Εισαγωγή στα Τεχνητά Νευρωνικά Δίκτυα και στη Μηχανική Μάθηση

(Διασκευή διαφανειών από ΕΑΠ-ΠΛΗ31)

Διδάσκων:

Ι. ΧΑΤΖΗΛΥΓΕΡΟΥΔΗΣ

Πανεπιστήμιο Πατρών, Τμήμα Μηχ/κών Η/Υ και Πληροφορικής

Τεχνητά Νευρωνικά Δίκτυα-ΤΝΔ (Artificial Neural Networks-ANN)

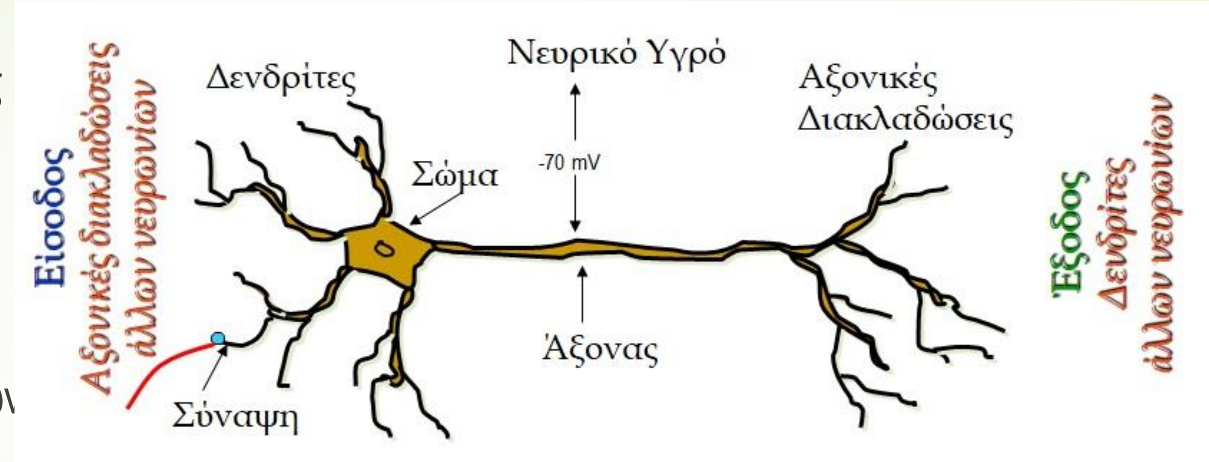
- ΤΝΔ: προέκυψαν από την ανάγκη να φτιάξουμε **υπολογιστικά μοντέλα** του ανθρώπινου εγκεφάλου (βιολογικά νευρωνικά δίκτυα)
- Πρόβλημα: δεν ξέρουμε ακόμα (με ακρίβεια) πώς λειτουργεί ο ανθρώπινος εγκέφαλος!
- 1950: απλουστευμένα μαθηματικά μοντέλα του εγκεφάλου.
- Τα πρώτα ΤΝΔ: προσομοίωση αυτών των μοντέλων σε υπολογιστή (επίλυση στοιχειωδών προβλημάτων)

Τεχνητά Νευρωνικά Δίκτυα-ΤΝΔ (Artificial Neural Networks-ANN)

- Ο εγκέφαλος αποτελείται από ένα τεράστιο αριθμό διασυνδεδεμένων **νευρώνων (neurons**), δηλαδή νευρικών κυττάρων.
- Κάθε νευρώνας
  - ✓ δέχεται ερεθίσματα (εισόδους) από άλλα κύτταρα μέσω συνδέσεων τα οποία επηρεάζουν την κατάστασή του και, ανάλογα με την κατάσταση στην οποία βρίσκεται
  - ✓ στέλνει ερεθίσματα (εξόδους) για να επηρεάσει με τη σειρά του την κατάσταση άλλων νευρώνων.
- Κάθε σύνδεση μεταξύ δύο νευρώνων χαρακτηρίζεται από μια **τιμή ισχύος (συναπτικό δυναμικό)** η οποία υποδηλώνει πόσο ισχυρή είναι η μεταξύ τους αλληλεπίδραση.

Ο Βιολογικός Νευρώνας

- **Δενδρίτες**, που αποτελούν τις γραμμές εισόδου των ερεθισμάτων (βιολογικών σημάτων)
- **Σώμα**, στο οποίο γίνεται η συσσώρευση των ερεθισμάτων και ο καθορισμός της διέγερσης του νευρώνα.
- **Νευροάξονας**, που αποτελεί τη γραμμή εξόδου του νευρώνα.
- **Σύναψη**, που είναι το σημείο διασύνδεσης μεταξύ δύο νευρώνων.



