

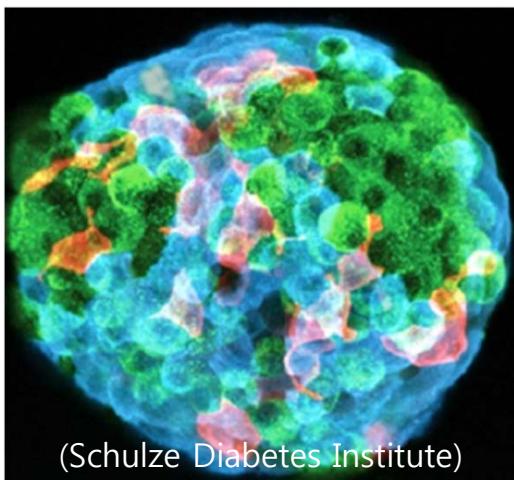
Machine Learning I

Junghyo Jo

(APCTP)

Dec 16, 2016

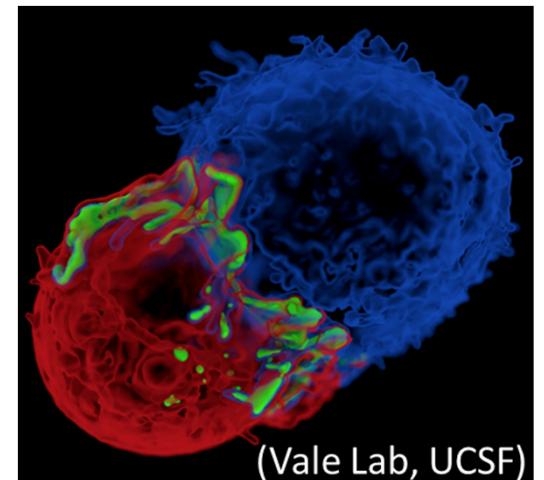
Design principles of cellular networks



(Schulze Diabetes Institute)



(Image: Visual Studio Magazine)



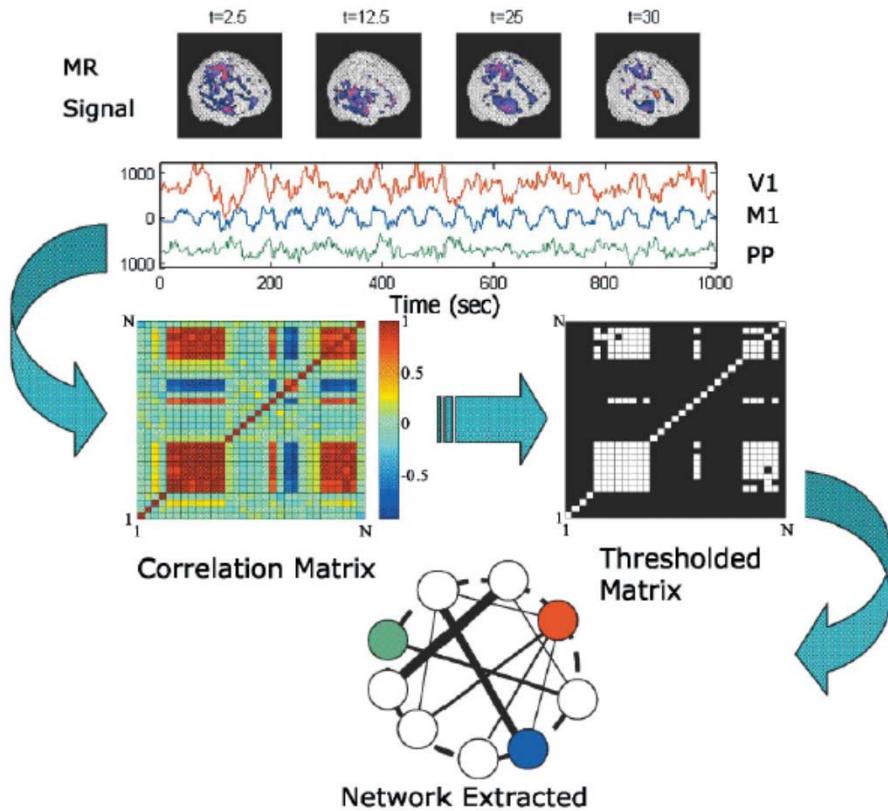
(Vale Lab, UCSF)

Metabolism

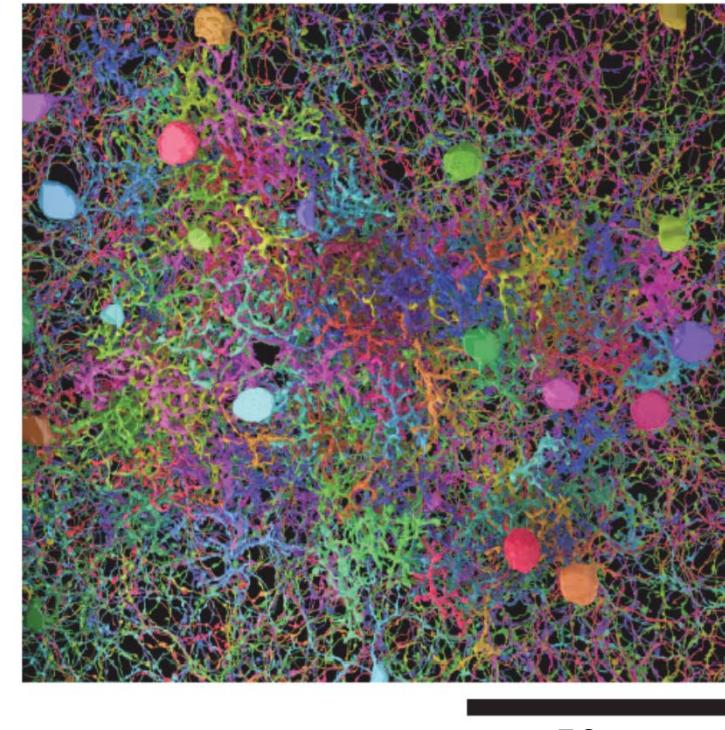
Learning

Immunity

Neural networks



(Eguiluz et al., PRL 2005)



