

Neural Networks Pt. 3 Hebbian Learning and Information Theory

Lecture 8 I400/I590

Artificial Life as an approach to Artificial Intelligence

Larry Yaeger

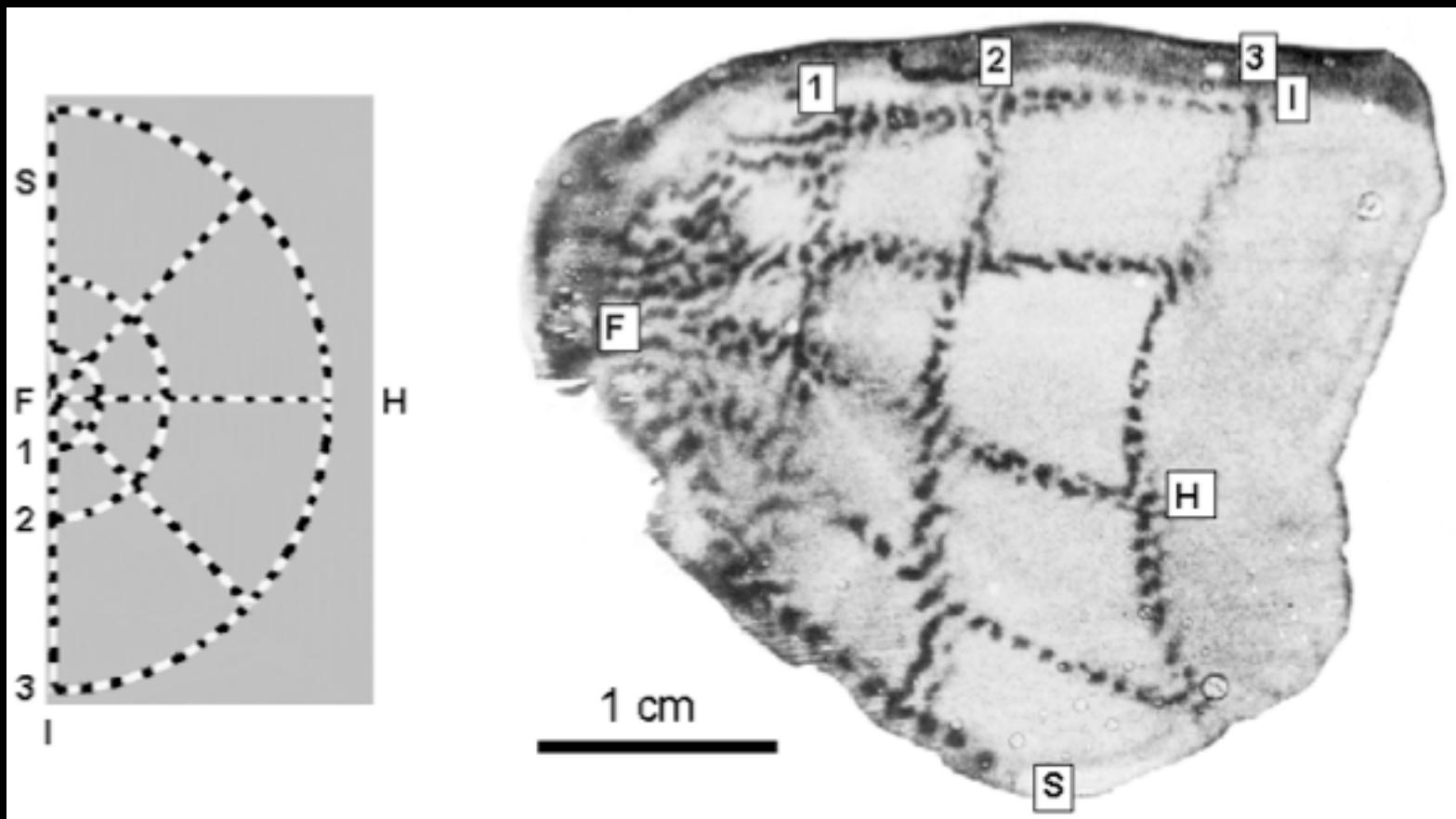
Professor of Informatics, Indiana University

