rightarrow\ \langle v, TCw \rangle_C = \lambda_w \langle v, w \rangle_C$$

Hence $\lambda_v \langle v, w\rangle_C = \lambda_w \langle v, w\rangle_C$. As $\lambda_v \neq \lambda_w$, it follows that $\langle v, w\rangle_C = 0$, hence v and w are orthogonal with respect to the given dot product.

A fixed point for f^T is a vector $w = (w_1 ... w_n)^t$ such that $TCw = (w^t C w)w$. In other words, w is fixed by f^T if and only if it is an eigenvector of TC (with corresponding eigenvalue μ), normalized such that $\parallel w \parallel_C = \mu$. Clearly, this is possible when $\mu > 0$ and only then.

Hence for any positive eigenvalue of TC, we can find a fixed vector defined as above. Consider then $\mu > 0$ an eigenvalue of TC (suppose TC has at least one positive eigenvalue) and call w its corresponding fixed vector w:

$$TCw = \mu w, \quad \|w\|_C = \mu$$

If the multiplicity of μ is one, then w is orthogonal in $\langle , \rangle_C$ to all other eigenvectors of TC.

Recall that

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

Take w to be a fixed point of f^T. w will hence be an eigenvector of TC, with eigenvalue $\lambda_w = (w^t C w)w > 0$. Calculate:

$$\begin{aligned} Df_w w &= w + \gamma[TCw - 2w(Cw)^t w - (w^t C w)w] = \\ &= w + \gamma[-2ww^t C w] = [1 - 2\gamma\lambda_w]w \end{aligned}$$

$$\begin{aligned} Df_w v &= v + \gamma[TCv - 2ww^t C v - \lambda_w v] = \\ &= v + \gamma[(\lambda_v - \lambda_w)v - 2\langle w, v\rangle_C w] = (1 - \gamma[\lambda_w - \lambda_v])v \end{aligned}$$

for any other eigenvector v of TC with eigenvalue $\lambda_v \neq \lambda_w$:

It is fairly easy to see that Df_w has all eigenvalues less than one in absolute value if and only if λ_w is the principal eigenvalue of TC and $\gamma < \frac{1}{\lambda_w}$.

In conclusion: Suppose the matrix TC has a positive eigenvalue μ with multiplicity one. Then its corresponding eigenvector w normalized as $\|w\|_C = \mu$ is the only hyperbolic attracting fixed vector of f^T, provided that γ is smaller than the inverse of the eigenvalue μ.

4 The error matrix and the output performance

Recall that, in the absence of error, the network set to learn the distribution with correlation C converges to the principal eigenvector of C, which is $w = (100..0)^t$.

Consider:

$$\cos(\theta) = \frac{\langle w^C, w^{TC}\rangle}{\|w^C\| \cdot \|w^{TC}\|}$$

the cosine of the angle between the attractor w^C in the zero error case and the attractor w^{TC} in the case described in this section. As mentioned before, $w^C = (10...0)^t$. Hence:

$$\cos(\theta) = \frac{\mid w_1^{TC} \mid}{\|w^{TC}\|}$$

can be considered a reasonable measure of the output error for a given error matrix T.

We are interested in seeing how this output error angle changes with the quality of information transfer, for various sizes of the network $n \geq 2$ and a fixed correlation $\lambda > 1$.

From now on, we will consider the inputs to be uncorrelated and such that one neuron in the network has a preferential treatment to all others, namely we take C to be a diagonal $n \times n$ matrix of the form:

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

where $\lambda > 1$.

The construction of T involves a number $q \in [0, 1]$ that we will refer to as the "quality" of the network. q measures, as its name suggests, the accuracy of information transfer in the network and is independent of the size n of the network. In the error-matrices T we consider, each change in the synaptic-strength vector "reaches" its target in a proportion of $Q = q^n$ and "misses" it in a proportion of $1 - Q = 1 - q^n$. The probability for a synaptic contact to be mistaken to any of its neighbors is the same, in other words the error $1 - Q$ is in average evenly distributed to all neighbors of the target. It remains to define how a network understands the notion of "*neighboring cell*". We will describe in this section three such examples, and plot the results we obtained by Matlab calculations in each case. The next section is dedicated to a detailed analytic analysis of the most plausible model.

<u>Error model 1:</u> Each synapse has two neighbours, hence a change directed to the component w_i of the weight vector will be in proportion $1 - Q$ due to leaks from w_{i-1} and w_{i+1}. (Here, we consider the indexes modulo n, so that $w_1 = w_{n+1}$ is a neighbour of w_n and $w_n = w_0$ is a neighbour of w_1.) The corresponding error matrix will be:

$$T = \begin{pmatrix} Q & \epsilon & 0 & \cdot & \cdot & \epsilon \\ \epsilon & Q & \epsilon & 0 & \cdot & 0 \\ 0 & \epsilon & Q & \epsilon & \cdot & 0 \\ . & . & . & . & . & . \\ 0 & \cdot & \cdot & \epsilon & Q & \epsilon \\ \epsilon & 0 & \cdot & \cdot & \epsilon & Q \end{pmatrix} = QI + \epsilon(P + P^{-1})$$

where

$$P = \begin{pmatrix} 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & 0 \\ 0 & 0 & 0 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & 0 & 0 & 1 \\ 1 & 0 & \cdot & \cdot & 0 & 0 \end{pmatrix}$$

and the leak to each neighbour $\epsilon = \frac{1-Q}{2}$.