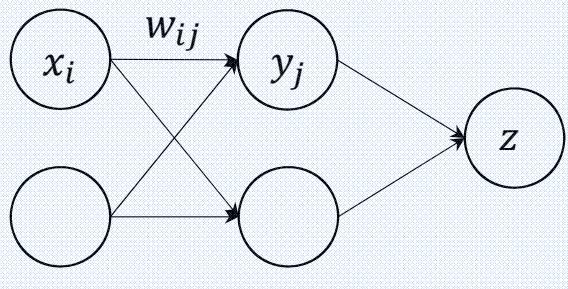
(Kim et al., Nature 2014)

Functional/structural connectivity

Machine learning

Backpropagation

- Rumelhart, Hinton, Williams, Nature 1986
- Gradient-descent algorithm
- input-output pairs

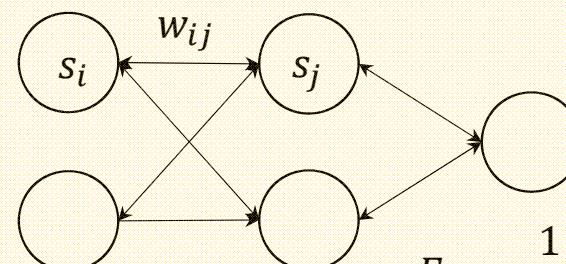


$$y_j = \frac{1}{1 + e^{-(\sum_i w_{ij}x_i - b_j)}}$$

$$\Delta w_{ij} \propto -\frac{dC(z, z')}{dw_{ij}}$$

Boltzmann machine

- Ackley, Hinton, Sejnowski, Cognitive Science 1985
- Stochastic searching algorithm
- Correlated states

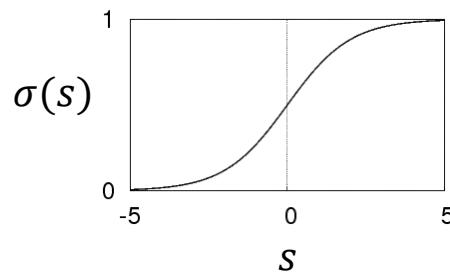
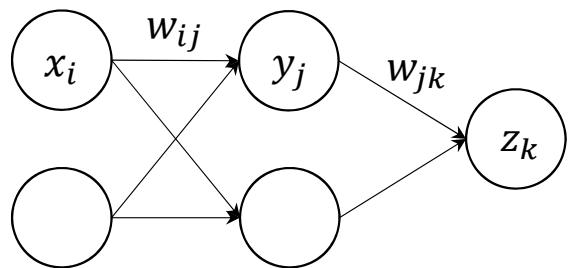


$$E = -\frac{1}{2} \sum_{ij} w_{ij}s_i s_j + \sum_j b_j s_j$$

$$P(s_j = 1) = \frac{e^{-E(s_j=1)}}{e^{-E(s_j=0)} + e^{-E(s_j=1)}} = \frac{1}{1 + e^{-(\sum_i w_{ij}s_i - b_j)}}$$

$$\Delta w_{ij} \propto \langle s_i s_j \rangle_{data} - \langle s_i s_j \rangle_{model}$$

Backpropagation



$$\sigma(s) = \frac{1}{1 + e^{-s}}$$

$$\frac{d\sigma}{ds} = \sigma(1 - \sigma)$$

$$y_j = \sigma \left(\sum_i w_{ij} x_i \right)$$

$$z_k = \sigma \left(\sum_j w_{jk} y_j \right)$$

$$C = \frac{1}{2} \sum_k (z_k - t_k)^2$$

$$\frac{\partial C}{\partial w_{jk}} = (z_k - t_k) \frac{\partial z_k}{\partial w_{jk}} = (z_k - t_k) \underbrace{\sigma'(s_k)}_{\delta_k} y_j = \delta_k y_j$$

$$\begin{aligned} \frac{\partial C}{\partial w_{ij}} &= \sum_k \frac{\partial C}{\partial z_k} \frac{\partial z_k}{\partial y_j} \frac{\partial y_j}{\partial w_{ij}} \\ &= \sum_k (z_k - t_k) \sigma'(s_k) w_{jk} \sigma'(s_j) x_i \\ &= \underbrace{\sum_k \delta_k w_{jk} \sigma'(s_j)}_{\delta_j} x_i = \delta_j x_i \end{aligned}$$

Network topology and machine learning

Can we use the topological features (sparsity, scale-freeness, small-worldness, modularity, assortativity, etc) of real neural networks to better design artificial neural networks for machine learning?