Self-Organization

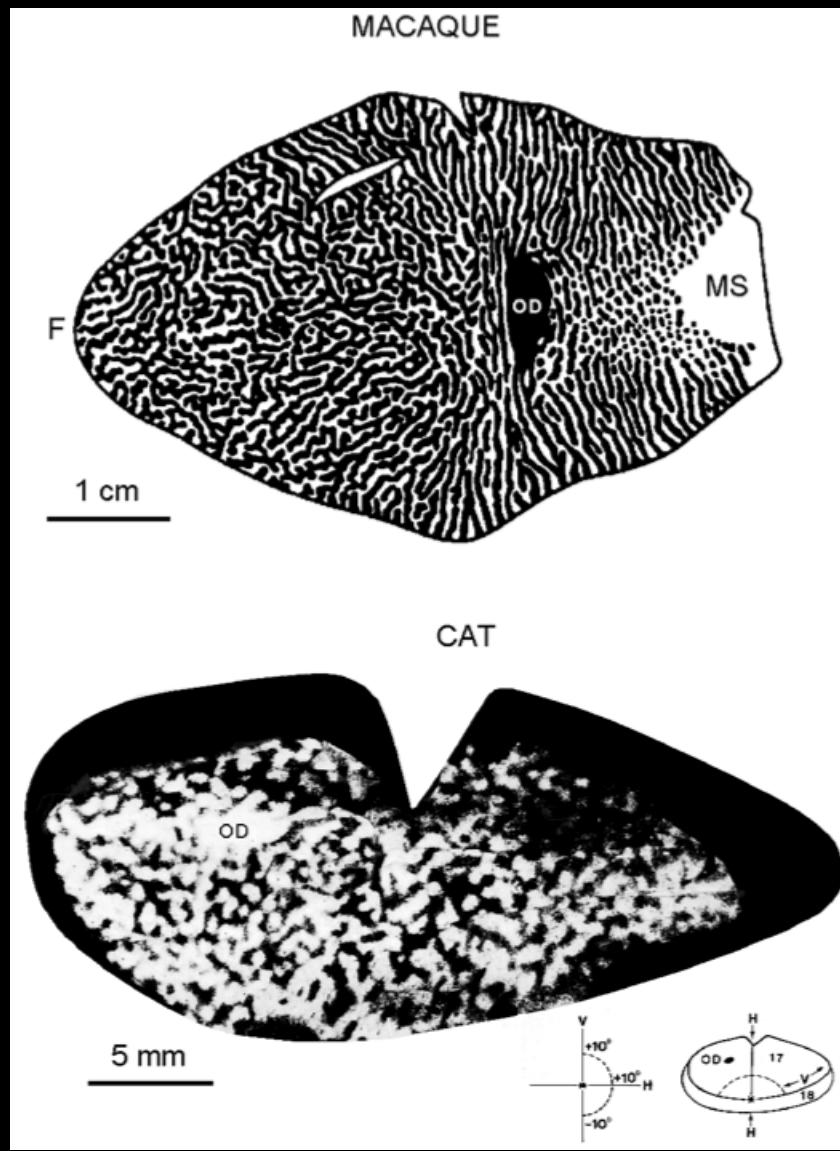
- To work with artificial neural networks that are biologically plausible, we require:
 - Unsupervised learning
 - Supervised algorithms unlikely to be present in the brain
 - Explicit training signals don't exist
 - Self-organization of brain maps
 - Consistent with the way processing is done in the brain
 - Should preferentially represent high-probability stimuli
 - Based on local interactions (no global control)
- Visual cortex is one of the most thoroughly studied brain regions, and makes an excellent source of inspiration and verification for brain models