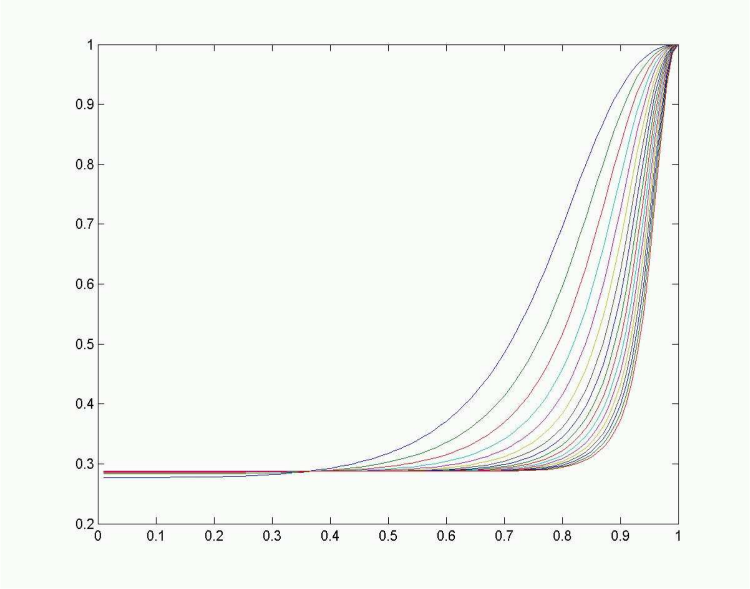

Figure 1: ***Model 1:*** *We plot* $\cos(\theta(q))$ *(vertical axis) against* q *(horizontal axis), for values of* n *from* 5 *to* 20 *neurons (each graph correspondes to a value of* n*). The maximal eigenvalue of* C *is fixed to* $\lambda = 10$*. For all* n*,* $\cos(\theta)$ *increases from about* 0.3 *to* 1*, as the quality increases from worst (*$q = 0$*) to perfect (*$q = 1$*). All graphs show a slow change for small* q *and a sudden increase on a small interval before* $q = 1$*. The plots are lower and the change is steeper as* n *gets larger. Note that the worst angle output error for each* n *(i.e. the value* $\cos(\theta(0))$*) does not show significant dependence on the size* n*.*

Let $\theta(q)$ be, as before, the angle between the principal component of C and the principal component of TC, for a given quality q. We use a Matlab computation to obtain the eigenspaces of TC and plot $\cos(\theta)$ with respect to q (see figure 1).

Error model 2: The incoming leaks from neighbours decrease exponentially with the distance to the correct synapse. In other words:

$$T = \begin{pmatrix} Q & \frac{\epsilon}{2} & \frac{\epsilon}{2^2} & \cdot & \frac{\epsilon}{2^2} & \frac{\epsilon}{2} \\ \frac{\epsilon}{2} & Q & \frac{\epsilon}{2} & \frac{\epsilon}{2^2} & \cdot & \frac{\epsilon}{2^2} \\ \frac{\epsilon}{2^2} & \frac{\epsilon}{2} & Q & \frac{\epsilon}{2} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{\epsilon}{2^2} & \cdot & \cdot & \frac{\epsilon}{2} & Q & \frac{\epsilon}{2} \\ \frac{\epsilon}{2} & \frac{\epsilon}{2^2} & \cdot & \cdot & \frac{\epsilon}{2} & Q \end{pmatrix}$$

