Unsupervised learning

Overview

- Hebbian learning rules
- Oja rule and PCA
- Single cell refinement; ON/OFF RGCs
- Network development:
  - Topographic maps and ocular dominance
- Feature-based models (if time permits)

Hebb’s postulate

- Supervised learning requires a teacher to provide error signals; how far can the nervous system get on its own?
- What general principles might the nervous system take advantage over?
- “When an axon of cell A is near enough to excite B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased” (Hebb, 1949).
- “cells that fire together wire together” or “out of sync lose the link”.
- Detect correlations [over some small time window, say 50ms] between firing of cells.
- Common to use firing-rates, rather than individual spike times.

\[ \Delta w_j = \epsilon y x_j \quad \text{where } y = f(w \cdot x) \]

Biological basis of learning

Left: EPSP recorded in response to presynaptic stimulation before and after paired stimulation. Long-term potentiation (LTP).

Right: in-vivo recordings from retinotectal synapses. Spike timing dependent plasticity (STDP).
Variants on Hebbian rule

activation rule \( y = \mathbf{w} \cdot \mathbf{x} \)

Hebb rule \( \frac{d\mathbf{w}_j}{dt} = yx_j \)
equivalently, for discrete update \( \Delta \mathbf{w}_j = \epsilon yx_j \)

This rule is unstable; for positive inputs and weights we get only growth of connections (examine \( \frac{d|\mathbf{w}|^2}{dt} \)). Can introduce decay of connections by thresholds either on input or output:

postsynaptic threshold \( \tau \frac{d\mathbf{w}_j}{dt} = (y - \theta y)x_j \quad \theta y = \langle y \rangle \)

presynaptic threshold \( \tau \frac{d\mathbf{w}_j}{dt} = y(x_j - \theta x_j) \quad \theta x_j = \langle x_j \rangle \)

However, these rules are still unstable.

Fast inputs, slow weight changes

Often we average over inputs \( \mathbf{x} \), assuming inputs change quicker than synaptic weights \( \mathbf{w} \).

\[
\tau \frac{d\mathbf{w}_j}{dt} = \langle yx_j \rangle , \quad y = \sum_i w_i x_i \\
= \langle \sum_i w_i x_i x_j \rangle \\
= \sum_i w_i \langle x_i x_j \rangle = \sum_i C_{ji} w_i \\
= \mathbf{C} \mathbf{w}
\]

Normalisation

Hebbian-based learning rules alone are unstable. Need some way to keep weights within bounds and introduce competition. Approaches:

1. Enforce limits on individual weights, e.g. \([0,1]\).
2. Renormalise weights periodically to rigidly satisfy some constraint \((\sum_j w_j = K \text{ or } \sum_j w_j^2 = K)\).

\[
\tau \frac{d\mathbf{w}}{dt} = y\mathbf{x} - \left[ \frac{y(\mathbf{n} \cdot \mathbf{x})}{\mathbf{n} \cdot \mathbf{n}} \right] \mathbf{n} \quad \mathbf{n} = \text{vector of 1s}
\]

subtractive normalisation: subtract \( k \) from every weight.

\[
\tau \frac{d\mathbf{w}}{dt} = y\mathbf{x} - \left[ \frac{y(\mathbf{n} \cdot \mathbf{x})}{\mathbf{n} \cdot \mathbf{w}} \right] \mathbf{w}
\]

divisive normalisation: divide every weight by \( k \).

Subtractive and divisive normalisation have different geometrical effects.

Correlation and covariance matrices

Correlation matrix \( Q_{ij} = \langle x_i x_j \rangle \) or \( Q = \langle \mathbf{x} \mathbf{x}^T \rangle \).

Covariance matrix \( C_{ij} = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle \quad \mu_i = \langle x_i \rangle \)

Properties:

Real, symmetric matrix \( \Rightarrow \) \( N \) orthogonal real eigenvectors.

Positive semi-definite: for any input \( \mathbf{u} \), \( \mathbf{u}^T \mathbf{C} \mathbf{u} \geq 0 \).

All eigenvalues of a positive semidefinite matrix are non-negative.
Eigenvectors of correlation matrices predict development

\[ \tau_w \frac{dw}{dt} = Qw \]

Following approach with dynamical systems, we can rewrite weight growth in terms of eigenvectors.

\[ w(t) = \sum_{i=1}^{N} \exp \left( \frac{\lambda_i t}{\tau_w} \right) (w(0) \cdot e_i) e_j \]

Since all eigenvalues are non-negative, all grow as long as \( e_i \cdot w(0) \neq 0 \). At large times, growth dominated by \( e_1 \) s.t. \( w \propto e_1 \).

What about saturation constraints?

**Oja rule as principal component analysis extractor**

Add terms to learning rule so that constraints are dynamically enforced:

\[ \tau \frac{dw}{dt} = yx - \alpha y^2 w \]

At s.s. \( w \) is maximal eigenvector of \( Q \). This is Principal components analysis (PCA). Rule finds vector such that projection onto that vector maximises the variance of the responses. (nth PCA = nth largest eigenvector of correlation matrix of inputs.)