Retinotopy

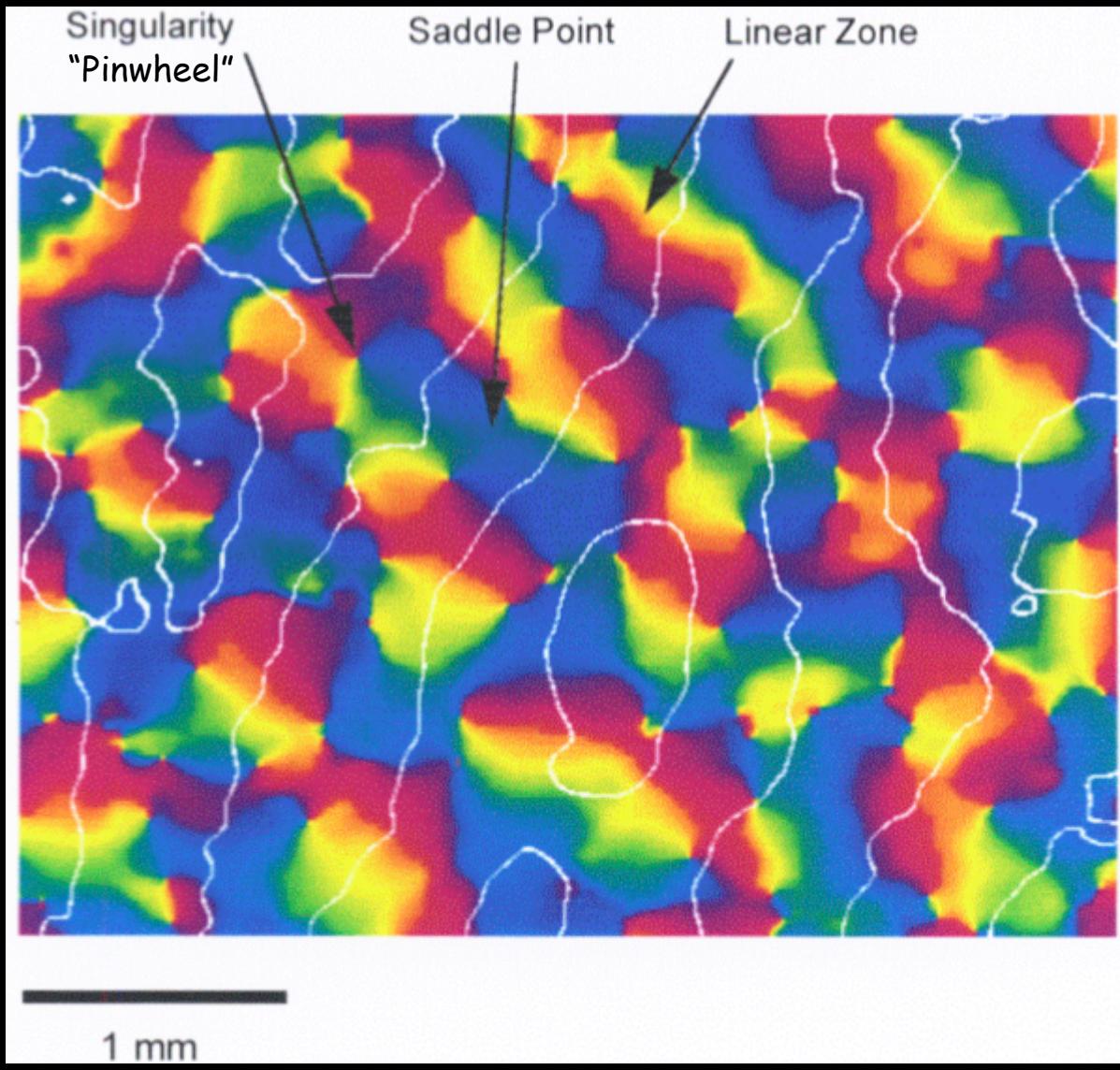
- Retinotopic map in a Macaque monkey



Occular Dominance



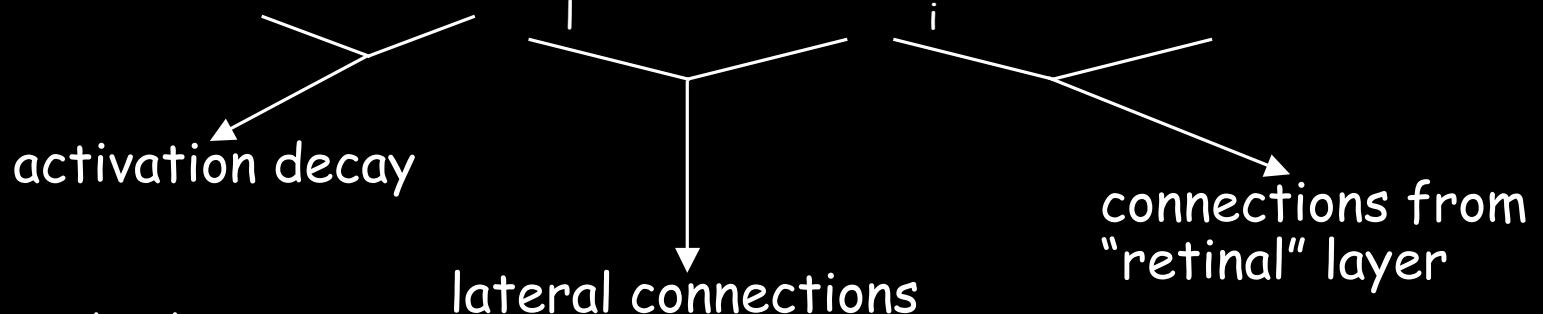
Orientation Columns



Modeling Visual Cortex

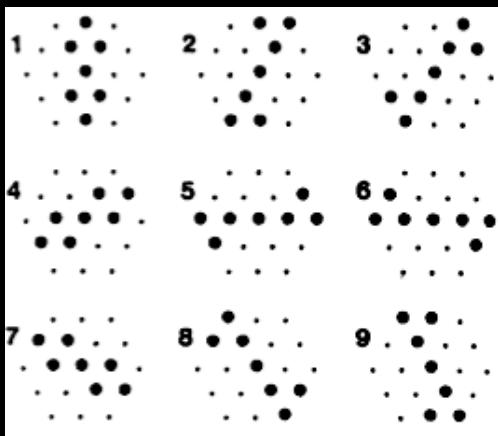
- von der Malsburg (1973) neuron activation rule:

