The Network of *Grossberg (1976)*

\[
\frac{dy_j}{dt} = -Ay_j + (B - y_j)(f(y_j) + I_j) - (C + y_j)\sum_{k \neq j} f(y_k)
\]

where \( I_j = \sum_i x_i w_{ij} \)

or \( \vec{x} \cdot \vec{w}_j = \|x\|\|\vec{w}_j\|\cos \theta \)

So the winner will be the cell that has incoming weights closest to the current input pattern

(same is true for *von der Malsburg network*)
The Network of *Grossberg (1976)*

\[
\frac{dy_j}{dt} = -Ay_j + (B - y_j)\left(f(y_j) + I_j\right) - (C + y_j)\sum_{k \neq j} f(y_k)
\]

In the F2 RCF of *Grossberg (1976)*, there are no inherent neighborhood relationships of the F2 cells.

I.e., All cells excite themselves and inhibit equally all other cells in the network.

As a result the network forms no topological relationships whatsoever.
In von der Malsburg model the competition is distance dependent, as a result feature detectors for similar patterns are closer together than neurons responding to less similar patterns. This gives a smoothly varying fuzzy classification where class nodes are also clustered.
The tendency for nearby cells to be tuned to similar orientations results because the cells tend to fire in clusters. This in turn is due to the nature of the on-center, off-surround interactions:

Cell at location \((i-1)\) becomes tuned to an orientation of 0 degrees.

Cell at \((i+1)\) becomes tuned to 20 degrees.

Cell at location \(i\), which fires often with both the cell at \((i-1)\) and the cell at \((i+1)\) due to the on-center off-surround connectivity, will tend to become tuned to the average of these orientations, i.e. 10 degrees.

\[
\begin{align*}
  &w_{ij} \\
  &i-j
\end{align*}
\]
The Network of *Grossberg* (1976)

One would not expect any tendency for the F2 cells to fire in clusters (even in the non-choice case), and therefore one would not expect that neighboring cells in F2 would become tuned to similar orientations.

\[
\frac{dy_j}{dt} = -Ay_j + (B - y_j) \left( \sum_i f(y_i)C_{ij} + I_j \right) - (D + y_j) \sum_{k \neq j} f(y_k)E_{kj}
\]

To explain the visual cortex topography, the F2 network would need to be modified as shown.

This kind of RCF is briefly discussed in *Grossberg* (1976) and was studied in *Levine and Grossberg* (1976).
Issue with Stability in SOM

The afferent weight vector becomes more parallel to the current F1 STM pattern

Recall that we can set up a sequence of inputs so that the weight vectors keep rotating around and never stabilize
Issue with Competitive Learning

Even with non-pathological data sequence it is possible that the weights will never stabilize, so the algorithm will never terminate.

One approach is to reduce the learning rate, but then
  – after some time we will not learn new patterns
  – we will not be able to track changes through time

This leads to stability-plasticity dilemma.
Kohonen’s Critique and Approach

Map in von der Malsburg’s model is patchy, local order exists, but there is no global order, this is good for orientation selectivity but not good for retinotopy.

Gradual reduction of learning rate is OK,

– when the system learns the complete mapping from input to output it does not need to learn more
– plasticity in the cortex is much higher during development than later in life

So he created a cross-breed between vector quantization and neural network to preserve a global mapping of the data.
Vector Quantization

Create a codebook of vectors \( w \) (similar to weights coming into neurons)

Check the similarity between the input \( x \) and each of these vectors and select a winner \( j \)

- Similarity can be based on dot product or simply on the difference between these two vectors

Do a gradient descent optimization

\[
\dot{w}_{ij} = \eta(t) \left( x_i - w_{ij} \right)
\]

where \( \eta(t) \) is monotonically decreasing learning rate

Note that this is very similar to postsynaptically gated decay

\[
\dot{w}_{ij} = \left( \eta x_i - \alpha w_{ij} \right) y_j
\]

except that learning rate is not constant and \( y_j \in \{0,1\} \)
The resulting codebook approximates the probability distribution of inputs, but in a disorganized way. Kohonen wanted to create feature detectors mapped in such a way that resulting map conveys information about the similarity between original data points on a global scale. That would also result in a smooth probability density surface. Representation of data in a lower dimensional space so that the distances between points correspond to similarity between original data items is called multidimensional scaling. Kohonen’s SOM is often used as a visualization tool for analysis of high-dimensional data.
Poverty Map

A variety of factors representing quality of life used as feature vectors

Colors represent the overall living quality
Preparations

Need to have some distance metric for each of the components of the input vector $x$

