Thus, instead of having a label for every individual weight w_i, a set of equally spaced *representative* weights r_n is used, with positions $\mathbf{x}_n$. The value of r_n is the average value of the represented weights. The number of weights represented by r_n is given by the local synaptic density $A_n = A(\mathbf{x}_n)$, where $A(\mathbf{x})$ is the synaptic density function.

Let the dynamics in the $\mathbf{w}$ representation be written $\frac{d}{dt} w_i = \sum_j C^{\mathbf{w}}_{ij} w_j$. Then the dynamics in the $\mathbf{r}$ representation are given by $\frac{d}{dt} r_m = \sum_n C^{\mathbf{r}}_{mn} A_n r_n$: here, $C^{\mathbf{r}}_{mn} \equiv C(\mathbf{x}_m, \mathbf{x}_n)$, and it is assumed that the function $C(\mathbf{x}_m, \mathbf{x}_n)$ varies slowly on the scale of the spacing between the representative synapses (further discussed in [20, Appendix A]). Define the matrix $\mathbf{A} \equiv \mathrm{diag}\{A_n\}$. Then these dynamics can be written $\frac{d}{dt}\mathbf{r} = \mathbf{L}^{\mathbf{r}}\mathbf{r}$ with $\mathbf{L}^{\mathbf{r}} = \mathbf{C}^{\mathbf{r}}\mathbf{A}$.

One may also consider two alternative bases for the representative weights. First, $\mathbf{v} = \mathbf{A}^{\frac{1}{2}}\mathbf{r}$ is the *symmetrized* basis (the reason for this name follows); here, $\mathbf{A}^{\frac{1}{2}} \equiv \mathrm{diag}\{A_n^{\frac{1}{2}}\}$. Second, $\mathbf{t} = \mathbf{A}\mathbf{r}$ is the *total* represented weight (this representation is used in [24]-[28], where $\mathbf{t}$ is called $\mathbf{S}$). The dynamics in these bases are given by: $\frac{d}{dt}\mathbf{v} = \mathbf{L}^{\mathbf{v}}\mathbf{v}$, with $\mathbf{L}^{\mathbf{v}} = \mathbf{A}^{\frac{1}{2}}\mathbf{C}^{\mathbf{r}}\mathbf{A}^{\frac{1}{2}}$; and $\frac{d}{dt}\mathbf{t} = \mathbf{L}^{\mathbf{t}}\mathbf{t}$, with $\mathbf{L}^{\mathbf{t}} = \mathbf{A}\mathbf{C}^{\mathbf{r}}$. If the operator $\mathbf{C}^{\mathbf{w}}$ is symmetric, the operator $\mathbf{L}^{\mathbf{v}}$ is also symmetric, hence the name "symmetrized" for the $\mathbf{v}$ representation.

The constrained dynamics $\frac{d}{dt}\mathbf{w} = \mathbf{P}^{\mathbf{w}}\mathbf{C}^{\mathbf{w}}\mathbf{w}$ can similarly be translated into constrained dynamics in each representative basis: $\frac{d}{dt}\mathbf{r} = \mathbf{P}^{\mathbf{r}}\mathbf{L}^{\mathbf{r}}\mathbf{r}$, etc. Each projection operator can be written $\mathbf{P} = \mathbf{1} - \frac{\mathbf{s}\mathbf{c}^{\mathrm{T}}}{\mathbf{s}\cdot\mathbf{c}}$ for some subtracted vector $\mathbf{s}$ and constraint vector $\mathbf{c}$ (section 2.2). To describe constraints in the representative bases, we will describe the vectors $\mathbf{s}$ and $\mathbf{c}$ that form the projection operator in each representation.

First, consider a type 1 constraint. Let $\mathbf{n}^{\mathbf{w}}$ by the constraint vector in the $\mathbf{w}$ representation. Define $\mathbf{n}^{\mathbf{r}}$ to be the vector corresponding to $\mathbf{n}^{\mathbf{w}}$ in the $\mathbf{r}$ representation, formed by averaging in the same way that $\mathbf{r}$ is formed from $\mathbf{w}$. Thus, if $\mathbf{n}^{\mathbf{w}}$ is the DC vector $(1, 1, \ldots, 1)^{\mathrm{T}}$, then $\mathbf{n}^{\mathbf{r}}$ is also $(1, 1, \ldots, 1)^{\mathrm{T}}$ but with smaller dimension. The constraint vector $\mathbf{c}$ is found by writing the constraint in the appropriate representation: $\mathbf{n}^{\mathbf{w}} \cdot \mathbf{w} = \mathbf{n}^{\mathbf{r}} \cdot \mathbf{A}\mathbf{r} = \mathbf{A}\mathbf{n}^{\mathbf{r}} \cdot \mathbf{r}$ (we have assumed the n_i are approximately constant over a spatial grid position). Thus, in the $\mathbf{r}$ representation, the constraint vector $\mathbf{c}$ is $\mathbf{A}\mathbf{n}^{\mathbf{r}}$. The subtracted vector $\mathbf{s}$ in the $\mathbf{r}$ representation is determined by averaging the vector of the $\mathbf{w}$ representation: thus, the S1 subtracted vector is $\mathbf{n}^{\mathbf{r}}$, while the M1 subtracted vector is $\mathbf{r}$. Equations in the $\mathbf{v}$ and $\mathbf{t}$ bases are found by a change of basis from the $\mathbf{r}$ equations.