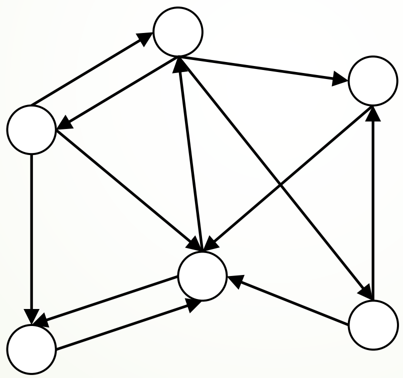
Έχει παρατηρηθεί ότι το σήμα που εξέρχεται από το νευροάξονα ενός νευρώνα και εισέρχεται στο δενδρίτη του άλλου νευρώνα **διαμορφώνεται** κατά ένα ποσοστό που σχετίζεται με την **ισχύ της σύναψης** που ονομάζεται **συναπτικό δυναμικό**.

Συναπτικό Δυναμικό (Synaptic Potential)

- Το συναπτικό δυναμικό μπορεί να ενισχύει (θετικό) ή να καταστέλλει (αρνητικό) το σήμα εξόδου.
- Η γνώση μας είναι ‘αποθηκευμένη’ στις τιμές των συναπτικών δυναμικών.
- **Μάθηση στα βιολογικά συστήματα είναι η μεταβολή των συναπτικών δυναμικών.**
- Όσο περισσότερο χρησιμοποιείται μια σύναψη τόσο ενισχύεται το δυναμικό της.

Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ)

- Αρχιτεκτονικές δομές (**δίκτυα**) αποτελούμενη από ένα πλήθος διασυνδεδεμένων **μονάδων επεξεργασίας** (τεχνητοί νευρώνες).
- Κάθε **σύνδεση** μεταξύ δύο μονάδων χαρακτηρίζεται από μια **τιμή βάρους**.



- Κάθε μονάδα επεξεργασίας χαρακτηρίζεται από **εισόδους** και **εξόδους**. Υλοποιεί τοπικά έναν **απλό υπολογισμό** με βάση τις εισόδους που δέχεται και μεταδίδει το αποτέλεσμα (έξοδος) σε άλλες μονάδες επεξεργασίας με τις οποίες συνδέεται.

Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ)

- Οι **τιμές των βαρών** των συνδέσεων αποτελούν τη **γνώση** που είναι αποθηκευμένη στο ΤΝΔ και καθορίζουν τη **λειτουργικότητά του**.
- Συνήθως ένα ΤΝΔ αναπτύσσει μία συνολική λειτουργικότητα μέσω μιας μορφής **εκπαίδευσης** (μάθησης).

Δυνατότητες των ΤΝΔ

- Βασικές ικανότητες του **ανθρώπινου εγκεφάλου**
  - ✓ **Μάθηση από παραδείγματα**
  - ✓ **Ικανότητα Γενίκευσης**
  - ✓ Αποθηκεύει εμπειρίες (κατανεμημένη αποθήκευση)
  - ✓ Αυτοοργάνωση
  - ✓ Ανοχή σε θόρυβο και ελλιπείς πληροφορίες
  - ✓ Ανοχή σε βλάβες
- Οι ικανότητες του εγκεφάλου συμπληρωματικές ως προς τους συμβατικούς υπολογιστές
- Τις παραπάνω δυνατότητες έχουν (σε κάποιο βαθμό) και τα Τεχνητά Νευρωνικά Δίκτυα

Μηχανική Μάθηση (Machine Learning)

- **Εκπαίδευση ενός ΤΝΔ**:
  - ✓ **καθορισμός των βαρών** των συνδέσεών του έτσι ώστε να επιτελείται μια **επιθυμητή λειτουργία** η οποία περιγράφεται με τη **χρήση παραδειγμάτων**
- **Ικανότητα Γενίκευσης:**
  - ✓ Ο αντικειμενικός στόχος της διαδικασίας εκπαίδευσης: να αποκτήσει δηλαδή το ΤΝΔ κατάλληλες τιμές βαρών ώστε να ‘δίνει σωστές απαντήσεις’ για παραδείγματα που ‘μοιάζουν’ σε αυτά με τα οποία εκπαιδεύτηκε
- Τα ΤΝΔ έχουν αποδειχθεί μια επιτυχημένη **τεχνολογία για την ανάπτυξη συστημάτων με καλή γενικευτική ικανότητα χρησιμοποιώντας ένα σύνολο από αντιπροσωπευτικά παραδείγματα εκπαίδευσης.**

