Derivation of Linear Hebbian Equations from a Nonlinear Hebbian Model of Synaptic Plasticity

Kenneth D. Miller
Department of Physiology
University of California
San Francisco, CA 94143 - 0444
ken@phy.ucsf.edu

Abstract

A linear Hebbian equation for synaptic plasticity is derived from a more complex, nonlinear model by considering the initial development of the difference between two equivalent excitatory projections. This provides a justification for the use of such a simple equation to model activity-dependent neural development and plasticity, and allows analysis of the biological origins of the terms in the equation. Connections to previously published models are discussed.

This article was published as: **Neural Computation 2,** 319-331 (1990).

Recently, a number of authors (e.g. [Miller *et al.*, 1986, Linsker, 1986, Miller *et al.*, 1989]) have studied linear equations modelling Hebbian or similar correlation-based mechanisms of synaptic plasticity, subject to nonlinear saturation conditions limiting the strengths of individual synapses to some bounded range. Such studies have intrinsic interest for understanding the dynamics of simple feedforward models. However, the biological rules for both neuronal activation and synaptic modification are likely to depend nonlinearly on neuronal activities and synaptic strengths in many ways. When are such simple equations likely to be useful as models of development and plasticity in biological systems?

One critical nonlinearity for biological modeling is rectification: Biologically, a synaptic strength cannot change its sign, because a given cell's synapses are either all excitatory or all inhibitory. Saturating or similar nonlinearities that bound the range of synaptic strengths may be ignored if one is concerned with the early development of a pattern of synaptic strengths, and if the initial distribution of synaptic strengths is well on the interior of the allowed region in weight space. However, if a model's outcome depends on a synaptic variable taking both positive and negative values, then the bound on synaptic strengths at zero must be considered.

Previous models make two proposals that avoid this rectification nonlinearity. One proposal is to study the difference between the strengths of two separate, initially equivalent excitatory projections innervating a single target structure [Miller *et al.*, 1986, Miller *et al.*, 1989, Miller, 1989a]. This difference in strengths is a synaptic variable that may take both positive and negative values. An alternative proposal is to study the sum of the strengths of two input projections, one excitatory, one inhibitory, that are statistically indistinguishable from one another in their connectivities and activities [Linsker, 1986].

The proposal to study the difference between the strengths of two equivalent excitatory projections is motivated by study of the visual system of higher mammals. Examples in that system include the projections from the lateral geniculate nucleus to the visual cortex of inputs serving the left and right eyes [Miller *et al.*, 1989] (reviewed in [Miller & Stryker, 1990]) or of inputs with on-center and off-center receptive fields [Miller, 1989a]. Examples exist in many other systems (briefly reviewed in [Miller, 1990]). Assuming that the difference between the two projections is initially small, the early development of the difference can be described by equations linearized about the uniform condition of complete equality of the two projections. This can allow linear equations to be used to study aspects of early development in the presence of more general nonlinearities.

This letter presents the derivation of previously studied simple, linear Hebbian equations, beginning from a nonlinear Hebbian model in the presence of two equivalent excitatory input projections. The outcome of this derivation is contrasted with that resulting from equivalent excitatory and inhibitory projections. Applications to other models are then discussed.

1 Assumptions

The derivation depends on the following assumptions:

A1 There are two modifiable input projections to a single output layer. The two input projections are equivalent in the following sense:

- There is a topographic mapping that is identical for the two input layers: Each of the two input projections represent the same topographic coordinates, and the two project in an overlapping, continuous manner to the output layer.
- The statistics of neuronal activation are identical within each projection (N.B., the correlations *between* the two projections may be quite different from those within each

projection);

A2 Synaptic modification occurs via a Hebb rule in which the roles of output cell activity and that of input activity are mathematically separable;

A3 The activity of an output cell depends (nonlinearly) only on the summed input to the cell.

In addition, the following assumptions are made for simplicity; they can be relaxed in a straightforward manner, at the cost of more complicated equations:

A4 The Hebb rule and the output activation rule are taken to be instantaneous, ignoring time delays. (Instantaneous rules follow from more complicated rules in the limit in which input patterns are sustained for long times compared to dynamical relaxation times. This limit appears likely to be applicable to visual cortex, where geniculate inputs typically fire in bursts sustained over many tens or hundreds of milliseconds [Creutzfeldt & Ito, 1968]);

A5 The statistics of neuronal activation are time-invariant;

A6 There are lateral interconnections in the output layer which are time-invariant;

A7 The input and output layers are two-dimensional and spatially homogeneous;

A8 The topographic mapping from input to output layers is linear and isotropic.

