# Objective Function Formulation of the BCM Theory of Visual Cortical Plasticity: Statistical Connections, Stability Conditions*

**Nathan Intrator and Leon N Cooper**†
Physics Department and
Center for Neural Science
Brown University
Providence, RI 02912

**Abstract**

In this paper, we present an objective function formulation of the BCM theory of visual cortical plasticity that permits us to demonstrate the connection between the unsupervised BCM learning procedure and various statistical methods, in particular, that of Projection Pursuit. This formulation provides a general method for stability analysis of the fixed points of the theory and enables us to analyze the behavior and the evolution of the network under various visual rearing conditions. It also allows comparison with many existing unsupervised methods. This model has been shown successful in various applications such as phoneme and 3D object recognition. We thus have the striking and possibly highly significant result that a biological neuron is performing a sophisticated statistical procedure.

*Keywords:* unsupervised learning, feature extraction, dimensionality reduction.

*In *Neural Networks* Vol. 5, pp. 3-17, 1992

†This work was supported in part by the National Science Foundation, the Office of Naval Research, and the Army Research Office.

# 1 Introduction

In the past decade, much work has been done on a theory of synaptic plasticity in visual cortex (Bienenstock, Cooper and Munro, 1982; BCM). This theory accounts in a precise and quantitative fashion for the modification of response properties of neurons in striate cortex obtained by manipulating the visual experience of the animal during a critical period of postnatal development. It allows a precise specification of theoretical equivalents of experimental situations and makes possible detailed and quantitative comparison of theory with experiment in what are called classical rearing conditions. These include normal rearing, monocular deprivation, reverse suture, strabismus, binocular deprivation, as well as the restoration of normal binocular vision after various forms of deprivation. In detailed simulations, Clothiaux et. al. (1991) find quantitative agreement of theory and experiment both for equilibrium states and the kinetics by which they are reached.

In this paper, we present an objective function formulation of the BCM theory of visual cortical plasticity. This permits us to demonstrate the connection between the unsupervised BCM learning procedure and various statistical methods, in particular, that of Projection Pursuit. This analysis has led us to modify slightly our learning rule resulting in improved stability and statistical properties. It also provides a general method for stability analysis of the fixed points of the theory and enables us to analyze the behavior and the evolution of the network under various visual rearing conditions. This new model has some advantages over the original exploratory projection pursuit model (Friedman, 1987). Due to its computational efficiency, it can extract several features in parallel, taking into account the interaction between the different extracted features via a lateral inhibition network. Feature extraction based on this model have been applied to various real-world problems such as phoneme recognition of a small-speaker database (Intrator, 1992), multi-speaker phoneme recognition from the TIMIT database (Intrator and Tajchman, 1991) using the Lyon's cochlear model (Slaney, 1988), and 3D object recognition (Intrator and Gold, 1993; Intrator et al., 1991). We thus have the striking and possibly highly significant result that a biological neuron is performing a sophisticated statistical procedure.

Section 2 reviews the evolution of the BCM theory, and its relevance to modeling of the primary visual cortex, area 17. Section 3 describes the statistical motivation behind unsupervised learning. This is used to motivate the objective function formulation of the modified BCM model given in Section 4. Based on statistical considerations, this formulation is further extended to a nonlinear neuron in a lateral inhibition network. In Section 5 we analyze the limiting behavior of the synaptic modification equations, using the formulation described in Section 4 and a connection established between the solution of the averaged deterministic differential equations and the solution of the random version of the equations (Appendix A). Analysis of this model in several situations related to visual experiments is given in Section 6.

# 2 Review of BCM Theory

In this section, we briefly review relevant experimental observations and the BCM theory. This will serve to introduce relevant notation and biological terms.

## 2.1 Visual Cortical Plasticity: Experimental Results

Neurons in the primary visual cortex, area 17, of normal adult cats are sharply tuned to the orientation of an elongated slit of light and most are activated by stimulation of either eye (Hubel

and Wiesel, 1959). Both of these properties – orientation selectivity and binocularity – depend on the type of visual environment experienced during a critical period of early postnatal development.

Monocular deprivation (MD) has profound and reproducible effects on the functional connectivity of striate cortex during the critical period, extending from approximately 3 weeks to 3 months of age in the cat (Frègnac and Imbert, 1984; Sherman and Spear, 1982). Brief periods of MD will result in a dramatic shift in the OD of cortical neurons so that most will be responsive exclusively to the open eye. The OD shift after MD is the best known, and most intensively studied type of visual cortex plasticity.

When MD is initiated late in the critical period (Presson and Gordon, 1982), or after a period of rearing in the dark (Mower et al., 1985), it will induce clear changes in cortical OD without a corresponding anatomic change in the geniculocortical projections. Long-term recordings from awake animals also indicate that OD changes can be detected within a few hours of monocular experience; this seems too rapid to be explained by the formation or elimination of axon terminals. Moreover, deprived-eye responses in visual cortex may be restored within minutes to hours under some conditions (Duffy et al., 1976), which suggests that synapses deemed functionally *disconnected* are nonetheless physically present. Therefore, it is reasonable to assume that changes in the functional binocularity may be explained by changes in the efficacy of individual cortical synapses.

The consequences of binocular deprivation (BD) on visual cortex stand in striking contrast to those observed after MD. First, MD leads to a loss of orientation selectivity in the deprived eye much faster than in BD. Second, although 7 days of MD during the second postnatal month leave few neurons in the striate cortex responsive to stimulation of the deprived eye, most cells remain responsive to stimulation through either eye after a comparable period of BD (Wiesel and Hubel, 1965). Thus it is not merely the absence of patterned activity in the deprived geniculate projection that causes the decrease in synaptic efficacy after MD.

The result of a reversed suture (RS) experiment is even more striking. In this experiment, the kitten is first exposed to normal visual environment, then one eye is sutured closed, for a few days until the sutured eye becomes functionally disconnected. At that time the sutured eye is opened and exposed to normal visual environment again, and the previously opened eye is closed. The result from this experiment is that the newly opened eye does not recover before the previously opened eye becomes disconnected.

## 2.2 Single Neuron Theory

A theoretical solution to the problem of visual cortical plasticity, was presented by Cooper, Liberman, and Oja (1979). According to this theory, the synaptic efficacy of active inputs increases when the postsynaptic target is concurrently depolarized beyond a *modification threshold*, $\Theta_M$. However, when the level of postsynaptic activity falls below $\Theta_M$, then the strength of active synapses decreases.

An important feature was added to this theory in 1982 by Bienenstock Cooper and Munro (BCM). They proposed that the value of the modification threshold is not fixed, but instead varies as a nonlinear function of the average output of the postsynaptic neuron. This provided stability properties and explained, for example, why the low level of postsynaptic activity during binocular deprivation does not drive the strengths of all cortical synapses to zero. Their form of synaptic modification can be written as:

$$\dot{m}_j = \phi(c, \Theta_M) d_j \tag{1}$$

where $m_j$ is the efficacy of the j$^{\text{th}}$ Lateral Geniculate Nucleus (LGN) synapse onto a cortical neuron, $d_j$ is the level of presynaptic activity of the j$^{\text{th}}$ LGN afferent, $c$ is the level of activation of the postsynaptic activity of the postsynaptic neuron, which is given (in the linear region), by $m \cdot d$, and $\Theta_M$ is a nonlinear function of some time averaged measure of cell activity that in the original BCM formulation was proposed as

$$\Theta_M = (\bar{\bar{c}})^2. \tag{2}$$

(In BCM, this time average is replaced, for simplicity, by a spatial average over the environmental inputs ($\bar{\bar{c}} \rightarrow m \cdot \bar{d}$). The shape of the function $\phi$ is given in Figure 1 for two different values of the threshold $\Theta_M$.

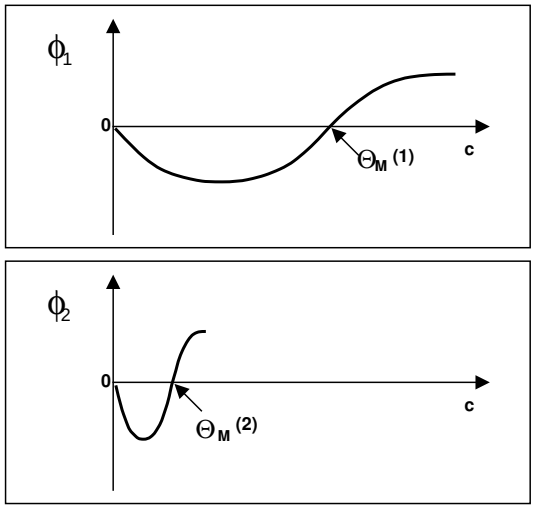


Figure 1: The $\phi$ function for two different $\Theta_M$'s

Further discussion of the biological relevance of the theory can be found in (Saul and Daniels, 1986; Bear et al., 1987; Bear and Cooper, 1990; Clothiaux et al., 1991).

### 2.3 Lateral Inhibition Network: Mean Field Theory

An extension of the single cell BCM neuron to a lateral inhibition network was presented by (Scofield and Cooper, 1985) and a mean field approximation of this network by (Cooper and Scofield, 1988).

The activity of neuron $j$ in such a network is affected by its input vector $d$ and by the adjacent neurons in the network and can be written

$$c_i = m_i \cdot d + \sum_j L_{ij} c_j. \tag{3}$$

In the context of visual cortex, the first term is due to the input from LGN and the second due to input from other cortical cells. Define $\bar{c}$ as the spatially averaged activity of all the cortical cells in the network: $\bar{c} = \frac{1}{N} \sum_i c_i$. The mean field approximation is obtained by replacing the inhibitory contribution of cell $j$, $c_j$ by its average value so that $c_i$ becomes:

$$c_i = m_i \cdot d + \bar{c} \sum_j L_{ij}. \tag{4}$$

From a consistency condition it follows that $\bar{c} = \bar{m} \cdot d + \bar{c} L_0 = (1 - L_0)^{-1} \bar{m} \cdot d$, where $\bar{m} = \frac{1}{N} \sum_i m_i$, and $L_0 = \frac{1}{N} \sum_{ij} L_{ij}$, so that $c_i = (m_i + (1 - L_0)^{-1} \bar{m} \sum_j L_{ij}) d$.