**Effect of saturation limits upon development** (TN Fig 8.3)

\[ Q = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix} \quad e_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \lambda_1 = 1.4 \]

**Extracting multiple principal components**

What happens when we wire up multiple output neurons and use the Oja rule? Sanger rule (1989):

\[ \Delta w_{ij} = \epsilon y_j (x_i - \sum_{k=1}^{j} y_k w_{ik}) \]

Non-local update rule but reliable extraction of PCs in order.
Orientation selective receptive fields (Hancock et al. 1991)

The retina (Helga Kolb)

Not just a camera, faithfully reproducing image . . .

On- and off-centre segregation

Cells in the retina can be classified as on-centre or off-centre depending on their response to light stimulation. The LGN (lateral geniculate nucleus) is primary projection for retinal cells. How do connections develop from the retina to the LGN? Synapse elimination is key.

Developing

Adult

Neural activity plays an instructive role in this process; block activity ⇒ block segregation.

Do cells that fire together wire together? (Lee et al., 2002).

On- and off-centre RGC firing patterns

(P16-24; no differences at P9-11)
Model for on- vs off-centre segregation

Basic network:
- RGCs
- LGN relay cell

Network with inhibition:
- Inhibitory neuron

Inputs: spike trains from paired extracellular recordings

Dynamics:
\[
\begin{align*}
    y &= \sum_j w_j x_j - \Gamma \sum_j x_j \quad \text{Output activity} \\
    \Delta w_j &= \eta (y(x_j - \theta)) \quad \text{Update rule}
\end{align*}
\]

Analysis of LGN model.

Dropping ' on \( w \), we have a correlation matrix equation, where e.g. for 2 inputs:
\[
\begin{align*}
    \frac{d}{dt} w &= Aw \\
    \frac{d}{dt} w &= \lambda_1 (e_1 \cdot w) e_1 + \lambda_2 (e_2 \cdot w) e_2
\end{align*}
\]
where \( \lambda_i \) and \( e_i \) are eigenvalues and eigenvectors of \( A \).

If \( \lambda_{\text{max}} \) -ve, all weights decay; else look at \( e_{\text{max}} \):
\[
\begin{align*}
    (+), (-) &\Rightarrow \text{no segregation} \\
    (-), (+) &\Rightarrow \text{segregation}
\end{align*}
\]

So, given real spike trains, compute \( \langle x_1 x_2 \rangle, \langle x_1 x_1 \rangle, \langle x_2 x_2 \rangle \) from data (assuming e.g. 500 msec bin widths), calc \( A \) and see if there is segregation.

Analytical approach: example 1

\[
\begin{align*}
    \langle x_1 x_2 \rangle &= 0.06 \\
    \langle x_1 x_1 \rangle &= 0.14, \langle x_2 x_2 \rangle = 0.58 \\
    \langle x_1 \rangle &= 0.15, \langle x_2 \rangle = 0.48 \\
    \theta &= 0 \quad \text{no punishment}
\end{align*}
\]

\[
A = \begin{pmatrix} 0.14 & 0.06 \\ 0.06 & 0.58 \end{pmatrix}, \quad e_1 = \begin{pmatrix} 0.13 \\ 0.99 \end{pmatrix}, \lambda_1 = 0.59 \quad \text{max: no segregation}
\]

\[
e_2 = \begin{pmatrix} 0.99 \\ -0.13 \end{pmatrix}, \lambda_2 = 0.13.
\]
Analytical approach: example 2

\[
\langle x_1 x_2 \rangle = 0.06 \quad \text{input correlations}
\]
\[
\langle x_1 x_1 \rangle = 0.14, \langle x_2 x_2 \rangle = 0.58
\]
\[
\langle x_1 \rangle = 0.15, \langle x_2 \rangle = 0.48
\]
\[
\theta = 0.7 \text{Hz} \quad \text{mild punishment}
\]

\[
A = \begin{pmatrix}
0.035 & -0.045 \\
-0.276 & 0.244
\end{pmatrix}
\]

\[
e_1 = \begin{pmatrix}
0.17 \\
-0.99
\end{pmatrix}, \lambda_1 = 0.29 \quad \text{max: segregation}
\]

\[
e_2 = \begin{pmatrix}
-0.68 \\
-0.73
\end{pmatrix}, \lambda_2 = -0.01.
\]

On-off segregation: model results

On-off segregation: phase plots

Summary of on- and off-centre segregation

- Although on- and off-centre inputs are correlated, they can segregate if inputs have mild punishment (\(\theta\)). If \(\theta\) is too small or too high, segregation does not occur.
- Segregation is in favour of off-centre cells, unless they are inhibited by \(\Gamma\).
- How might a group of output units develop? Allow interactions between neighbouring output neurons . . .
Ocular dominance: synapse elimination and neural activity

- Correlated activity between pre- and postsynaptic cells thought to drive refinement of connections (Hebb).
- Competition (e.g. for limited resources) driving development.
- Synapse elimination before eye-opening ...