2 Notation

We let Roman letters $(x, y, z, \ldots)$ label topographic location in the output layer, and Greek letters $(\alpha, \beta, \gamma, \ldots)$ label topographic location in each of the input layers. We use the labels 1 and 2 to refer to the two input projections.

We define the following functions

1. $o(x,t)$: activity (e.g., firing rate, or membrane potential) of output cell at location x at time t;

2. $i^1(\alpha,t), i^2(\alpha,t)$: activity of input of type 1 or 2, respectively, from location α at time t.

3. $A(x-\alpha)$: synaptic density or "arbor" function, describing connectivity from the input layer to the output layer. This tells the number of synapses from an input with topographic location α onto the output cell with topographic location x. This is assumed time-independent and independent of projection type;

4. $s^1_k(x,\alpha,t), s^2_k(x,\alpha,t)$: strength of the k^{th} synapse of type 1 or 2, respectively, from the input at α to the output cell at x at time t. There are $A(x-\alpha)$ such synapses of each type;

5. $S^1(x,\alpha,t)$, $S^2(x,\alpha,t)$: total synaptic strength at time t from the input of type 1 or 2, respectively, at location α, to the output cell at x. $S^1(x,\alpha,t) = \sum_k s^1_k(x,\alpha,t)$ (and similarly for $S^2(x,\alpha,t)$).

6. $B(x-y)$: intracortical connectivity function, describing total (time-invariant) synaptic strength from the output cell at y to the output cell at x. B depends only on $x-y$ by assumption A7 of spatial homogeneity.

3 Derivation of linear Hebbian equations from a nonlinear Hebbian rule

The Hebbian equation for the development of a single type 1 synapse s_k^1 from α to x can, by assumptions A2 and A4, be written

$$\begin{aligned} \frac{ds_k^1(x,\alpha,t)}{dt} &= \lambda h_o\Big[o(x,t)\Big] h_i\left[i^1(\alpha,t)\right] - \epsilon - \gamma s_k^1(x,\alpha,t) \\ &\qquad\qquad \text{subject to } 0 \le s_k^1 \le s_{\max}. \end{aligned} \tag{1}$$

We assume that h_o is a differentiable function, but h_o and h_i are otherwise arbitrary functions incorporating nonlinearities in the plasticity rule. λ, ϵ and γ are constants. Summing over all type 1 synapses from α to x yields

$$\begin{aligned} \frac{dS^1(x,\alpha,t)}{dt} &= \lambda A(x-\alpha) h_o\Big[o(x,t)\Big] h_i\left[i^1(\alpha,t)\right] - \epsilon A(x-\alpha) - \gamma S^1(x,\alpha,t) \\ &\qquad\qquad \text{subject to } 0 \le S^1(x,\alpha,t) \le s_{\max} A(x-\alpha) \end{aligned} \tag{2}$$

(and similarly for S^2). There are small differences between Eq. 1 and Eq. 2 when some but not all synapses s^k have reached saturation. We will be concerned with the early development of a pattern, before synapses saturate, and so ignore these differences. We will omit explicit mention of the saturation limits hereafter.

Define the direct input to a cell as $\theta(x,t) \equiv \sum_\beta \{S^1(x,\beta,t) f_i[i^1(\beta,t)] + S^2(x,\beta,t) f_i[i^2(\beta,t)]\}$. The nonlinear activation rule is, by assumptions A3 and A4,

$$o(x,t) = g\left\{\theta(x,t) + \sum_y B(x-y) f_o\Big[o(y,t)\Big]\right\} \tag{3}$$

f_o and g are assumed to be differentiable functions, but they and f_i are otherwise arbitrary functions incorporating the nonlinearities in the activation rules.

We make the following nontrivial assumption:

A9 For each input vector $\theta(t)$, Eq. (3) defines a unique output vector $\mathbf{o}(t)$.

Biologically, this is the assumption that the inputs determine the state of the outputs. Mathematically, this can be motivated by studies of the Hartline-Ratliff equation [Hadeler & Kuhn, 1987].[1] With this assumption, $o(x,t)$ can be regarded as a function of the variables $\theta(y,t)$ for varying y.