If we assume that the lateral connection strengths are function only of the relative distance $i - j$, then $L_{ij}$ becomes circular matrix so that $\sum_i L_{ij} = \sum_j L_{ij} = L_0$, and

$$c_i = (m_i + L_0 (1 - L_0)^{-1} \bar{m}) d. \tag{5}$$

In the mean field approximation, one can therefore write $c_i(\alpha) = (m_i - \alpha) d$, with $\alpha = |L_0|(1 + |L_0|)^{-1} \bar{m}$.

When analyzing the position and stability of the fixed points using this approximation, it follows under some mild assumption on the evolution of the average synaptic weights, that there is a mapping

$$m_i' \leftrightarrow m_i(\alpha) - \alpha$$

such that for every neuron in such a network with synaptic weight vector $m_i$ there is a corresponding neuron with weight vector $m_i'$ that undergoes the same evolution (around the fixed points) subject to a translation $\alpha$.

## 3 Extraction of Optimal Unsupervised Features

When a classification of high dimensional vectors is sought, the *curse of dimensionality* (Bellman, 1961) becomes the main factor affecting the classification performance. The curse of dimensionality is due to the inherent sparsity of high dimensional spaces; thus the amount of training data needed to get reasonably low variance estimators becomes ridiculously high. This has led many researchers in recent years to construct methods that specifically avoid this problem. In those cases in which important structure in the data actually lies in a much smaller dimensional space, it becomes reasonable to try to reduce the dimensionality before attempting the classification. This approach can be successful if the dimensionality reduction/feature extraction method loses as little information as possible in the transformation from the high dimensional space to the low dimensional one.

At a first glance, it seems that a supervised feature extraction method, such as multiple discriminant analysis (see review in Bryan, 1951; Sebestyen, 1962) will always be superior to an unsupervised one, because if one has more information about the problem, it is natural to expect that finding the solution is easier. However, due to the global constraint imposed by the supervision, when the number of parameters (i.e. the dimensionality and number of nodes) is large, the network often will get stuck in a local minimum which is far from an optimal solution. Unsupervised methods however, use local objective functions which may lead to less sensitivity to the number of parameters in the estimation, and therefore have the potential to avoid the curse of dimensionality (Barron and Barron, 1988).

For the purpose of pattern classification, it is important to devote our attention to those dimensionality reduction methods that allow discrimination between classes and not faithful representations of the data. This leaves out the class of methods such as factor analysis (see review in Harman, 1967) which tend to combine features that seem to have high correlation.

A general class of unsupervised dimensionality reduction methods, called exploratory projection pursuit is based on seeking *interesting* projections of high dimensional data points (Kruskal, 1969; Switzer, 1970; Kruskal, 1972; Friedman and Tukey, 1974; Friedman, 1987; Huber, 1985, for review). The notion of interesting projections is motivated by an observation made by Diaconis

and Freedman (1984), that for most high-dimensional clouds, most low-dimensional projections are approximately normal. This finding suggests that the important information in the data is conveyed in those directions whose single dimensional projected distribution is far from Gaussian. Various projection indices differ on the assumptions about the nature of deviation from normality, and in their computational efficiency. Friedman (1987) argues that the most computationally efficient measures are based on polynomial moments. However although many synaptic plasticity models are based on second order statistics and lead to extraction of the principal components (Sejnowski, 1977; von der Malsburg, 1973; Oja, 1982; Miller et al., 1989; Linsker, 1988), second order polynomials are not sufficient to characterize the important features of a distribution (see examples in Duda and Hart (1973) p. 212, and the example in Figure 2). This suggests that in order to use polynomials for measuring deviation from normality, higher order polynomials are required, and care should be taken in order to avoid their over-sensitivity to outliers. From our earlier discussion it follows that these polynomial moments should be of higher order than two. In some special cases where the data is known in advance to be bi-modal, it is relatively straightforward to define a good projection index (Hinton and Nowlan, 1990), however, when the structure is not known in advance, it is still valid to seek multi-modality in the projected data.

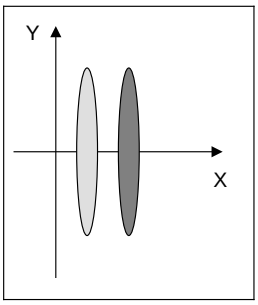


Figure 2: Two data clusters which can be separated by projecting to the $x$ axis, can not be separated by projecting to the $y$ axis, although the variance in the $y$ axis is larger.

Despite the computational attractiveness, projection indices based on polynomial moments are not directly applicable, since they very heavily emphasize departure from normality in the tails of the distribution (Huber, 1985). Friedman (1987) addresses this issue by introducing a nonlinear transformation that compresses the projected data from $R$ to $[-1, 1]$ using a normal distribution function. We address the problem by applying a sigmoidal function to the projections, and then applying an objective function based on polynomial moments.

# 4 Formulation of the BCM Theory Using an Objective Function

With the intuitive idea discussed above, we now present an objective function formulation of the synaptic modification theory of Bienenstock, Cooper and Munro (BCM). This yields a statistically plausible objective function whose minimization finds those projections having a single dimensional projected distribution that is far from Gaussian.

This formulation allows us to interpret the biological neuron's behavior from a statistical point of view. In addition, it provides a more powerful means of investigating the kinetics of synaptic development as well as the location and stability of the fixed points under various environmental

conditions.

### 4.1 Single Neuron

We first informally describe the statistical formulation that leads to this objective function. Using a metaphor motivated by statistical decision theory, a neuron is considered as capable of deciding whether to fire or not for a given input and vector of synaptic weights. A loss function is attached to each decision. The neuron's task is then to choose the decision that minimizes the loss. Since the loss function depends on the synaptic weight vector in addition to the input vector, it is natural to seek a synaptic weight vector that will minimize the sum of the losses associated with every input, or more precisely, the average loss (also called the risk). The search for such a vector, which yields an optimal synaptic weight vector under this formulation, can be viewed as learning or parameter estimation. In those cases where the risk is a smooth function, its minimization can be accomplished by gradient descent.

The ideas presented so far make no specific assumptions regarding the loss function, and it is clear that different loss functions will yield different learning procedures. For example, if the loss function is related to the inverse of the projection variance (including some normalization) then minimizing the risk will yield directions that maximize the variance of the projections, i.e. will find the principal components.

Before presenting a loss function, let us more precisely define the neuronal input, and two useful functions: We consider a neuron with input vector $x = (x_1, \ldots, x_n)$, synaptic weight vector $m = (m_1, \ldots, m_n)$, both in $R^n$, and activity (in the linear region) $c = x \cdot m$. The input $x$ is assumed to be a bounded, and piecewise constant stochastic process. We allow some time dependency in the presentation of the training patterns, by requiring that $x$ is of Type II mixing[1]. These assumptions are plausible, since they represent the closest continuous approximation to the usual training algorithms, in which training patterns are presented at random. They are needed for the approximation of the resulting deterministic gradient descent by a stochastic one (Intrator, 1990). For this reason we use a *learning rate* $\mu$ that has to decay in time so that this approximation is valid. Define the threshold $\Theta_M = E[(x \cdot m)^2]$, and the functions $\hat{\phi}(c, \Theta_M) = c^2 - \frac{1}{2}c\Theta_M$, $\phi(c, \Theta_M) = c^2 - c\Theta_M$.

Our projection index is aimed at finding directions for which the projected distribution is far from Gaussian; more specifically, since high dimensional clusters have a multimodal projected distribution, our aim is to find a projection index (loss function) that emphasizes multimodality. For computational efficiency, we would like to base the projection index on polynomial moments of low degree. Using second degree polynomials, one can get measures of the mean and variance of the distribution; these, however, do not give information on multimodality; therefore, higher order polynomials are necessary. Further, the projection index should exhibit the fact that bimodal distribution is already interesting, and any additional mode should make the distribution even more interesting.

With this in mind, consider the following family of loss functions that depend on the synaptic weight vector and on the input $x$

$$L_m(x) \quad = \quad -\mu \int_0^{(x \cdot m)} \hat{\phi}(s, \Theta_M) ds$$

[1] The mixing property specifies the dependency of the future of the process on its past.

$$= -\mu\{\frac{1}{3}(x \cdot m)^3 - \frac{1}{4}E[(x \cdot m)^2](x \cdot m)^2\} \quad (1)$$

The motivation for this loss function can be seen in Figure 3, which represents the $\phi$ function and the associated loss function $L_m(x)$. For simplicity the loss for a fixed threshold $\Theta_M$ and synaptic vector $m$ can be written as $L_m(c) = -\mu c^2(\frac{c}{3} - \frac{\Theta_M}{4})$, where $c$ represents the linear projection of $x$ onto $m$.

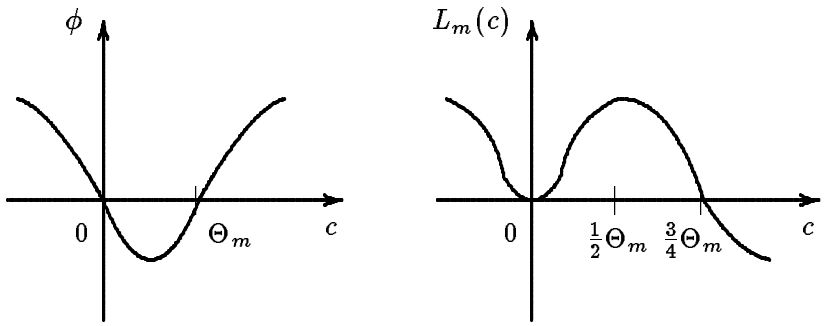


Figure 3: The function $\phi$ and the loss functions for a fixed $m$ and $\Theta_M$.

