Training Techniques of ANN

(Module 2.2 of syllabus)

Outline

Hebbian learning,
Perceptron Learning,
Delta learning rule,
Widrow Hoff learning,
Winner take all Learning Rule ,
Out star learning

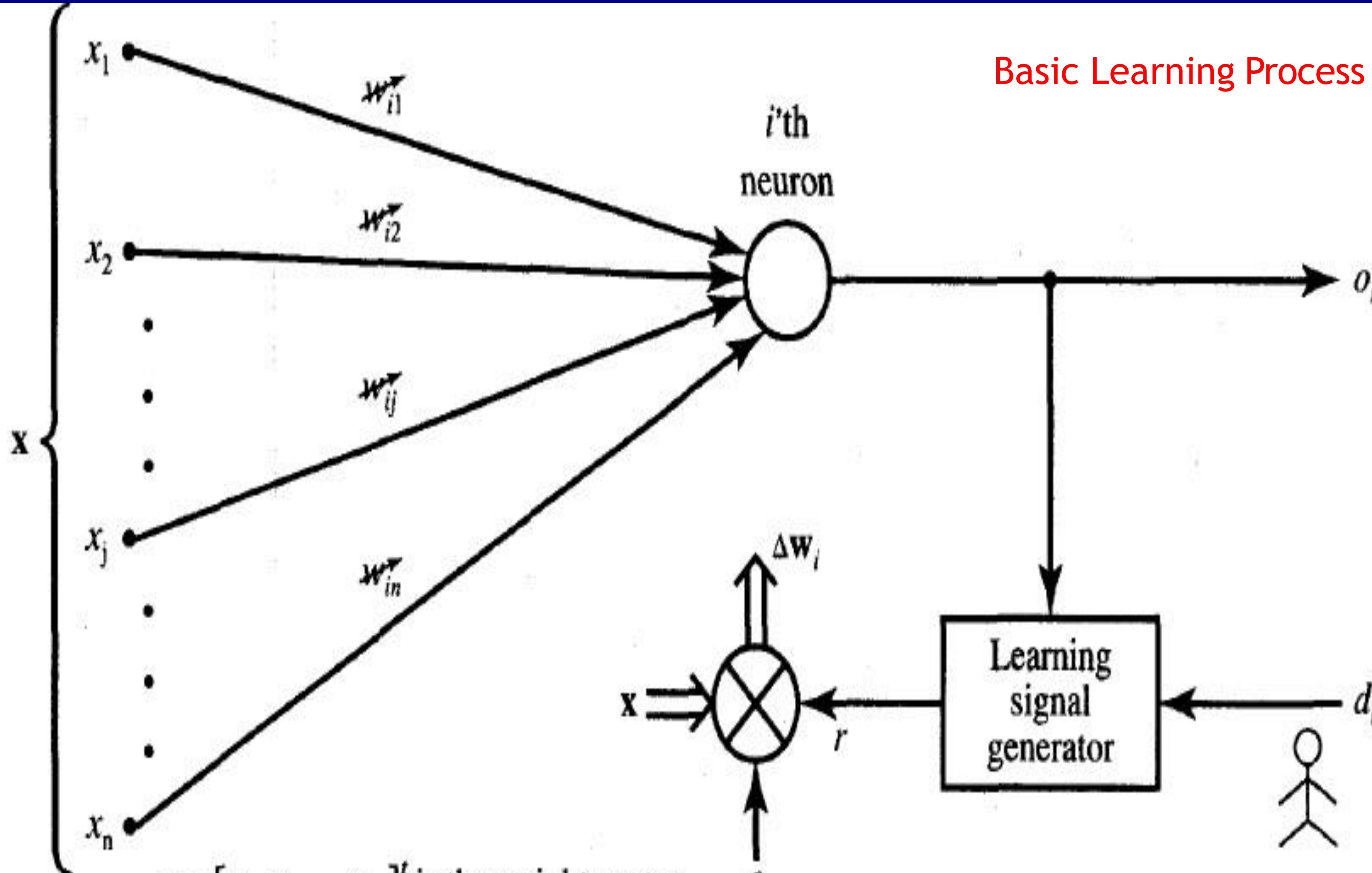
Neural Network Learning Rules

- A neuron →adaptive element.
- Weights are modifiable depending on the input signal it receives, its output value and the associated teacher response.

- In some cases the teacher signal is not available and no error information can be used, thus the neuron will modify its weights based only on the input/output as per unsupervised learning.
- A general rule adapted in neural network studies: the weight vector i.e.

$$\mathbf{w}_i = \begin{bmatrix} w_{i1} & w_{i2} & \cdots & w_{in} \end{bmatrix}^t$$

Basic Learning Process



$\mathbf{w}_i = [w_{i1} w_{i2} \ldots w_{in}]^t$ is the weight vector undergoing training

- Increase in weights proportion to the input x and learning signal r.
- Input learning signal r is a function of w_i, x and summations of the teacher's signal d_i.

$$r = r(\mathbf{w}_i, \mathbf{x}, d_i)$$

- The increment of the weight vector w_i produced by the learning step at time t according to the general rule is given by:

$$\Delta \mathbf{w}_i(t) = cr\left[\mathbf{w}_i(t), \mathbf{x}(t), d_i(t)\right] \mathbf{x}(t)$$

- C → positive number called the learning constant that determines rate of learner.