We now transform from the variables $\mathbf{S}^1$ and $\mathbf{S}^2$ to sum and difference variables. Define the following:

$$\begin{aligned} \mathbf{S}^D &\equiv \mathbf{S}^1 - \mathbf{S}^2, \\ \mathbf{S}^S &\equiv \mathbf{S}^1 + \mathbf{S}^2, \\ f_i^S(\alpha,t) &\equiv f_i\left[i^1(\alpha,t)\right] + f_i\left[i^2(\alpha,t)\right], \\ f_i^D(\alpha,t) &\equiv f_i\left[i^1(\alpha,t)\right] - f_i\left[i^2(\alpha,t)\right], \\ h_i^D(\alpha,t) &\equiv h_i\left[i^1(\alpha,t)\right] - h_i\left[i^2(\alpha,t)\right], \end{aligned} \tag{4}$$

[1] The Hartline-Ratliff equation is Eq. (3) for $g(x) = \{x,\ x \ge 0;\ 0,\ x < 0\}$ and $f_o(x) = x$. That equation has a unique output for every input, for symmetric $\mathbf{B}$, iff $\mathbf{1} - \mathbf{B}$ is positive definite; a more general condition for $\mathbf{B}$ nonsymmetric can also be derived [Hadeler & Kuhn, 1987].

$$\theta^S(x,t) \equiv \frac{1}{2}\sum_\beta S^S(x,\beta,t)f_i^S(\beta,t),$$
$$\theta^D(x,t) \equiv \frac{1}{2}\sum_\beta S^D(x,\beta,t)f_i^D(\beta,t),$$

Note that $\theta(x,t) = \theta^S(x,t) + \theta^D(x,t)$. The Hebb rule for the difference, $\mathbf{S}^D \equiv \mathbf{S}^1 - \mathbf{S}^2$ is, from Eq. 2,

$$\frac{dS^D(x,\alpha,t)}{dt} = \lambda A(x-\alpha)h_o\Big[o(x,t)\Big]h_i^D(\alpha,t) - \gamma S^D(x,\alpha,t) \tag{5}$$

S^D is a synaptic variable that can take on both positive and negative values, and whose initial values are near zero. We will develop a linear equation for $\mathbf{S}^D$ by linearizing Eq. 5 about the uniform condition $\mathbf{S}^D = \mathbf{0}$. We will accomplish this by expanding Eq. 5 about $\theta^D = \mathbf{0}$ to first order in θ^D.

Let $o^S(x,t)$ be the solution of

$$o^S(x,t) = g\left\{\theta^S(x,t) + \sum_y B(x-y)f_o\Big[o^S(y,t)\Big]\right\}. \tag{6}$$

Then, letting $'$ signify the derivative of a function,

$$\frac{dS^D(x,\alpha,t)}{dt} = \lambda A(x-\alpha)h_i^D(\alpha,t)\left\{h_o\left[o^S(x,t)\right] + h_o'\left[o^S(x,t)\right]\sum_y \frac{do(x,t)}{d\theta(y,t)}\Big|_{\theta^S}\theta^D(y,t)\right\}$$
$$-\gamma S^D(x,\alpha,t) + O\left[(\theta^D)^2\right]. \tag{7}$$

Letting ${g'}^S(x,t) \equiv g'\left\{\theta^S(x,t) + \sum_y B(x-y)f_o[o^S(y,t)]\right\}$, the derivative of $o(x,t)$ is

$$\frac{do(x,t)}{d\theta(y,t)}\Big|_{\theta^S} = {g'}^S(x,t)\left[\mathbf{1} + \tilde{\mathbf{B}} + (\tilde{\mathbf{B}})^2 + \cdots\right]_{xy} \tag{8}$$

where $\mathbf{1}$ is the identity matrix, $\tilde{\mathbf{B}}$ is the matrix with elements $\tilde{B}_{xy} = B(x-y)f_o'[o^S(y,t)]{g'}^S(y,t)$, and $[\cdots]_{xy}$ means the xy element of the matrix in brackets. Letting

$$I(x,y,t) = \left[\mathbf{1} + \tilde{\mathbf{B}} + (\tilde{\mathbf{B}})^2 + \cdots\right]_{xy} \tag{9}$$
$$C^D(\alpha,\beta,t) = \frac{1}{2}h_i^D(\alpha,t)f_i^D(\beta,t) \tag{10}$$
$$M(x,t) = h_o'\left[o^S(x,t)\right]{g'}^S(x,t) \tag{11}$$