M2 constraints cannot be directly translated from the $\mathbf{w}$ representation without knowing the variance in the synaptic weights represented at each spatial grid position. So, we instead consider a similar constraint that conserves $\mathbf{t} \cdot \mathbf{t} = \mathbf{A}^2\mathbf{r} \cdot \mathbf{r}$. Then the $\mathbf{s}$ and $\mathbf{c}$ vectors for each constraint in each representation and basis are:

	S1				M1				M2			
	$\mathbf{w}$	$\mathbf{r}$	$\mathbf{v}$	$\mathbf{t}$	$\mathbf{w}$	$\mathbf{r}$	$\mathbf{v}$	$\mathbf{t}$	$\mathbf{w}$	$\mathbf{r}$	$\mathbf{v}$	$\mathbf{t}$
$\mathbf{s}$	$\mathbf{n}^{\mathbf{w}}$	$\mathbf{n}^{\mathbf{r}}$	$\mathbf{A}^{\frac{1}{2}}\mathbf{n}^{\mathbf{r}}$	$\mathbf{A}\mathbf{n}^{\mathbf{r}}$	$\mathbf{w}$	$\mathbf{r}$	$\mathbf{v}$	$\mathbf{t}$	$\mathbf{w}$	$\mathbf{r}$	$\mathbf{v}$	$\mathbf{t}$
$\mathbf{c}$	$\mathbf{n}^{\mathbf{w}}$	$\mathbf{A}\mathbf{n}^{\mathbf{r}}$	$\mathbf{A}^{\frac{1}{2}}\mathbf{n}^{\mathbf{r}}$	$\mathbf{n}^{\mathbf{r}}$	$\mathbf{n}^{\mathbf{w}}$	$\mathbf{A}\mathbf{n}^{\mathbf{r}}$	$\mathbf{A}^{\frac{1}{2}}\mathbf{n}^{\mathbf{r}}$	$\mathbf{n}^{\mathbf{r}}$	$\mathbf{w}$	$\mathbf{A}^2\mathbf{r}$	$\mathbf{A}\mathbf{v}$	$\mathbf{t}$

Now we determine whether $\mathbf{D}$ has positive eigenvalues. If $\mathbf{e}^a$ is not the principal eigenvector of $\mathbf{C}$, then $\mathbf{D}$ has a positive eigenvalue and the fixed point is unstable. If $\mathbf{e}^a$ is the principal eigenvector of $\mathbf{C}$, and it is nondegenerate (no other eigenvector has the same eigenvalue), then all eigenvalues of $\mathbf{D}$ except perhaps a zero corresponding to $\mathbf{e}^a$ are negative; so the fixed point is stable. If $\mathbf{C}$ has N degenerate principal eigenvectors, and $\mathbf{e}^a$ is one of them, then $\mathbf{D}$ has $N-1$ zeros corresponding to perturbations within the degenerate subspace: the principal eigenvector fixed points are thus marginally stable (eigenvalue 0) to perturbations within this subspace, and stable to other perturbations. •

Theorem 4 *Under a subtractively enforced constraint, if* $\mathbf{C}$ *has at least two eigenvectors with positive eigenvalues, then any interior fixed point is unstable.*

Proof. We consider a subtractively constrained equation (Eq. 4): $\frac{d}{dt}\mathbf{w} = \mathbf{C}\mathbf{w} - \epsilon(\mathbf{w})\mathbf{n}$. The linearized operator is $\mathbf{D} = \mathbf{C} - \mathbf{n}[\nabla\epsilon(\mathbf{w}^{\mathrm{FP}})]^{\mathrm{T}}$. Let $\mathbf{e}^{\perp}$ be a linear combination of the two eigenvectors of $\mathbf{C}$ with positive eigenvalues, such that $\mathbf{e}^{\perp}\cdot\mathbf{n} = 0$. Then $\mathbf{e}^{\perp}\mathbf{D}\mathbf{e}^{\perp} > 0$; so $\mathbf{D}$ must have a positive eigenvalue. •

Theorem 5 *Let* i *and* j *be indices in the synaptic basis. Suppose, for all* i *and* j, $C_{ii} > 0$ *and* $C_{ii}C_{jj} > C_{ij}C_{ji} > 0$. *Then under a subtractively enforced constraint, either all synapses or all but one are saturated in a stable final condition.*

