Hebbian Learning

Outline:
• kinds of plasticity
• Hebbian learning rules
• weight normalization
• intrinsic plasticity
• spike-timing dependent plasticity
• developmental models based on Hebbian learning
A Taxonomy of Learning Settings

- Unsupervised
- Self-supervised
- Reinforcement
- Imitation
- Instruction
- Supervised

Increasing amount of “help” from the environment
Unsupervised Hebbian Learning

Hebb’s Idea (1949):
If firing of neuron A contributes to firing of neuron B, strengthen connection from A to B.

Possible uses: (after Hertz, Krogh, Palmer)
• Familiarity detection
• Principal Component Analysis
• Clustering
• Encoding

Donald Hebb
Network Self-organization

Three ingredients for self-organization:
• positive feedback loops (self-amplification)
• limited resources leading to competition between elements
• cooperation between some elements possible

- correlated activity leads to weight growth (Hebb)
- weight growth leads to more correlated activity
- weight growth limited due to competition
Long term potentiation (LTP) and Long term depression (LTD)

- observed in neocortex, cerebellum, hippocampus, …
- requires paired pre- and postsynaptic activity
Local Learning

Biological plausibility of learning rules: want “local learning” rule. Weight change should be computed only from information that is locally available at the synapse:

- pre-synaptic activity
- post-synaptic activity
- strength of this synapse
- …

But not:

- any other pre-synaptic activity
- any other individual weight
- …

Also: weight change only on information available at that time. Knowledge of detailed history of pre- and post-synaptic activity is not plausible. (Locality in time domain)

These are problems for many learning rules derived from information theoretic ideas.

Dayan & Abbott, 2001
Single linear unit

\[ u \rightarrow v = w^T u \]

\[ \Delta w_i = \eta v u_i \]

\[ \Delta w = \eta v u \]

simple Hebbian learning, \( \eta \) is learning rate

- inputs drawn from some probability distribution
- simple Hebb rule moves weight vector in direction of current input
- frequent inputs have bigger impact on resulting weight vector: familiarity
- Problem: weights can grow without bounds, need competition

Called correlation based learning, because average weight change proportional to correlation between pre- and post-synaptic activity:

\[ \langle \Delta w_i \rangle = \langle \eta v u_i \rangle = \eta \langle v u_i \rangle \]
Reminder: Correlation and Covariance

**Correlation:**
\[ C_{ij} \equiv \langle x_i x_j \rangle \]

**Covariance:**
\[ \Sigma_{ij} \equiv \langle (x_i - \mu_i)(x_j - \mu_j) \rangle \]

**Mean:**
\[ \mu_i \equiv \langle x_i \rangle \]

\[ C \equiv \langle x^T x \rangle \]
\[ \Sigma \equiv \langle (x - \mu)^T (x - \mu) \rangle \]
\[ \mu \equiv \langle x \rangle \]
Continuous time formulation

\[ \tau \frac{dw}{dt} = v u \quad \text{averaging across stimulus ensemble} \]

\[ \tau \frac{dw}{dt} = \langle vu \rangle = \langle u (u^T w) \rangle = \langle (uu^T)w \rangle = \langle uu^T \rangle w = Qw \]

another good reason to call this a correlation based rule!

Relation to Eigenstructure of correlation matrix: weight vector will align with the Eigenvector of the correlation matrix with the biggest Eigenvalue.
(See blackboard).
Covariance rules

Simple Hebbian rule only allows for weight growth (pre- and post-synaptic firing rates are non-negative numbers: no account of LTD. Covariance rules are one way of fixing this.

\[ \tau \frac{d w}{d t} = (v - \theta_v) u, \text{ with post - synaptic threshold } \theta_v \]

\[ \tau \frac{d w}{d t} = v(u - \theta_u), \text{ with pre - synaptic threshold } \theta_u \]

Averaging RHS over the stimulus ensemble and using the mean input as the threshold gives: \( \theta_u = \langle u \rangle \)

\[ \tau \frac{d w}{d t} = \langle v(u - \theta_u) \rangle = \langle (w^T u)(u - \theta_u) \rangle \]

\[ = \langle uu^T \rangle w - (\theta_u \theta_u^T) w \]

\[ = \Sigma w \]

justifying the name “covariance” rule!
Illustration of correlation and covariance rules:

A. either rule on zero mean data
B. correlation rule for non-zero mean
C. covariance rule for non-zero mean

Note: weight vector will grow without bounds in both cases
The need to limit weight growth

• simple Hebb and covariance rules are unstable: lead to unbounded weight growth

• Remedy 1: weight clipping: don’t let weight grow above/below certain limit $w_{\text{max}}/w_{\text{min}}$

• Remedy 2: some form of weight normalization
  • Multiplicative: subtract something from each weight $w_i$ that is proportional to that weight $w_i$
  • Subtractive: subtract something from each weight $w_i$ that is the same for all weights $w_i$. This leads to strong competition between weights, typically resulting in all weights going to $w_{\text{max}}$ or $w_{\text{min}}$
Weight normalization

Two most popular schemes:
• sum of weights equals one (all weights assumed positive)
• sum of squared weights equals one

Idea: force weight vector to lie on a “constraint surface”

\[ w_1 + w_2 = 1 \]

\[ w_1^2 + w_2^2 = 1 \]
Subtractive vs. Multiplicative

Different ways of going back to constraint surface

\[ w_1 + w_2 = 1 \]

\[ w_1^2 + w_2^2 = 1 \]
Synaptic Scaling

A

Control
Reduced activity
Enhanced activity

10 pA ______________ 10 pA ______________
1s 10 ms

B

Percentage of events

Control
Reduced activity
Enhanced activity
Reduced activity / 2.68
Enhanced activity \times 1.58

mEPSC amplitude (pA)

C

Prosynaptic terminal

Postsynaptic terminal

Potentiation

Scaling

Target activity
High activity
Target activity
Oja’s rule

**Idea:** subtract term proportional to $v^2$ to limit weight growth

$$\Delta w_i = \eta vu_i - \eta v^2 w_i = \eta v(u_i - vw_i), \text{ or similarly}$$

$$\tau_w \frac{dw}{dt} = v u - v^2 w$$

- special form of multiplicative normalization
- leads to extraction of first **principal component of input correlation matrix**
- weight vector converges to unit length
Yuille’s rule

subtract term proportional to squared norm of weight vector

\[ \Delta w_i = \eta (V \xi_i - w_i |\vec{w}|^2) \quad |\vec{w}| \rightarrow \sqrt{\lambda_+} \]

“non-local”

weight vector’s length becomes square root of eigenvalue of principal eigenvector of input correlation matrix
The BCM Rule

\[ \tau_w \frac{dw}{dt} = \nu u(v - \Theta_v) \]

post-synaptic activity

\[ \tau_\Theta \frac{d\Theta_v}{dt} = v^2 - \Theta_v \]

competition between weights due to sliding threshold

some biological evidence for sliding threshold