- The weight vector adapted at time t becomes at the next instant or learning step.

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + cr\left[\mathbf{w}_i(t), \mathbf{x}(t), d_i(t)\right]\mathbf{x}(t)$$

- For the k^{th} step we thus have

$$\mathbf{w}_i^{k+1} = \mathbf{w}_i^k + cr(\mathbf{w}_i^k, \mathbf{x}^k, d_i^k)\mathbf{x}^k$$

- Now continuous time learning can be expressed as

$$\frac{d\mathbf{w}_i(t)}{dt} = cr\mathbf{x}(t)$$

Learning Rules

Hebbian Learning Rule

Hebbian Learning Rule

- Unsupervised learning
- Learning signal = neuron' output (Hebb.- 1949).
- We have

$$r \stackrel{\Delta}{=} f(\mathbf{w}_i^t \mathbf{x})$$

- The increment Δw_i of the weight vector becomes

$$\Delta \mathbf{w}_i = cf(\mathbf{w}_i^t \mathbf{x})\mathbf{x}$$

$$f(net_i) = f(w_i^t . x) = o_i$$

$$net_i = w_i^t . x$$

- Single weight adapted using following increment

$$\Delta w_{ij} = cf(\mathbf{w}_i^t \mathbf{x})x_j$$

$$\Delta w_{ij} = co_i x_j, \quad \text{for } j = 1, 2, \ldots, n$$

Hebbian Learning Rule

- Requires the weight initialization at small random values around $w_i = 0$
- Represents a purely feed forward unsupervised learning.
- The rule states that "If the cross product of output and input or oscillation term $o_i x_j$ is positive then this will result in an increase of weight w_{ij} otherwise the weight decreases"

- Example:- Hebbian learning with binary and continuous activation functions of a very simple network is to be trained using an initial weight vector and three inputs

$$\mathbf{w}^1 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0.5 \end{bmatrix}$$

needs to be trained using the set of three input vectors as below

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -2 \\ 1.5 \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 1 \\ -0.5 \\ -2 \\ -1.5 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 1.5 \end{bmatrix}$$

for an arbitrary choice of learning constant $c = 1$. Since the initial weights are of nonzero value, the network has apparently been trained beforehand. Assume first that bipolar binary neurons are used, and thus $f(net) = \mathrm{sgn}(net)$.

Step 1 Input $\mathbf{x}_1$ applied to the network results in activation net^1 as below:

$$net^1 = \mathbf{w}^{1t}\mathbf{x}_1 = \begin{bmatrix} 1 & -1 & 0 & 0.5 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \\ 1.5 \\ 0 \end{bmatrix} = 3$$

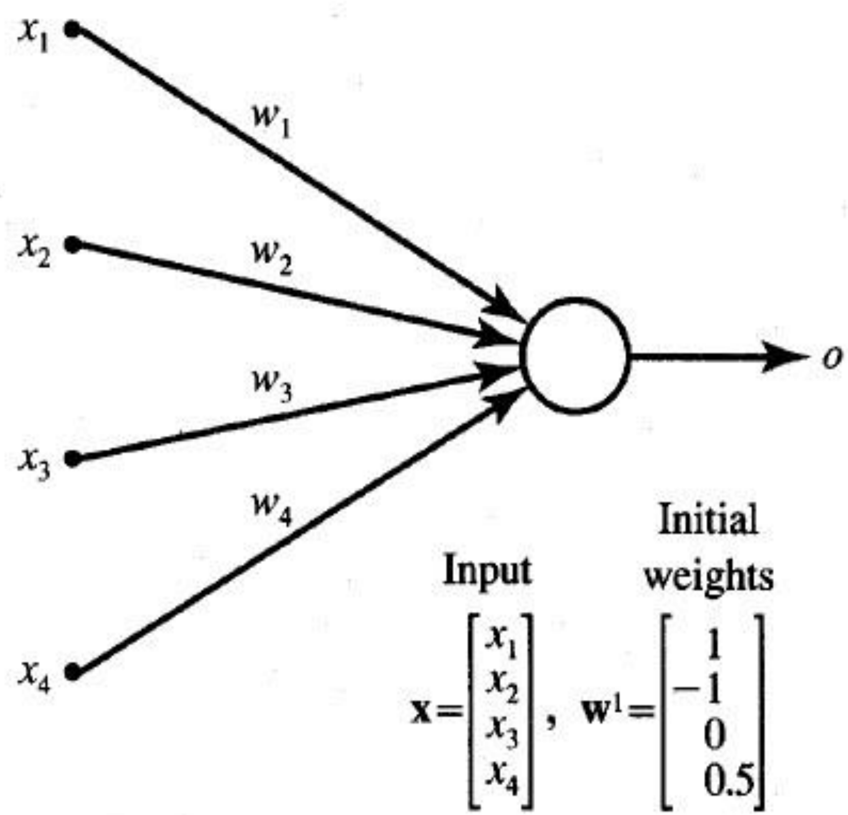


Figure 2.22 Network for training in Examples 2.4 through 2.6.