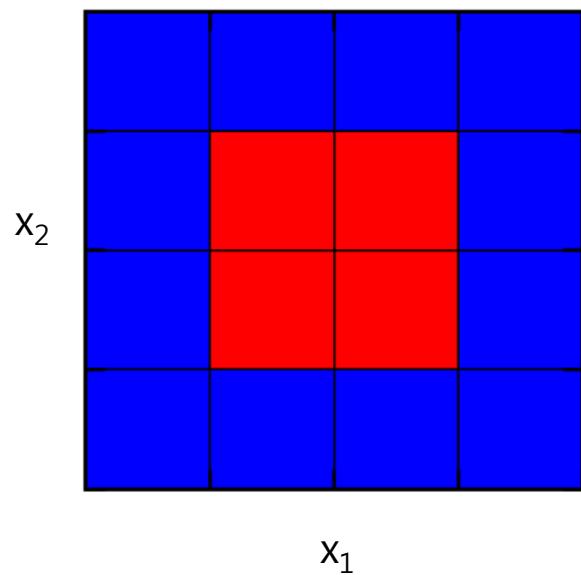
- Task dependence
- Learning-algorithm dependence

I. Pattern recognition with online learning (Jo and Periwal, unpublished)

II. Minimal perceptrons for memorizing complex patterns

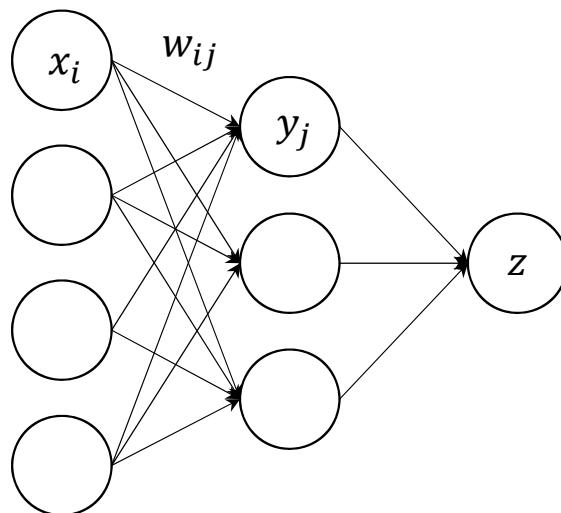
(Pastor, Song, Hoang, and Jo, Physica A 2016)

I. Online learning



e.g., (1,2) → 1 (red)

01 10 → 1



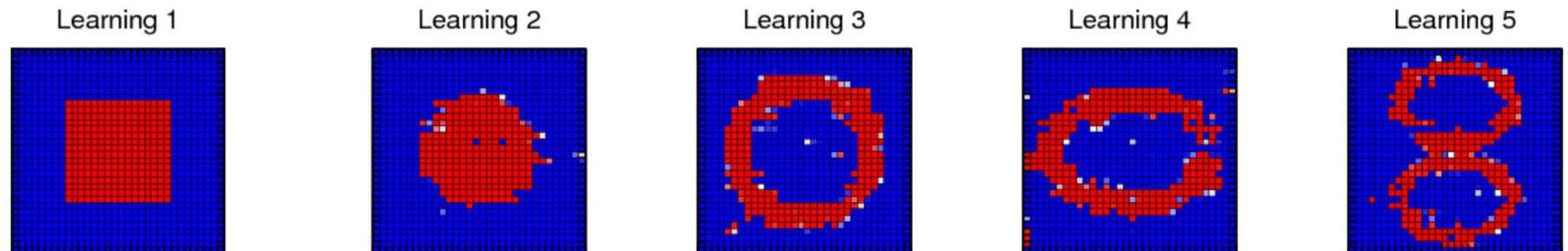
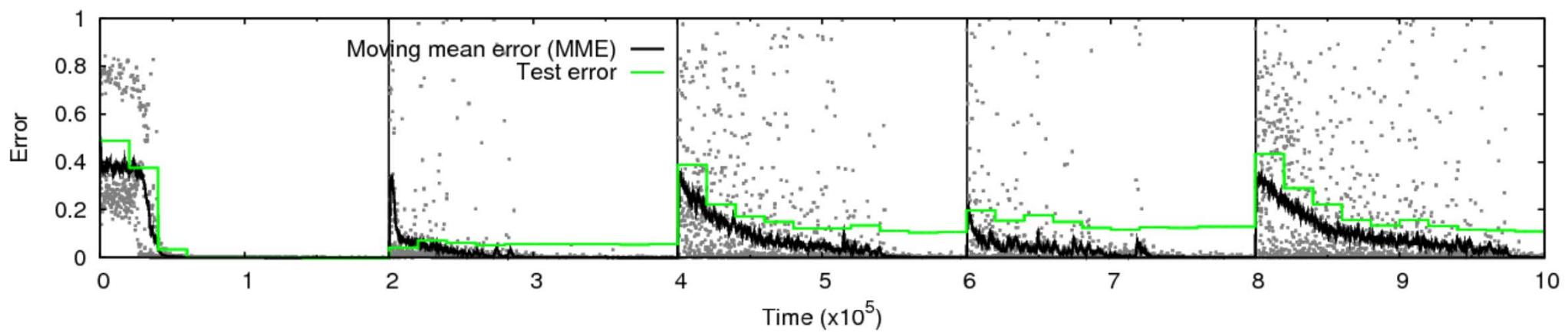
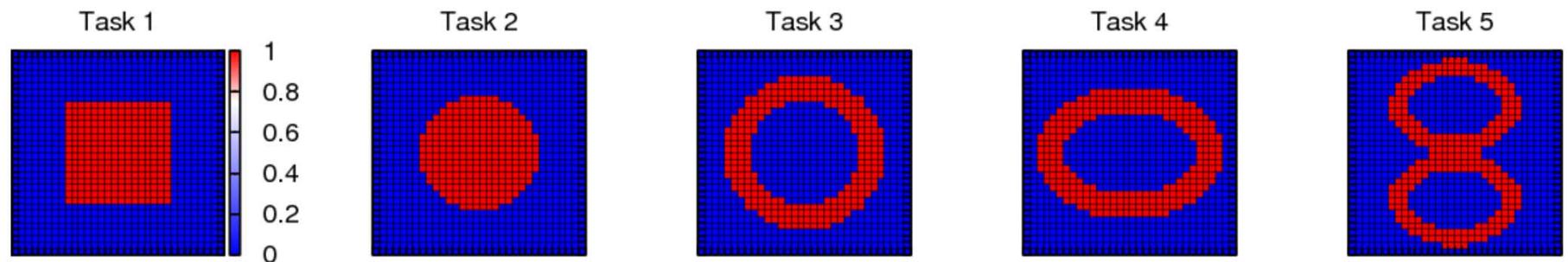
Forward propagation