we find that Eq. (7) becomes, to first order in θ^D,

$$\frac{dS^D(x,\alpha,t)}{dt} = \lambda A(x-\alpha)h_o\left[o^S(x,t)\right]h_i^D(\alpha,t) - \gamma S^D(x,\alpha,t)$$
$$+\lambda A(x-\alpha)M(x,t)\sum_{y,\beta} I(x,y,t)C^D(\alpha,\beta,t)S^D(y,\beta,t). \tag{12}$$

This equation can be interpreted intuitively. The first term is the Hebbian term of Eq. 5 in which the output cell's activity has been replaced by the activity it would have if $\theta^D = \mathbf{0}$, that is, if $\mathbf{S}^D = \mathbf{0}$. The last term is the Hebbian term with the output cell's activity replaced by the first order change in that activity due to the fact that $\theta^D \neq \mathbf{0}$. In this term, $M(x,t)$ measures the degree to which, near $\theta^D = \mathbf{0}$, the activity of the output cell at x can be significantly modified, for purposes of the Hebb rule, by changes in the total input it receives. $I(x,y,t)$ measures the change in the total input to the cell at x due to changes in the direct input to the cell at y. $C^D(\alpha,\beta,t)S^D(y,\beta,t)$ incorporates both the change in the direct input to the cell at y due to the fact that $\theta^D \neq \mathbf{0}$, and the difference in the activities of the inputs from α that are being modified.

4 Averaging

Given some statistical distribution of input patterns $i(\alpha,t)$, Eq. 12 is a stochastic differential equation. To transform it to a deterministic equation, we average it over input activity patterns. The result is an equation for the mean value $\langle \mathbf{S}^D \rangle$, averaged over input patterns. The right hand side of the equation consists of an infinite series of terms, corresponding to the various cumulants of the stochastic operators of Eq. 12 [Keller, 1977, Miller, 1989b]. However, when λ and γ are sufficiently small that $\mathbf{S}^D$ can be considered constant over a period in which all input activity patterns are sampled, only the first term is significant. We restrict attention to that term.

After averaging, the first term on the right side of Eq. 12 yields zero, by equality of the two input projections. The lowest order term resulting from averaging of the last term is

$$\lambda A(x-\alpha) \sum_{y,\beta} \left\langle M(x,t) I(x,y,t) C^D(\alpha,\beta,t) \right\rangle S^D(y,\beta,t).$$

where we retain the notation $\mathbf{S}^D$ for $\langle \mathbf{S}^D \rangle$. We now assume:

A10 We can approximate $\left\langle M(x) I(x,y) C^D(\alpha,\beta) \right\rangle$ by $\left\langle M(x) I(x,y) \right\rangle \left\langle C^D(\alpha,\beta) \right\rangle$.

Assumption A10 will be true if the sum of the two eyes' inputs is statistically independent of the difference between the two eyes' inputs. By equivalence of the two input projections the sum and difference are independent at the level of two-point interactions: $\langle S^S S^D \rangle = \langle S^1 S^1 \rangle - \langle S^2 S^2 \rangle = 0 = \langle S^S \rangle \langle S^D \rangle$. By assumption A7 of spatial homogeneity, $\langle M(x) I(x,y) \rangle$ can only depend on $x-y$, while $\langle C^D(\alpha,\beta) \rangle$ can only depend on $\alpha - \beta$. With these assumptions, then, the linearized version of this nonlinear model becomes

$$\frac{dS^D(x,\alpha,t)}{dt} = \lambda A(x-\alpha) \sum_{y,\beta} I(x-y) C^D(\alpha-\beta) S^D(y,\beta,t) - \gamma S^D(x,\alpha,t) \tag{13}$$

where

$$\begin{aligned} I(x-y) &\equiv \left\langle M(x) I(x,y) \right\rangle \\ &= \left\langle h_o' \left[o^S(x,t) \right] {g'}^S(x,t) \left[\delta_{xy} + B(x-y) f_o' \left[o^S(y,t) \right] {g'}^S(y,t) + \ldots \right] \right\rangle \end{aligned} \tag{14}$$

and

$$C^D(\alpha-\beta) \equiv \left\langle C^D(\alpha,\beta) \right\rangle = \frac{1}{2} \left\langle h_i^D(\alpha,t) f_i^D(\beta,t) \right\rangle. \tag{15}$$

Note that the nonlinear functions referring to the output cell, h_o, f_o and g, enter into Eq. 13 only in terms of their derivatives. This reflects the fact that the base level of output activity, $\mathbf{o}^S$, makes no contribution to the development of the difference $\mathbf{S}^D$ because the first term of Eq. 12 averages to 0. Only the alterations in output activity induced by θ^D contribute to the development of $\mathbf{S}^D$.