$$dH_k(t)/dt = -\alpha_k H_k(t) + \sum_l p_{lk} H_l^*(t) + \sum_i s_{ik} A_i^*(t) \quad k = 1, \dots, N$$



$$\Delta s_{ik} \propto A^* H^*$$

input patterns



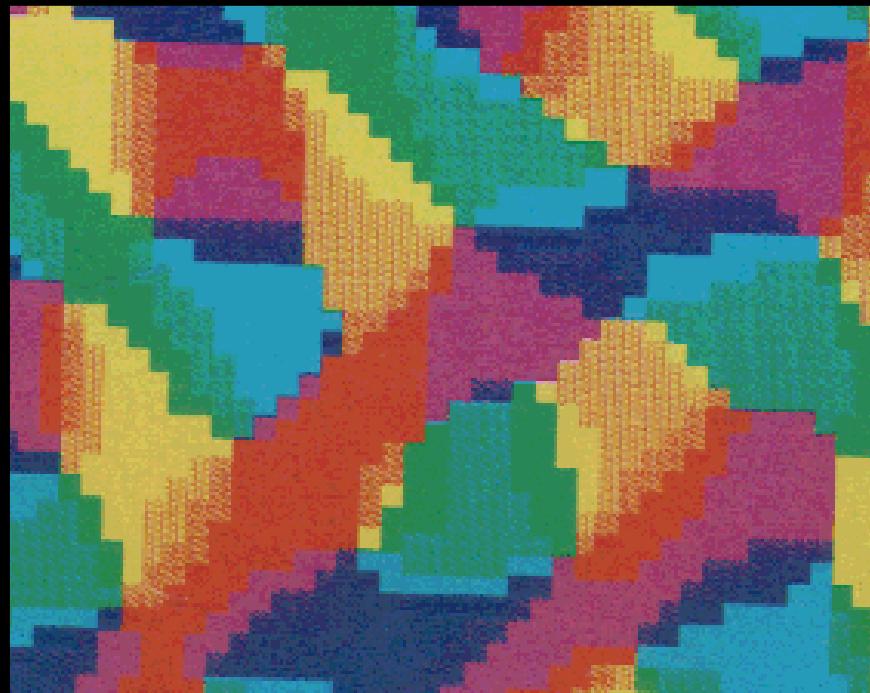
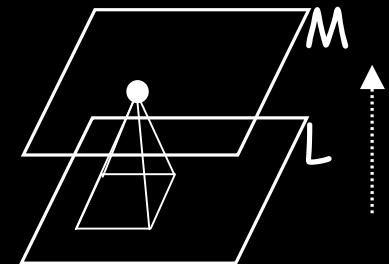
orientation preferences