$$y_j = \sigma \left(\sum_{i=1}^4 w_{ij} x_i - b_j \right)$$

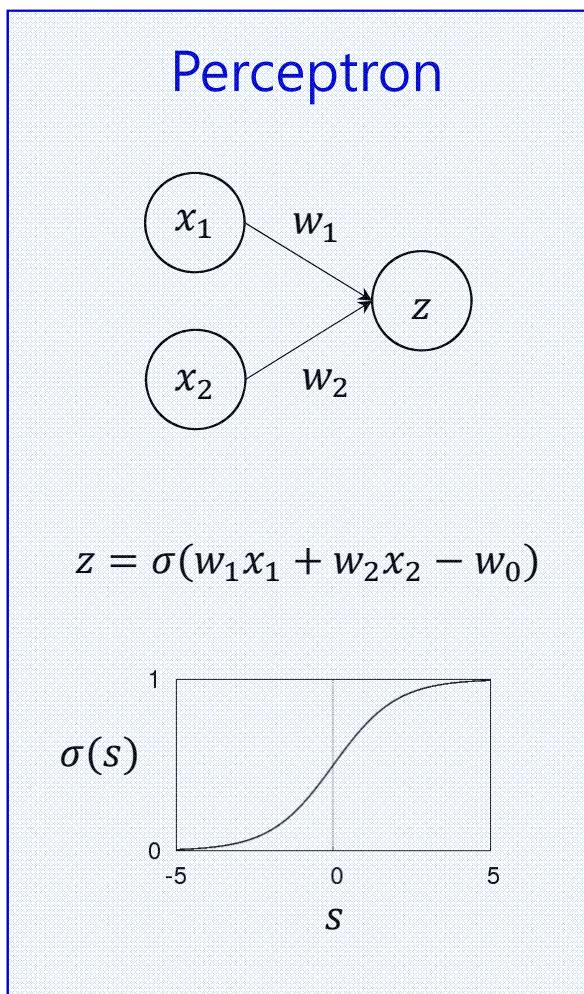
$$\text{Error} = |z - z'| \quad C = \frac{\text{Error}^2}{2}$$

Backward propagation

$$w_{ij} = w_{ij} - \alpha \frac{\partial C}{\partial w_{ij}}$$



II. Minimal perceptrons for binary patterns

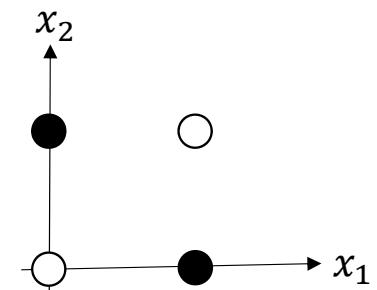
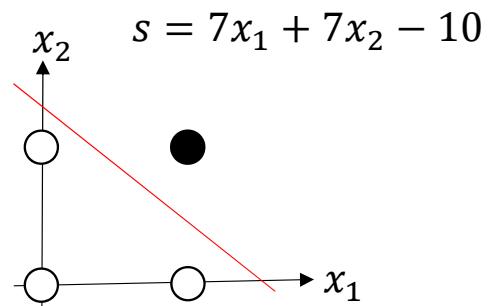


AND problem

x_1	x_2	z'
0	0	0
0	1	0
1	0	0
1	1	1

XOR (parity) problem

x_1	x_2	z'
0	0	0
0	1	1
1	0	1
1	1	0

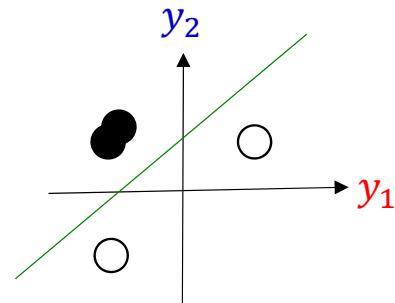
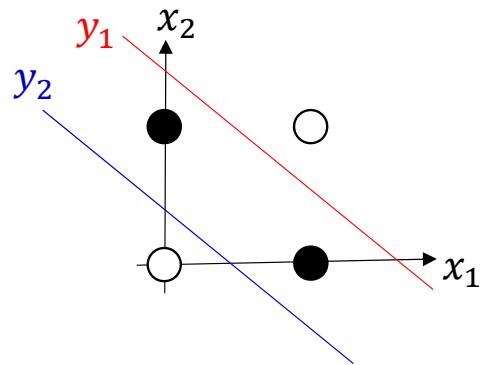


(Marvin Minsky and Seymour Papert, 1969)