The graph of the loss function shows that for any fixed $m$ and $\Theta_m$, the loss is small for a given input $x$, when either $c = x \cdot m$ is close to zero, or when $x \cdot m$ is larger than $\Theta_m$. Moreover, the loss function remains negative for $(x \cdot m) > \Theta_M$, therefore any kind of distribution at the right hand side of $\Theta_M$ is possible, and the preferred ones are those which are concentrated further from $\Theta_M$.

It remains to show why it is not possible that a minimizer of the average loss will be such that all the mass of the distribution will be concentrated on one side of $\Theta_M$. This can not happen because the threshold $\Theta_M$ is dynamic and depends on the projections in a nonlinear way, namely, $\Theta_M = E(x \cdot m)^2$. This implies that $\Theta_M$ will always move itself to a position such that the distribution will never be concentrated at only one of its sides.

The risk (expected value of the loss) is given by:

$$\begin{aligned} R_m &= -\mu E\{\frac{1}{3}(x \cdot m)^3 - \frac{1}{4}E[(x \cdot m)^2](x \cdot m)^2\} \\ &= -\mu\{\frac{1}{3}E[(x \cdot m)^3] - \frac{1}{4}E^2[(x \cdot m)^2]\}. \end{aligned} \quad (2)$$

Since the risk is continuously differentiable, its minimization can be achieved via a gradient descent method with respect to $m$, namely:

$$\begin{aligned} \frac{dm_i}{dt} = -\frac{\partial}{\partial m_i}R_m &= \mu\ \{E[(x \cdot m)^2 x_i] - E[(x \cdot m)^2]E[(x \cdot m)x_i]\} \\ &= \mu\ E[\phi(x \cdot m, \Theta_M)x_i]. \end{aligned} \quad (3)$$

The resulting differential equations give a somewhat different version of the law governing synaptic weight modification of the BCM theory. The difference lies in the way the threshold $\Theta_M$ is determined. In the original form this threshold was $\Theta_M = E^p(c)$ for $p > 1$, while in the current form $\Theta_M = E(c^p)$ for $p > 1$. The latter takes into account the variance of the activity (for $p = 2$) and therefore is always positive, this ensures stability even when the average of the inputs is zero.

It should be noted here, that the original theory (1982) assumed that the inputs were positive, whereas the present threshold relaxes this assumption and yields stability for a larger class of bounded inputs.

Either form seems consistent with presently available experimental results (Clothiaux et al., 1991) but imply quite different underlying physiological mechanisms. The original BCM form requires that a history of activity (likely cell depolarization) be stored and then via a non-linear process produce the modification threshold. The present form of $\Theta_M$ requires that the non-linear process occur first. When the existence of the moving threshold is established by observation[2], the next question of great interest will be its precise dependence on cell parameters.

### 4.2 Extension to a Nonlinear Neuron

The fact that the distribution has part of its mass on both sides of $\Theta_M$ makes it a plausible projection index that seeks multi- modalities. However, this projection index will be more general if, in addition, the loss is insensitive to outliers and if we allow any projected distribution to be shifted so that the part of the distribution that satisfies $c < \Theta_M$ will have its mode at zero. The over-sensitivity to outliers is addressed by considering a nonlinear neuron in which the neuron's activity is defined to be $c = \sigma(x \cdot m)$, where $\sigma$ usually represents a smooth sigmoidal function. The ability to shift the projected distribution so that one of its modes is at zero is achieved by introducing a threshold $\beta$ so that the projection is defined to be $c = \sigma(x \cdot m + \beta)$. From the biological viewpoint, $\beta$ can be considered as spontaneous activity. The modification equations for finding the optimal threshold $\beta$ are easily obtained by observing that this threshold effectively adds one dimension to the input vector and vector of synaptic weights so that $x = (x_1 \ldots, x_n, 1)$, $m = (m_1, \ldots, m_n, \beta)$, and therefore, $\beta$ can be found by using the same synaptic modification equations. For the rest of the paper we shall assume that this threshold is added to the projection, without specifically writing it.

For the nonlinear neuron, $\Theta_M$ is defined to be $\Theta_M = E[\sigma^2(x \cdot m)]$. The loss function is given by:

$$\begin{aligned} L_m(x) &= -\mu \int_0^{\sigma(x \cdot m)} \hat{\phi}(s, \Theta_M) ds \\ &= -\mu \{\frac{1}{3}\sigma^3(x \cdot m) - \frac{1}{4}E[\sigma^2(x \cdot m)]\sigma^2(x \cdot m)\} \end{aligned} \tag{4}$$

The gradient of the risk becomes:

$$\begin{aligned} -\nabla_m R_m &= \mu \{E[\sigma^2(x \cdot m)\sigma' x] \\ &\quad -E[\sigma^2(x \cdot m)]E[\sigma(x \cdot m)\sigma' x]\} \\ &= \mu\, E[\phi\Big(\sigma(x \cdot m), \Theta_M\Big)\sigma' x], \end{aligned} \tag{5}$$

where $\sigma'$ represents the derivative of $\sigma$ at the point $(x \cdot m)$. Note that the multiplication by $\sigma'$ reduces sensitivity to outliers of the differential equation since for outliers $\sigma'$ is close to zero. The gradient decent procedure is valid, provided that the risk is bounded from below (see Section 5).

[2]Some indications of a moving threshold has been already found (Yang and Faber, 1991)

### 4.3 Extension to a Network with Feed-Forward Inhibition

We now define a network with feed-forward inhibition. The activity of neuron $k$ in the network is $c_k = x \cdot m_k$, where $m_k$ is the synaptic weight vector of neuron $k$. The *inhibited* activity and threshold of the $k$'th neuron is given by

$$\tilde{c}_k = c_k - \eta \sum_{j \neq k} c_j, \qquad \tilde{\Theta}_M^k = E[\tilde{c}_k^2]. \tag{6}$$

This feed-forward network should be contrasted with a lateral inhibition network (used for example by Cooper and Scofield, 1988) in which the inhibited activity is given by $c_k = c_k(0) + \sum L_{ij} c_j$. The relation between these two networks will be discussed in the next section.

For the feed-forward network the loss function is similar to the one defined in a single feature extraction with the exception that the activity $c = x \cdot m$ is replaced by $\tilde{c}$. Therefore the risk for node $k$ is given by:

$$R_k = -\mu\{\frac{1}{3}E[\tilde{c}_k^3] - \frac{1}{4}E^2[\tilde{c}_k^2]\}, \tag{7}$$

and the total risk is given by

$$R = \sum_{k=1}^{N} R_k. \tag{8}$$

To find the gradient of $R$ we write:

$$\begin{aligned}
\frac{\partial \tilde{c}_k}{\partial m_j} &= -\eta x, \quad \frac{\partial \tilde{c}_k}{\partial m_k} = x, \\
\frac{\partial R_k}{\partial m_k} &= \frac{\partial R_k}{\partial \tilde{c}_k}\frac{\partial \tilde{c}_k}{\partial m_k} = -\mu\{E[\tilde{c}_k^2 x] - E[\tilde{c}_k^2]E[\tilde{c}_k x]\}, \\
\frac{\partial R_j}{\partial m_k} &= \frac{\partial R_j}{\partial \tilde{c}_j}\frac{\partial \tilde{c}_j}{\partial m_k} = -\eta\frac{\partial R_j}{\partial m_j}, \\
\Rightarrow \quad \frac{\partial R}{\partial m_k} &= \frac{\partial R_k}{\partial m_k} - \eta\sum_{j\neq k}\frac{\partial R_j}{\partial m_j} \\
&= \mu\,[E[\phi(\tilde{c}_k, \tilde{\Theta}_M^k)x] - \eta\sum_{j\neq k} E[\phi(\tilde{c}_j, \tilde{\Theta}_m^j)x]].
\end{aligned} \tag{9}$$

The equation performs a constraint minimization in which the derivative with regard to one neuron can become orthogonal (when $\eta \to 1$) to the sum over the derivatives of all other synaptic weights. Nevertheless, the coupling is very simple to calculate, and does not require any matrix inversion. The equation therefore, demonstrates the ability of the network to perform exploratory projection pursuit in parallel, since the minimization of the risk involves minimization of nodes $1, \ldots, N$, which are loosely coupled.

When the nonlinearity of the neuron is included, the inhibited activity is defined (as in the single neuron case) as $\tilde{c}_k = \sigma(c_k - \eta\sum_{l\neq k} c_l)$. $\tilde{\Theta}_M^k$, and $R_k$ are defined as before. However, in this case

$$\frac{\partial \tilde{c}_k}{\partial m_j} = -\eta\sigma'(\tilde{c}_k)x, \quad \frac{\partial \tilde{c}_k}{\partial m_k} = \sigma'(\tilde{c}_k)x. \tag{10}$$

Therefore the total gradient becomes:

$$\dot{m}_k = \frac{\partial R}{\partial m_k} = \mu\{E[\phi(\tilde{c}_k, \tilde{\Theta}_M^k)\sigma'(\tilde{c}_k)x] - \eta \sum_{j \neq k} E[\phi(\tilde{c}_j, \tilde{\Theta}_m^j)\sigma'(\tilde{c}_j)x]\}. \tag{11}$$

The lateral inhibition network performs a search of k-dimensional projections together; thus may find a richer structure that a stepwise approach might miss (e.g. see example 14.1, Huber, 1985).