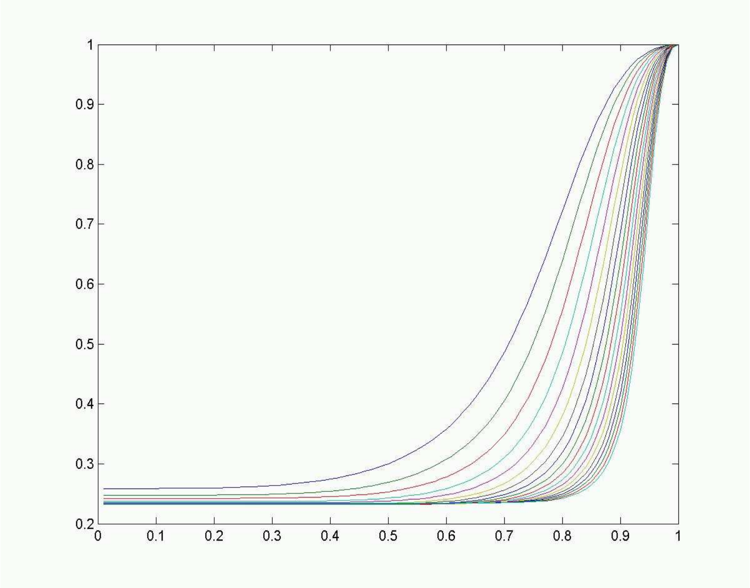

Figure 2: ***Model 2:*** *We plot* $\cos(\theta)$ *against* q*, for fixed correlation factor* $\lambda = 10$ *and* n *from* 5 *to* 20 *neurons (each curve corresponds to a size* n*, lower curves correspond to larger* n*). For all* n*,* $\cos(\theta)$ *increases from small values to* 1*, as the quality increases from* 0 *to* 1*. The change is slow for small values of* q*. For* q *approaching* 1*, there is a sudden raise to* $\cos(\theta) = 1$ *(perfect output). This raise gets steeper as* n *gets larger. The small starting values* $\cos(\theta(0))$ *are slightly more scattered than in* ***model 1*** *for various sizes* n*.*

$$T = \begin{cases} QI + \frac{\epsilon}{2}(P + P^{-1}) + ... + \frac{\epsilon}{2^k}(P^k + P^{-k}), \text{ if } n = 2k+1 \\ \\ QI + \frac{\epsilon}{2}(P + P^{-1}) + ... + \frac{\epsilon}{2^{k-1}}(P^{k-1} + P^{-(k-1)}) + \frac{\epsilon}{2^k}P^k, \text{ if } n = 2k \end{cases}$$

for $k \geq 1$, where ϵ is such that the sum of all leaks into a synaptic contact from other synapses is $1 - Q$:

$$2(\frac{\epsilon}{2} + ... + \frac{\epsilon}{2^k}) = 1 - Q, \text{ if } n = 2k+1$$
$$2(\frac{\epsilon}{2} + ... + \frac{\epsilon}{2^{k-1}}) + \frac{\epsilon}{2^k} = 1 - Q, \text{ if } n = 2k$$

<u>Error model 3:</u> Each synapse in the network receives leaks in equal proportion from all others, hence the individual error $\epsilon = \frac{1-Q}{n-1}$ and the matrix T is given by:

$$T = \begin{pmatrix} Q & \epsilon & \epsilon & \cdot & \cdot & \epsilon \\ \epsilon & Q & \epsilon & \epsilon & \cdot & \epsilon \\ \epsilon & \epsilon & Q & \epsilon & \cdot & \epsilon \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \epsilon & \cdot & \cdot & \epsilon & Q & \epsilon \\ \epsilon & \epsilon & \cdot & \cdot & \epsilon & Q \end{pmatrix} = QI + \epsilon[P + P^2 + ... + P^{(n-1)}]$$

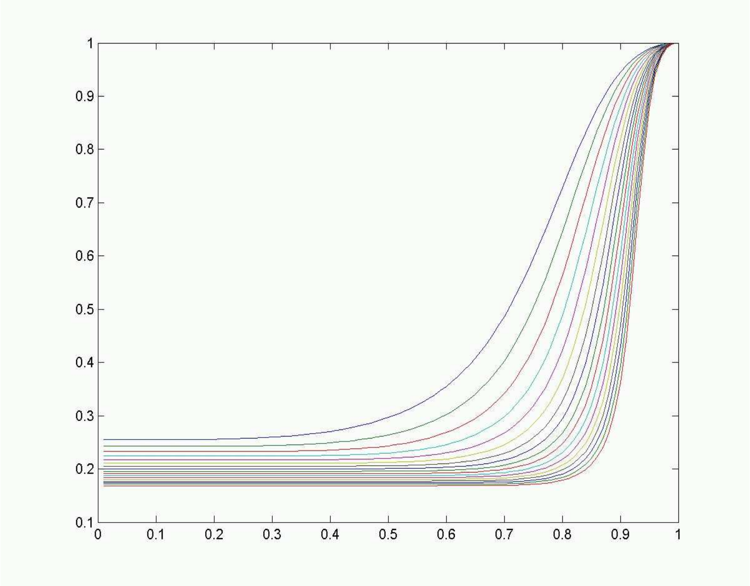

Figure 3: ***Model 3:*** *The plot of cosine of the error angle against the quality q has mainly the same shape (we ahow plots for $\lambda = 10$ and n from 5 to 20 neurons). However, the values $\cos(\theta(0))$ are clearly more scattered for various sizes n than in the previous plots. In fact, as n gets arbitrarily large, $\cos(\theta(0))$ approaches zero (see the proof in section 5 and the discussion in section 6).*

5 Plotting the output performance

The purpose of this section is to give an analytical explanation for the plots obtained for model 3. Appendix A should give some indication of why it is interesting to look at the case where all input cells are equal neighbors. The appendix also shows why our particular choice of the dependence $Q = q^n$ is not arbitrary, but motivated by a detailed analysis of the geometry of the network and of the synaptic mechanisms.

