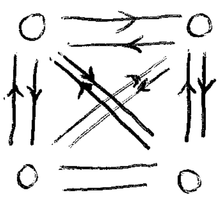
Hopfield networks

A fully connected feedback network, symmetric weights:

(no loops)



$w_{ji} = w_{ij}$

Associative memories.

Hebbian learning: Donald Hebb, 1949

weights between neurons

$$\frac{dw_{ij}}{dt} = \frac{dw_{ji}}{dt} \sim \mathrm{Corr}(x_i, x_j) \qquad (*)$$

neuron activities

If a stimulus is present $\Rightarrow$ neuron i activated

If another stimulus is present $\Rightarrow$ neuron j activated

If the two stimuli are correlated $\Rightarrow$

$\Rightarrow$ according to $(*)$, w_{ij} & w_{ji} will increase with time & become large.

Now, if neuron i is stimulated, neuron j is activated too $\Rightarrow$ associative memory.

Binary Hopfield network

$\overset{j}{\circ} \xrightarrow[w_{ij}]{} \overset{i}{\circ}$ $\quad I$ nodes (neurons)

Fully connected, $w_{ij} = w_{ji}$; $w_{ii} = 0$, $\forall i$

Biases may be included as weights w_{i0} from neuron $\emptyset$ with $\underbrace{x_0 = 1}_{\text{permanently on}}$.

x_i = activity of neuron i

(1.) Activity rule: (state update rule) $\quad x_i(\overset{\text{sum of inputs}}{a_i}) = \begin{cases} 1 & a_i \geq 0 \\ -1 & a_i < 0 \end{cases}$

Synchronous updates: compute $a_i = \sum_j w_{ij} x_j$, $\forall i$

Then update all neuron states: $x_i = x_i(a_i)$, $\forall i$

Asynchronous updates: compute a_i & x_i for one neuron at a time, continue in a fixed or random sequence of neurons.

(2.) Learning rule: goal $\Rightarrow$ make a set of memories $\{\vec{x}^{(n)}\}$ stable states of the Hopfield network. $\vec{x}^{(n)} = (\underbrace{-1, -1, +1, \ldots, +1}_{I \text{ entries}})$

N memories

Use Hebbian learning: sum over memories

$$w_{ij} = \eta \sum_{n=1}^{N} x_i^{(n)} x_j^{(n)},$$

$\eta > 0$ is a constant [e.g. $\eta = \frac{1}{N}$]

Biological motivation:

1. Biological memories are associative, recalled spontaneously
   Moscow $\leftrightarrow$ Russia , Oslo $\leftrightarrow$ Norway
2. Bio. memories are error-tolerant & robust   Ozlo $\Leftrightarrow$ Narway $\Rightarrow$ Oslo $\leftrightarrow$ Norway
   O.lo $\Leftrightarrow$ N.rway $\Rightarrow$ Oslo $\leftrightarrow$ Norway
3. Bio. memories are distributed

Continuous Hopfield network

Same as binary but with

$$x_i = \tanh(a_i) \quad \text{or} \quad x_i = \tanh(\beta a_i)$$

Binary Hopfield network is the $\beta \to \infty$ limit of the continuous Hopfield network.

Hopfield networks are basically spin glasses:

$$\underset{\substack{\uparrow \\ \text{Energy}}}{E} = -\frac{1}{2}\sum_{ij} J_{ij} x_i x_j - \sum_i h_i x_i$$

$$J_{ij} \Leftrightarrow w_{ij} \,, \quad h_i \Leftrightarrow w_{i0}$$

[although x_i are not necessarily ± 1]

Spin glasses can be treated using a mean-field approach.

Mean-field approximation

Consider $P(\vec{x}) = \frac{1}{Z} e^{-\overset{\text{inverse } T}{\beta} E(\vec{x})}$, where $\vec{x}$ ← state

$$E(\vec{x}) = -\frac{1}{2}\sum_{i,j} \underset{\text{couplings}}{J_{ij}}\, x_i x_j - \sum_i \underset{\text{fields}}{h_i}\, x_i \quad \&$$

$$\underset{\text{part'n f'n}}{Z} = \sum_{\vec{x}} e^{-\beta E(\vec{x})}$$

We want to approximate $P(\vec{x})$ with $Q(\vec{x}, \vec{\theta})$ which is simpler to compute.