We have not yet achieved a linear equation for development. $I(x-y)$ depends on $\mathbf{S}^S$ through the derivatives of h_o, f_o, and g. Because the equation for $\mathbf{S}^S$ remains nonlinear, Eq. 13 is actually part of a coupled nonlinear system. Intuitively, the sum $\mathbf{S}^S$ is primarily responsible for the activation of output cells when $\mathbf{S}^D$ is small. $\mathbf{S}^S$ therefore serves to "gate" the transmission of influence across the output layer: the cells at x and at y must both be activated within their dynamic range, so that small changes in their inputs cause changes in their responses or in their contribution to Hebbian plasticity, in order for $I(x-y)$ to be nonzero. To render the equation linear, we must assume

A11 The shape of $I(x-y)$ does not vary significantly during the early, linear development of $\mathbf{S}^D$.

Changes in the amplitude of $I(x-y)$ will only alter the speed of development, not its outcome, and can be ignored. Assumption A11 can be loosely motivated by noting that $\mathbf{S}^S$ is approximately spatially uniform, so that $B(x-y)$ should be the primary source of spatial structure in $I(x-y)$;[2] and that cortical development may act to keep cortical cells operating within their dynamic range.

5 Comparison to the Sum of an Excitatory and an Inhibitory Projection

An alternative proposal to that developed here is to study the sum of the strengths of two indistinguishable input projections, one excitatory, one inhibitory [Linsker, 1986]. This case is mathematically distinct from the sum of two equivalent excitatory projections, because the Hebb rule does not change sign for the inhibitory population relative to the excitatory population. That is, in response to correlated activity of the pre- and postsynaptic cells, inhibitory synapses become weaker, not stronger, by a Hebbian rule.

To understand the significance of this distinction, let $\mathbf{S}^2$ now represent an inhibitory projection, so that $S^2 \leq 0$. Then the variable that is initially small, and in which we expand in order to linearize, is the synaptic sum $\mathbf{S}^S$, rather than the difference $\mathbf{S}^D$. Define $\mathbf{o}^D$ analogously to the definition of $\mathbf{o}^S$ in Eq. 6, with θ^D in place of θ^S. Let $h_i^S(\alpha,t) \equiv h_i\left[i^1(\alpha,t)\right] + h_i\left[i^2(\alpha,t)\right]$. Then one finds in place of Eq. 12

$$\begin{aligned}
\frac{dS^S(x,\alpha,t)}{dt} &= \lambda A(x-\alpha)h_o\left[o^D(x,t)\right]h_i^S(\alpha,t) - \gamma S^S(x,\alpha,t) - 2\epsilon A(x-\alpha) \\
&\quad + \lambda A(x-\alpha)M^S(x,t)\sum_{y,\beta} I^S(x,y,t)C^S(\alpha,\beta,t)S^S(y,\beta,t).
\end{aligned} \tag{16}$$

where $C^S = \frac{1}{2}h_i^S f_i^S$, and M^S and I^S are defined like M and I except that derivatives are taken at θ^D and $\mathbf{o}^D$ rather than at θ^S and $\mathbf{o}^S$.

Unlike Eq. 12, the first term of Eq. 16 does not disappear after averaging.[3] This means that the development of $\mathbf{S}^S$ depends upon a Hebbian coupling between the summed input activities, and the

[2] The correlation structure of the summed inputs can also contribute to $I(x-y)$, since cortical cells with separation $x-y$ must be coactivated for $I(x-y)$ to be nonzero. Arguments can be made that the relevant lengths in $I(x-y)$ appear smaller than an arbor diameter (e.g. see [Miller, 1990, Miller & Stryker, 1990]), and thus are on a scale over which cortical cells receive coactivated inputs regardless of input correlation structure.