The updated weights are

$$\mathbf{w}^2 = \mathbf{w}^1 + \mathrm{sgn}\,(net^1)\mathbf{x}_1 = \mathbf{w}^1 + \mathbf{x}_1$$

and plugging numerical values we obtain

$$\mathbf{w}^2 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0.5 \end{bmatrix} + \begin{bmatrix} 1 \\ -2 \\ 1.5 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \\ 1.5 \\ 0.5 \end{bmatrix}$$

where the superscript on the right side of the expression denotes the number of the current adjustment step.

Step 2 This learning step is with $\mathbf{x}_2$ as input:

$$net^2 = \mathbf{w}^{2t}\mathbf{x}_2 = \begin{bmatrix} 2 & -3 & 1.5 & 0.5 \end{bmatrix} \begin{bmatrix} 1 \\ -0.5 \\ -2 \\ -1.5 \end{bmatrix} = -0.25$$

The updated weights are

$$\mathbf{w}^3 = \mathbf{w}^2 + \operatorname{sgn}(net^2)\mathbf{x}_2 = \mathbf{w}^2 - \mathbf{x}_2 = \begin{bmatrix} 1 \\ -2.5 \\ 3.5 \\ 2 \end{bmatrix}$$

Step 3 For input $\mathbf{x}_3$, we obtain in this step

$$net^3 = \mathbf{w}^{3t}\mathbf{x}_3 = \begin{bmatrix} 1 & -2.5 & 3.5 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -1 \\ 1.5 \end{bmatrix} = -3$$

The updated weights are

$$\mathbf{w}^4 = \mathbf{w}^3 + \operatorname{sgn}(net^3)\mathbf{x}_3 = \mathbf{w}^3 - \mathbf{x}_3 = \begin{bmatrix} 1 \\ -3.5 \\ 4.5 \\ 0.5 \end{bmatrix}$$

- Revisiting the Hebbian learning example with continuous bipolar activation function f(net), using input x_1 and initial weight w^1 we obtain neuron output values and updated weights for λ=1.
- The only difference compared with the previous case is that instead of f(net)=sigm(net). Now the neurons response is computed-

Step 1

$$f(net^1) = 0.905$$

$$\mathbf{w}^2 = \begin{bmatrix} 1.905 \\ -2.81 \\ 1.357 \\ 0.5 \end{bmatrix}$$

Subsequent training steps result in weight vector adjustment as below:

Step 2

$$f(net^2) = -0.077$$

$$\mathbf{w}^3 = \begin{bmatrix} 1.828 \\ -2.772 \\ 1.512 \\ 0.616 \end{bmatrix}$$

Step 3

$$f(net^3) = -0.932$$

$$\mathbf{w}^4 = \begin{bmatrix} 1.828 \\ -3.70 \\ 2.44 \\ -0.783 \end{bmatrix}$$

Comparison of learning using discrete and continuous activation functions indicates that the weight adjustments are tapered for continuous $f(net)$ but are generally in the same direction. ■

Perceptron Learning Rule

Perceptron Learning Rule

- Learning signal = ***difference*** between the ***desired and the actual neuron's response*** (ROSENBLATT-1958).
- **Supervised** learning
- Learning signal=