$\vec{\theta}$ ← adjustable prms

Introduce
$$\underset{\text{mf free energy}}{\beta \tilde{F}(\vec{\theta})} = \beta \underbrace{\langle E(\vec{x})\rangle_Q}_{\text{average energy}} - \underbrace{S_Q}_{\text{entropy}} =$$

$$= \beta \sum_{\vec{x}} Q(\vec{x}, \vec{\theta}) E(\vec{x}) + \sum_{\vec{x}} Q(\vec{x}, \vec{\theta}) \log Q(\vec{x}, \vec{\theta}) =$$

$$= \sum_{\vec{x}} Q \log \frac{Q}{\underbrace{e^{-\beta E}}_{ZP}} = \underbrace{\sum_{\vec{x}} Q \log \frac{Q}{P}}_{D(Q\|P),\ \text{KL distance} \geq 0} - \underbrace{\log Z}_{\sum_{\vec{x}} Q(\vec{x}) = 1} =$$

$$= D(Q\|P) + \beta F.$$

$Z = e^{-\beta F}$

So, $\tilde{F}$ is bounded below by exact free energy $F \Rightarrow$ vary $\vec{\theta}$ to minimize $\tilde{F}$.

Consider $$Q(\vec{x},\vec{a}) = \frac{1}{Z_Q} e^{\sum_n a_n x_n}$$ [uncoupled spins]

Then $$S_Q = -\sum_{\vec{x}} Q \log Q = -\underset{\substack{\uparrow \\ \text{sum over spins}}}{\sum_n} \Big[\overbrace{q_n \log q_n + (1-q_n)\log(1-q_n)}^{-S_n,\ \text{entropy of spin } n} \Big],$$ where

$$\begin{cases} q_n = \dfrac{e^{a_n}}{e^{a_n}+e^{-a_n}}, & \Leftarrow \text{prob. that spin } n = +1 \\ 1-q_n = \dfrac{e^{-a_n}}{e^{a_n}+e^{-a_n}}. & \Leftarrow \text{prob. that spin } n = -1 \end{cases}$$

Furthermore,

$$\langle E(\vec{x})\rangle_Q = \sum_{\vec{x}} Q\Big[-\frac{1}{2}\sum_{i,j} J_{ij} x_i x_j - \sum_i h_i x_i\Big] =$$

$$= -\frac{1}{2}\sum_{i,j} J_{ij}\langle x_i\rangle\langle x_j\rangle - \sum_i h_i \langle x_i\rangle,$$

where $$\langle x_i \rangle = \frac{e^{a_i} - e^{-a_i}}{e^{a_i}+e^{-a_i}} = \tanh(a_i).$$

Finally,

$$\beta \tilde{F}(\vec{a}) = -\beta \left[\frac{1}{2} \sum_{\substack{i,j \\ i \neq j}} J_{ij} \langle x_i \rangle \langle x_j \rangle + \sum_i h_i \langle x_i \rangle \right] - \sum_n S_n .$$
$(J_{ii} = 0)$

Minimize $\tilde{F}$:

$$\frac{\partial}{\partial q_n} S_n(q_n) = \cancel{-1} - \log q_n \cancel{+1} + \log(1-q_n) =$$

$$= \log \frac{1-q_n}{q_n} = -2a_n .$$

Note that $q_n = \frac{1}{1+e^{-2a_n}}$,

$$e^{-2a_n} = \frac{1}{q_n} - 1 = \frac{1-q_n}{q_n} .$$

Moreover,

$$\langle x_i \rangle = \frac{1-e^{-2a_i}}{1+e^{-2a_i}} = \frac{1-\left(\frac{1}{q_i}-1\right)}{1+\left(\frac{1}{q_i}-1\right)} = 2q_i - 1,$$

so that $\frac{\partial}{\partial q_n} \langle x_n \rangle = 2$.

Then
$$\beta \frac{\partial \tilde{F}}{\partial a_m} = -\beta \left[\sum_j J_{mj} \underbrace{\frac{\partial \langle x_m \rangle}{\partial q_m}}_{2} \langle x_j \rangle + h_m \overbrace{\frac{\partial \langle x_m \rangle}{\partial q_m}}^{2} \right] \left(\frac{\partial q_m}{\partial a_m} \right) - \left(\frac{\partial q_m}{\partial a_m} \right) \underbrace{\frac{\partial S_m}{\partial q_m}}_{-2a_m} =$$

$$= 2\underbrace{\left(\frac{\partial q_m}{\partial a_m}\right)}_{\neq 0}\left\{-\beta\left[\sum_j J_{mj}\langle x_j\rangle + h_m\right] + a_m\right\}.$$

$$\frac{\partial \tilde{F}}{\partial a_m} = 0 \;\Rightarrow\; a_m = \beta\left[\sum_j J_{mj}\langle x_j\rangle + h_m\right].$$

Recall that $\langle x_j\rangle = \tanh(a_j)$.