[3] Footnote added to this electronic reprint, not in the *Neural Computation* article: the reason for this difference is simple. For the case of two equivalent excitatory projections, there is a fixed point around which to linearize, in the following sense: if the two projections are exactly equal, they will remain exactly equal forever. For the sum of an excitatory and an inhibitory projection, there is no such fixed point, because the dynamics distinguishes between the two projections. For example, when the two projections are exactly equal, one projection may grow while the other shrinks. The first term corresponds to the derivative when the two projections are exactly equal; this vanishes only when that point is a fixed point.

output cell's activity in response to $\mathbf{S}^D$ (the activity the cell would have if $\mathbf{S}^S = \mathbf{0}$). Thus, direct Hebbian couplings to both $\mathbf{S}^D$ and $\mathbf{S}^S$ drive the initial development of $\mathbf{S}^S$, rendering it difficult to describe the dynamics by a simple linear equation like Eq. 13.

In [Linsker, 1986], two assumptions were made that together lead to the disappearance of this first term. First, the output functions h_o, f_o and g were taken to be linear. This causes the first term to be proportional to C^D. To present the second assumption, we define correlation functions C^{11}, C^{12}, C^{21}, C^{22} among and between the two input projections by $C^{jk}(\alpha-\beta) = \langle h_i[i^j(\alpha,t)] f_i[i^k(\beta,t)]\rangle$. By equivalence of the two projections, $C^{11} = C^{22}$ and $C^{12} = C^{21}$. Then $C^D = C^{11} - C^{12}$. The second assumption was that correlations between the two projections are identical to those within the two projections; that is, $C^{12} = C^{11}$. This means that $C^D \equiv 0$, and so the first term disappears. These assumptions also ensure that $\mathbf{S}^D$ does not change in time, prior to synaptic saturation.

Equation 16 also differs from Eq. 12 in implicitly containing two additional parameters that Linsker named k_1 and k_2. k_1 is the decay parameter ϵ. The parameter k_2 arises as follows. One can reexpress the "correlation functions" C^{jk} in terms of "covariance functions" Q^{jk}: $C^{jk} = Q^{jk} + k_2$, where $Q^{jk}(\alpha - \beta) = \Big\langle (h_i[i^j(\alpha,t)] - \langle h_i[i^j(\alpha,t)]\rangle)(f_i[i^k(\beta,t)] - \langle f_i[i^k(\beta,t)]\rangle)\Big\rangle$ and $k_2 = \langle h_i[i^j(\alpha,t)]\rangle\langle f_i[i^k(\beta,t)]\rangle$. k_2 is independent of the choice of j and k. The Q's have the advantage that $\lim_{(\alpha-\beta)\to\infty} Q^{jk}(\alpha-\beta) = 0$; if f_i and h_i are linear, the Q's are true covariance functions. The correlation function relevant to the sum of an inhibitory and an excitatory projection is $C^S = C^{11} + C^{12} = Q^{11} + Q^{12} + 2k_2$. In contrast, the correlation function relevant to the difference between two excitatory projections is $C^D = C^{11} - C^{12} = Q^{11} - Q^{12}$, which has no k_2 dependence. Thus, the parameters k_1 and k_2 do not arise in considering the difference between two excitatory input projections, because they are identical for each input projection and thus disapear from the equation for the difference; whereas these parameters do arise in considering the sum of an excitatory and an inhibitory projection. In [MacKay & Miller, 1990], it was shown that these parameters can significantly alter the dynamics, and play crucial roles in many of the results of [Linsker, 1986].

In summary, the proposal to study equivalent excitatory and inhibitory projections does not robustly yield a linear equation in the presence of nonlinearities in the output functions h_o, f_o and g. Even in the absence of such additional nonlinearities, it can lead to different dynamical outcomes than the proposal studied here. It also is biologically problematic. It would not apply straightforwardly to such feedforward projections as the retinogeniculate and geniculocortical projections in the mammalian visual system, which are exclusively excitatory. Where both inhibitory and excitatory populations do exist, the two are not likely to be equivalent. For example, inhibitory neurons are often interneurons that, when active, inhibit nearby excitatory neurons, potentially rendering the three correlation structures C^{11}, C^{12}, and C^{22} quite distinct; connectivity of such interneurons is also distinct from that of nearby excitatory cells [Toyama *et al.*, 1981, Singer, 1977]. Similarly, while there is extensive evidence that excitatory synapses onto excitatory cells may be modified in a Hebbian manner [Nicoll *et al.*, 1988], current evidence suggests that there may be little modification of inhibitory synapses, or of excitatory synapses onto the aspinous inhibitory interneurons, under the same stimulus paradigms [Abraham *et al.*, 1987, Griffith *et al.*, 1986, Singer, 1977].