$$r \overset{\Delta}{=} d_i - o_i$$

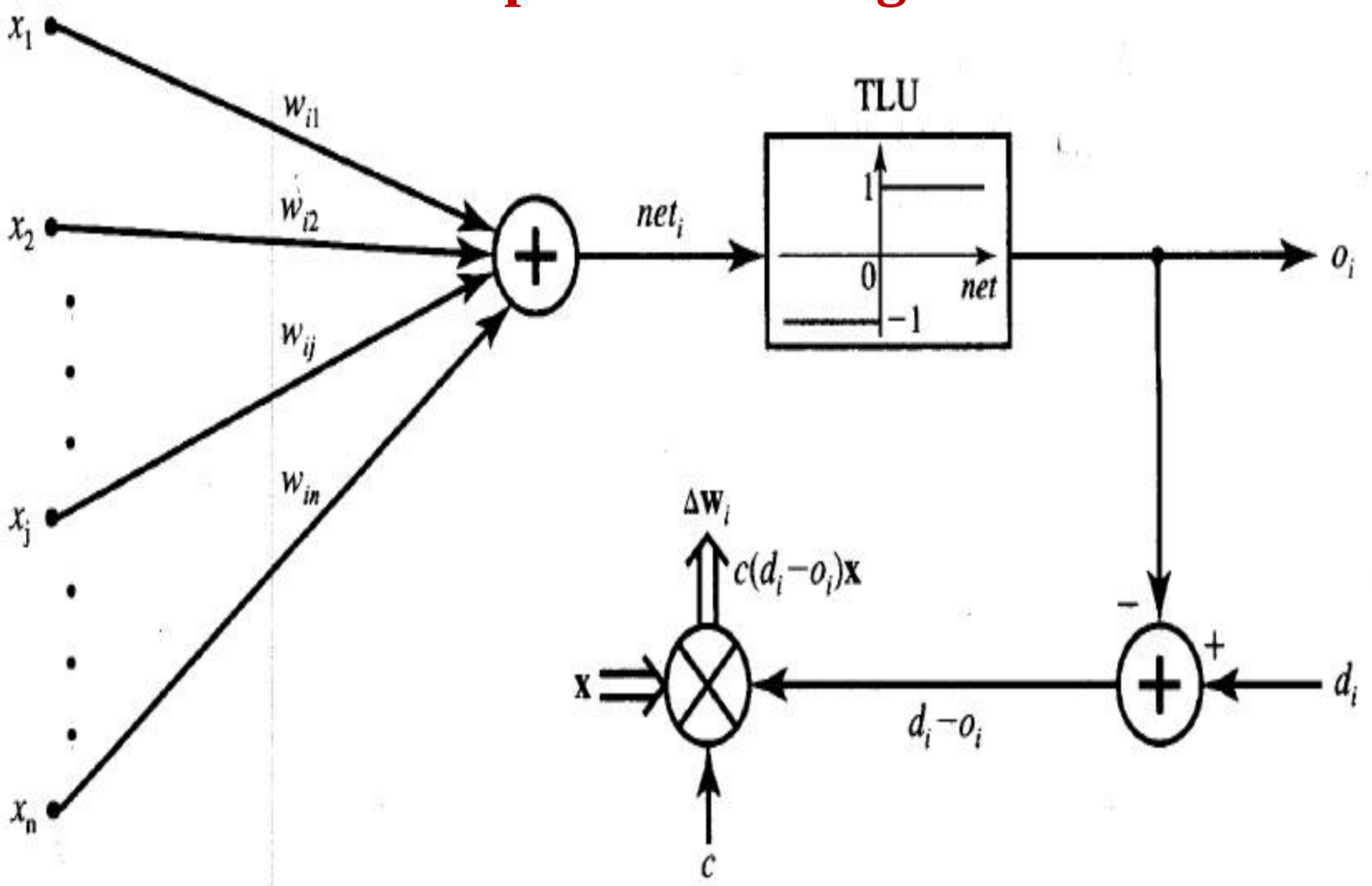
where $o_i = \text{sgn}(\mathbf{w}_i^t\mathbf{x})$, and d_i is the desired response

Weight adjustments in this method, $\Delta\mathbf{w}_i$ and Δw_{ij}, are obtained as follows

$$\Delta\mathbf{w}_i = c\left[d_i - \text{sgn}(\mathbf{w}_i^t\mathbf{x})\right]\mathbf{x}$$

$$\Delta w_{ij} = c\left[d_i - \text{sgn}(\mathbf{w}_i^t\mathbf{x})\right]x_j, \quad \text{for } j = 1, 2, \ldots, n$$

Perceptron Learning Rule



[Perceptron Learning Rule]

Perceptron Learning Rule

- ***Note-1:-*** This rule is applicable only for binary neuron response and the relationship express the rule for the binary case.
- ***Note-2:-*** Under this rule weight is only adjusted if and only if o_i is incorrect.
- ***Note-3:-*** Weight can be initialized at any value since the desired response is either +1 or -1, the weight adjustment reduce to

$$\Delta \mathbf{w}_i = \pm 2c\mathbf{x}$$

- +ve sign is applicable when

$$d_i = 1, \text{ and } \operatorname{sgn}(\mathbf{w}^t\mathbf{x}) = -1,$$

- -ve sign is applicable when

$$d_i = -1, \text{ and } \operatorname{sgn}(\mathbf{w}^t\mathbf{x}) = 1.$$

Delta Learning Rule

Continuous Perceptron Learning rule

Delta Learning Rule

- Valid for the continuous activation functions i.e.