Hidden layer

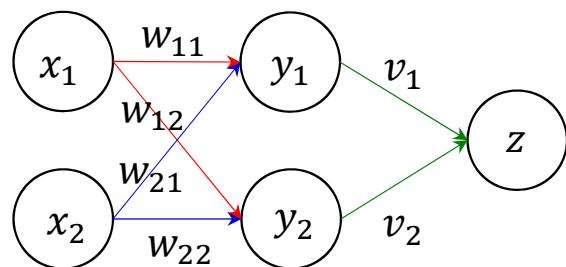
$$w_{11}x_1 + w_{21}x_2 - w_{01} = 0$$

x_1	x_2	z'
0	0	0
0	1	1
1	0	1
1	1	0



$$w_{12}x_1 + w_{22}x_2 - w_{02} = 0$$

$$v_1y_1 + v_2y_2 - v_0 = 0$$



$$N = 2 \rightarrow H = 2$$

$$H = \begin{cases} \frac{N}{2} + 1 & (\text{even } N) \\ \frac{N+1}{2} & (\text{odd } N) \end{cases}$$

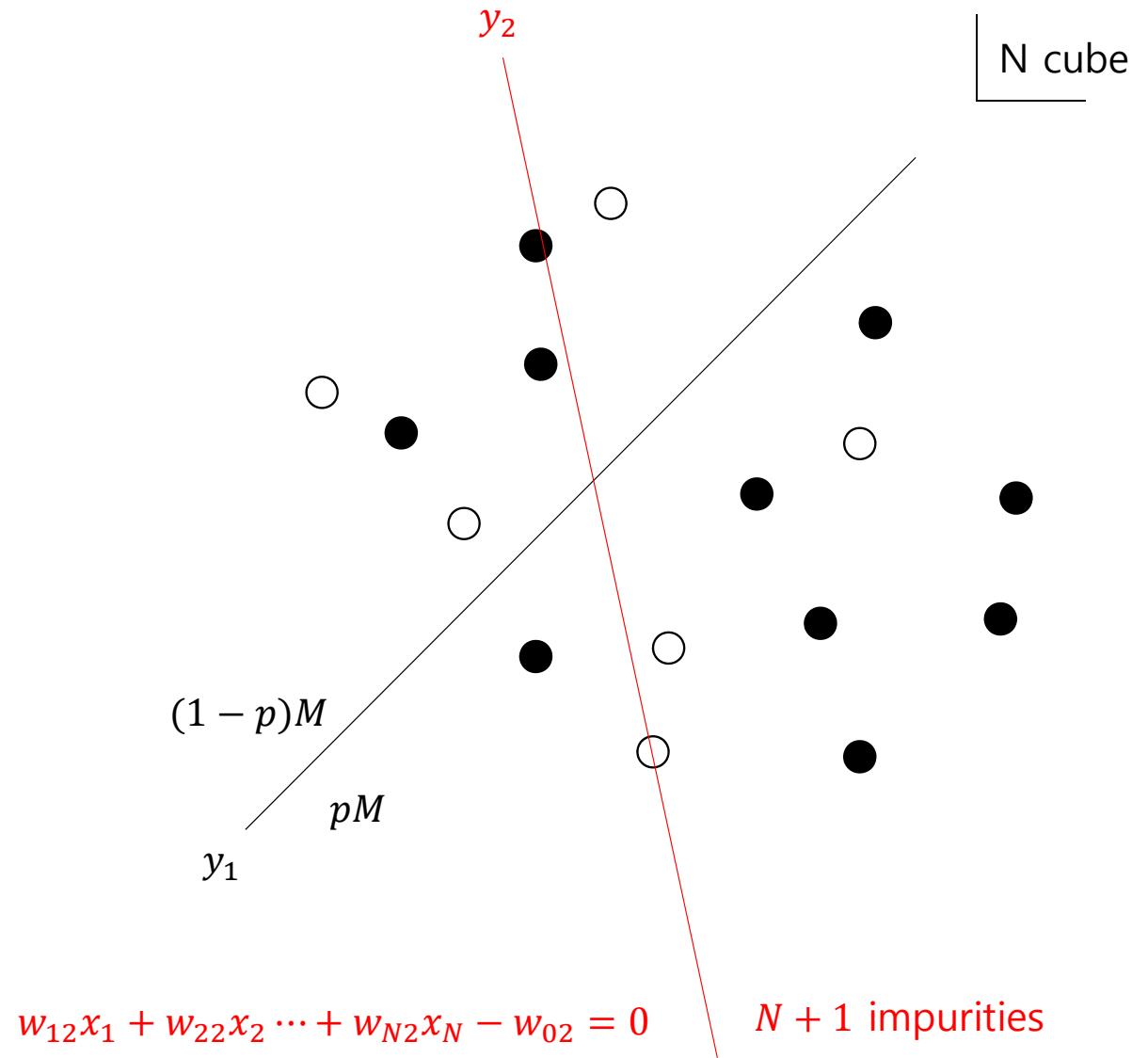
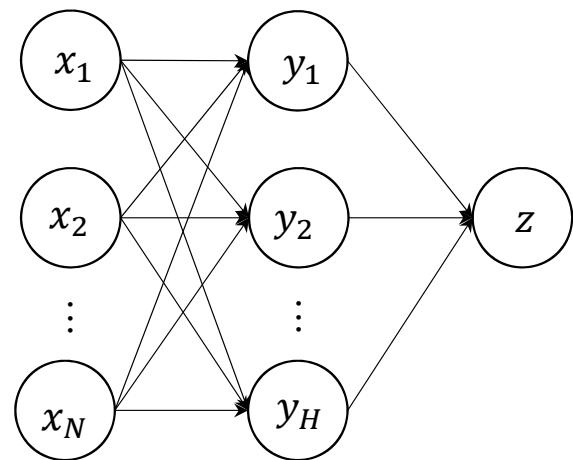
(Songtag, J. Comput. System Sci. 1992)