The characteristic polynomial of T is easy to calculate:

$$P^T(x) = det(T - xI) = (Q - \epsilon - x)^{(n-1)}(1 - x)$$

Hence T is invertible, except when $Q - \epsilon = 0$, i.e. when $q = \frac{1}{\sqrt[n]{n}}$.

Similarly, the characteristic polynomial of TC is:

$$P^{TC}(x) = det(TC - xI) = (Q - \epsilon - x)^{n-2}[x^2 - x(\lambda + 1 + \epsilon(\lambda - 1 - n\lambda)) + \lambda - n\lambda\epsilon]$$

Hence TC has three distinct eigenvalues: $Q - \epsilon = \frac{nq^n - 1}{n-1}$ with multiplicity $n-2$ and the two real roots of the quadratic equation:

$$S(x) = x^2 - x(\lambda + 1 + \epsilon(\lambda - 1 - n\lambda)) + \lambda - n\lambda\epsilon = 0$$

Call μ the larger solution of the two. μ is also the maximal eigenvalue of TC (as shown in lemma 5.2):

$$\mu = \frac{1}{2}[B + \sqrt{\Delta}]$$

where $B = \lambda + 1 + \epsilon(\lambda - 1 - n\lambda)$ and $\Delta = B^2 - 4\lambda(1 - n\epsilon)$. Recall that $\epsilon = \epsilon(q)$ and so $\mu = \mu(q)$. Clearly $Q - \epsilon = \frac{nq^n - 1}{n-1} < 1$, hence:

$$1 < \mu < \lambda$$

The hyperbolic attractor of the dynamical system defined by f^T is an eigenvector of TC with eigenvalue μ. So we are interested in emphasizing briefly some properties of μ, as a start.

Lemma 5.1.

$$\frac{\lambda n}{\lambda(n-1) + 1} < \mu < \lambda$$

Proof. A simple calculation shows that:

$$S(\frac{\lambda n}{\lambda(n-1) + 1}) = \frac{-\lambda(\lambda - 1)^2(n-1)}{[\lambda(n-1) + 1]^2} < 0$$

and

$$S(\lambda) = \epsilon\lambda(\lambda - 1)(n - 1) > 0$$

Therefore the largest root μ of $S(x) = 0$ is situated between these two values. □

Lemma 5.2. *$\mu = \mu(q)$ is the maximal eigenvalue of TC and has multiplicity one, for any $0 < q < 1$.*

Proof. Clearly $Q - \epsilon < 1 < \mu$, so all we need to show is that the equation $S(x) = 0$ has two distinct roots. Indeed:

$$\Delta = B^2 - 4(\lambda - \lambda n\epsilon) = [\lambda + 1 + \epsilon(\lambda - 1 - \lambda n)]^2 - 4(\lambda - \lambda n\epsilon) =$$

$$= [(\lambda - 1) + \epsilon(\lambda - 1 - \lambda n)]^2 + 4\epsilon(\lambda - 1)$$

Hence μ is an eigenvalue of multiplicity one. □

Let's take a brief look at how the analysis in section 3 works in this case. Note that TC is not symmetric, may not be positive definite, or even invertible. But its maximal eigenvalue is strictly positive and has multiplicity one.

A fixed point for f^T is a vector $w = (w_1 ... w_n)$ such that $TCw = (w^t Cw)w$. With our present choice for C:

$$w^t Cw = \lambda w_1^2 + w_2^2 + ... + w_n^2 \ \geq 0$$

So, if ν is a strictly positive eigenvalue of TC, then the eigenvector corresponding to ν, normalized such that $\lambda w_1^2 + w_2^2 + ... + w_n^2 = \nu$ is a fixed vector for f^T.

$$det(TC - xI) = (Q - \epsilon - x)^{n-2} S(x)$$

and $Q - \epsilon \leq 0$ for $q \leq \frac{1}{\sqrt[n]{n}}$. So we may not have n distinct fixed vectors for f^T. But the maximal eigenvalue $\mu > 0$. We apply the results in section 3 and we obtain that the principal eigenvector w , normalized such that $\lambda w_1^2 + ... + w_n^2 = \mu$ is the only attractor for f^T.

We will be looking for the *direction* of the principal component of TC:

$$\begin{pmatrix} \lambda Q & \epsilon & \epsilon & \cdot & \cdot & \epsilon \\ \lambda\epsilon & Q & \epsilon & \epsilon & \cdot & \epsilon \\ \lambda\epsilon & \epsilon & Q & \epsilon & \cdot & \epsilon \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \lambda\epsilon & \cdot & \cdot & \epsilon & Q & \epsilon \\ \lambda\epsilon & \epsilon & \cdot & \cdot & \epsilon & Q \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{pmatrix} = \mu \begin{pmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ \cdot \\ w_n \end{pmatrix}$$

If we fix $w_1 = \frac{1}{\lambda}$, we can calculate:

$$w_j = \frac{1}{\lambda} \frac{\mu - \lambda(Q - \epsilon)}{\mu - (Q - \epsilon)}$$