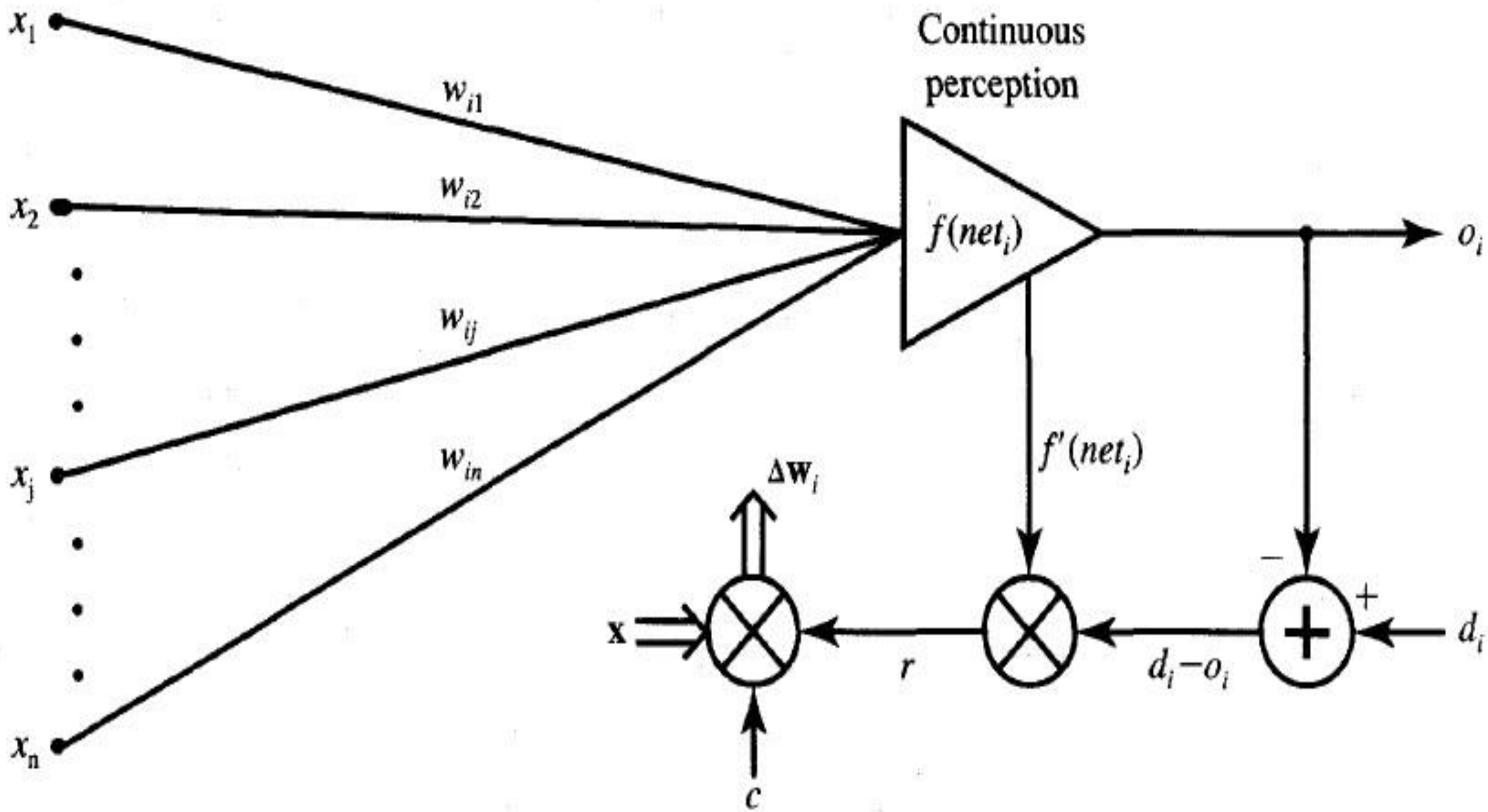
$$f(net) \triangleq \frac{2}{1+\exp(-\lambda net)} - 1 \quad \text{...Eqn. (A)}$$

$$f(net) \triangleq \operatorname{sgn}(net) = \begin{cases} +1, & net > 0 \\ -1, & net < 0 \end{cases} \quad \text{..Eqn. (B)}$$

- Supervised learning
- Learning signal for this rule called delta and is defined as:

$$r \triangleq [d_i - f(\mathbf{w}_i^t \mathbf{x})] f'(\mathbf{w}_i^t \mathbf{x})$$

Delta Learning Rule



Delta Learning Rule

- $f'(\mathbf{w}_i^t\mathbf{x})$ → derivative of activation function f(net) for $net = \mathbf{w}_i^t\mathbf{x}$.
- Readily derived from the condition of least-square error between o_i and d_i.
- Calculating the gradient vector w.r.t. w_i of square error defined as:

$$E \triangleq \frac{1}{2}(d_i - o_i)^2$$

which is equivalent to

$$E = \frac{1}{2}\left[d_i - f(\mathbf{w}_i^t\mathbf{x})\right]^2$$

we obtain the error gradient vector value

$$\nabla E = -(d_i - o_i)f'(\mathbf{w}_i^t\mathbf{x})\mathbf{x}$$

Delta Learning Rule

- -ve sign is present because we want to move the weight vector in the direction that decrease (E).