Random patterns

	x_1	x_2	...	x_N	z'
$M = 2^N$	0	0	...	0	1
	1	0	...	0	0

	1	1	...	1	1

pM : # of $z'=1$ (black)



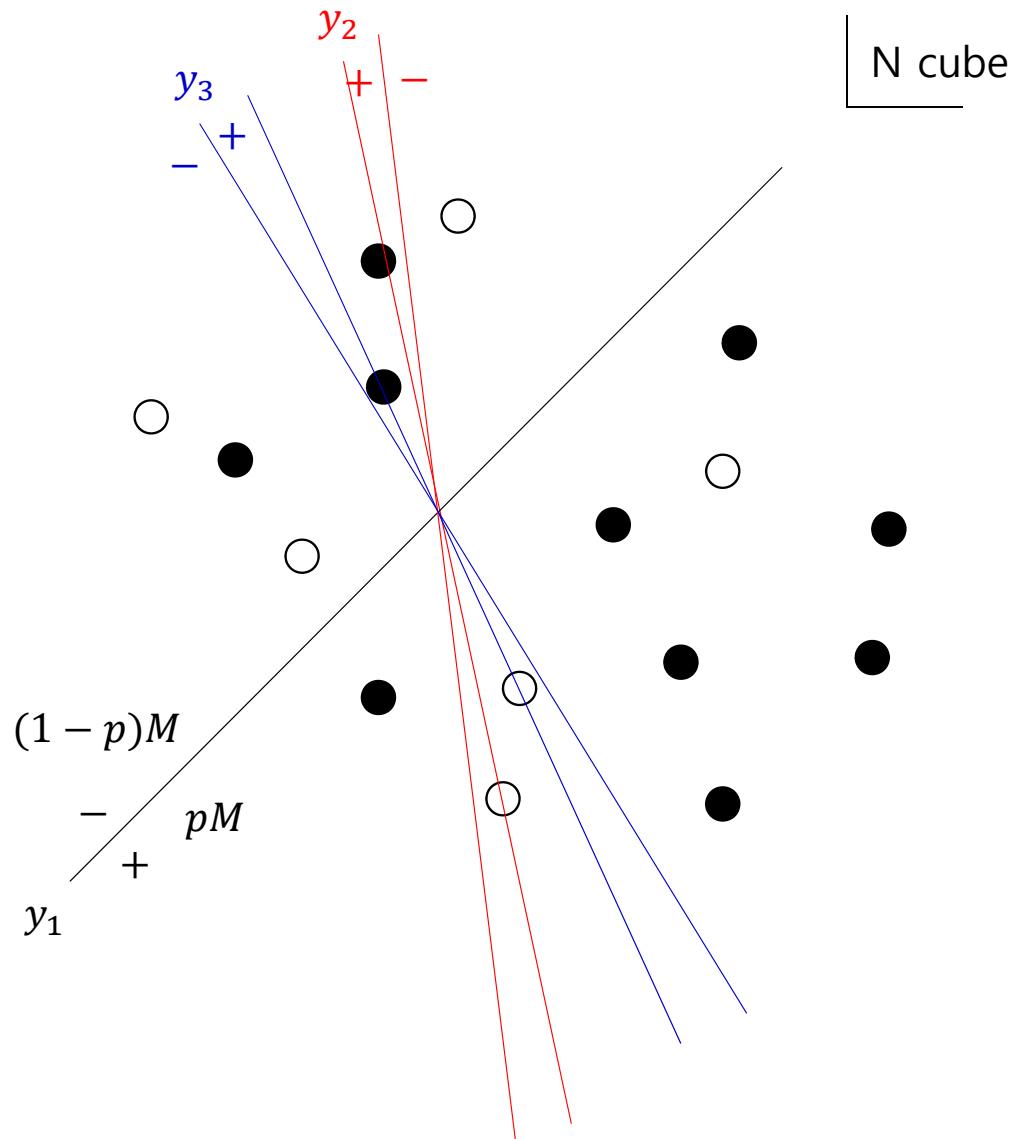
	y_1	y_2	y_3	s
black	1	-1	1	$\textcolor{green}{1}$
	1	1	-1	$\textcolor{green}{1}$
white	-1	-1	1	$\textcolor{green}{-1}$
	-1	1	-1	$\textcolor{green}{-1}$
black impurity	-1	1	1	$\textcolor{green}{1}$
white impurity	1	-1	-1	$\textcolor{green}{-1}$

$$s = y_1 + y_2 + y_3$$

$$\bar{H} = \frac{2p(1-p)M}{N+1} + 1$$

$$\bar{H}' = \frac{2p(1-p)M}{N+1} \left[1 - \frac{p(1-p)(N-1)}{N} \right] + 1$$

$$\mathbf{x} + \mathbf{x}^* = (1, 1, \dots, 1)$$



Machine Learning II

Boltzmann Machine

Learning data

Data: $S = (S_{in}, S_{out})$

$$S^\mu, \mu \in \{1, 2, \dots, M\}$$

Learning:

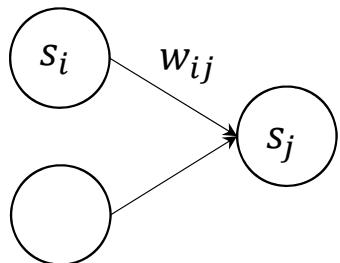
(1) Discriminative model (BP, BM) $S_{in} \rightarrow S_{out}$