Proof. This is a generalization of a similar proof in [18]. Suppose $\mathbf{C}\mathbf{w}^{\mathrm{FP}} - \epsilon(\mathbf{w})\mathbf{n} =$ "$\mathbf{0}$" when synapses w_i^{FP} and w_j^{FP} are both not saturated. By "$\mathbf{0}$", we mean that each component of the vector is either 0, or else of a sign that would take an already saturated synapse beyond its limiting value. Let $\hat{\mathbf{u}}^i$ be the unit vector with i^{th} weight 1 and all other elements 0, and similarly for $\hat{\mathbf{u}}^j$. The action of $\mathbf{C}$ in the $\hat{\mathbf{u}}^i/\hat{\mathbf{u}}^j$ plane is given by the matrix $\begin{pmatrix} C_{ii} & C_{ij} \\ C_{ji} & C_{jj} \end{pmatrix}$. The eigenvalues of this matrix are both real and positive when the conditions of the theorem are met. $\mathbf{D}$ when acting in the $\hat{\mathbf{u}}^i/\hat{\mathbf{u}}^j$ plane is as given in proof of theorem 4 (except that the function $\epsilon(\mathbf{w})$ might be altered when some synapses are saturated). So if $\mathbf{e}^{\perp}$ is a linear combination of $\hat{\mathbf{u}}^i$ and $\hat{\mathbf{u}}^j$ with $\mathbf{e}^{\perp}\cdot\mathbf{n} = 0$, then $\mathbf{e}^{\perp}\mathbf{D}\mathbf{e}^{\perp} > 0$. •

B Constraints in Alternative Synaptic Representations

Suppose that the correlation C_{ij} between two inputs is a function of their spatial positions: $C_{ij} = C(\mathbf{x}_i, \mathbf{x}_j)$ where $\mathbf{x}_i$ is the position of the i^{th} input. The $\mathbf{x}_i$ may not have a regular spatial arrangement: for example, a cell may receive a greater density of inputs from the center of its receptive field than from the periphery. In this case, it is often convenient to use an alternative representation in which the synaptic variables are arrayed in a regular spatial grid, and the density of synapses at each grid position is described by a synaptic density function or 'arbor function'.

A Proofs of Mathematical Results

We study dynamics confined to a constraint surface and governed by Eqs. 3 or 4. As in the text, we use indices $a, b, \ldots$ to refer to the eigenvector basis of $\mathbf{C}$. We assume that $\mathbf{C}$ has a complete set of orthonormal eigenvectors $\mathbf{e}^a$ with corresponding eigenvalues λ^a.

Write the constrained equation (Eq. 3 or 4) as $\frac{d}{dt}\mathbf{w} = \mathbf{f}(\mathbf{w})$. To determine the stability of a fixed point $\mathbf{w}^{\mathrm{FP}}$ (where $\mathbf{f}(\mathbf{w}) = \mathbf{0}$), we linearize $\mathbf{f}(\mathbf{w})$ about the fixed point. Call this linearized operator $\mathbf{D}$; in the eigenvector basis of $\mathbf{C}$, it is a matrix with elements $D_{ab} = \frac{\partial f_a(\mathbf{w})}{\partial w_b}|_{\mathbf{w}=\mathbf{w}^{\mathrm{FP}}}$. We define the constraint plane to be the hyperplane tangent to the constraint surface at the fixed point, and the constraint vector $\mathbf{c}$ to be the vector normal to the constraint plane. $\mathbf{c}$ is a left eigenvector of $\mathbf{D}$.[13]

We assume that multiplicative or subtractive confinement of the dynamics to the constraint surface means two things. First, $\mathbf{D}$ has one zero or negative eigenvalue corresponding to the enforcement of the constraint, with associated left eigenvector $\mathbf{c}$. Therefore any right eigenvector of $\mathbf{D}$ with positive eigenvalue is parallel to the constraint plane. Second, the constraint plane is not parallel to the subtracted vector that enforces the constraint ($\mathbf{w}$ for multiplicative enforcement, or $\mathbf{n}$ for subtractive).

The stability of a fixed point is determined by the eigenvalues of $\mathbf{D}$ [12]. If one eigenvalue is positive, the fixed point is unstable; if all eigenvalues are negative, the fixed point is stable. In assessing the outcome of the constrained dynamics, we are only concerned with stability of the fixed point to perturbations within the constraint surface. Thus, if all eigenvalues are negative except one zero eigenvalue corresponding to a direction outside the constraint surface, then the fixed point is stable.

We hereafter use eigenvector to mean right eigenvector. We will use the gradient operator, defined by $\nabla g(\mathbf{x}) \equiv \sum_a \mathbf{e}^a \frac{\partial g(\mathbf{w})}{\partial w_a}|_{\mathbf{w}=\mathbf{x}}$.