### 4.4 Some Related Statistical and Computational Issues

The proposed method uses low order polynomial moments which are computational efficient, yet it does not suffer from the main draw back of polynomial moments – sensitivity to outliers. It naturally extends to multi-dimensional projection pursuit using the feed-forward inhibition network. The number of calculations of the gradient grows linearly with the dimensionality and *linearly* with the number of projections sought. The projection index contains a single dimensional scaling (see the contribution of Hastie and Tibshirani to the discussion in Jones and Sibson, 1987), therefore, removing the need for a sphering transformation to the data, however, a sphering transformation will result in a type III projection index (see Huber, 1985). The projection index has a natural stochastic gradient descent version which further accelerates the calculation by eliminating the need to calculate the empirical expected value of the gradient. All the above lead to a fully parallel algorithm that may be implemented on a multi-processor machine, and produce a practical feature extractor for very high dimensional problems.

Although, the projection index is motivated by the desire to search for clusters in the high dimensional data, the resulting feature extraction method is quite different from other pattern recognition methods that search for clusters. Since the class labels are not used in the search, the projection pursuit is not biased to the class labels. This is in contrast with classical methods such as discriminant analysis (Fisher, 1936; Sebestyen, 1962, and numerous recent publications). The issue of using an unsupervised method vs. supervised for revealing structure in the data has been discussed extensively elsewhere. We would only like to add that it is striking that in various low-dimensional examples (Friedman and Tukey, 1974; Jones, 1983; Friedman, 1987) the exploratory capabilities of PP were not worse than those of supervised method such as discriminant analysis and factor analysis in discovering structure, thus suggesting that in high dimensions where supervised methods may fail, still PP can find useful structure.

The resulting method concentrates on projections that allow discrimination between clusters and not faithful representation of the data, which is in contrast to principal components analysis, or factor analysis which tend to combine features that have high correlation (see review in Harman, 1967).

The method differs from cluster analysis by the fact that it searches for clusters in the low dimensional projection space, thus avoiding the inherent sparsity of the high dimensional space. The search for multi-modality is further constrained by the desire to seek those projections that are orthogonal to all but one of the clusters (or have a mode at zero). This constraint simplifies the search, since it implies that a set of $K$ linearly independent clusters may have at most $K$ optimal projections as opposed to at most $\binom{K}{2}$ separating hyperplanes.

### 4.5 Comparison of Linear Feed-Forward with Lateral Inhibition Network: Mean Field Approximation

For the linear case, using the notation of Cooper and Scofield, (1988), neuron activity in the lateral inhibition network is given by

$$c = Md + Lc. \tag{12}$$

In the mean field approximation this becomes

$$c = Md + L\bar{c}, \tag{13}$$

where $M$ is the synaptic matrix for $N$ neurons, $c = (c_1, \ldots, c_N)^T$, $L$ is the inhibitory connection matrix with norm less than 1 and $\bar{c}$ is the averaged activity over all neurons in the network. In the context of visual cortex, the first term is due to the input from LGN and the second due to input from other cortical cells. If we define $c(0) = Md$, then the averaged inhibited activity can be written as

$$\bar{c} = \bar{c}(0) + L\bar{c}, \tag{14}$$

or

$$\bar{c} = (I - L)^{-1}\bar{c}(0) = (I + L + L^2 + L^3 + \ldots)\,\bar{c}(0). \tag{15}$$

Using this notation, the activity of a neuron in the feed-forward network as defined in section 4.3 can be written

$$\tilde{c} = c(1) = c(0) + Lc(0), \tag{16}$$

which leads to an averaged activity of the form

$$\bar{c}(1) = (I + L)\bar{c}(0), \tag{17}$$

which is a a first order approximation of (14); it is useful primarily because it removes the need to invert a matrix, which becomes impossible in the nonlinear neuronal case. In addition, successive approximations

$$\begin{aligned} \bar{c}(1) = \bar{c}(0) + L\bar{c}(0) &= (I + L)\bar{c}(0), \\ \bar{c}(2) = \bar{c}(0) + L\bar{c}(1) &= (I + L + L^2)\bar{c}(0), \\ &\vdots \end{aligned} \tag{18}$$

can be thought of as including mono-synaptic, bi-synaptic, tri-synaptic etc. events and thus follow the time course of the post-synaptic potentiation. It follows that $\bar{c}(k) \to \bar{c}$, as $k \to \infty$, thus recapturing the lateral inhibition network. Within a scaling factor, the first order feed-forward network, as will be shown below, generates the same synaptic modification equations as the lateral inhibition network in the Cooper and Scofield mean field approximation.

For a feed-forward network with neuron activity given by $\tilde{c}_i = m_i \cdot d + \sum_j L_{ij}c_j$, as in section 4,

$$\dot{m}_k = -\frac{\partial R}{\partial m_k} = -[\frac{\partial R_k}{\partial m_k} + \sum_j L_{kj}\frac{\partial R_j}{\partial m_j}]$$

$$= \mu[E[\phi(\tilde{c}_k, \tilde{\Theta}^k_M)x] + \sum_j L_{kj} E[\phi(\tilde{c}_j, \tilde{\Theta}^j_m)x]]. \quad (19)$$

Let $\bar{m} = \frac{1}{N}\sum_j m_j$ ($m_j \in R^n$). We assume that the inhibitory contributions are a function only of the $i-j$ (not dependent on the absolute position of a cell in the network), so that $\sum_i L_{ij} = L_0 = \sum_j L_{ij}$, and that $\sum_i L_{ij} E[\phi(\tilde{c}_i, \tilde{\Theta}^i_m)x] = \sum_j L_{ij} E[\phi(\tilde{c}_j, \tilde{\Theta}^j_m)x]$. Then we get:

$$\begin{aligned} N\dot{\bar{m}} = \sum_j \dot{m}_j &= \mu[\sum_k E[\phi(\tilde{c}_k, \tilde{\Theta}^k_M)x] + \sum_j L_0 E[\phi(\tilde{c}_j, \tilde{\Theta}^j_m)x]] \\ &= \frac{(1+L_0)}{L_0}\mu \sum_k \sum_j L_{kj} E[\phi(\tilde{c}_k, \tilde{\Theta}^k_M)x]. \end{aligned} \quad (20)$$

This implies that

$$\frac{L_0}{1+L_0}\dot{\bar{m}} = \mu \sum_k L_{jk} E[\phi(\tilde{c}_k, \tilde{\Theta}^k_M)x, \quad (21)$$

and hence,

$$\dot{m}_k = \mu[E[\phi(\tilde{c}_k, \tilde{\Theta}^k_M)x] + \frac{L_0}{1+L_0}\dot{\bar{m}}]. \quad (22)$$

Compare this with Cooper and Scofield, (1988) (eq. A3):

$$\dot{m}_k = \mu[E[\phi(c_k, \Theta^k_m)x] + L_0\dot{\bar{m}}], \quad (23)$$

equations 22 and 23 differ only in the constant of inhibition. Thus the mean field approximation of the feed-forward network yields the lateral inhibition mean field result merely by scaling the average inhibition.

One result of the mean field approximation (Cooper and Scofield, 1988) is that there is a transformation such that

$$m(\alpha) = m' - \alpha \quad (24)$$

and so the gradient with respect to the weights yields two terms $\dot{m}$ and $\dot{\alpha}$. In the adiabatic case, we assume that $\alpha$ varies slowly with respect to each individual $m$, so that $\dot{\alpha} = 0$ is a reasonable approximation. In this situation the analysis of section 5 applies for $m(\alpha)$ (the mean field network) as well as for $m(0)$. In addition, the argument given in the appendix of (Cooper and Scofield, 1988) regarding the non adiabatic case holds here as well.

From the system 9 we can get:

$$\dot{\bar{m}} = \frac{\mu}{N} E[\sum_k [\phi(\tilde{c}_k, \tilde{\Theta}^k_M)x] - \eta(\frac{N}{N-1})\sum_j E[\phi(\tilde{c}_j, \tilde{\Theta}^j_m)x]], \quad (25)$$

which implies

$$\dot{\alpha} = \dot{\bar{m}} = \frac{\mu}{N}[1 - \eta(\frac{N}{N-1})]E[\sum_k [\phi(\tilde{c}_k, \tilde{\Theta}^k_M)x]]. \quad (26)$$

Therefore, at a fixed point, when all of the cells of the network have reached their respective fixed points, $\dot{m}'_i = 0$ implies that $\dot{\bar{m}} = 0$, meaning $\dot{\alpha} = 0$. Thus the position and stability of the fixed points (as given in section 5 ) apply for the mean field network with no additional approximations.

# 5 Analysis of the Fixed Points in High Dimensional Space

In appendix A we show using a general result on random differential equations (Intrator, 1990) that the solution of the random differential equations remains as close as we like, in the $L^2$ sense, to the solution of the deterministic equations. We have shown in section 4 that the deterministic equation converges to a local minimum of the risk. This implies that the solution of the random differential equation converges to a local minimum of the risk in $L^2$. Based on the statistical formulation, we can say that the local minima of the risk are *interesting features* extracted from the data, which correspond to directions in which the single dimensional distribution of the projections is far from a Gaussian distribution, by means of penalized skewness measure.

In the following, we attempt to analyze the shape of the high dimensional risk function, under specific inputs, namely, we look for the location of the critical points of the risk, and locate those which have a local minima given a specific training set. This completely characterizes the solution of the synaptic modification equations, and sheds some more light on the power of the risk functional in finding interesting directions in the data. In doing so we gain some detailed information on the behavior of the solution of the random differential equations, as a model for learning in visual cortex, under various rearing conditions.