- The gradient rectifies the direction that produces the steepest increase in E. The –ve of this vector therefore gives the direction of steepest decrease.

The components of the gradient vector are

$$\frac{\partial E}{\partial w_{ij}} = -(d_i - o_i) f'(\mathbf{w}_i^t \mathbf{x}) x_j, \quad \text{for } j = 1, 2, \ldots, n$$

Delta Learning Rule

- Minimization of error requires the weight changes to be in the gradient direction, therefore

$$\Delta \mathbf{w}_i = -\eta \nabla E$$ [Where η is a +ve constant]

- From the above equations

$$\Delta \mathbf{w}_i = \eta(d_i - o_i) f'(net_i)\mathbf{x}$$

- Or for the single weight the adjustment will be

$$\Delta w_{ij} = \eta(d_i - o_i) f'(\mathrm{net}_i) x_j, \quad \text{for } j = 1, 2, \ldots, n$$

- Note: weight adjustment is computed based on minimization of the squared error

Delta Learning Rule

- Considering the general learning rule and plugging in the learning signed as defined in

$$\Delta \mathbf{w}_i = c(d_i - o_i)f'(\mathrm{net}_i)\mathbf{x}$$

$$\Delta w_i = \frac{1}{2}(d_i - o_i)(1 - o_i^2)c.x$$

- Since c and η have been assumed to be arbitrary constant.
- Weights are initialized at any value for this method of training.

$$f'(net) = \frac{1}{2}(1 - o^2)$$

For Bipolar activation function

$$f'(net) = o(1 - o)$$

For Unipolar activation function

Winner-Take-All Learning Rule

Competitive Learning Rule

Winner-Take-All Learning Rule

- Winner-Take-All learning rule differs from the other rules.
- Used for unsupervised learning, rather it is used for learning statistical properties of inputs.
- Learning is based on the concept that m^{th} neuron has the maximum response due to input x. That neuron is declared as winner.

$$W_m = [w_{m1}, w_{m2}, \ldots\ldots\ldots\ldots\ldots w_{mn}]^t$$

- In the following figure
  - 1) neurons are arranged in a layer of permits.
  - 2) adjusted weights are highlighted.