The norm of w is:

$$\|w\|^2 = \frac{1}{\lambda^2} \left[1 + (n-1)(\frac{\mu - \lambda(Q - \epsilon)}{\mu - (Q - \epsilon)})^2 \right]$$

Hence the error angle we defined in the last section is given by:

$$cos(\theta) = \frac{w_1}{\| w \|}$$

Only for computation simplicity, we prefer to look instead at:

$$| \tan(\theta) |= \sqrt{n-1} \frac{\mu - \lambda(Q - \epsilon)}{\mu - (Q - \epsilon)} = \sqrt{n-1} \frac{\mu - \lambda(1 - n\epsilon)}{\mu - (1 - n\epsilon)}$$

But μ is solution of $S(x) = 0$, which gives us:

$\mu - \lambda(1 - n\epsilon) = \frac{\mu(\lambda-1)\epsilon}{\mu-1}$ and $\mu - (1 - n\epsilon) = \frac{\mu\epsilon(n-1)(\lambda-1)}{\lambda-\mu}$

Hence:

$$| \tan(\theta) |= \frac{1}{\sqrt{n-1}} \frac{\lambda - \mu}{\mu - 1}$$

where $\mu = \mu(q)$, for $q \in [0, 1]$.

From now on, for a fixed $n \geq 2$ and $\lambda > 1$, this will be the function of q we will study:

$$h(q) =\mid \tan(\theta(q)) \mid= \frac{1}{\sqrt{n-1}} \frac{\lambda - \mu(q)}{\mu(q) - 1}$$

where $\mu = \mu(q)$ is the larger root of the equation $S(x) = 0$ (recall that the coefficients of $S(x)$ depend on q).

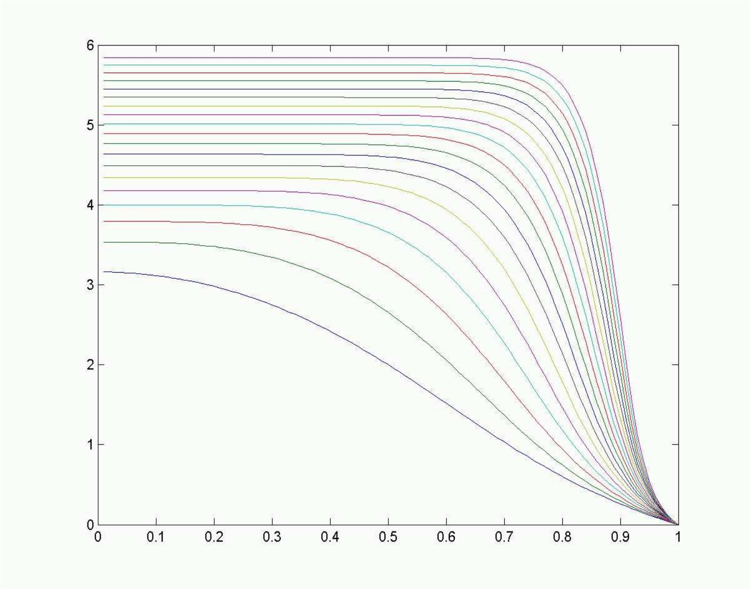

Figure 4: *The plot shows the behavior of the function* $h(q) =\mid \tan(\theta(q)) \mid= \frac{1}{\sqrt{n-1}} \frac{\lambda-\mu(q)}{\mu(q)-1}$ *for values of n from 5 to 20. For each n, $h(q)$ decreases slowly for q between $q = 0$ and $q = q_0$, then has a sudden drop from $h(q_0)$ to zero for q between $q = q_0$ and $q = 1$. The value $h(0) =\mid \tan(\theta(0)) \mid$ approaches ∞ (i.e. $\cos(\theta(0))$ gets close to zero) when n goes to ∞. Also, the drop gets arbitrarily steep and narrow as n gets arbitrarily large.*

Step 1: We would like first to evaluate h at $q = 0$ (maximum error, all information transfer goes "wrong") and at $q = 1$ (quality 1, no error).

If $q = 1$, then $\epsilon(q) = 0$, so $S(x) = x^2 - (\lambda + 1)x + \lambda$. Hence $\mu(1) = \lambda$, so $h(1) = 0$. This is the expected result: the output error angle θ should be zero when the quality is perfect.

If $q = 0$, then $\epsilon(q) = \frac{1}{n-1}$. This gives us:

$$\mu(0) = \frac{n - 2 + \sqrt{(n-2)^2 + 4\lambda(n-1)}}{2(n-1)}$$

Clearly:

$\mu(0) \longrightarrow 1$ as $n \longrightarrow \infty$ and

$\sqrt{n-1}(\mu(0)-1) = \frac{4(\lambda-1)\sqrt{n-1}}{n+\sqrt{(n-2)^2+4\lambda(n-1)}} \longrightarrow 0$ as $n \longrightarrow \infty$

Hence $h(0) = \dfrac{1}{\sqrt{n-1}}\dfrac{\lambda-\mu}{\mu-1}$ approaches ∞ as n goes to ∞.