(2) Generative model (BM)

$$P(S) = P(S_{in}, S_{out})$$

$$P(S(t+1)|S(t))$$

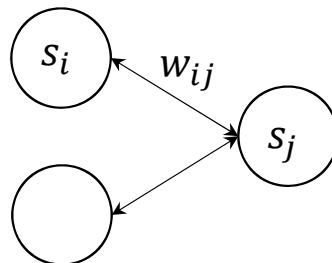
Back Propagation (BP)



$$s_j = \frac{1}{1 + e^{-\sum_i w_{ij}s_i}}$$

deterministic

Boltzmann Machine (BM)



$$P[s_j(t+1) = 1|S(t)] = \frac{e^{\sum_i w_{ij}s_i(t)}}{e^{\sum_i w_{ij}s_i(t)} + e^{-\sum_i w_{ij}s_i(t)}}$$

$$P[s_j(t+1) = -1|S(t)] = \frac{e^{-\sum_i w_{ij}s_i(t)}}{e^{\sum_i w_{ij}s_i(t)} + e^{-\sum_i w_{ij}s_i(t)}}$$

$$\bar{s}_j(t+1) = \tanh \left[\sum_i w_{ij}s_i(t) \right]$$

stochastic

Hopfield model (1982)

Memorize a vector $S = (s_1, s_2, \dots, s_N)$, $s_i = \pm 1$

$$s_j = \operatorname{sgn} \left(\sum_i w_{ij} s_i \right)$$

$w_{ij} = \frac{1}{N} s_i s_j$ Guarantees that S becomes a stable fixed point.

Proof)

$$s_j = \operatorname{sgn} \left(\frac{1}{N} \sum_i s_i s_j s_i \right) = \operatorname{sgn}(s_j) \quad \text{fixed point}$$

Consider S' which is different from S by one bit

$$s'_j = \operatorname{sgn} \left(\frac{1}{N} \sum_i s_i s_j (s_i + \delta s_i) \right) = \operatorname{sgn}(s_j \pm \frac{2}{N} s_j) \approx \operatorname{sgn}(s_j) \quad \text{stable}$$

Hopfield model (1982)

Can we memorize multiple vectors? $S^\mu, \mu \in \{1, 2, \dots, M\}$

$$w_{ij} = \frac{1}{N} \sum_{\mu=1}^M s_i^\mu s_j^\mu$$

Hebb's rule

$$X_1 + X_2 + \dots + X_{N(M-1)} \sim \sqrt{N(M-1)}$$

$$s_j^\alpha = \text{sgn} \left(\frac{1}{N} \sum_i \sum_\mu s_i^\mu s_j^\mu s_i^\alpha \right) = \text{sgn} \left(s_j^\alpha + \underbrace{\frac{1}{N} \sum_i \sum_{\mu \neq \alpha} s_i^\mu s_j^\mu s_i^\alpha}_{\sim \sqrt{\frac{M}{N}}} \right)$$

memorable ($M \ll N$)

Problems of additional attractors

$$-S^\mu, S^\alpha \pm S^\beta, \dots$$

Boltzmann Machine (1985)

(1) Independent data

$$S^\mu, \mu \in \{1, 2, \dots, M\}$$

Data distribution $P(S) = \frac{1}{M} \sum_{\mu=1}^M \delta(S - S^\mu)$

Model distribution $Q(S) = \frac{e^{-E(S)}}{Z}$

$$E(S) = - \sum_{ij} w_{ij} s_i s_j$$

$$Z = \sum_S e^{-E(S)}$$

Kullback-Leibler divergence

$$\begin{aligned} D(P||Q) &= \sum_S P(S) \log \frac{P(S)}{Q(S)} \\ &= \frac{1}{M} \sum_{\mu=1}^M \log Q(S^\mu) + \text{constant} \end{aligned}$$

$$\delta w_{ij} \propto -\frac{\partial D(P||Q)}{\partial w_{ij}}$$

$$D(P||Q) = \sum_S P(S) \log \frac{P(S)}{Q(S)}$$

$$Q(S) = \frac{e^{-E(S)}}{Z}$$

$$E(S) = - \sum_{ij} w_{ij} s_i s_j$$

$$Z = \sum_S e^{-E(S)}$$

$$\delta w_{ij} \propto -\frac{\partial D(P||Q)}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \sum_S P(S) \log Q(S) = \sum_S \frac{P(S)}{Q(S)} \frac{\partial Q(S)}{\partial w_{ij}}$$

$$= \sum_S \frac{P(S)}{Q(S)} \left(\frac{s_i s_j e^{-E(S)} Z - e^{-E(S)} \sum_{S'} s'_i s'_j e^{-E(S')}}{Z^2} \right)$$

$$= \sum_S s_i s_j P(S) - \sum_S P(S) \langle s_i s_j \rangle_Q$$

$$= \langle s_i s_j \rangle_P - \langle s_i s_j \rangle_Q$$

$$\langle s_i s_j \rangle_P = \frac{1}{M} \sum_{\mu=1}^M s_i^\mu s_j^\mu$$

Data correlation

$$\langle s_i s_j \rangle_Q = \sum_S s_i s_j \frac{e^{-E(S)}}{Z}$$

Model correlation

Physical meaning of $D(P||Q)$?

$$\begin{aligned} D(P||Q) &= \sum_S P(S) \log \frac{P(S)}{Q(S)} \quad Q(S) = \frac{e^{-E(S)/T}}{Z} \\ &= \sum_S P(S) \log P(S) - \sum_S P(S) \log Q(S) \\ &= \sum_S P(S) \log P(S) + \sum_S P(S) \frac{E(S)}{T} + \sum_S P(S) \log Z \\ &= -S_P + \frac{\langle E \rangle_P}{T} - \frac{F_Q}{T} \end{aligned}$$

$$T \cdot D(P||Q) = F_P - F_Q \quad F_P \equiv \langle E \rangle_P - T \cdot S_P$$