We can see that there is an analogy between Hopfield networks & mean-field approach to spin glasses:

Hopfield	Mean-field
$x_i = \tanh(\beta a_i)$	$\langle x_i\rangle = \tanh(a_i)$
$a_i = \sum_j w_{ij} x_j + w_{i0}$	$a_i = \beta\left[\sum_j \underbrace{J_{ij}}_{w_{ij}}\langle x_j\rangle + \underbrace{h_i}_{w_{i0}}\right]$

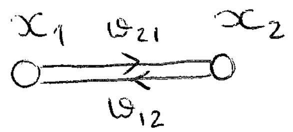
Thus $\tilde{F}$ plays the role of free energy in the Hopfield network. On the free energy landscape defined by $\tilde{F}$, there are distinct basins of attraction corresponding to each memory. Asynchronous update: changing 1 spin at a time. Synchronous update: changing <u>all</u> spins (a global move, is <u>not</u> guaranteed to decrease $\tilde{F}$).

It can be shown that asynchronous updates are guaranteed to decrease $\tilde{F}$ monotonously, similarly to steepest descent.

Moreover, F is convex in the vicinity of the minima $\Rightarrow$ the dynamics converges to a stable fixed point (which one depends on the ~~boundary~~ initial conditions) & there are no limit cycles (& no chaotic behavior).

Examples

① Binary Hopfield network



no biases

$w_{12} = w_{21} = 1$

$\underline{t=0}$: $x_1 = 1$, $x_2 = -1$ $\qquad E = -x_1 x_2 = 1$

Synchronous update: $\begin{cases} a_1 = w_{12} x_2 = x_2, \\ a_2 = w_{21} x_1 = x_1. \end{cases}$

$\underline{t=1}$: $\begin{cases} a_1 = -1 \\ a_2 = 1 \end{cases} \Rightarrow \begin{cases} x_1 = -1 \\ x_2 = 1 \end{cases}$ both spins flipped $\quad E = 1$

$\underline{t=2}$: $\begin{cases} a_1 = 1 \\ a_2 = -1 \end{cases} \Rightarrow \begin{cases} x_1 = 1 \\ x_2 = -1 \end{cases}$ both spins flipped again, etc.

$E = 1$

$\swarrow$ no convergence, limit cycle

Asynchronous update:

Update spin 1, then spin 2 (& continue)

$\underline{t=1}$: $a_1 = -1 \Rightarrow \begin{cases} x_1 = -1, \\ x_2 = -1 \end{cases}$ spin 1 flipped $\quad E = -1$

$\underline{t=2}$: $a_2 = -1 \Rightarrow \begin{cases} x_1 = -1, \\ x_2 = -1 \end{cases}$ spin 2 not flipped $\quad E = -1$

$(-1,-1)$ is a stable fixed point

[(1,1) is another]

2. Continuous Hopfield network

2 nodes, no biases : $h_1 = h_2 = 0$

$\beta = 1$, $J_{12} = J_{21} = 1$:

$$\tilde{F} = -\underbrace{\bar{x}_1}_{2q_1-1}\underbrace{\bar{x}_2}_{2q_2-1} + q_1 \log q_1 + (1-q_1)\log(1-q_1) + q_2 \log q_2 + (1-q_2)\log(1-q_2).$$

t=0: start with $\bar{x}_1 = 1$, $\bar{x}_2 = -1$:

$$\begin{cases} a_1 = +\infty, \\ a_2 = -\infty \end{cases} \qquad \begin{cases} q_1 = 1, \\ q_2 = 0 \end{cases}$$

Then $\tilde{F} = 1$

t=1:

(asynchronous 121212... updates)

$$a_1 = \bar{x}_2 = -1 \Rightarrow \begin{cases} \bar{x}_1 = \tanh(a_1) = \dfrac{e^{-1}-e}{e^{-1}+e} \simeq -0.76 \\ \bar{x}_2 = -1 \end{cases}$$

$\Downarrow$

$$\begin{cases} q_1 = \left(\dfrac{e^{-1}-e}{e^{-1}+e} + 1\right)\dfrac{1}{2} = \underbrace{\dfrac{1}{1+e^2}}_{\simeq 0.12}, \\ q_2 = 0 \end{cases}$$

Then $\tilde{F} = \dfrac{e^{-1}-e}{e^{-1}+e} +$

$$+ \frac{1}{1+e^2}\log\frac{1}{1+e^2} + \frac{e^2}{1+e^2}\log\frac{e^2}{1+e^2} \approx -1.13$$

t=2: $a_2 = \bar{x}_1 \simeq -0.76 \Rightarrow \begin{cases} \bar{x}_1 = -0.76, \\ \bar{x}_2 = \tanh(a_2) \simeq -0.64 \end{cases}$

$$\begin{cases} q_1 = 0.12 \\ q_2 = \dfrac{1+\bar{x}_2}{2} \simeq 0.18 \end{cases} \Rightarrow \tilde{F} \simeq -1.32$$

etc.