Theorem 3 *Under a multiplicatively enforced constraint, if the principal eigenvector of* $\mathbf{C}$ *is an interior fixed point it is stable. Interior fixed points that are non-principal eigenvectors are unstable.*

Proof. We consider a multiplicatively constrained equation (Eq. 3): $\frac{d}{dt}\mathbf{w} = \mathbf{C}\mathbf{w} - \gamma(\mathbf{w})\mathbf{w}$. A fixed point is an eigenvector of $\mathbf{C}$: $\mathbf{w}^{\mathrm{FP}} = w_a \mathbf{e}^a$ for some a, $\gamma(\mathbf{w}^{\mathrm{FP}}) = \lambda^a$. The linearized operator is $\mathbf{D} = \mathbf{C} - \lambda^a \mathbf{1} - w_a \mathbf{e}^a [\nabla\gamma(\mathbf{w}^{\mathrm{FP}})]^{\mathrm{T}}$. In the eigenvector basis of $\mathbf{C}$, $\mathbf{D}$ is a diagonal matrix with the addition of one row of off-diagonal elements; such a matrix has the same eigenvalues as the diagonal matrix alone (because the characteristic equation, $\det(\mathbf{D} - \lambda\mathbf{1}) = 0$, is unchanged by the additional row). The diagonal part of $\mathbf{D}$ is the matrix $\mathbf{C} - \lambda^a \mathbf{1} - \kappa \mathbf{e}^a [\mathbf{e}^a]^{\mathrm{T}}$, where $\kappa \equiv w_a \frac{\partial\gamma(w)}{\partial w_a}|_{\mathbf{w}=w_a\mathbf{e}^a}$. This is the operator $\mathbf{C}$ with eigenvalue λ^a reduced to $-\kappa$ and all other eigenvalues reduced by λ^a. Note that $\mathbf{e}^a$ is the eigenvector of $\mathbf{D}$ with eigenvalue $-\kappa$; since $\mathbf{e}^a$ is not parallel to the constraint plane, $-\kappa \leq 0$.

[13] *Proof*: For any $\Delta\mathbf{w}$ in the constraint plane, $\mathbf{D}\Delta\mathbf{w}$ must remain within the constraint plane; that is, $\mathbf{c}^{\mathrm{T}}\mathbf{D}\Delta\mathbf{w} = 0$ for all $\Delta\mathbf{w}$ satisfying $\mathbf{c}^{\mathrm{T}}\Delta\mathbf{w} = 0$. Therefore, $\mathbf{c}^{\mathrm{T}}\mathbf{D} \propto \mathbf{c}^{\mathrm{T}}$.

learning equation for which the unconstrained time development operator has a complete set of eigenvectors and is linear over a significant region. The instability of subtractive constraints also depends on the unconstrained operator having two positive eigenvalues, which is typical for Hebbian learning. An interesting alternative is *anti*-Hebbian learning [29]: in this case, all unconstrained eigenvalues are reversed in sign from the Hebbian case, so typically no eigenvalue is positive. Our analysis applies to this case also: multiplicatively constrained dynamics flow to the principal eigenvector, which is the vector that would have the *smallest* eigenvalue under Hebbian dynamics [29]; while subtractively constrained dynamics flow *to* the fixed point, which is stable.

Multiplicative and subtractive constraints represent two fundamentally different methods of controlling the size of the weight vector. Multiplication equally rescales all weight patterns, while subtraction directly acts on only a single weight pattern. Because this difference is general, many of the results we have found for the linear case may generalize to cases involving nonlinear rules.

Biologically, there is as yet little evidence as to the mechanisms that lead activity-dependent plasticity to be competitive or to achieve selectivity. Among the two choices of subtractive and multiplicative constraints, subtractive seem to resemble biology more closely in systems where sharp receptive fields are achieved, and in visual cortex where ocular dominance columns are likely to develop without requiring anticorrelation between the eyes; while multiplicative constraints might resemble biology more closely in situations like adult cortical plasticity where continually moving and graded representations may occur [15]. We do not advocate that one or the other of these is the biologically correct choice. Rather, we wish: (1) to point out that different choices of competitive mechanism can yield different outcomes, so it is important for the modeler to know whether and how their results depend on these choices; and (2) to begin to distinguish and characterize different classes of such mechanisms, which might then be compared to biology.

$\sum_j w_j = k_1$. Assuming $\langle x \rangle > x_0$, the condition $k_2 < 0$ implies that $\langle x \rangle < x_\theta$. That is, average or weakly activated inputs should lose synaptic strength when the postsynaptic cell is highly activated. This is not the case in at least one system, LTP in hippocampus [7].