Κατηγορίες Μάθησης από Παραδείγματα

- **Επιβλεπόμενη Μάθηση** ή μάθηση με επίβλεψη (supervised learning)
- **Μη Επιβλεπόμενη Μάθηση** ή μάθηση χωρίς επίβλεψη ή (unsupervised learning)
- **Ενισχυτική Μάθηση** ή μάθηση με ενίσχυση (reinforcement learning)
- **Ημιεπιβλεπόμενη Μάθηση** ή μάθηση με ημιεπίβλεψη (semi-supervised learning)

Επιβλεπόμενη Μάθηση

- Τα στοιχεία του συνόλου των παραδειγμάτων είναι **ζεύγη** της μορφής: **(είσοδος, επιθυμητή έξοδος)** **($X=\{x^i, t^i)\}$, $i=1,\ldots,N$)**.
- Κάθε $\mathbf{x^i}$ είναι της μορφής $<v_{i1}, v_{i2}, \ldots, v_{in}>$ όπου v_{ij} $j = 1, n$ είναι τιμές που αντιστοιχούν σε χαρακτηριστικά/ιδιότητες του προβλήματος που αντιπροσωπεύει το σύνολο δεδομένων.
- Ποιοτικά μπορούμε να θεωρήσουμε κάθε ζεύγος ως: (ερώτηση/x^i, σωστή απάντηση/t^i).
- Το σύστημα μάθησης **υλοποιεί συσχετίσεις εισόδου – εξόδου**
- Όταν κάποιο δεδομένο $\mathbf{x^i}$ εμφανίζεται ως είσοδος θέλουμε το ΤΝΔ να παρέχει στην έξοδο την αντίστοιχη επιθυμητή τιμή $\mathbf{t^i}$.

Επιβλεπόμενη Μάθηση

- Ο όρος επιβλεπόμενη μάθηση προκύπτει από το ανάλογο του ‘επιβλέποντος’:
  - ✓ επιβλέπει την διαδικασία μάθησης, θέτοντας ερωτήσεις και παρέχοντας ταυτόχρονα και τις σωστές απαντήσεις.
- Κατάλληλη για δύο μεγάλες κατηγορίες προβλημάτων:
  - ✓ **ταξινόμησης ή κατηγοριοποίησης (classification)**

    t : ετικέτα κατηγορίας (class label)
  - ✓ **συναρτησιακής προσέγγισης ή παλινδρόμησης (regression ή function approximation)**

    t : αριθμός (value)

Μη Επιβλεπόμενη Μάθηση

- Τα παραδείγματα εκπαίδευσης **δεν** περιλαμβάνουν την επιθυμητή έξοδο αλλά μόνο τα δεδομένα εισόδου **($X=\{x^i\}$, $i=1,\ldots,N$).**
- Στόχος είναι η εξαγωγή κάποιων **βασικών δομικών ιδιοτήτων** των δεδομένων εκπαίδευσης (π.χ. εύρεση ομάδων).
- Κατηγορίες Προβλημάτων:
  - ✓ **Ομαδοποίηση (clustering): χωρισμός των δεδομένων εκπαίδευσης σε ομάδες** έτσι ώστε δεδομένα στην ίδια ομάδα να 'μοιάζουν' αρκετά μεταξύ τους και να είναι αρκετά 'διαφορετικά' από αυτά των άλλων ομάδων.
  - ✓ **Μείωση της διάστασης των δεδομένων (dimensionality reduction):** προβολή των δεδομένων σε ένα χώρο μικρότερης διάστασης στον οποίο να διατηρούνται κατά το δυνατόν οι σχετικές αποστάσεις μεταξύ των δεδομένων στον αρχικό πολυδιάστατο χώρο
  - ✓ Αν η διάσταση του χώρου προβολής είναι **δύο** τότε είναι δυνατή η **οπτικοποίηση (visualisation)** των αρχικών πολυδιάστατων δεδομένων.
  - ✓ **Τοπογραφικός Χάρτης Δεδομένων (topographic data map)**

Ενισχυτική Μάθηση

- Στο σύστημα μάθησης δεν παρέχεται η επιθυμητή έξοδος για κάθε είσοδο, αλλά μόνο η τιμή μιας ποσότητας που ονομάζεται **σήμα ενίσχυσης** (**reinforcement signal**) **($X=\{x^i, r^i)\}$, $i=1,\ldots,N$),**
- Το σήμα ενίσχυσης **r** δηλώνει εάν το σύστημα παρείχε απόκριση προς τη σωστή ή την λάθος κατεύθυνση χωρίς όμως να παρέχει λεπτομέρειες για το **ποια είναι** η σωστή απόκριση
- Εφαρμογές σε ρομποτική, παιχνίδια