Modeling Visual Cortex

- Ralph Linsker (1986)

$$X^M = \sum_{i \in L} w_i X_i^L + a_1$$

$$\Delta w_i = a_2 (X^M - a_3) (X_i^L - a_4) + a_5$$

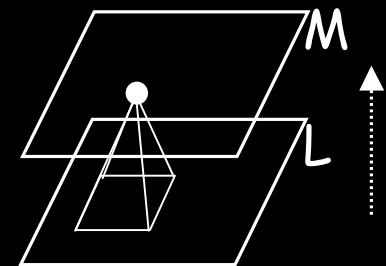


General Organizing Principle

- Linsker (1987)
"This principle of 'maximum information preservation' states that the signal transformation that is to be realized at each stage is one that maximizes the information that the output signal values (from that stage) convey about the input signal values (to that stage), subject to certain constraints and in the presence of processing noise. The quantity being maximized is a Shannon information rate."
- In fact, what Linsker maximizes in his analysis is the Mutual Information between the ensemble of input signals and the ensemble of output signals

Maximum Information Preservation

- Connections should develop in such a way that the signals transferred from L to M (in the presence of noise) produce signals in M that retain maximal information about L
 - Given M, want to be able to make the best guess about L
 - Minimize the amount of information conveyed by measuring the input values L, if you already know the output values M
 - Minimize the uncertainty in L, given M



Information Rate

- Regard the set of input signal values L (at a given time) as an input "message"
- This message is processed (by the connections between L and M) to produce an output message M
- The Shannon information rate (the average information transmitted from L to M per message) is:

$$R = \sum_L \sum_M P(L,M) \log[P(L,M) / P(L)P(M)]$$

- This is just the Mutual Information between the ensemble of input signals L and output signals M
 - $P(L)$ is the probability of the input message being L
 - $P(M)$ is the probability of the output message being M
 - $P(L,M)$ is the joint probability of the input being L and the output being M

Reverse: Minimize Uncertainty

- This information rate can be rewritten as

$$R = I_L - I_{L|M} = H(L) - H(L|M)$$

- Since I_L is fixed by the properties of the input ensemble, maximizing R means minimizing $I_{L|M}$
 - Minimize the uncertainty in input signal L , given output signal M

Forward: Maximize Uncertainty, Minimize Conditional Uncertainty

- This information rate can also be rewritten as

$$R = I_M - I_{M|L} = H(M) - H(M|L)$$

- To maximize R , want to simultaneously (if possible)

- Maximize I_M

- Make every possible message M equiprobable

- Minimize $I_{M|L}$

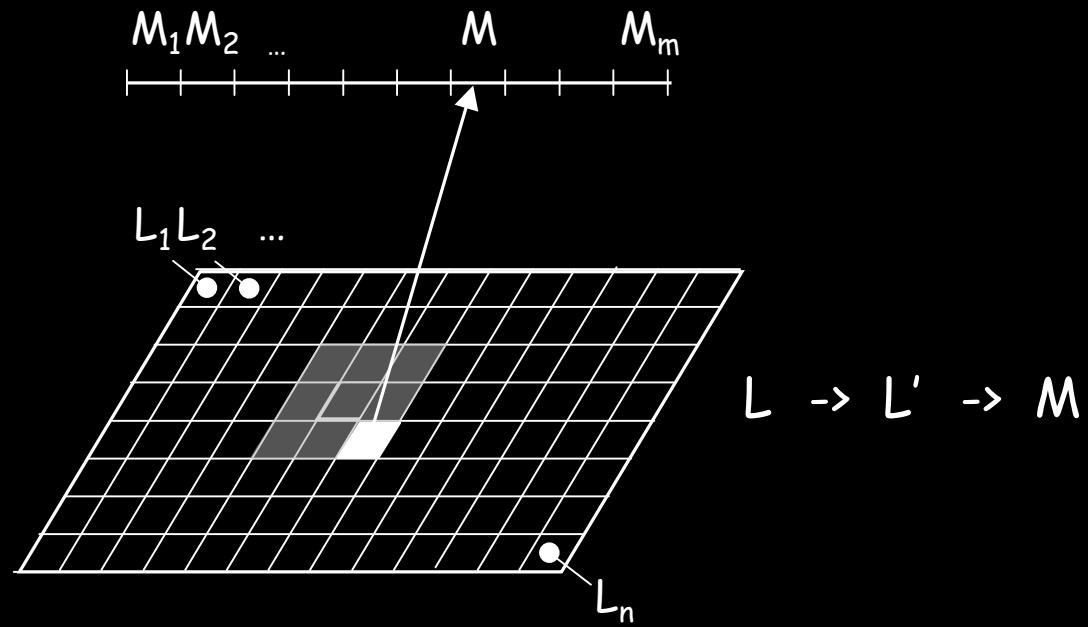
- Transform each unique L into a unique M (e.g., $M = L$)