Bienenstock *et al.* [1] proposed that the postsynaptic threshold for synaptic plasticity y_θ (Eq. 12) should depend nonlinearly on the average postsynaptic activity $\langle y \rangle$: $y_\theta = \langle y \rangle^2 / y_{\rm set}$ where $y_{\rm set}$ is a cell's "preset" desired activity level. They refer to this form of postsynaptic threshold as a "sliding threshold". It has the effect of adjusting synaptic strengths to achieve $\langle y \rangle \approx y_{\rm set}$. They set $y = \mathbf{w} \cdot \mathbf{x}$, so this sliding threshold conserves $\mathbf{w} \cdot \langle \mathbf{x} \rangle \approx y_{\rm set}$. It thus is an alternative method of achieving a constraint like that of S1 or M1, $\mathbf{w} \cdot \mathbf{n}$ =constant, but with constraint vector $\mathbf{n} = \langle \mathbf{x} \rangle$. Bienenstock *et al.* used a nonlinear function $\phi(y - y_\theta)$ in place of $(y - y_\theta)$ in Eq. 12. A more recent formulation [14] sets $y_\theta = \langle y^2 \rangle / y_{\rm set}$. It will be interesting to determine how each of these rules compares to S1 and M1 on simple problems such as those of Fig. 1. Recent results both in the peripheral auditory system [45] and in hippocampus [13] are suggestive that increased neural activity may elevate a threshold for modification in a manner specific to those inputs whose activity was increased; this is suggestive of an increase in x_θ that is specific to activated inputs, rather than an increase in y_θ, which would elevate thresholds for all inputs.

Covariance models [37, 38] have been proposed to solve the problem that chance coincidences drive synapses to saturate at $w_{\rm max}$. This problem was known to occur in Hebb models where modification simply depends on the product of input and output activities, $\frac{d}{dt} w_i \propto y x_i$. In a covariance model, one instead takes $\frac{d}{dt} w_i$ proportional to the covariance of y and x_i, $\frac{d}{dt} w_i \propto (y - \langle y \rangle)(x_j - \langle x \rangle)$. This is a Hebb model with linearly sliding thresholds: $y_\theta = \langle y \rangle$, $x_\theta = \langle x \rangle$. With these choices and no intrinsic decay in the Hebb rule, we obtain $k_2 = 0$ and $n = 0$ in Eq. 13, so synaptic growth is driven by the simple Hebbian equation $\frac{d}{dt} \mathbf{w} = \mathbf{Q}\mathbf{w}$ without decays. Thus, the problem of loss of selectivity under a Hebbian rule is not avoided by a linear covariance rule.

5 Discussion

We have demonstrated that multiplicative and subtractive constraints can lead to fundamentally different outcomes in linear learning. Under multiplicative constraints, the weight vector tends to the principal eigenvector of the unconstrained time development operator. This is a "graded" receptive field in which most mutually correlated inputs are represented. Thus, when two equally active eyes compete, both retain equal innervation unless the two eyes are anticorrelated. Under subtractive constraints, the weight vector tends to a receptive field that is "sharpened" to a subset of maximally-correlated inputs: the weights of these inputs reach the maximum allowed strength, while all other weights reach the minimum allowed strength. When two eyes compete, subtractive constraints can lead to domination by one eye (ocular dominance segregation) provided only that correlations within one eye are stronger than those between the eyes.

Though we have worked in the context of linear Hebbian learning, these results apply to any

in [39]). Increased activation might also increase decay by increasing a threshold for synaptic modification; this is discussed in section 4.3.

As simple-minded models of decay, we propose several pictures. First, a synapse's strength might be proportional to its number of postsynaptic receptors. If loss of each of those receptors is an independent event, then decay will occur proportionate to the synapse's strength, yielding multiplicative decay. Replacement would occur at a rate determined by the Hebb rule. Second, a synapse's strength might depend upon the portion of a fixed number of molecules that have been phosphorylated or otherwise activated . Such a model has been proposed in which the molecules are calcium/calmodulin-dependent protein kinase, which is concentrated at the postsynaptic density [19]. Suppose there are m unactivated and m^* activated molecules, $m + m^* = M$. Suppose that Hebbian modification results in promotion of m to m^*,[12] while decay of m^* to m is activity independent and occurs at a rate $-\kappa m^*$. If synaptic strength is linear in m and m^*, $w_i = am_i + bm_i^*$, then this decay is $-\kappa(w_i - aM)/(b-a)$. This is a mixture of subtractive and multiplicative decay. Third, decay might occur when either postsynaptic or presynaptic activity fails to reach a threshold level for positive Hebbian modification. This results in subtractive decay. We examine this in more detail in the next section.

4.3 Use of thresholds to achieve selectivity

Consider a linear activation rule and a linear Hebbian rule for a single output cell:

$$y(t) = \sum_i w_i(t)\,[x_i(t) - x_0] + y_0 \qquad \text{(Activation)} \tag{11}$$

$$\frac{d}{dt} w_i(t) = \lambda[y(t) - y_\theta][x_i(t) - x_\theta] - \gamma w_i - \epsilon \qquad \text{(Hebb)} \tag{12}$$

Here, $y(t)$ is the activation of the output cell, $\mathbf{x}(t)$ the vector of input activations. x_0 and y_0 are baseline levels of input and output activity, respectively. x_θ and y_θ are threshold activity levels for Hebbian plasticity. ϵ and γw_i are synaptic decay terms.