Winner-Take-All Learning Rule

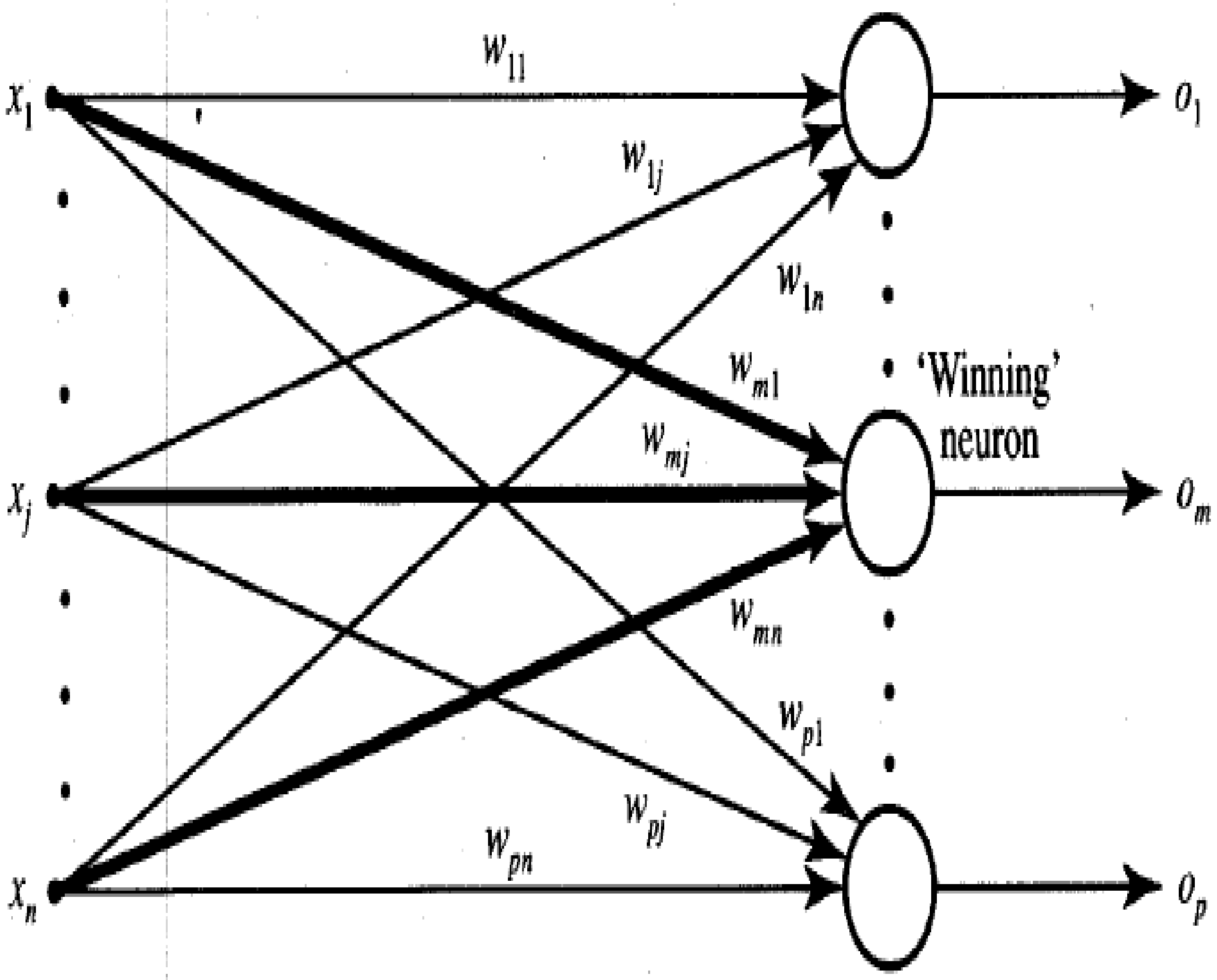


Fig: Winner Take All Learning Rule

Winner-Take-All Learning Rule

containing weights highlighted in the figure is the only one adjusted in the given unsupervised learning step. Its increment is computed as follows

$$\Delta \mathbf{w}_m = \alpha(\mathbf{x} - \mathbf{w}_m)$$

or, the individual weight adjustment becomes

$$\Delta w_{mj} = \alpha(x_j - w_{mj}), \quad \text{for } j = 1, 2, \ldots, n$$

where $\alpha > 0$ is a small learning constant, typically decreasing as learning progresses. The winner selection is based on the following criterion of maximum activation among all p neurons participating in a competition:

$$\mathbf{w}_m^t \mathbf{x} = \max_{i=1,2,\ldots,p} (\mathbf{w}_i^t \mathbf{x})$$

Outstar Learning Rule

Outstar Learning Rule

- Works for layer of neurons.
- Supervised learning.
- Allows the network to extract statistical properties of the input and output signal.
- The adjustment weight computed as follows:

$$\Delta \mathbf{w}_j = \beta(\mathbf{d} - \mathbf{w}_j)$$

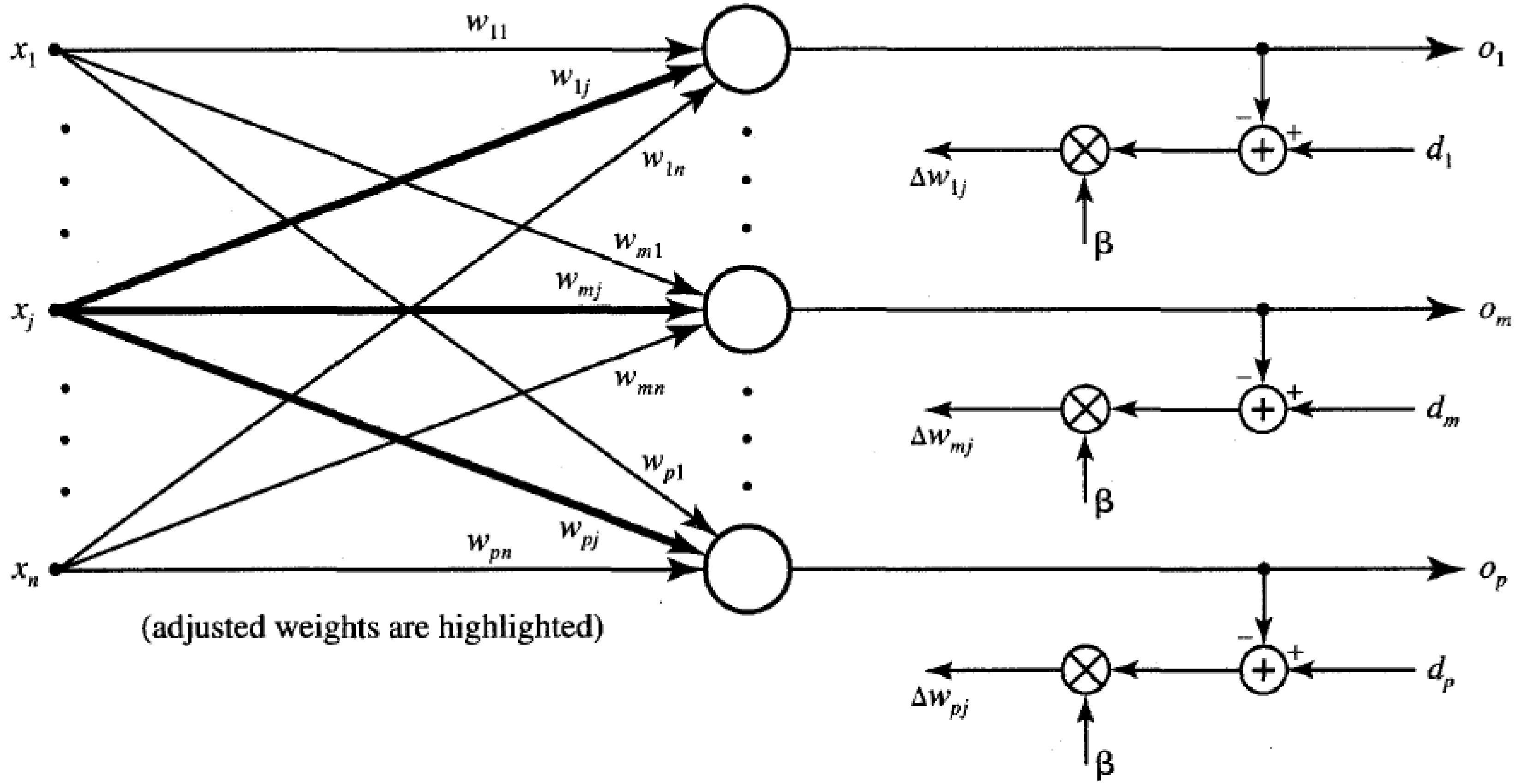
or, the individual weight adjustments are