Step 2: We analyze the behavior of the derivative $h'(q)$. We calculate:

$$h'(q) = -\frac{\lambda-1}{\sqrt{n-1}}\frac{\mu'(q)}{(\mu(q)-1)^2}$$

By implicit differentiation with respect to q of the equation $S(\mu(q)) = 0$ we get that:

$$\mu'(q)[2\mu(q)-B] = -\epsilon'(q)[\mu(q)(\lambda n-\lambda+1)-\lambda n]$$

Recall that we called $B = B(q) = \lambda+1+\epsilon(\lambda-1-\lambda n)$ and we showed that $\mu > \frac{B}{2}$, hence $2\mu - B > 0$, $\forall\, q \in [0,1]$. We have also showed that $\mu > \frac{\lambda n}{\lambda n-\lambda+1}$, $\forall\, q \in [0,1]$. Moreover, $\epsilon'(q) = -\frac{n}{n-1}q^{n-1} \leq 0$, where equality happens for $q = 0$.

In conclusion $\mu'(q) \geq 0$, $\forall q \in [0,1]$, with equality when $q = 0$. This means that $h'(0) < 0$ for $0 < q \leq 1$ and $h'(q) = 0$ if $q = 0$, which tells us that, as q goes from 0 to 1, the function h decreases from the initial value $h(0)$ to zero.

Moreover: $\mu(1) = \lambda$, hence $\mu'(1) = n\lambda$.

$$h'(1) = -\frac{\lambda}{\lambda-1}\frac{n}{\sqrt{n-1}} \longrightarrow \infty \quad \text{as} \quad n \longrightarrow \infty$$

So the graph of h is flat at $q = 0$ for any value of n, but decreases with a steeper slope at $q = 1$ as n gets larger.

Step 3: The second derivative of h is (the prime symbol means derivative with respect to q):

$$h''(q) = -\frac{\lambda-1}{\sqrt{n-1}}\frac{\mu''(\mu-1)-2(\mu')^2}{(\mu-1)^2}$$

Again by implicit differentiation,we can calculate:

$$\mu''(2\mu-B) = -2(\mu')^2 - 2\epsilon'\mu'(\lambda n-\lambda+1) - \epsilon''[\mu(\lambda n-\lambda+1)-\lambda n]$$

Clearly at $q = 0$ we have an inflection point.

At $q = 1$:

$\mu''(1) = \dfrac{2n^2\lambda}{(n-1)(\lambda-1)} + n\lambda(n-1) > 0$

and

$h''(1) = \dfrac{\lambda-1}{\sqrt{n-1}}\dfrac{n\lambda}{n-1}[(n-2)^2+\lambda(n^2-1-3] < 0 \quad \text{for} \quad n \leq 2$

It follows that h is concave up at $q = 1$.

It is fairly easy to show that h has at most one inflection point between $q = 0$ and $q = 1$. We do not aim for its exact value, but for a good estimate that would permit us to compare the behavior of the function $h(q)$ for various values of the size n.

An interesting point to test is $q_0 = \frac{1}{\sqrt[n]{n}}$. This is the value of q that makes the matrix T noninvertible, in other words it is the quality of the network of size n where the information is distributed evenly among all neurons (i.e. $Q = \epsilon$). We would expect the learning process to perform very poorly for values of q less than q_0, so we foresee a steep drop in the graph between $q_0 = \frac{1}{\sqrt[n]{n}}$ and $q = 1$ (from bad performance to perfect performance $h(1) = 0$). If $n \longrightarrow \infty$, then $q_0 \longrightarrow 1$, so the transition from $h(q_0) = \sqrt{n-1}$ to $h(1) = 0$ gets more sudden with the size increase. Indeed:

$$\epsilon(q_0) = \tfrac{1}{n}, \quad B = 1 + \tfrac{\lambda-1}{n}$$

$$\mu(q_0) = 1 + \frac{\lambda - 1}{n}$$

$$h(q_0) = \sqrt{n-1}$$

$$\mu'(q_0) = -\frac{\epsilon'(n-1)(\lambda-1)^2}{n+\lambda-1}$$

$$h'(q_0) = -\frac{n\sqrt[n]{n}(\lambda-1)^2}{(n+\lambda-1)\sqrt{n-1}}$$

$$\mu''(q_0)(\mu(q_0)-1) - 2(\mu'(q_0))^2 =$$

$$= \frac{(n-1)(\lambda-1)^3}{n+\lambda-1}\left(\frac{2(\epsilon'(q_0))^2[n^2+(\lambda-1)^2]}{(n+\lambda-1)^2} - \frac{\epsilon''(q_0)}{n}\right) > 0$$

Hence:

$$h''(q_0) = -\frac{\lambda-1}{\sqrt{n-1}}\,\frac{\mu''(q_0)(\mu(q_0)-1) - 2(\mu'(q_0))^2}{(\mu'(q_0)-1)^2} < 0$$

So $h''(q_0) < 0$ and $h''(1) > 0$, hence there is an inflection point between q_0 and 1. Also, note that the value of h at $q = q_0$ is still quite large, and the derivative $h'(q_0)$ quite small (in fact, $h'(q_0) \longrightarrow 0$ as $n \longrightarrow \infty$). So, after decreasing fairly slowly for $0 < q < q_0$, the graph has indeed a steep drop from $h(q_0) = \sqrt{n-1}$ to $h(1) = 0$ between $q = q_0$ and $q = 1$, passing through an inflection stage.