After substituting Eq. 11 into Eq. 12 and averaging over input patterns, and assuming that all inputs have identical mean activity, the thresholds lead to decay terms that are identical for all synapses (subtractive decay). The resulting equation is

$$\frac{d}{dt} w_i(t) = \sum_j Q_{ij} w_j + k_2 \sum_j w_j - n - \gamma w_i \tag{13}$$

where $Q_{ij} = \lambda \left\langle (x_i - \langle x \rangle)(x_j - \langle x \rangle) \right\rangle$, $k_2 = \lambda\,(\langle x \rangle - x_\theta)\,(\langle x \rangle - x_0)$, and $n = \epsilon + \lambda\,(\langle x \rangle - x_\theta)\,(y_\theta - y_0)$. Both the k_2 term and the n term are subtractive decay terms.

Eq. 13 can be rewritten $\frac{d}{dt} w_i(t) = \sum_j Q_{ij} w_j + k_2(\sum_j w_j - k_1) - \gamma w_i$ where $k_1 = n/k_2$. As discussed in section 4.1, for k_2 large and negative this becomes equivalent to S1 constraints conserving

[12] Actually this should make Hebbian modification proportional to m, which will complicate the plasticity equation.

transient in which the constraint surface is reached, has been shown in [20, Appendix E]. This form of S1 constraint was used by Linsker in [18]. Multiplicative constraints can be similarly formulated: M1 constraints are achieved by setting $\gamma(\mathbf{w}) = k_2(\mathbf{n} \cdot \mathbf{w} - k_1)$ provided $k_2 \mathbf{n} \cdot \mathbf{w} > 0$ always (examine $\mathbf{n} \cdot \frac{d}{dt}\mathbf{w}$). M2 constraints are achieved by setting $\gamma(\mathbf{w}) = |k_2|(\mathbf{w} \cdot \mathbf{w} - k_1)$.

Dynamics that converge to a multiplicative constraint can also be formulated by substituting a constant $k > 0$ for the denominators in Eqs. 5 or 7. Let $\mathbf{c}$ be the constraint vector: for M1, $\mathbf{c} = \mathbf{n}$; for M2, $\mathbf{c} = \mathbf{w}$. Then such convergent dynamics are achieved by setting $\gamma(\mathbf{w}) = \mathbf{c} \cdot \mathbf{C}\mathbf{w}/k$. If the initial condition and dynamics maintain $\mathbf{c} \cdot \mathbf{C}\mathbf{w} > 0$, then the dynamics will flow to the constraint surface $\mathbf{c} \cdot \mathbf{w} = k$ and remain stable to perturbations off it thereafter (as can be seen by examining $\mathbf{c} \cdot \frac{d}{dt}\mathbf{w}$). Oja [31] used such M2 constraints with $k = 1$ and proved convergence to the principal eigenvector. The self-organizing feature maps of Kohonen [17] also are stabilized in a convergent, multiplicative manner.[11]

4.2 Biological implementations of constraints

Multiplicative or subtractive constraints in their general form (Eqs. 3–4) can be implemented as follows: (1) each of a cell's synaptic weights w_i undergoes decay: the rate of its decay, relative to that of other synapses on the cell, is either proportional to its strength w_i (multiplicative constraints) or given by a constant n_i (subtractive constraints); (2) the overall gain of this decay, $\gamma(\mathbf{w})$ or $\epsilon(\mathbf{w})$, is set for the cell as a whole to achieve the constraint. The separation of the relative decay rate, determined at each synapse, from the overall gain, determined cellularly, allows constraints to be implemented without a cellular assessment of the individual strength of each synapse.

We propose two simple models by which the overall gain of the decay might increase with the total synaptic strength supported by a cell. First, a cell might have a limited capacity to metabolically supply its synapses, so that greater total synaptic strength would mean less supply and thus faster decay for each synapse. This could apply either to the total synaptic strength projected by a cell, or to the total strength received by a cell. Second, the overall rate of decay might increase with a cell's average degree of activation, which in turn would increase with the total synaptic strength received by a cell. Increased activation could lead to greater overall decay by increasing expression or release of a molecule that destabilizes or removes synapses, such as a protease; or by decreasing expression or release of a molecule that stabilizes or adds synapses, such as a trophic, adhesion, or sprouting factor. Many such mechanisms have been proposed (reviewed

[11] Kohonen's equation for the weights $\mathbf{w}$ onto one output cell is $\frac{d}{dt}\mathbf{w} = y[\mathbf{x} - \mathbf{w}]$, where $\mathbf{x}$ is the vector of input activities, y the output cell activity, and y is nonzero only for the most-activated postsynaptic cell and its neighbors. The second, stabilizing term is proportional to $\mathbf{w}$, so the stabilization is multiplicative. The nature of the weight stabilization depends in part on the normalization of input activities. If $\mathbf{n} \cdot \mathbf{x}$=constant, then eventually $\mathbf{n} \cdot \mathbf{w} = \mathbf{n} \cdot \mathbf{x}$ (examine $\mathbf{n} \cdot \frac{d}{dt}\mathbf{w}$). A more common choice is $\mathbf{x} \cdot \mathbf{x}$=constant. Then eventually $\mathbf{w} \cdot \mathbf{w} \leq \mathbf{x} \cdot \mathbf{x}$ (if $\mathbf{w} \cdot \mathbf{w} > \mathbf{x} \cdot \mathbf{x}$, then $\frac{d}{dt}(\mathbf{w} \cdot \mathbf{w}) < 0$); and we expect that eventually $\mathbf{w} \cdot \mathbf{w} \approx \mathbf{x} \cdot \mathbf{x}$ since the winner-take-all activation rule eventually leads only one or a small range of $\mathbf{x}$ to activate the cell, and $\mathbf{w}$ comes to match that $\mathbf{x}$.