$$\Delta w_{mj} = \beta(d_m - w_{mj}), \quad \text{for } m = 1, 2, \ldots, p$$

Outstar Learning Rule

Note that in contrast to any learning rule discussed so far, the adjusted weights are fanning out of the j'th node in this learning method and the weight vector is defined accordingly as

$$\mathbf{w}_j \triangleq \begin{bmatrix} w_{1j} & w_{2j} & \cdots & w_{pj} \end{bmatrix}^t$$



LMS or Widrow-Hoff

LMS or Widrow-Hoff

- Least Mean Square error (LMS or Widrow-Hoff)
- Supervised training
- Widrow and Hoff had the insight that they could estimate the mean square error by using the squared error at each iteration
- An approximate steepest descent algorithm, in which the performance index is mean square error.
- Widely used today in many signal processing applications.
- Precursor to the back propagation algorithm for multilayer networks.

LMS or Widrow-Hoff

- Special case of Delta Learning rule

$$net_i = w_i^t x$$

$$\Delta w_i = C\left(d_i - o_i\right) f' net_i \ x$$

$$net_i = 1$$

$$\Delta w_i = C\left(d_i - o_i\right) x$$

$$\Delta w_i = C\left(d_i - w_i^t x\right) x$$

LMS or Widrow-Hoff

Mean Square Error

- Network output compared to the target.
- **The error** is calculated as the difference between the target output and the network output.
- Try to minimize the average of the sum of these errors.

$$mse = \frac{1}{Q}\sum_{k=1}^{Q} e(k)^2 = \frac{1}{Q}\sum_{k=1}^{Q} (t(k) - a(k))^2$$

- The **LMS** algorithm adjusts the weights and biases of the linear network so as to **minimize this mean square error.**

Summary of learning rules and their properties.

Learning rule	Single weight adjustment Δw_{ij}	Initial weights	Learning	Neuron characteristics	Neuron / Layer
Hebbian	$co_i x_j$ $j = 1, 2, \ldots, n$	0	U	Any	Neuron
Perceptron	$c\left[d_i - \operatorname{sgn}(\mathbf{w}_i^t \mathbf{x})\right] x_j$ $j = 1, 2, \ldots, n$	Any	S	Binary bipolar, or Binary unipolar*	Neuron
Delta	$c(d_i - o_i)f'(net_i)x_j$ $j = 1, 2, \ldots, n$	Any	S	Continuous	Neuron
Widrow-Hoff	$c(d_i - \mathbf{w}_i^t \mathbf{x})x_j$ $j = 1, 2, \ldots, n$	Any	S	Any	Neuron
Correlation	$cd_i x_j$ $j = 1, 2, \ldots, n$	0	S	Any	Neuron
Winner-take-all	$\Delta w_{mj} = \alpha(x_j - w_{mj})$ m-winning neuron number $j = 1, 2, \ldots, n$	Random Normalized	U	Continuous	Layer of p neurons
Outstar	$\beta(d_i - w_{ij})$ $i = 1, 2, \ldots, p$	0	S	Continuous	Layer of p neurons

c, α, β are positive learning constants
S — supervised learning, U — unsupervised learning
* — Δw_{ij} not shown