- More generally, prefer any “sharpening” of the $P(M|L)$ distribution, so for each L , $P(M|L)$ is near zero except for a small set of messages M

A Geometric Model

- Let each L be a message in a low-dimensional space
 - For now, make it just two-dimensional
 - Extension to higher dimensionality is straightforward
- Let M be one of a number of discrete states
- Let the transformation $L \rightarrow M$ consist of two steps:
 - A noise process transforms $L \rightarrow L'$, with transformed messages lying within a radius of ν centered on L
 - The altered message L' is mapped deterministically onto one of the output messages M (this is $L' \rightarrow M$)

A Simplified Geometric Model



Geometric Partitioning

- A given $L' \rightarrow M$ mapping corresponds to a partitioning of the L space into regions labeled by the output states M
- Let A denote the total area of the L state space
- For each M , let $A(M)$ denote the area of L space that is labeled by M
- Let $s(M)$ denote the total border length that regions labeled M share with regions labeled other than M
 - Note that a point L lying within a distance ν of a border might end up being mapped into either M -value – call this a “borderline” L
 - Note that a point L that is more than a distance of ν from every border can only be mapped to the M -value of the region containing it

Geometric Interpretation

- Suppose ν is sufficiently small that the area occupied by borderline L states is small compared to the total area of the L space
- For the case in which $P(L)$ is uniform over L, the information rate R is given by (through terms of order ν)

$$R = - \sum_M [A(M)/A] \log[A(M)/A] - (\gamma\nu/A) \sum_M s(M)$$

- $P(M) = A(M) / A$
- $P(M|L) \log P(M|L)$ is zero except for borderline L
 - for L squarely in the M-labeled region, we get $1 \log 1 = 0$
 - for L squarely in a non-M-labeled region, we get $0 \log 0 = 0$
- γ is a positive number whose value depends on the details of the noise process, and which determines $P(M|L)$ for borderline L as a function of distance from the border

Geometric Interpretation

- Again, $R = \underbrace{-\sum_M [A(M)/A] \log[A(M)/A]}_{I_M} - \underbrace{(\gamma v/A) \sum_M s(M)}_{I_{M|L}}$
- For small v (low noise), I_M dominates
 - It is maximized when the $A(M)$ [and hence the $P(M)$] values are equal for all M
- The second term, $I_{M|L}$, is minimized (and thus R is maximized) when the sum of the border lengths of all M regions is minimized
 - This corresponds to the "sharpening" of $P(M|L)$ discussed previously
- The "infomax" solution is obtained by approximately equal partitions with near-minimum border length

Nonuniform Input Message Probabilities

- If $P(L)$ is nonuniform, the same result (equal areas, minimum border lengths) is obtained, except that both the area and border-length elements are weighted by local values of $P(L)$
- As a result, the infomax principle tends to produce maps in which regions of the input space that are activated more frequently are given greater representation in the output space
 - “Neuronal competition”
 - More neurons are automatically recruited to map input signals that are more common

Lateral Connections

- To consider a situation in which there are lateral connections in M , the problem can be logically decomposed into one more stage: $L \rightarrow L' \rightarrow M' \rightarrow M$
 - The $M' \rightarrow M$ stage is a result of lateral interactions
 - Previous deterministic output is now M'
 - M' is mapped into a neighborhood of states M
- Essentially the same result holds, with M being optimally represented by regions of equal size in L , with minimum length borders

Maximizing Variance

- Remember that one of the terms to be maximized in maximizing $I(M,L)$ is $I(M)$
- Since output messages M may be described by the mean of M plus the variance of M , maximizing $\text{Var}(M)$ also maximizes $I(M)$

A Hebb-type Rule

- Consider multiple layers of cells ($A, B, C, D\dots$) connected with local receptive fields
- In each layer, consider an output cell M and the cells $L_1, L_2, \dots L_N$ that provide input to it
- The linear response rule is then:

$$M = a_1 + \sum_j L_j c_j$$

where c_j is the strength of the j^{th} input connection to M and a_1 is an arbitrary constant