6 Connections to Previous Models

Equation (13) is that studied in [Miller *et al.*, 1986, Miller, 1989a, Miller *et al.*, 1989]. It is also formally equivalent to that studied in [Linsker, 1986] except for the absence of the two parameters k_1 and k_2.[4]

[4] Also, lateral interactions in the output layer were not introduced until the final layer in [Linsker, 1986]. They were then introduced perturbatively, so that I was approximated by $\mathbf{1} + \mathbf{B}$. $\mathbf{B}$ was referred to as f in that paper.

The current approach allows the analysis of other previous models. For example, in the model of [Willshaw & von der Malsburg, 1976], g was take to be a linear threshold function ($g(x) = x - \delta$ for $x > \delta$, where δ is a constant threshold; $g(x) = 0$ otherwise); this can be approximated by a differentiable function. The functions h_o and f_o were taken to be the identity, while h_i and f_i were taken to be step functions: 1 if the input was active, 0 if it was not. A time-dependent activation rule was used, but input activations were always sustained until a steady state was reached so that this rule is equivalent to Eq. 3. These rules were only applied to a single input projection, but the present analysis allows examination of the case of two input projections. From Eq. 14, it can be seen that choosing g to be a linear threshold function has two intuitively obvious effects: (1) on the average, patterns for which θ^S would fail to bring the output cell at x above threshold do not cause any modification of $\mathbf{S}^D$ onto that cell; (2) Such patterns also make no average contribution to $I(y - x)$ for all y, i.e. if the cell at x is not above threshold it cannot influence plasticity on the cell at y. More generally, given an ensemble of input patterns and the initial distribution of S^S, the functions $I(x - y)$ and $C^D(\alpha - \beta)$ could be calculated explicitly from Eqs. 14 and 15, respectively. Similarly, [Hopfield, 1984] proposed a neuronal activation rule in which f_i and f_o are taken to be sigmoidal functions and g is the identity; while many current models (i.e. [Rumelhart *et al.*, 1986]) take f_i and f_o to be the identity, but take g to be sigmoidal. Again, such activation rules can be analyzed within the current framework.

7 Conclusions

It is intuitively appealing to think that activity-dependent neural development may be described in terms of functions A, I, and C that describe, respectively, connectivity from input to output layer ("arbors"), intralaminar connectivity within the output layer, and correlations in activity within the input layer. I have shown that formulation of linear equations in terms of such functions can be sensible for modeling aspects of early neural development in the presence of nonlinearities in the rules governing cortical activation and Hebbian plasticity. The functions I and C can be expressed in terms of the ensemble of input activities and the functions describing cortical activation and plasticity. This gives a more general relevance to results obtained elsewhere characterizing the outcome of development under Eq. 13 in terms of these functions [Miller *et al.*, 1989, Miller, 1989a, MacKay & Miller, 1990].

The current formulation is of course extremely simplified. Notable simplifications include the lack of plasticity in intralaminar connections in the output layer; the instantaneous nature of the equations; the assumption of spatial homogeneity; and more generally, the lack of any attempt at biophysical realism. The derivation requires several additional assumptions whose validities are difficult to evaluate. The current effort provides a unified framework for analyzing a large class of previous models. It is encouraging that the resulting linear model is sufficient to explain many features of cortical development [Miller *et al.*, 1989, Miller, 1989a]; it will be of interest, as more complex models are formulated, to see the degree to which they force changes in the basic framework analyzed here.

Acknowledgements

I thank J.B. Keller for suggesting to me long ago that the ocular dominance problem should be approached by studying the early development of an ocular dominance pattern and linearizing about the nearly uniform initial condition. I thank M.P. Stryker for supporting this work, which was performed in his laboratory. I am supported by an N.E.I. Fellowship and by a Human Frontiers Science Program Grant to M.P. Stryker (T. Tsumoto, Coordinator). I thank M.P. Stryker, D.J.C. MacKay, and especially the action editor for helpful comments.