Those metrics should be weighted to reflect the importance of each individual component

Relative magnitude of two first principal components of the input set can hint for the height/width ratio of the resulting map

Resolution of the map is up to you, larger maps achieve better topology but require more time to converge
Algorithm

For every data point $x$:
Pick a winning node $j_w$ with weight $w$ closest to $x$
Update the weight to all nodes according to

$$
\dot{w}_{ij} = \eta(t) h(j_w, j, t) \left( x_i - w_{ij} \right)
$$

where $h()$ is a neighborhood function, for example

$$
h(j_w, j, t) = h_o e^{-\frac{(j_w - j)^2}{\sigma(t)^2}}
$$

or some other function that favors the winner and its neighbors
Critical Point

Both the learning rate and the neighborhood function shall decrease through time

Starting with large neighborhood ensures global mapping, decreasing it allows to fine tune the details
Issues, Tips and Tricks

As nothing favors a particular orientation of the resulting map, any mirror or point-symmetric inversion can result:

– Does not match the data on retinotopic map consistency

500 times the number of neurons is a good iteration count:

– Full scale visual cortex model will need too many iterations

Learning rate shall be large (0.9) in the beginning, go down to about 0.01-0.1 in the first 1000 iterations, and continue to slowly decrease afterwards

– Might be reasonable considering developmental cycle

Neighborhood size shall start with more than a half of network diameter, and decrease during first 1000 iterations to about nearest-neighbor size

– Not too reasonable, there is no such drastic connectivity reduction in the brain
Problems / Features

Generally slow convergence

Possible initializations that will not allow the network to converge

Like VQ, SOM is sensitive to input distribution (can it be a reason for cortical magnification?)
Comparison of Three Models

All three models have

– Competition at the second layer that allows only a subset of cells to learn for each input presentation

– Associative learning law

Differences:

– Grossberg (1976) has WTA competition with global inhibition of all other cells by the winner, while von der Malsburg (1973) and Kohonen (1982) have distance-dependent inhibition

This gives the latter two the ability to learn smooth topological representation of the input features
Comparison of Three Models

Differences:

– Kohonen (1982) does a gradual reduction of the neighborhood size as the learning progresses, the other two models do not

This gives his network the ability to learn global topology of the input features during initial phase of learning and fine tune it later

This also makes his model less realistic, in biology the global topology (like retinotopy) is probably genetically prewired

– Kohonen (1982) does a gradual reduction of the learning rate, the other two do not

It is unclear if it is necessary without neighborhood reduction, but it sure leads to a much slower convergence
Comparison of Three Models

Differences:

- *Grossberg (1976)* has shunting neuronal dynamics, *von der Malsburg (1973)* has additive dynamics, and *Kohonen (1982)* does not have any dynamics.

This gives the latter the simplicity for use outside of biological domain.


Both methods appear biologically realistic, but the latter is non-local.

Similarity between postsynaptically gated decay and gradient descent can lead to local minima, but without global topology it does not matter.
Comparison of Three Models

Differences:

– Kohonen (1982) has shown how the input distribution can lead to features similar to cortical magnification. von der Malsburg (1973) network probably can do this too, as well as distance dependent Levine and Grossberg (1976) network.

– Grossberg (1976) is stable and converging for all parameter choices, von der Malsburg (1973) and Kohonen (1982) are not.
In general, systems of equations of a type

\[ \dot{x}_i = f(x_1, \ldots, x_n) \]

are not guaranteed to converge

Both Grossberg’s and von der Malsburg’s RCFs are of this type

But if the equations are of certain specific types the stability can be proven
Stability of a Critical Point

A critical point is stable if all the trajectories that start near the critical point stay near this point as time evolves.

The system in the middle displays stable oscillations.

If the trajectories converge to critical point, it is called asymptotically stable (point on the left).

If the starting point does not matter than the critical point is globally asymptotically stable (right).

CN 510 Lecture 19
Stability of a Linear System

If any eigenvalue has a positive real part the point is unstable and so is the system.

If real parts are negative or zero then things get complicated:
  – If imaginary eigenvalues are non-repeating the point will show stable oscillations, but we will not consider the system stable.