In all cases, saturation of all synapses is expected without constraints or under subtractive constraints, but not under multiplicative constraints.

Several cautions must be emphasized about this approach. First, it predicts only the wavelength of weight alternation, and does not distinguish between different weight structures with similar wavelength. Second, the approach is heuristic: its validity must be checked in any particular case. The final weight pattern is expected to be one in which the dominant PF and RF patterns are "knitted together" into a compatible overall pattern. In some cases such a "knitting" may not be possible: the PF and RF patterns predicted above may be incompatible, in which case the proposed heuristics will fail.

We can apply this analysis to understand the effects of subtractive input constraints on ocular dominance segregation. Consider the development of the difference $\mathbf{w}^{\mathrm{D}}$ between projections from the two eyes (section 2.5). An RF across which $\mathbf{w}^{\mathrm{D}}$ is all-positive or all-negative is a monocular receptive field. When receptive fields are monocular, an oscillation across PFs of $\mathbf{w}^{\mathrm{D}}$ corresponds to the oscillation between ocular dominance columns [24] (Fig. 4). Subtractive *output* constraints have no effect on the development of $\mathbf{w}^{\mathrm{D}}$: such constraints affect only the sum, not the difference, of the two projections (section 2.5). So, a monocular, DC receptive field pattern of $\mathbf{w}^{\mathrm{D}}$ can develop despite subtractive constraints on output cells, provided that the corresponding correlation function ($C^{\mathrm{D}} \equiv C^{\mathrm{LL}} - C^{\mathrm{LR}}$) has the peak of its Fourier transform at frequency 0. Subtractive *input* constraints do affect $\mathbf{w}^{\mathrm{D}}$: such constraints separately conserve the total strength from the left-eye input and from the right-eye input at each position, and thus conserve the total difference as well as the total sum from each position. So subtractive input constraints ensure that there is an oscillation of $\mathbf{w}^{\mathrm{D}}$ across PFs with wavelength no larger than an arbor diameter. Thus, subtractive input constraints determine the width of a left-eye plus a right-eye ocular dominance column to be an arbor diameter when the unconstrained dynamics would lead to larger columns, but have little effect on ocular dominance segregation otherwise [27].

4 How Can Constraints Be Implemented?

4.1 Learning rules that converge to constrained dynamics

The formulations in Eqs. 5-7 confine the dynamics to the constraint surface that contains the initial weight vector. Alternatively, constraints may be formulated so that the dynamics converge from an arbitrary initial weight vector to one particular constraint surface, and remain on that constraint surface thereafter. In this case the dynamics are described by Eqs. 5–7 after an initial transient in which the constraint surface is reached.

Such a formulation of S1 constraints is obtained by setting $\epsilon(\mathbf{w}) = |k_2|(\mathbf{n} \cdot \mathbf{w} - k_1)$ for constants k_1 and k_2. When $|k_2|$ is large, this term enforces the constraint $\mathbf{n} \cdot \mathbf{w} = k_1$. Equivalence of this formulation to that in terms of projection operators, in the limit of large $|k_2|$ and after an initial