References

[Abraham *et al.*, 1987] Abraham, W.C., B. Gustafsson and H. Wigstrom (1987). "Long-term potentiation involves enhanced synaptic excitation relative to synaptic inhibition in guinea-pig hippocampus." *J. Physiol. (London)* **394**, 367-380.

[Creutzfeldt & Ito, 1968] Creutzfeldt, O. and Ito, M. (1968). "Functional synaptic organization of primary visual cortex neurones in the cat", *Exp. Brain Res.* **6**, 324-352.

[Griffith *et al.*, 1986] Griffith, W.H., T.H. Brown and D. Johnston (1986). "Voltage-clamp analysis of synaptic inhibition during long-term potentiation in hippocampus." *J. Neurophys.* **55**, 767-775.

[Hadeler & Kuhn, 1987] Hadeler, K.P. and D. Kuhn (1987). "Stationary states of the Hartline-Ratliff model", *Biol. Cybern.* **56**, 411-417.

[Hopfield, 1984] Hopfield, J.J. (1984). " Neurons with graded response have collective computational properties like those of two-state neurons", *Proc. Natl. Acad. Sci. USA* **81**, 3088-3092.

[Keller, 1977] Keller, J.B. (1977). "Effective behavior of heterogeneous media," pp. 631–644 in *Statistical Mechanics and Statistical Methods in Theory and Application*, U. Landman, Ed. (Plenum).

[Linsker, 1986] Linsker, R. (1986). "From basic network principles to neural architecture" (series), *Proc. Nat. Acad. Sci. USA*, **83**, 7508-7512, 8390-8394, 8779-8783.

[MacKay & Miller, 1990] MacKay, D.J.C. and K.D. Miller. "Analysis of Linsker's simulations of Hebbian rules," *Neural Computation* **2**, 169-182.

[Miller, 1990] Miller, K.D. (1990). "Correlation-based mechanisms of neural development," in *Neuroscience and Connectionist Theory*, M.A. Gluck and D.E. Rumelhart, Eds. (Lawrence Erlbaum Associates, Hillsdale NJ), pp. 267-353.

[Miller, 1989a] Miller, K.D. (1989). "Orientation-selective cells can emerge from a Hebbian mechanism through interactions between ON- and OFF-center inputs," *Soc. Neurosc. Abst.* **15,** 794.

[Miller, 1989b] Miller, K.D. (1989). *Correlation-based mechanisms in visual cortex: Theoretical and experimental studies*, Ph.D. Thesis, Stanford University Medical School (University Microfilms, Ann Arbor).

[Miller *et al.*, 1986] Miller, K.D., J.B. Keller and M.P. Stryker (1986). "Models for the formation of ocular dominance columns solved by linear stability analysis," *Soc. Neurosc. Abst.* **12**, 1373.

[Miller *et al.*, 1989] Miller, K.D., J.B. Keller and M.P. Stryker (1989). "Ocular dominance column development: analysis and simulation," *Science* **245**, 605-615.

[Miller & Stryker, 1990] Miller, K.D. and M.P. Stryker (1990). "Ocular dominance column formation: mechanisms and models," in *Connectionist Modeling and Brain Function: The Developing Interface*, S.J. Hanson and C.R. Olson, Eds. (MIT Press/Bradford), pp. 255-350.

[Nicoll *et al.*, 1988] Nicoll, R.A., J.A. Kauer and R.C. Malenka (1988). "The current excitement in long-term potentiation," *Neuron* **1**, 97-103.

[Rumelhart *et al.*, 1986] Rumelhart, D.E., G.E. Hinton and R.J. Williams (1986). "Learning representations by back-propagating errors," *Nature* **323**, 533-536.

[Singer, 1977] Singer, W. (1977) "Effects of monocular deprivation on excitatory and inhibitory pathways in cat striate cortex," *Exp. Brain Res.* **30**, 25-41.

[Toyama *et al.*, 1981] Toyama, K., M. Kimura and K. Tanaka (1981). "Organization of cat visual cortex as investigated by cross-correlation techniques," *J. Neurophysiol.* **46**, 202-213.

[Willshaw & von der Malsburg, 1976] Willshaw, D.J. and C. von der Malsburg (1976). "How patterned neural connections can be set up by self-organization," *Proc. R. Soc. Lond. B.* **194**, 431-445.