We consider linear neurons under the mean field assumptions. Furthermore, since the introduction of the threshold $\beta$ does not pose any mathematical difficulty as was described before, we omit it in the analysis.

First, we analyze the limiting behavior of the solution in the case where we have $n$ linearly independent inputs (not necessarily orthogonal). The introduction of noise into the system will be done in the next sections.

## 5.1 $n$ linearly independent inputs

The random differential equation is given by

$$\dot{m}_\epsilon = \epsilon\ \mu(t)\phi(x \cdot m, \Theta_m)x, \quad m_\epsilon(0) = m_0, \tag{1}$$

the averaged (batch) version of the gradient descent is given by:

$$\dot{\bar{m}}_\epsilon = \epsilon\ \mu(t)E[\phi(x \cdot \bar{m}, \Theta_{\bar{m}})x] \quad \bar{m}_\epsilon(0) = m_0. \tag{2}$$

The main tool in establishing the following results is the connection between the solution to the deterministic differential equation 2 and the solution of the random differential equation 1. A general result which yields this connection is given in (Intrator, 1990) and will be discussed in the appendix. When applied to this specific differential equation, the result says that

$$\sup_{t>T} E|m_\epsilon - \bar{m}_\epsilon|^2 \underset{\epsilon \to 0}{\longrightarrow} 0. \tag{3}$$

**Proposition 5.1** *Let $x^{(1)}, \ldots, x^{(n)}$ be $n$ linearly independent bounded vectors in $R^n$. Let $D$ be the random process so that $P[D = x^{(i)}] = p_i, \quad p_i > 0,\ i = 1, \ldots, n,\ \sum p_i = 1$.*

*Then the critical points of equation 1 are the $2^n$ weight vectors $m^{(i)} \in R^n$, each a solution to one of the equations: $Am^{(i)} = v^{(i)}$, $i = 0, \ldots, 2^n - 1$, where $A$ is the matrix whose i'th row is the*

*input vector* $x^{(i)}$, *and* $\{v^{(i)}, \quad i = 0, \ldots, 2^n - 1\}$, *is the* $n$ *dimensional set of vectors of the form:*

$$\begin{aligned} v^{(0)} &= (0, \ldots, 0), \\ v^{(1)} &= (\frac{1}{p_1}, 0, \ldots, 0), \\ v^{(2)} &= (0, \frac{1}{p_2}, 0, \ldots, 0), \\ v^{(3)} &= (\frac{1}{p_1 + p_2}, \frac{1}{p_1 + p_2}, 0, \ldots, 0), \\ v^{(4)} &= (0, 0, \frac{1}{p_3}, 0, \ldots, 0), \\ &\ldots, \\ v^{(2^n - 1)} &= (1, \ldots, 1). \end{aligned}$$

*Proof* Rewrite equation 1 in the form

$$\dot{m}_\epsilon = \epsilon \mu(t)(x \cdot m_\epsilon)[x \cdot m_\epsilon - \Theta_M]x, \tag{4}$$

where $\Theta_M = E[(x \cdot m_\epsilon)^2]$. Since $\epsilon\mu(t) > 0$, then $m_\epsilon$ is a critical point if either $x^{(i)} \cdot m_\epsilon = 0$, or $x^{(i)} \cdot m_\epsilon = \Theta_M \neq 0, \ \ i = 1, \ldots, n$.

There are exactly $2^n$ possibilities for the set of $n$ numbers $x^{(i)} \cdot m_\epsilon$ to be either 0, or nonzero. Therefore there are only $2^n$ possible solutions.

Let $m_\epsilon^{(1)}$ be such that $m_\epsilon^{(1)} \cdot x^{(1)} \neq 0$, and $m_\epsilon^{(1)} \cdot x^{(i)} = 0, \ i > 1$. Then for $m_\epsilon^{(1)}$,

$$\Theta_M = E[(x \cdot m_\epsilon)^2] = \sum_{i=1}^{N} p_i (x^{(i)} \cdot m_\epsilon^{(1)})^2 = p_1 (x^{(1)} \cdot m_\epsilon^{(1)})^2. \tag{5}$$

When we combine the condition $\Theta_M = x^{(1)} \cdot m_\epsilon^{(1)}$, we get $x^{(1)} \cdot m_\epsilon^{(1)} = \frac{1}{p_1}$.

Now suppose $m = m_\epsilon^{(3)}$, is such that $x^{(1)} \cdot m_\epsilon^{(3)} = x^2 \cdot m_\epsilon^{(3)} = \Theta_M$, and $x^{(i)} \cdot m_\epsilon^{(3)} = 0, \quad i > 2$. In this case $\Theta_M = p_1 (x^{(1)} \cdot m_\epsilon^{(3)})^2 + p_2 (x^2 \cdot m_\epsilon^{(3)})^2$, which yields, $x^j \cdot m_\epsilon^{(3)} = \frac{1}{p_1 + p_2}, \quad j = 1, 2$.

The other cases are treated similarly. ♢

### 5.1.1 Stability of the solution

Let $m(t)$ be a solution of a random differential equation, then $m_0$ is said to be a stable point if for any $\delta > 0$ there is a $\tau(\delta)$ such that for any $K > 0$, $t \geq \tau$

$$P\{|m(t) - m_0|^2 > K\} \leq \frac{\delta}{K}. \tag{6}$$

This roughly says that the if $m_0$ is a stable point, then the probability of finding the solution *far* from this point is *small*.

**Lemma 5.1** *Let* $m$ *be a critical point for the random differential equation. Then* $m$ *is a stable (unstable) critical point, if it is a stable (unstable) critical point for the averaged deterministic version. The stability of the stochastic equation is in the* $L^2$ *sense.*

*Proof* From equation 3 follows that if $m_\epsilon$ is a critical point of the random version, then it is a critical point of the averaged deterministic equation. If this point is a stable (unstable) point of the deterministic equation, then perturbing both equations (i.e. starting from an initial condition that is close to $m_\epsilon$), will yield that the deterministic equation will converge back to (diverge from) the original critical point. This is independent of $\epsilon$ and with probability one since it is a deterministic equation. Consequently, the random solution must stay close to the deterministic solution which in this case, implies the stability (instability) of the random solution. $\Diamond$

**Theorem 5.1** *Under the conditions of proposition 5.1, the critical points $m_\epsilon^{(i)}$ that are stable are only those in which the corresponding vector $v^i$, has one and only one nonzero element in it.*

*Proof* From the above two lemmas it follows that it is enough to check the stability on the deterministic version of the equations, at the critical points of the random version.

The gradient is then given by:

$$-\nabla_m R_m = E[(x \cdot m)^2 x] - E[(x \cdot m)^2]E[(x \cdot m)x], \tag{7}$$

and the second order derivative is given by:

$$-\nabla_m^2 R_m = 2E[(x \cdot m)x \times x] - E[(x \cdot m)^2]E[x \times x] - 2E[(x \cdot m)x] \times E[(x \cdot m)x]. \tag{8}$$

The critical point $m = 0$ is clearly unstable, since the second derivative matrix is zero, and changes sign around $m = 0$. For selective solution, we can choose without loss of generality, $m^{(1)}$ which is the solution for $v^{(1)}$. Putting $m^{(1)}$ into the gradient equation gives:

$$\begin{aligned} -\nabla_m R|_{m=m^{(1)}} &= p_1(x^{(1)} \cdot m^{(1)})^2 x^{(1)} - p_1(x^{(1)} \cdot m^{(1)})^2 p_1(x^{(1)} \cdot m^{(1)})x^{(1)} \\ &= p_1(x^{(1)} \cdot m^{(1)})^2[1 - p_1](x^{(1)} \cdot m^{(1)})x^{(1)}. \end{aligned} \tag{9}$$

Since $m^{(1)}$ is a critical point and $x^{(1)}$ is the preferred input, we get from the fact that the gradient is equal to zero at $m^{(1)}$: $(x^{(1)} \cdot m^{(1)}) = \frac{1}{p_1}$, $E(x \cdot m)^2 = \frac{1}{p_1}$.

Define the matrix $B$ to be

$$B = E[x \times x] = \sum_{i=1}^{N} p_i x^{(i)} \times x^{(i)}, \tag{10}$$

since the inputs are independent and span the whole space, it follows that $B$ is positive definite. Putting $(x^{(1)} \cdot m^{(1)})$ into 8 gives:

$$-\nabla_m^2 R|_{m=m^{(1)}} = p_1(x^{(1)} \cdot m^{(1)})\Big(2x^{(1)} \times x^{(1)} - \frac{1}{p_1}B - 2x^{(1)} \times x^{(1)}\Big), \tag{11}$$

which is negative definite, thus leading to a stable critical point.

Now, assume, without loss of generality, that $m = m^{(3)}$, then

$$\begin{aligned} -\nabla_m R|_{m=m^{(3)}} &= [p_1(x^{(1)} \cdot m^{(3)})^2 x^{(1)} + p_2(x^{(2)} \cdot m^{(3)})^2 x^{(2)}] \\ &\quad -[p_1(x^{(1)} \cdot m^{(3)})^2 + p_2(x^{(2)} \cdot m^{(3)})^2] \\ &\quad [p_1(x^{(1)} \cdot m^{(3)})x^{(1)} + p_2(x^{(2)} \cdot m^{(3)})x^{(2)}]. \end{aligned} \tag{12}$$

Since $m^{(3)}$ is a critical point, we have from proposition 5.1 that $(x^{(1)} \cdot m^{(3)}) = (x^{(2)} \cdot m^{(3)}) = \frac{1}{p_1+p_2}$, and $E(x \cdot m)^2 = \frac{1}{p_1+p_2}$.