- Define a potential range of Hebb-type rules:

$$\Delta c_i = a_2 L_i M + a_3 L_i + a_4 M + a_5$$

where a_{2-5} are arbitrary constants ($a_2 > 0$)

A Hebb-type Rule

- Assuming the weights c change slowly, relative to L and M , allows us to average the previous equation over an ensemble of many presentations. Using the first equation to express M in terms of the $\{L_j\}$ and rearranging, Linsker obtains

$$\dot{c}_i = \sum_j Q_{ij} c_j + [k_1 + (k_2/N) \sum c_j]$$

where $k_{1,2}$ are particular combinations of a_{1-5} and

$$Q_{ij} = \langle (L_i - \bar{L}) \times (L_j - \bar{L}) \rangle$$

- Both $\langle \rangle$ and the overbar denote an ensemble average
- \bar{L} , the ensemble average of the input activity at a synapse, can be taken to be the same value for all synapses i and j

A Hebb-type Rule

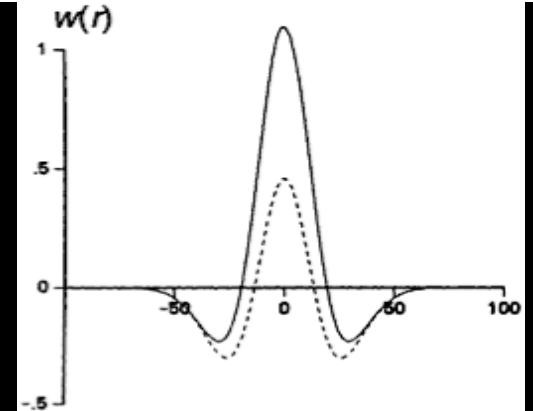
- The c values are constrained to lie between c_- and c_+ .
 - A more biologically plausible model would have some excitatory synapses ranging between zero and c_+ , and some inhibitory synapses ranging between c_- and zero, but the analysis of this more complicated case produces the same results
- For simplicity, allow each layer of connections to fully mature before analyzing the subsequent layer
 - Compute Q_{ij} for the ensemble of input patterns L
 - For random snow activity in the first layer, Q_{ij} is 1.0 when i and j are the same cell and 0.0 otherwise
 - Initialize the c values randomly
 - Solve the equation for \dot{c} and numerically integrate, producing mature c values for that layer
 - Compute the output patterns, which become the input patterns L to the next layer

Simulation Results

- Values of k_1 , k_2 , and the size of the receptive field determine the mature c values in each layer
 - The $k_{1,2}$ values determine the total connection strength $\sum c_j$ of the inputs to the M cell
- There turn out to be a limited number of ways each layer can develop

Layer Types

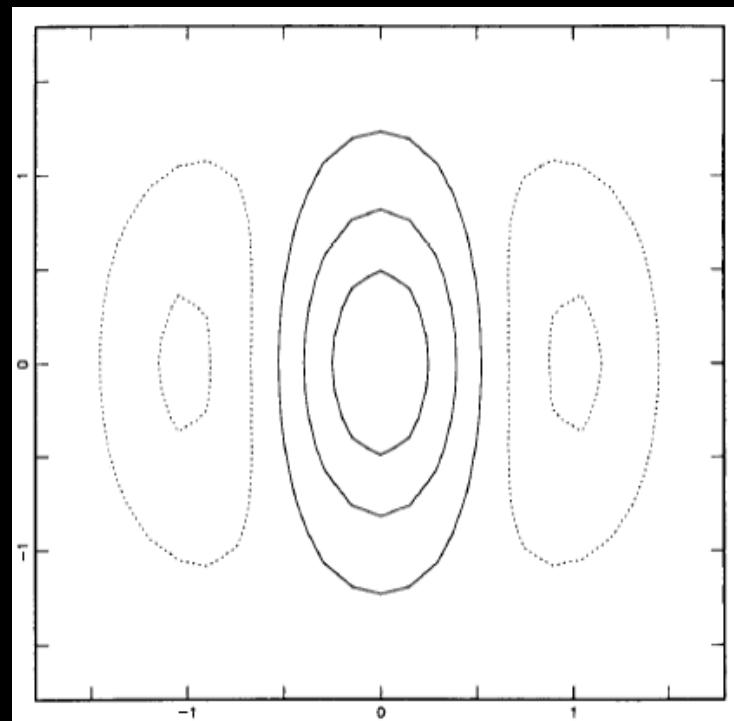
- The first cell type emerges in Layer B
- There is a parameter regime in which each c value reaches its excitatory limit c_+ .
- In this case, each B cell, once it has matured, computes the local average of the underlying area in layer A from which it receives input
 - Nearby B cells have correlated activity
- Each cell in layer C then matures into a *center-surround* cell that acts as a contrast-sensitive filter
 - Some respond maximally to a dark spot on a white background, some the inverse