6 Discussion and conclusions

The molecular quasispecies model of Eigen describes the stationary distribution of polynucleotides maintained by chemical reactions (affecting error-prone replication) and by transportation processes. This distribution shows a sharp transition between a drifting population of essentially random macromolecular sequences and a localized population of close relatives. This transition at a threshold value of the error was found to depend on sequence length, distributions of selective values and population sizes. The error threshold seems to set a limit to the genome length of several classes of RNA viruses.

Comparing Oja's evolution equation for synaptic weights with Eigen's equation for prebiotic evolution, [S1] observed the formal equivalence of the two. which was not too surprising. We hope that this paper underlines the deeper structural similarities between the two models, as well as their relevant differences. The conclusion to our discussion is aimed in the same direction as Eigen's: could we find a threshold that sets a limit for the size of a functional learning network?

All examples we considered in this paper have a common feature: the accuracy of information transfer within the network Q depends exponentially on the size n of the network. The base of the exponential is a parameter $q \in [0, 1]$, characteristic of the network alone and size-independent. Due to this dependence, the plots of $\cos(\theta)$ against q have similar shape in all three cases. Moreover, this particular property makes our models fairly similar to Eigen's model of self-replicating symbol-sequences. The plots of our results show therefore a notably similar error catastrophe.

However, there are considerable differences between examples 1,2 and 3, that reflect the degree of scattering of the error within our linear system. The notion of "neighbor" of a synapse is totally distinct in each of the three cases. In **model 1**, each synapse has two equivalent neighbors. In **model 3**, each synapse has all others as neighbors equally participating to the inward information leak. In **model 2**, we are looking at an intermediate case: for a fixed synapse, all other synapses are neighbors of different degrees, the degree being given by the distance to the fixed cell. The leak into the given synapse decreases exponentially with the distance. Hence all cells in the network are mutual neighbors, but less related if more distant.

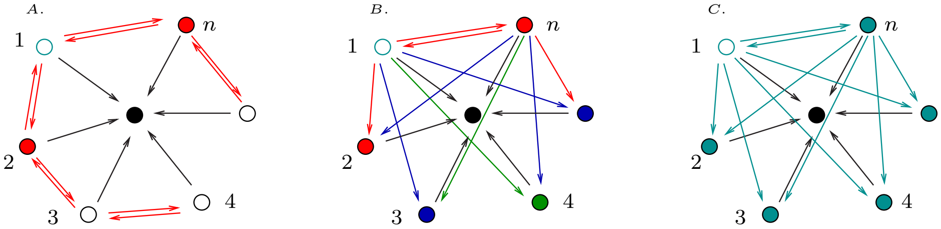

Figure 5: *The meaning of neighbours in out three models. I the three schemata, we emphasize the neighbors of cell* 1*: In figure* ***A.*** *(model 1) the cell has two equal neighbours (in red). In figure* ***B.*** *(model 2), the cell has all others as neighbours of different degrees: degree 1 (closest, red), degree 2 (blue), degree 3 (green). Figure* ***C.*** *shows the schema for model 3, in which all cells are equal neighbors (cyan).*

The pictures (see figure 6) show distinctly that the larger the error diffusion, the more significant the scatter of the values of the maximal output error $\cos(\theta(0))$. To clarify, let's fix a correlation factor $\lambda > 1$. In the models with more error diffusion (2 and 3), the maximal output angle error for a given n can get arbitrary small with increasing n, as opposed to model 1, where the maximal output error does not seem to depend on the value n of the size.

In conclusion: Suppose the quality q is very poor, i.e. close to zero. If leaks are fairly localized (model 1), the answer obtained through learning does not get significantly further from the truth with an increase in the size of the network. On the contrary, if leaks have high diffusion, large values of n lead to a totally wrong answer $(\cos(\theta(0)) \simeq 0)$.

An equally interesting remark concerns the high-quality interval, i.e. values of q close to 1. All three families of graphs present a sudden rise in the very high-quality range, rise that becomes almost vertical for large values of n. Recall that for model 3 we have an analytic interpretation for this phenomenon: for values of q smaller than $q_0 = \frac{1}{\sqrt[n]{n}}$, the synapse gets more information from each individual leak than from its source, so a network with quality poorer than q_0 is unable to learn anything close to the correct answer in this case.

This sudden change from *bad answer* to *good answer* within a very narrow quality range close to 1 is the remarkable similarity with the error catastrophe found by Eigen in the replication of molecules.