(Jörg Lücke)

Boltzmann Machine (1985)

(2) Time-sequence data

$$S(t=1), S(t=2), \dots, S(t), \dots$$

Data probability

$$Q(S) = \prod_{t=1} Q[S(t+1)|S(t)] = \prod_{t=1} \prod_{j=1}^N Q[s_j(t+1)|S(t)]$$

$$\log Q = \sum_{t=1} \sum_{j=1}^N [w_{ij} s_i(t) s_j(t+1) - \log(e^{H_j} + e^{-H_j})]$$

$$\delta w_{ij} \propto \frac{\partial \log Q}{\partial w_{ij}} = \sum_{t=1} \left[s_i(t) s_j(t+1) - \frac{e^{H_j} - e^{-H_j}}{e^{H_j} + e^{-H_j}} \frac{\partial H_j}{\partial w_{ij}} \right]$$

$$= \sum_{t=1} [s_i(t) s_j(t+1) - s_i(t) \tanh H_j(t)]$$

$$= \langle s_i(t) s_j(t+1) \rangle_t - \langle s_i(t) \bar{s}_j(t+1) \rangle_t$$

Kullback-Leibler divergence

$$\begin{aligned} D(P||Q) &= \sum_S P(S) \log \frac{P(S)}{Q(S)} \\ &= \frac{1}{M} \sum_{\mu=1}^M \log Q(S^\mu) + \text{constant} \end{aligned}$$

$$Q[s_j(t+1)|S(t)] = \frac{e^{\sum_i w_{ij} s_i(t) s_j(t+1)}}{e^{\sum_i w_{ij} s_i(t)} + e^{-\sum_i w_{ij} s_i(t)}}$$

$$H_j \equiv \sum_i w_{ij} s_i(t)$$

Boltzmann Machine (1985)

(3) Hidden data

$$S = (v, h)$$

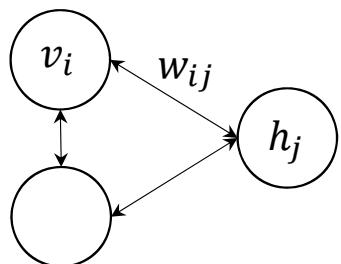
Data distribution $P(v)$

Model distribution

$$Q(v) = \sum_h Q(v, h)$$

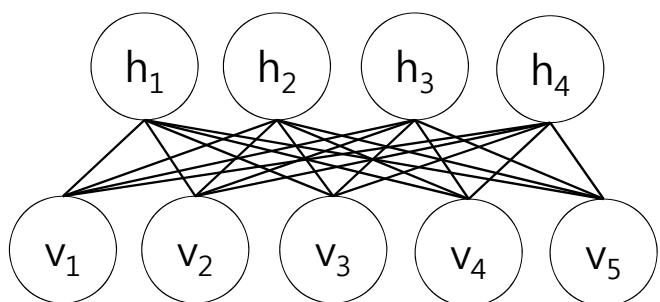
$$Q(s) = Q(v, h) = \frac{e^{-E(s)}}{Z} \quad E(s) = - \sum_{ij} w_{ij} s_i s_j$$

$$D(P||Q) = \sum_v P(v) \log \frac{P(v)}{Q(v)}$$



$$\begin{aligned} \delta w_{ij} &\propto -\frac{\partial D(P||Q)}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \sum_v P(v) \log Q(v) \\ &= \sum_v \frac{P(v)}{Q(v)} \frac{\partial}{\partial w_{ij}} \left(\sum_h \frac{e^{-E(s)}}{Z} \right) \\ &= \sum_v \frac{P(v)}{Q(v)} \left(\sum_h \frac{s_i s_j e^{-E(s)} Z - e^{-E(S)} \sum_{s'} s'_i s'_j e^{-E(s')}}{Z^2} \right) \\ &= \sum_v \frac{P(v)}{Q(v)} \left(\sum_h s_i s_j Q(v, h) - Q(v, h) \langle S_i S_j \rangle_Q \right) \\ &= \sum_{v,h} s_i s_j Q(h|v) P(v) - \langle S_i S_j \rangle_{Q(v,h)} \\ &= \langle s_i s_j \rangle_{P(v,h)} - \langle s_i s_j \rangle_{Q(v,h)} \end{aligned}$$

Restricted Boltzmann Machine (Smolensky, 1986)



$$E(v, h) = - \sum_{ij} x_{ij} v_i v_j - \sum_{ij} y_{ij} h_i h_j - \sum_{ij} w_{ij} v_i h_j$$

$$x_{ij} = y_{ij} = 0$$

$$\delta w_{ij} \propto \langle v_i h_j \rangle_{P(v,h)} - \langle v_i h_j \rangle_{Q(v,h)}$$

$$\langle v_i h_j \rangle_{P(v,h)} = \sum_{v,h} v_i h_j Q(h|v) P(v) = \sum_{v,h} v_i h_j P(v) \prod_j Q(h_j|v) = \sum_{v,h_j} v_i h_j P(v) Q(h_j|v)$$

$$\langle v_i h_j \rangle_{Q(v,h)} = \sum_{v,h} v_i h_j Q(h|v) Q(v) = \sum_{v,h} v_i h_j Q(v) \prod_j Q(h_j|v) = \sum_{v,h_j} v_i h_j Q(v) Q(h_j|v)$$

$$\delta w_{ij} \propto v_i(0)\bar{h}_j(0) - v_i(1)\bar{h}_j(1) \quad (\text{Hinton, 2006})$$