Putting this into equation 8 gives:

$$\begin{aligned}
-\nabla_m^{(3)} R_{|_{m=m^{(3)}}} &= 2p_1(x^{(1)} \cdot m^{(3)})x^{(1)} \times x^{(1)} + 2p_2(x^{(2)} \cdot m^{(3)})x^{(2)} \times x^{(2)} \\
&\quad -\frac{1}{p_1+p_2}B \\
&\quad -2\Big([p_1(x^{(1)} \cdot m^{(3)})x^{(1)} + p_2(x^{(2)} \cdot m^{(3)})x^{(2)}] \\
&\quad \times [p_1(x^{(1)} \cdot m^{(3)})x^{(1)} + p_2(x^{(2)} \cdot m^{(3)})x^{(2)}]\Big) \\
&= \Big(\frac{2p_1}{p_1+p_2} - \frac{2p_1^2}{(p_1+p_2)^2}\Big)x^{(1)} \times x^{(1)} \\
&\quad + \Big(\frac{2p_2}{p_1+p_2} - \frac{2p_2^2}{(p_1+p_2)^2}\Big)x^{(2)} \times x^{(2)} \\
&\quad -\frac{1}{p_1+p_2}B \\
&\quad -\frac{2p_1p_2}{(p_1+p_2)^2}(x^{(1)} \times x^{(2)} + x^{(2)} \times x^{(1)}).
\end{aligned} \tag{13}$$

Denote the above gradient matrix by $G$. Without loss of generality we may assume that $p_1 \geq p_2$. Then consider a vector, $y$ which is orthogonal to all but $x^{(2)}$. Then

$$y^T G y = y^T x^{(2)} \times x^{(2)} y \frac{p_2}{p_1+p_2}\Big(6 - \frac{6p_2}{p_1+p_2} - 3\Big) \geq 0, \tag{14}$$

since $\frac{p_2}{p_1+p_2} \leq \frac{1}{2}$. It is easy to see, by replacing $m^{(3)}$ with $\lambda m^{(3)}$, that the second derivative along $m^{(3)}$ changes sign at $\lambda = 1$, which implies instability.

The proof for the other critical points follows in the exactly same way. ♢

## 5.2 Noise with no Patterned Input

This is a special case, which is related to the binocular deprivation environment discussed in 2.1, and hence is analyzed separately. In general, we consider input as being composed of pattern and noise. The patterned input represents a highly correlated set of patterns that appear at random, and are supposed to mimic key features in visual environment such as edges with different orientation etc. The noise in an uncorrelated type of input, which is assumed to exist in large network of neurons receiving inputs from several parts of cortex. Patterned input is associated with open eyes, pure noise with closed eyes.

When the input contains only noise, the averaged deterministic solution has a stable critical point, and the random solution stays close to the deterministic one as is shown in the appendix. When the input is composed of noise with zero mean only, we find that the averaged version has a stable zero solution (as opposed to the case with patterned input). This implies that the solution of the random version wanders about the origin but stays close to zero in $L^2$ norm.

### 5.2.1 Noise with Zero Mean

The crucial property of white noise $x$ is the fact that it is symmetric around zero, this implies that $E(x \cdot m)^3 = 0$, and the risk,

$$R_m = -\{E[(x \cdot m)^3] - E^2[(x \cdot m)^2]\} = E^2[(x \cdot m)^2] \geq 0. \tag{15}$$

It is easy to see that only for $m = 0$, $R_m = 0$, and this is the only critical point in this case. Since this result is related to binocular deprivation experiments, it should be emphasized again that the solution to the stochastic version of the differential equations will wander around zero in a random manner but with a small magnitude that is controlled by the learning rate $\mu$.

In view of the properties of the risk, we can say that when the distribution of $x$ has zero skewness in every direction, the only stable minima of the risk is $m = 0$. This is not true when the noise has a positive or a negative average as analyzed in the next section.

### 5.2.2 Noise with Positive Mean

We assume that $x$ is now bounded random noise, with $\bar{x} > 0$, and $x$ has the same single dimensional distribution in all directions, which implies that $\bar{x}_i = \bar{x}_1 > 0, \quad i \geq 0$. Let $x = \bar{x} + y$, where $y$ is random noise with zero average. Denote $\mathrm{Var}(y_1) = \lambda$. The following identities can easily be verified:

$$\begin{aligned} E(x \cdot m)^2 &= (\bar{x} \cdot m)^2 + \mathrm{Var}(y \cdot m), \\ E(y \cdot m)y &= \lambda m, \\ E(y \cdot m)^2 y &= 0. \end{aligned} \tag{16}$$

Putting these identities in the first and second gradient (eq. 7,8) we get:

$$-\nabla_m R_m = [(\bar{x} \cdot m)^2 + \mathrm{Var}(y \cdot m)]\bar{x} - [(\bar{x} \cdot m)^2 + \mathrm{Var}(y \cdot m)][(\bar{x} \cdot m)\bar{x} + \lambda m]. \tag{17}$$

We are looking for critical points of the gradient,

$$\nabla_m R_m = 0 \Rightarrow \quad m_i = [\frac{1}{\lambda} - \frac{(\bar{x} \cdot m)}{\lambda}]\bar{x}_i. \tag{18}$$

Equation 18 suggests a consistency condition that has to be filled, namely, if we multiply both sides of this equation by $\bar{x}_i$ and sum over all $i's$ we get:

$$(\bar{x} \cdot m) = [\frac{1}{\lambda} - \frac{(\bar{x} \cdot m)}{\lambda}] \parallel \bar{x} \parallel^2, \tag{19}$$

therefore,

$$(\bar{x} \cdot m) = \frac{\parallel \bar{x} \parallel^2}{\lambda + \parallel \bar{x} \parallel^2}. \tag{20}$$

When substituting 20 into 18 we get the explicit ratio between $m_i$ and $\bar{x}_i$, namely,

$$m_i = [\frac{1}{\lambda + \parallel \bar{x} \parallel^2}]\bar{x}_i. \tag{21}$$

The second derivative is given by:

$$\begin{aligned} -\nabla_m^2 R_m &= [2(\bar{x}\cdot m)^2(\bar{x}\cdot m)^2 - \mathrm{Var}(y\cdot m)](\bar{x}\times\bar{x}) \\ &+ E[\{2(y\cdot m) - 2(y\cdot m)(\bar{x}\cdot m)\}(y\times\bar{x})] \\ &- 2\lambda(\bar{x}\cdot m)(\bar{x}\times m) \\ &- 2\lambda(y\cdot m)(y\times m) \\ &- \lambda[(\bar{x}\cdot m)^2 + \mathrm{Var}(y\cdot m)]I. \end{aligned} \tag{22}$$

Using relations 16 and 20 of the critical points, we get to a gradient in terms of $\bar{x}\times\bar{x}$, and $\lambda$ – the variance of the noise. Let $\tau = \frac{1}{\lambda+\|\bar{x}\|^2}$, then

$$\begin{aligned} -\nabla_m^2 R_m &= \tau^2(\bar{x}\times\bar{x})[(2-4) \\ &\quad -\lambda \parallel \bar{x} \parallel^2 + 2\lambda^2\tau \\ &\quad -2\lambda^2 \parallel \bar{x} \parallel^2 -2\lambda^2]. \end{aligned} \tag{23}$$

It follows that the gradient is positive definite for any noise with variance $\lambda > 0$. This implies stability of the averaged version, and stability in the $L^2$ sense of the random version.

### 5.3 Patterned Input with Noise

We now explore the change in the position of critical points under small noise. The result relies on the smoothness of the projection index, and on the fact that noise can be presented as a small perturbation.

Let the input $x = d + h$, where $d$ is the patterned input and $h$ is a small random noise with zero mean. If the mean of the noise is non-zero it can always be obsorbed in the patterned input and the resulting noise will have a zero mean. Let $\lambda = \mathrm{Var}(h\cdot m)$ which is small as well. Consider the projection index

$$\begin{aligned} R_m &= -\mu\{\frac{1}{3}E[(x\cdot m)^3] - \frac{1}{4}E^2[(x\cdot m)^2]\} \\ &= -\mu\{\frac{1}{3}E[(d\cdot m)^3] - \frac{1}{4}E^2[(d\cdot m)^2] \\ &\quad +\lambda\{E(d\cdot m) - \frac{\lambda}{4}[1 + E(d\cdot m)^2]\}\}. \end{aligned} \tag{24}$$

Thus $R_m(d+h) = R_m(d) + O(\lambda)$, yielding robustness to small noise.

## 6 Application to Various Rearing Conditions

In the following section, we relate the analysis described above to some visual cortical plasticity experiments. Extensive simulation, using the complete set of known experimental results on visual cortical plasticity, have shown that the the modified version of $\Theta_M$ is consistent with the current experimental results.

### 6.1 Normal Rearing(NR)

This case has been covered by the theorem 5.1 from which it follows that a neuron will become selective to one of the inputs. Note that it also follows that the synaptic weights of both eyes become selective to the same orientation.

### 6.2 Monocular Deprivation (MD)

From theorem 5.1 we can get an explicit expression to $\Theta_M$ in the case of $n$ linearly independent inputs. Recall that the only stable points in such case are those in which the synaptic weight $m$ is orthogonal to all but one of the inputs. Assuming that all the $K$ inputs have the same probability $\frac{1}{K}$, we get: $\Theta_M = E(x \cdot m)^2 = \frac{1}{K}\sum_{i=1}^{K}(x_i \cdot m)^2 = \frac{1}{K}(x_{i_0} \cdot m)^2$ where $x_{i_0}$ is the input which is not orthogonal to $m$. Putting that into the deterministic version of the gradient descent it follows immediately that $x_{i_0} \cdot m = K$, which implies that $\Theta_M = \frac{1}{K}(x_{i_0} \cdot m)^2 = K$, and $E(x \cdot m) = \frac{1}{K}(x_{i_0} \cdot m) = 1$. This result will be used in the following MD analysis.