To relate this discussion to the properties of the distribution to be learned, let's tune the correlation factor $\lambda > 1$.We notice that the curves get lower with a decrease of λ. The meaning of this is the following: the more similar the inputs are on the n neurons, the larger the effect of the error (on a learning system with fixed quality q). In other words,the same network learns easier an uncorrelated distribution on inputs if the predominant input variance λ is larger. This is hardly a surprise.

In conclusion: A linear network of size n with information leaks is still capable of learning an input distribution in a quite large variety of settings. However, we were able to show that, if three distinct plausible contexts, if the network is too large it needs almost perfect quality of information transfer in order to provide an outcome reasonably close to the correct one. Otherwise, what it learns is, depending on the case, partially of totally wrong.

This could equip us with a reason why networks are never larger than a universal upper-bound $N \simeq 10,000$ neurons, although there would be no strict geometric constrains for obtaining larger ones. A system with an intrinsic, built-in quality shows a sharper learning capability if it does not exceed a certain size. *Bigger is not always better.*

Appendix

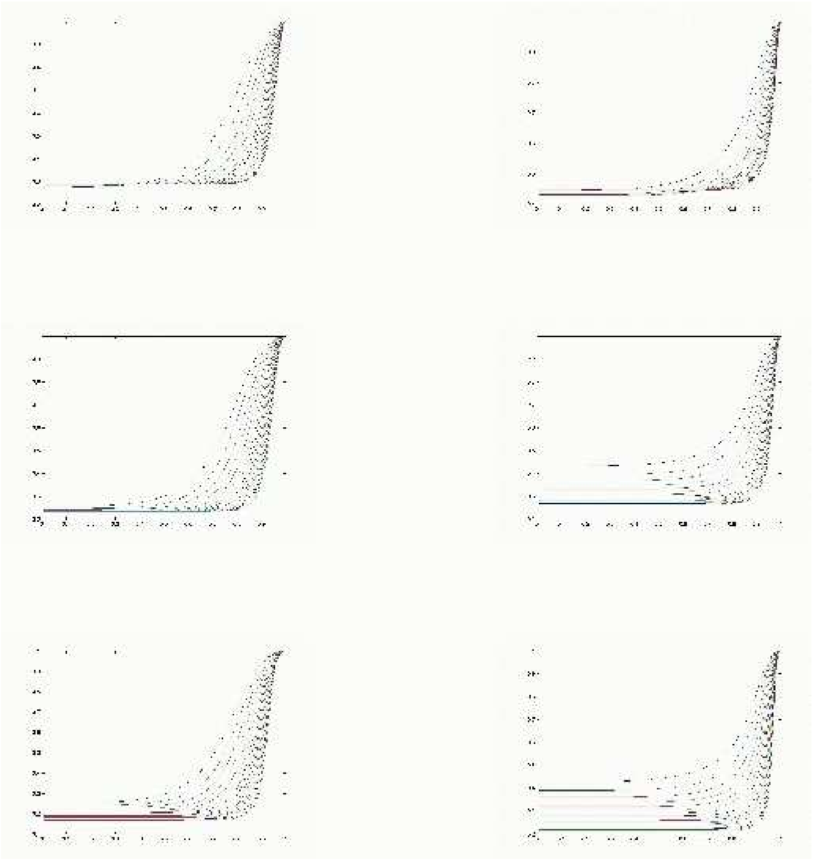

Figure 6: *The plots of* $\cos(\theta)$ *against* q *for* ***model 1*** *(top row),* ***model 2*** *(middle row) and* ***model 3*** *(bottom row), for the correlation factor* $\lambda = 10$ *(left column) and* $\lambda = 2$ *(right column).*

References

[E] M. Eigen, *Selforganization of matter and the evolution of biological macromolecules*, Naturwissenschaften 58, nr.10, pp 465-523 (1971)

[E1] M. Eigen, J. McCaskill,P. Schuster *Molecular Quasi-Species*, J. Phys. Chem., 1988, 92, 6881-6891

[HKP] J. Herz, A. Krogh, R. Palmer, *Introduction to the theory of neural computation*, Lecture notes volume in the Santa Fe Insitutue Studies in the Sciences of Complexity

[MM] K. Miller, D. MacKay, *The role of constraints in Hebbian learning*, Neural Computation 6, pp 100-126 (1994)

[O] E. Oja, *A simplified neuron model as a principal component analyzer*, Journal of Mathematical Biology 15, pp 267-273 (1982)

[OK] E. Oja, J. Karhunen, *On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix*, Journal of Mathematical Analysis and Applications 106, pp 69-84 (1985)

[R] C. Robinson, *Dynamical systems: stability, symbolic dynamics and chaos*, CRC Press, INC.

[S] H.G. Schuster, *Learning by maximizing the information transfer through nonlinear noisy neurons and noise breakdown*, Physical Review 46, nr.4, pp 2131-2138 (1992)

[S1] H.G. Schuster, *Complex Adaptive Systems - Prebiotic and artificial evolution*, Scator Verlag