Given adaptable weights $W$ and fixed lateral weights $M$, with mexican hat profile, we can get ocular dominance columns.

\[
\frac{dv}{dt} = -v + Wu + Mv
\]

\[
\frac{dW}{dt} = \langle vu \rangle = (I - M)^{-1}WQ
\]

Processing the visual scene
Processing the visual scene

Retinal photoreceptor distribution (Ground squirrel; Galli-Resta et al., 1999)

Green = rods; orange = S cones; unlabelled = M cones.
Field of view: \(\approx 400 \, \mu\text{m} \) wide.

Topographic maps are common in sensory areas

Kaas & Catania, 2002

Somatosensory maps

Broadly topographic, with over-representation of hands and face.

http://cogsci.bme.hu/~ikovacs/latas2005/prepI_4_2_files/fig5.jpg
http://faculty.washington.edu/chudler/flash/hom.html
Network development

- Developmental processes include: cell birth, death, fate selection, axon growth + targeting, synapse elimination.

Two major principles of sensory map formation

2. Activity-dependent processes to refine local connectivity.

Principles of topographic map formation

Key elements of map formation, demonstrated by Willshaw & von der Malsburg (1976):

1. Neighbouring presynaptic neurons fire in synchrony (retinal waves).
2. Cells that fire together wire together.
3. Neighbouring postsynaptic neurons should develop similar connections.
4. Constraints on synaptic growth (normalisation).
Example schematic of topographic map formation

Goodhill (1993) model of OD and topographic map development

Key model combining the development of ocular dominance and topography.

Initial connections random, but include topographic bias to ensure overall map layout is consistent.

Modelling visual inputs

To make one pair of left/right inputs.
- For each eye, generate 0/1 randomly for each pixel. Then, blur (convolve) activity independently in each eye.
- To introduce correlations between the eyes, find the pixel at the same position in each eye and combine them to a degree $h$:

$$x_i^{\text{left}} \leftarrow (1 - h)x_i^{\text{left}} + hx_i^{\text{right}}$$

$h = 0.0 \Rightarrow$ activity remains independent between eyes.
$h = 0.5 \Rightarrow$ activity same in each eye.

Many pairs of inputs created.

Output unit activation and weight update

- Activation:

$$y_j = \sum_i w_{ij}^{\text{left}} x_i^{\text{left}} + \sum_i w_{ij}^{\text{right}} x_i^{\text{right}}$$

- “Winner take all”: find output unit $g$ with highest firing rate; update weights of output units physically close to $g$:

$$g = \arg\max_j y_j$$

$$\Delta w_{ij}^{L/R} = \alpha x_i^{L/R} s(j, g)$$

$s(j, g)$ Gaussian function of distance between unit $j$ and winner $g$.

- Postsynaptic subtractive normalisation used.
Emergence of topography

(a) initial weights; b: 6000 iterations; c: 100,000 iterations

Ocular dominance and topography

\( h = 0.0 \) \hspace{1cm} \( h = 0.1 \) \hspace{1cm} \( h = 0.2 \)

Key prediction: stronger between-eye correlations ⇒ narrower stripes.

Feature-based models

Each element of an input vector codes for a feature of a stimulus, rather than e.g. a 2-d vector of pixel intensities.

\[ \mathbf{u} = (x, y, o, a \cos \theta, a \sin \theta)^T \]

where \((x, y)\) is the centre of mass of retinal activity, \(o\) is the ocularity [-1,1] and \((a, \theta)\) indicate orientation selectivity.

This allows two \(N \times N\) images to be concisely represented by a low-dimensional vector.

Weight modification in feature-based models

No need for normalisation, since weights become similar to input vectors.

e.g. elastic net methodology:

“softmax” output of cortical unit \(a\):

\[
 x_a = \exp \left( - \sum_b (u_b - W_{ab})^2 / (2\sigma_b^2) \right) \\
 v_a = \frac{x_a}{\sum_{a'} x_{a'}} \\
 \tau_w \frac{dW_{ab}}{dt} = \langle v_a (u_b - W_{ab}) \rangle + \beta \sum_{a' \in \text{neighs}(a)} (W_{a'b} - W_{ab})
\]

Self-organising map (SOM) works in similar way, but restricts weight updates to a “winner” and nearby neighbours.
Elastic net: TSP results for 100 cities (Durbin & Willshaw, 1987)

Joint orientation domain and ocular dominance column development

Summary

• Hebbian learning must be augmented by some form of competition.
• PCAs can be computed using unsupervised networks.
• Topographic map and OD development.
• Reading: Lee et al. (2002); Goodhill (1993); Hertz, Krogh and Palmer (1991) Introduction to the theory of neural computation for general background.

MPhil in Computational Biology

• Cambridge Computational Biology Institute (CCBI)
  http://www.ccbi.cam.ac.uk
• 1 Year MPhil course in Computational Biology.
• Currently accepting applications for 2009.
• Open afternoon, Wed 4th February 2–5pm. See website to register.