The assumptions in the monocular deprivation case are that the input to the left (right) eye is composed of noise only, namely $d^r$ represents patterned input plus noise, and $d^l = n$. We also assume that the noise has zero average and has a symmetric distribution uniform in all directions. We relax the assumption that $d^r$ has zero mean, and instead assume that $E(d^r \cdot m^r) < \frac{1}{2}E(d^r \cdot m^r)^2$, this is easily achieved when the dimensionality is larger than 2 (following from the calculation at the begining of this section). We have:

$$\begin{aligned}
R &= -\{\frac{1}{3}E(x \cdot m)^3 - \frac{1}{4}E^2(x \cdot m)^2\} \\
&= -\{\frac{1}{3}E[(d^r \cdot m^r + n \cdot m^l)^3] - \frac{1}{4}E^2[(d^r \cdot m^r + n \cdot m^l)^2]\} \\
&= -\{\frac{1}{3}E[(d^r \cdot m^r)^3] + \frac{1}{3}E[(n \cdot m^l)^3] + E[(d^r \cdot m^r)^2(n \cdot m^l)] + E[(d^r \cdot m^r)(n \cdot m^l)^2] \\
&\quad -\frac{1}{4}[E^2[(d^r \cdot m^r)^2] + E^2[(n \cdot m^l)^2] + 2E[(d^r \cdot m^r)^2]E[(n \cdot m^l)^2] \\
&\quad +4E[d^r \cdot m^r]E[n \cdot m^l]\left(E[(d^r \cdot m^r)^2] + E[(n \cdot m^l)^2] + E[d^r \cdot m^r]E[n \cdot m^l]\right)]\} \\
&= -\{\frac{1}{3}E[(d^r \cdot m^r)^3] - \frac{1}{4}E^2[(d^r \cdot m^r)^2]\} \\
&\quad +\frac{1}{4}\mathrm{Var}(n \cdot m^l)\left(\mathrm{Var}(n \cdot m^l) + 2E[(d^r \cdot m^r)^2] - 4E[d^r \cdot m^r]\right). \qquad (1)
\end{aligned}$$

The first term of the risk is due to the open eye and is therefore minimized when the neuron becomes selective as in the regular normal rearing case. The second term is non negative due to the previous assumption, and therefore can be minimized only if $m^l = 0$. Note that in a mean field, this means that $m^l \to \alpha$. It can also be seen that when the right eye becomes selective (implying that the term $2E[(d^r \cdot m^r)^2] - 4E[d^r \cdot m^r]$ becomes larger), then the driving force for $\mathrm{Var}(n \cdot m^l)$ to go to zero becomes larger. This is consistent with the experimental observation which suggests that the synapses of the closed eye do not go down until the open eye becomes selective.

### 6.3 Binocular Deprivation (BD)

This case has been analysed in section 5.2. During BD we assume that the input is noise; the conclusion was that either synaptic weights perform a random walk around zero, or in case of positive average noise, a random walk about a positive weight that is a function of the average of the noise and its variance.

### 6.4 Reversed Suture (RS)

The limiting behavior of RS is similar to that of MD, described above. Computer simulations show that it is possible to achieve a disconnection of the newly closed eye before the newly open eye

becomes selective (Clothiaux et al., 1991).

### 6.5 Strabismus

From theorem 5.1 we infer that a stable fixed point is such that its projection to one of the inputs is positive, and it is orthogonal to all the other inputs. Under strabismus we assume that the input to both eyes is uncorrelated, therefore this situation is possible only if the vector of synaptic weights of one eye is orthogonal to all but one of the inputs; thus the vector of synaptic weights of the other eye is orthogonal to all the inputs. Since the inputs span the whole space this vector must be zero.

## 7 Discussion

We have presented an objective function formulation of the BCM theory of visual cortical plasticity. This permits us to demonstrate the connection between the unsupervised BCM learning procedure and the statistical method of projection pursuit and provides a general method for stability analysis of the fixed points. Relating this unsupervised learning to statistical theory enables comparison with various other statistical and unsupervised methods for feature extraction.

Analysis of the behavior and the evolution of the network under various visual rearing conditions is in agreement with experimental results. We thus have the result that a biological neuron may be performing a sophisticated statistical procedure. An experimental question of great interest is posed: how does the modification threshold depend on the average activity of the cell $\Theta_M \simeq \bar{c}^2$ as in the original BCM versus $\Theta_M \simeq \bar{c^2}$ as presented here.

### Acknowledgements

We wish to thank Geoff Hinton for improving the clarity of the statistical part. Charles Bachmann, Eugene Clothiaux and Mike Perrone provided many helpful comments.

Research was supported by the National Science Foundation, the Army Research Office, and the Office of Naval Research.

## References

Barron, A. R. and Barron, R. L. (1988). Statistical learning networks: A unifying view. In Wegman, E., editor, *Computing Science and Statistics: Proc. 20th Symp. Interface*, pages 192–203. American Statistical Association, Washington, DC.

Bear, M. F. and Cooper, L. N. (1990). Molecular mechanisms for synaptic modification in the visual cortex: Interaction between theory and experiment. In Gluck, M. and Rumelhart, D., editors, *Neuroscience and Connectionist Theory*, pages 65–94. Lawrence Erlbaum, Hillsdale, New Jersey.

Bear, M. F., Cooper, L. N., and Ebner, F. F. (1987). A physiological basis for a theory of synapse modification. *Science*, 237:42–48.

Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

Bienenstock, E. L., Cooper, L. N., and Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal Neuroscience*, 2:32–48.

Bryan, J. G. (1951). The generalized discriminant function: mathematical foundations and computational routines. *Harvard Educational Review*, 21:90–95.

Clothiaux, E. E., Cooper, L. N., and Bear, M. F. (1991). Synaptic plasticity in visual cortex: Comparison of theory with experiment. *Journal of Neurophysiology*, 66:1785–1804.

Cooper, L. N., Liberman, F., and Oja, E. (1979). A theory for the acquisition and loss of neurons specificity in visual cortex. *Biological Cybernetics*, 33:9–28.

Cooper, L. N. and Scofield, C. L. (1988). Mean-field theory of a neural network. *Proceedings of the National Academy of Science*, 85:1973–1977.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12:793–815.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. John Wiley, New York.

Duffy, F. H., Sonodgrass, S. R., Burchfiel, J. L., and Conway, J. L. (1976). Bicuculline reversal of deprivation amblyopia in the cat. *Nature*, 260:256–257.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Frègnac, Y. and Imbert, M. (1984). Development of neuronal selectivity in primary visual cortex of cat. *Physiol. Rev*, 64:325–434.

Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C(23):881–889.

Harman, H. H. (1967). *Modern Factor Analysis*. University of Chicago Press, Second Edition, Chicago and London.

Hinton, G. E. and Nowlan, S. J. (1990). The bootstrap widrow-hoff rule as a cluster-formation algorithm. *Neural Computation*, 2(3):355–362.

Hubel, D. H. and Wiesel, T. N. (1959). Integrative action in the cat's lateral geniculate body. *J. Physiol*, 148:574–591.

Huber, P. J. (1985). Projection pursuit. (with discussion). *The Annals of Statistics*, 13:435–475.

Intrator, N. (1990). An averaging result for random differential equations. Technical Report 54, Center For Neural Science, Brown University.

Intrator, N. (1992). Feature extraction using an unsupervised neural network. *Neural Computation*, 4:98–107.

Intrator, N. and Gold, J. I. (1993). Three-dimensional object recognition of gray level images: The usefulness of distinguishing features. *Neural Computation*, 5:61–74.

Intrator, N., Gold, J. I., Bülthoff, H. H., and Edelman, S. (1991). Three-dimensional object recognition using an unsupervised neural network: Understanding the distinguishing features. In Feldman, Y. and Bruckstein, A., editors, *Proceedings of the 8th Israeli Conference on AICV*, pages 113–123. Elsevier.

Intrator, N. and Tajchman, G. (1991). Supervised and unsupervised feature extraction from a cochlear model for speech recognition. In Juang, B. H., Kung, S. Y., and Kamm, C. A., editors, *Neural Networks for Signal Processing – Proceedings of the 1991 IEEE Workshop*, pages 460–469. IEEE Press, New York, NY.

Jones, M. C. (1983). The projection pursuit algorithm for exploratory data analysis. Unpublished Ph.D. dissertation, University of Bath, School of Mathematics.

Jones, M. C. and Sibson, R. (1987). What is projection pursuit? (with discussion). *J. Roy. Statist. Soc.*, Ser. A(150):1–36.

Kruskal, J. B. (1969). Toward a practical method which helps uncover the structure of the set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'. In Milton, R. C. and Nelder, J. A., editors, *Statistical Computation*, pages 427–440. Academic Press, New York.

Kruskal, J. B. (1972). Linear transformation of multivariate data to reveal clustering. In Shepard, R. N., Romney, A. K., and Nerlove, S. B., editors, *Multidimensional Scaling: Theory and Application in the Behavioral Sciences, I, Theory*, pages 179–191. Seminar Press, New York and London.

Linsker, R. (1988). Self-organization in a perceptual network. *IEEE. Computer*, 88:105–117.

Miller, K. D., Keller, J., and Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. *Science*, 240:605–615.

Mower, G. D., Caplan, C. J., Christen, W. G., and Duffy, F. H. (1985). Dark rearing prolongs physiological but not anatomical plasticity of the cat visual cortex. *J. Comp. Neurol.*, 235:448–466.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Math. Biology*, 15:267–273.

Presson, J. and Gordon, B. (1982). The effects of monocular deprivation on the physiology and anatomy of the kitten's visual system. *Soc. Neurosci. Abstracts*, 8:5. 10.