Layer Types

- The next new type of cell to emerge, as we move to succeeding layers, is an *orientation-selective cell*
 - Responds maximally to a bright edge or bar against a dark background (or the reverse) when the edge or bar has a particular orientation

Here is the receptive field of a computed orientation-selective cell. A point of illumination at any position in the plane evokes an output response from the model cell that is proportional to the contour value at that location.



Layer Types

- Each orientation-selective cell will develop to favor an arbitrary orientation if the network contains only feed-forward connections
- However, if lateral connections between nearby cells in this layer are included in the simulation, then orientation preferences become more organized
 - Cells with similar orientation tuning develop to occupy irregular banded regions akin to mammalian visual cortex



Layer Types

- Remember, all these cell types emerge with random noise at the input layer
- Structured input to the first layer, A, can accelerate the appearance of these cell types (into earlier layers), but the same cell types appear
- If nearby pixels have correlated intensity values, and this is the only important input correlation present, then Q in layer A resembles the Gaussian Q previously found in layer B, and each feature-analyzing cell type develops as before, only one layer earlier
- If layer A is shown an ensemble of sinusoidal stripes with arbitrary phase and orientation, then orientation selectivity can be developed as early as layer B

Hebb Rules and Optimization

- Define the function

$$E \equiv E_Q + E_k$$

where

$$\begin{aligned} E_Q &= -\frac{1}{2} \langle (M - \bar{M})^2 \rangle \\ &= -\frac{1}{2} \sum_i \sum_j Q_{ij} c_i c_j \end{aligned}$$

and

$$E_k = -k_1 \sum c_j - (k_2/2N) (\sum c_j)^2$$

- E is constructed so as to have the property that
 - $\delta E / \delta c_i = \dot{c}_i$ for all i
- So E is always at a (local) minimum when the c_i mature

Hebb Rules and Optimization

- For any given value of total connection strength $\sum c_j$, E is minimized when $\langle (M - \bar{M})^2 \rangle$ is maximized; i.e., when the statistical variance in the output layer is maximized
- Therefore the Hebb-type learning rule we have been working with causes a cell to maximize the variance of its output activity
 - Subject to a $k_{1,2}$ parameter-driven constraint on total connection strength
 - Subject to saturation limits c_- and c_+
- Recall that maximizing the output variance maximizes one of the terms in our Shannon information rate

Maximizing Variance and Information

- Intuitively, if c values matured in such a way that M was constant (its variance was zero), regardless of the input signal L , then M would carry no information about L whatsoever
- Conversely, if the c values mature in such a way that $\text{Var}(M)$ is maximized, then each unique output signal M is capable of carrying a maximal amount of information about L
- Subject to certain constraints, our Hebb-type rule results in c values that produce exactly this maximum $\text{Var}(M)$
- This Hebb-type rule can also be shown to maximize information in M about L , by showing that it minimizes uncertainties in L given M , in an LMSE (Least Mean Squared Error) sense

Credits

- Images on slides 3-6 and 25 from Swindale, N.V. 1996. "The development of topography in the visual cortex: a review of models". *Network* 7:161-247
- Images on slides 7, 26, and 27 from the reading assignment, Linsker, R., "Self-Organization in a Perceptual Network", *Computer* 21(3), 105-117, March 1988 and the cover of that issue

References

- Extensive use was made of the two reading assignments:
 - Linsker1: Linsker, R., "Towards an Organizing Principle for a Layered Perceptual Network" in Neural Information Processing Systems, ed. by D. Z. Anderson. American Institute of Physics, New York, 1988
 - Linsker2: Linsker, R., "Self-Organization in a Perceptual Network", Computer 21(3), 105-117, March 1988
- Some context derived from the excellent review article (in the reading extras area):
 - Swindale, N.V. 1996. "The development of topography in the visual cortex: a review of models". Network 7:161-247