The bigger issue is that neither Grossberg’s RCF nor von der Malsburg’s model are linear due to sigmoidal, quadratic, or threshold linear signal functions.
Stability of a Non-linear System

General approach is:

- determine all critical points
- approximate the system with a linear one through Jacobian in the neighborhood of each point
- analyze the local stability around each point
- if possible build a phase plane for a global picture

There are also theorems that prove stability of different specific systems of non-linear ODEs

Cohen-Grossberg theorem is one of these and it proves absolute stability of global pattern formation for RCF-like general systems

Absolute stability of global pattern formation means any parameter or input choice will not break it
Cohen-Grossberg Theorem

Applies to a particular but quite general class of systems:

\[
\dot{x}_i = a_i(x_i) \left( b_i(x_i) - \sum_{j} c_{ij} d_j(x_j) \right)
\]

Proves that this system converges to a stable equilibrium point if several conditions are met.

Four of those conditions:

- \(c\) is symmetric and non-negative: \(c_{ij} = c_{ji} \geq 0\)
- \(b\) is continuous: \(\lim_{x_i \to g_i} (b_i(x_i)) = b_i(g_i)\)
- \(a\) is positive and continuous: \(a_i(x_i) > 0;\)
  \[\lim_{x_i \to g_i} (a_i(x_i)) = a_i(g_i)\]
- \(d\) is non-negative and monotone: \(d_i(x_i) \geq 0;\)
  \[\text{sign}(\dot{d}_i(x_i)) = \text{const}\]
von der Malsburg and C-G Theorem

Compare

\[ \begin{align*}
\dot{E}_i &= -aE_i + \sum_j p_{ji} f(E_j) + \sum_k s_{ki} R_k - \sum_l q_{li} f(I_l) \\
\dot{I}_i &= -aI_i + \sum_j r_{ji} f(E_j)
\end{align*} \]

with

\[ \dot{x}_i = a_i(x_i) \left( b_i(x_i) - \sum_j c_{ij} d_j(x_j) \right) \]

\( a = 1 \), which is continuous and positive.

Reduced C-G becomes

\[ \dot{x}_i = b_i(x_i) - \sum_j c_{ij} d_j(x_j) \]

CN 510 Lecture 19
von der Malsburg and C-G Theorem

Compare

\[ \dot{x}_i = b_i(x_i) - \sum_j c_{ij} d_j(x_j) \]

with

\[ \dot{E}_i = -a_i E_i + \sum_k s_{ki} R_k + \sum_j p_{ji} f(E_j) - \sum_l q_{li} f(I_l) \]

and

\[ \dot{I}_i = -a_i I_i + \sum_j r_{ji} f(E_j) \]

\( b \) is linear and therefore continuous

\( d \) is threshold-linear and therefore monotone and non-negative

The final requirement is \( c \) is symmetric and non-negative:

\[ c_{ij} = c_{ji} \geq 0 \]

To ensure it we need to check \(-p_{ij}, q_{ij}, \) and \(-r_{ij}\)

CN 510 Lecture 19
von der Malsburg and C-G Theorem

(1) E-cell to E-cell connections ($p_{ij}$) –

E-cells excite their immediate neighbors in the array

These are symmetric, but positive, making $c$ negative and breaking the requirement

Note, that this requirement eliminates all networks with on-center excitation from the list of eligible networks for C-G theorem
**von der Malsburg and C-G Theorem**

(2) **E-cell to I-cell connections** \( (r_{ij}) \) –

E-cells excite the corresponding I-cell and its immediate neighbors in the array.

(3) **I-cell to E-cell connections** \( (q_{ij}) \) –

I-cells inhibit the next-to-immediate neighbor E-cells.

These are not even symmetric, also breaking the requirement.
von der Malsburg and C-G Theorem

Note that the previous slide reasoning would be true for any network where excitation and inhibition is separated into principal cells and interneurons with different weight kernels.

Thus von der Malsburg’s model does not fit C-G requirements.

Does it automatically makes it unstable? Of course not…

But: “not every set of parameters $p$, $q$, $r$, and $s$ lead to stable solutions. those finally employed were found partly by trial and error”

Can we set any realistic on-center off-surround map meeting C-G requirements?
Is Proven Stability Absolutely Necessary?

Although a very nice thing to have, a guarantee of stability is not always a necessary thing
e.g., *von der Malsburg*’s model is very useful and insightful even without a guarantee of stability
Furthermore, many apparently stable networks are far too complicated to allow rigorous mathematical proof of stability
However, knowing when your system will remain stable makes the job of parameter searching significantly less troublesome and can insure reliable system performance for untested input domains

Are networks in the brain stable for all “parameter choices”?
Next Time

Diffusely projecting transmitter systems of the CNS and the implications of these systems in terms of a variety of drug effects

Readings:
None required, but any facts related to drug effects in the brain are welcome