Saul, A. and Daniels, J. D. (1986). Modeling and simulation I: Introduction and guidelines. *J. of Electrophysiological Techniques*, 13:95–109.

Scofield, C. L. and Cooper, L. N. (1985). Development and properties of neural networks. *Contemp. Phys.*, 26:125–145.

Sebestyen, G. (1962). *Decision Making Processes in Pattern Recognition*. Macmillan, New York.

Sejnowski, T. J. (1977). Storing covariance with nonlinearly interacting neurons. *Journal of Mathematical Biology*, 4:303–321.

Sherman, S. M. and Spear, P. D. (1982). Organization of visual pathways in normal and visually deprived cats. *Physiol. Rev.*, 62:738–855.

Slaney, M. (1988). Lyon's cochlear model. Technical report, Apple Corporate Library, Cupertino, CA 95014.

Switzer, P. (1970). Numerical classification. In Barnett, V., editor, *Geostatistics*. Plenum Press, New York.

von der Malsburg, C. (1973). Self-organization of orientation sensitivity cells in the striate cortex. *Kybernetik*, 14:85–100.

Wiesel, T. N. and Hubel, D. H. (1965). Comparison of the effects of unilateral and bilateral eye closure on cortical unit responses in kittens. *J. Neurophysiol.*, 28:1029–1040.

Yang, X. and Faber, D. S. (1991). Initial synaptic efficacy influences induction and expression of long-term changes in transmission. *Proceedings of the National Academy of Science*, 88(10):4299–4303.

# Appendix

## A Convergence of the Solution of the Random Differential Equations

To show the explicit dependency on the learning rate, we rewrite the random modification equations in the form:

$$\dot{m}_\epsilon = \epsilon\ \mu(t)\phi(x \cdot m_\epsilon, \Theta_M)x, \quad m_\epsilon(0) = m_0, \tag{1}$$

and the deterministic differential equations,

$$\dot{\bar{m}}_\epsilon = \epsilon\ \mu(t)E[\phi(x \cdot \bar{m}_\epsilon, \Theta_M)x], \quad \bar{m}_\epsilon(0) = m_0, \tag{2}$$

The convergence of the solution will be shown in two steps; First we show that the solution of the averaged deterministic equation converges, and then we use theorem A.1 to show the convergence of the solution of the random differential equation to the solution of its averaged deterministic equation.

## A.1 Convergence of the Deterministic Equation

The deterministic differential equations represent a negative gradient of the risk. Therefore, in order to show convergence of the solution, we only need to show that the risk is bounded from below. This will assure that the solution converges to a local minimum of the risk.

We can assume that $m$ the synaptic weight vector lies in the space spanned by the random variable $x$. When we replace the random variable $x$ with a training set $x^1, \ldots, x^n$, this assumption says that $m \in \text{Span}\{x^1, \ldots, x^n\}$. This implies that there is a $\lambda > 0$, so that $\forall m \ \text{Var}(x \cdot m) \geq \lambda \| m \|^2 > 0$.

To show that the vector $\bar{m}_\epsilon$ is bounded we assume that none of its components is zero (since zero is definitely bounded), and multiply both sides of the above equation by $\bar{m}_\epsilon$, this implies:

$$\begin{aligned} \frac{1}{2}\frac{d}{dt} \| \bar{m}_\epsilon \|_{l^2}^2 &= E[(x \cdot \bar{m}_\epsilon)^3] - E^2[(x \cdot \bar{m}_\epsilon)^2] \\ &\leq \| \bar{m}_\epsilon \|^3 - \text{Var}^2(x \cdot \bar{m}_\epsilon) \\ &\leq \| \bar{m}_\epsilon \|^3 - \lambda^2 \| \bar{m}_\epsilon \|^4 \\ &= \| \bar{m}_\epsilon \|^3 \{1 - \lambda^2 \| \bar{m}_\epsilon \|\}, \end{aligned} \tag{3}$$

which implies that $\| \bar{m}_\epsilon \| \leq \frac{1}{\lambda^2}$.

Using this fact we can now show the convergence of $\bar{m}_\epsilon$. We observe that $\dot{\bar{m}}_\epsilon = -\nabla R$, where $R(\bar{m}_\epsilon) = -\mu\{\frac{1}{3}E[(x \cdot \bar{m}_\epsilon)^3] - \frac{1}{4}E^2[(x \cdot \bar{m}_\epsilon)^2]\}$ is the risk. $R$ is bounded from below since $\| \bar{m}_\epsilon \|$ is bounded, therefore $\bar{m}_\epsilon$ converges to a local minimum of $R$ as a solution to the gradient descent.

## A.2 Convergence of the Random Equation

Using the fact that the averaged deterministic version convergence we shall now show the convergence of the random version. For this we need a general result on random differential equations (Intrator, 1990) which is cited below. This result is an extension of a result by Geman (1977) and roughly says that under some smoothness conditions on the second order derivatives of the differential equations, the solution of the random differential equation remains close (in the $L^2$ sense) to the deterministic solution *for all times.*

We start with some preliminary notation. let $H(x, \omega, t)$ be a continuous and mixing $R^m$ valued random process for any fixed $x$ and $t$, where $\omega$ is a sample point in a probability space. Define $G(x,t) = E[H(x, \omega, t)]$, the expected value with respect to $\omega$. Let $\mu(t)$ be a continous monotone function decreasing to zero, and let $\epsilon > 0$ be arbitrary. Consider the following random differential equation together with its associated averaged version,

$$\begin{aligned} \dot{x}_\epsilon(t,\omega) &= \epsilon \ \mu(t) H(x_\epsilon(t,\omega), \omega, t), \quad x_\epsilon(0,\omega) = x_0 \in R^n. \\ \dot{y}_\epsilon(t) &= \epsilon \ \mu(t) G(y_\epsilon(t), t), \qquad y_\epsilon(0) = x_0 \in R^n. \end{aligned} \tag{4}$$

$\epsilon$ generates a family of solutions $x_\epsilon$, and $y_\epsilon$.

**Theorem A.1** *Given the above system of random differential equations, assume:*

1. *$H \in R^n$ is jointly measurable with respect to its three arguments, and is of Type II $\varphi$ mixing.*
2. *$G(x,t) = E[H(x(s,\omega), t)]$, and for all $i$ and $j$*

$$\frac{\partial}{\partial x_j} G_i(x,t) \text{ exists, and is continuous in } (x,t).$$

3. *(a) There exists a unique solution, $x(t,\omega)$, on $[0,\infty)$ for almost all $\omega$; and*
   *(b) A solution to*
   $$\frac{\partial}{\partial t}g(t,s,x) = G(g(t,s,x),t), \quad g(s,s,x) = x,$$
   *exists on $[0,\infty)\times[0,\infty)\times R^n$.*

4. *There exist continuous functions $B_1(r), B_2(r)$, and $B_3(r)$, such that for all $i,j,k,\tau \geq 0$, and $\omega$:*
   *(a)* $\mid H_i(x,\omega,t) \mid \leq B_1(\mid x \mid)$;
   *(b)* $\mid (\partial/\partial x_j)H_i(x,\omega,t) \mid \leq B_2(\mid x \mid)$;
   *(c)* $\mid (\partial^2/\partial x_j \partial x_k)H_i(x,\omega,t) \mid \leq B_3(\mid x \mid)$.

5. $\sup_{\epsilon>0,t} \mid y_\epsilon(t) \mid \leq B_4$ *for some* $B_4$.

6. $\exists\ \gamma > 0,\ c > 0$, *such that* $\varphi(\delta) \leq \delta^{-\gamma}$, *and* $\mu(t) \leq t^{-(\frac{1}{\gamma}+1+c)}$, *for a monotone decreasing* $\mu$.

*Then under conditions 1-6:*

$$\lim_{\epsilon\to 0} \sup_{t\geq 0} E \mid x_\epsilon - y_\epsilon \mid^2 = 0, \tag{5}$$

To use this result, we need only to show that the deterministic and the random solutions are bounded, which will ensure conditions 2-5. Then under the mixing conditions 1 and 6 on the input $x$, we get the desired result.

Verifying that the random solution is bounded for every $\omega$ can be done by multiplying both sides of the random differential equations by $m_\epsilon$, assuming its components are not zero, and applying the assumptions made above on $\mathrm{Var}(x \cdot m_\epsilon)$, we get

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\| m_\epsilon \|^2 &= (x\cdot m_\epsilon)^3 - (x\cdot m_\epsilon)^2 E[(x\cdot m_\epsilon)^2] \\
&= (x\cdot m_\epsilon)^2\{(x\cdot m_\epsilon) - E[(x\cdot m_\epsilon)^2]\} \\
&\leq (x\cdot m_\epsilon)^2\{(x\cdot m_\epsilon) - \mathrm{Var}(x\cdot m_\epsilon)\} \\
&\leq (x\cdot m_\epsilon)^2\{(x\cdot m_\epsilon) - \lambda\| m_\epsilon \|^2\} \\
&\leq (x\cdot m_\epsilon)^2\{\| m_\epsilon \| - \lambda\| m_\epsilon \|^2\},
\end{aligned} \tag{6}$$

which implies that the derivative of the norm will become negative whenever $\| m_\epsilon \| > \lambda$, therefore $\| m_\epsilon \| \leq \frac{1}{\lambda}$.

Finally, since the random solution remains close to a converging deterministic solution, it remains close (in the $L^2$ sense) to its limit for large enough $t$.

$\delta$ is arbitrary, which implies that

$$E|m_\epsilon(t) - \tilde{m} \mid^2 \underset{\epsilon\to 0}{\longrightarrow} 0 \tag{7}$$

$\Diamond$