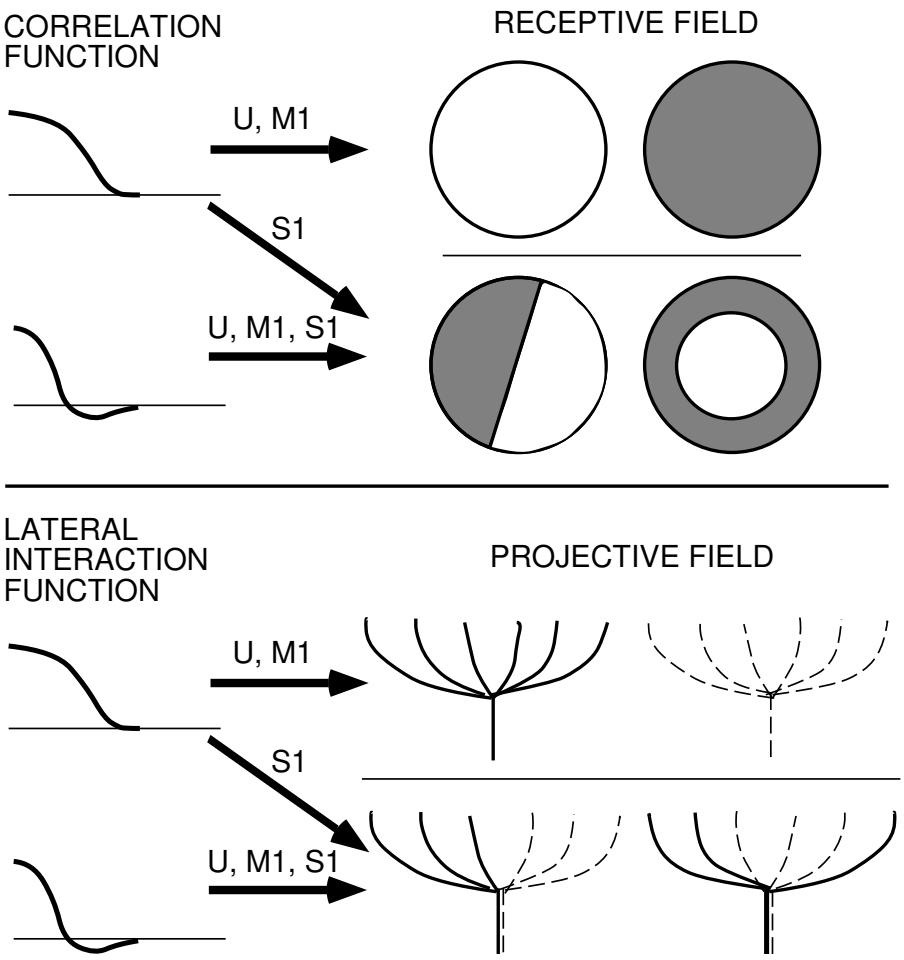

Figure 4: The Role of Constraints on Input and Output Cells: a Heuristic Approach
TOP: Output cell receptive fields (RFs) expected to develop under unconstrained dynamics ('U'), or under M1 or S1 constraints on output cells. White regions in receptive fields indicate positive weights; dark regions indicate zero or negative weights, depending on whether $w_{\min}$ is zero or negative. Correlations between input activities are shown as a function of input separation. Without constraints or under M1, the weights vary in sign to match the oscillation of the correlation function. Under S1, the weights always alternate in sign, with wavelength no larger than an arbor diameter. Note that this approach does not distinguish between different weight structures with similar wavelength of alternation, such as the two lower RFs. BOTTOM: Input cell projective fields (PFs) are determined in the same manner as RFs, except that: (1) the determining function is the output layer lateral interaction function; and (2) the determining constraints are those on input cells. Here, solid lines indicate positive weights, dotted lines indicate zero or negative weights.

In thinking about constraints on both input and output cells, the following heuristic approach appears to be useful. We refer to the projection to a single output cell as a receptive field or RF, and the projection from a single input location as a projective field or PF. The eigenvectors are patterns of weights across the entire network, not just across individual RFs or PFs. In simple Hebbian models, the dominant, fastest-growing patterns can often be characterized as follows.

First, in the absence of constraints, the RFs of the dominant patterns are primarily determined by a particular input correlation function, and the PFs of the dominant patterns are similarly determined by the output layer lateral interaction function [24]. If the correlations are all positive, the RFs have a single sign; if correlations oscillate in sign with input separation, the RFs oscillate in sign with a similar wavelength. A single-signed RF can be regarded as one that oscillates with an infinite wavelength, so we may summarize: in the absence of constraints, the RFs of the dominant patterns vary between positive and negative values with a wavelength corresponding to the peak of the Fourier transform of the appropriate input correlation function. Similarly, the PFs of the dominant patterns vary between positive and negative values with a wavelength corresponding to the peak of the Fourier transform of the output layer lateral interaction function.

Second, constraints on output cells appear only to affect the form of the individual RFs, while constraints on input cells only affect the form of the individual PFs. Consider the case of two layers that are topographically connected: each input cell initially makes synapses onto cells over a certain diameter ("arbor diameter") in the output layer, and adjacent input cells project adjacent arbors. Then output cells also receive connections over an arbor diameter from the input layer. Suppose that output or input cell constraints conserve total synaptic strength over the cell on which they act. Then an RF or PF that alternates in sign with a wavelength less than or equal to the arbor diameter is approximately AC, that is, it has summed synaptic strength near 0. An RF or PF that alternates with longer wavelength is not AC. Subtractive constraints selectively suppress the growth of non-AC patterns, whereas multiplicative constraints stabilize the growth of a dominant non-AC pattern. Thus, we arrive at the following heuristic rules for the wavelength with which RFs or PFs alternate in sign (Fig. 4):

1. If the dominant pattern in the absence of constraints has (RF,PF) wavelength larger than an arbor diameter, then

   (a) Subtractive (output,input) constraints suppress this pattern in favor of a pattern with (RF,PF) wavelength of an arbor diameter;

   (b) Multiplicative (output,input) constraints do not alter the (RF,PF) wavelength of this dominant pattern, but only stabilize its amplitude.

2. If the dominant pattern in the absence of constraints has (RF,PF) wavelength smaller than an arbor diameter, then (output,input) constraints, whether enforced multiplicatively or subtractively, will have little effect.