Ο Τεχνητός Νευρώνας

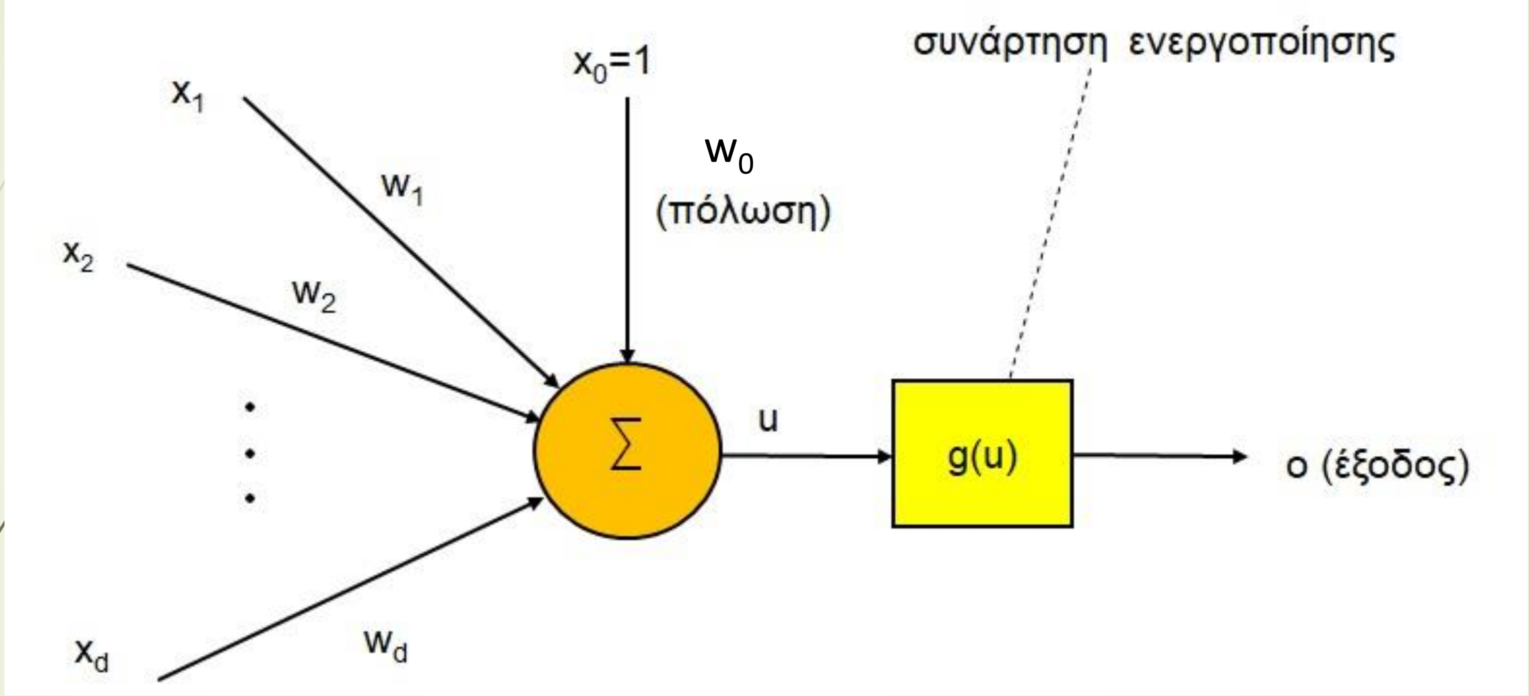
(Διασκευή διαφανειών από ΕΑΠ-ΠΛΗ31)

Διδάσκων:

Ι. ΧΑΤΖΗΛΥΓΕΡΟΥΔΗΣ

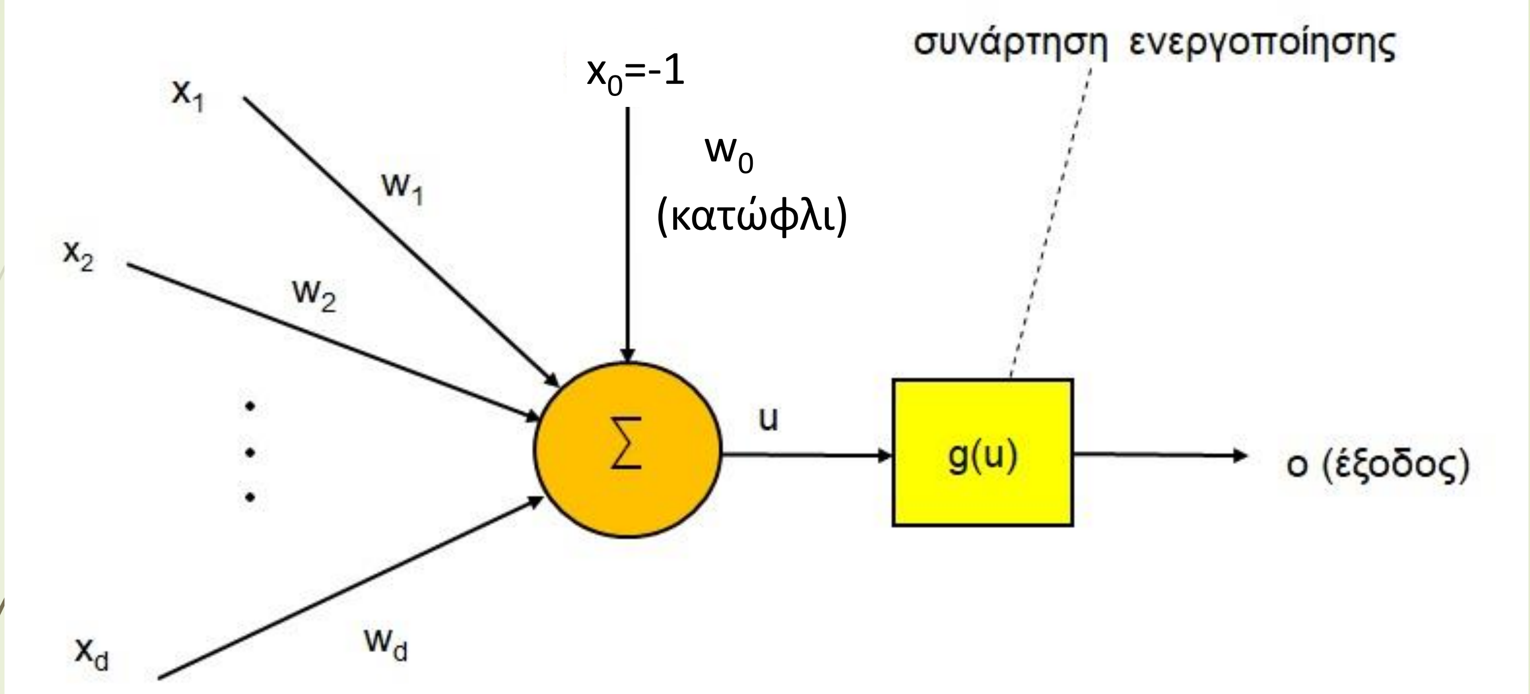
Πανεπιστήμιο Πατρών, Τμήμα Μηχ/κών Η/Υ και Πληροφορικής

Ο Τεχνητός Νευρώνας-Υπολογιστικό Μοντέλο



- d είσοδοι,
- σήμα εισόδου x_i (i=1,…,d)
- βάρη εισόδων w_i, (i=1,…,d)
- πόλωση w_0 → τιμή βάρους μιας σύνδεσης που η είσοδός της είναι μόνιμα στην τιμή 1

Ο Τεχνητός Νευρώνας-Εναλλακτικό Μοντέλο



- d είσοδοι,
- σήμα εισόδου x_i (i=1,…,d)
- βάρη εισόδων w_i, (i=1,…,d)
- κατώφλι w_0 → τιμή βάρους μιας σύνδεσης που η είσοδός της είναι μόνιμα στην τιμή -1

Τεχνητός Νευρώνας (neuron)

Ο υπολογισμός σε **δύο στάδια**:

- ✓ υπολογισμός της **συνολικής εισόδου (ενεργοποίηση)**:

$$u(x)=\sum_{i=1}^{d} w_i x_i + w_0$$

- ✓ υπολογισμός της εξόδου o(x) του νευρώνα περνώντας την συνολική είσοδο u(x) από μια **συνάρτηση ενεργοποίησης** (**activation function**)

$$o(x)=g(u)$$

Νευρώνας εσωτερικού γινομένου

$$u(x)=w^T x + w_0$$

Τεχνητός Νευρώνας (neuron)

Εναλλακτική διατύπωση:

- ✓ Διάνυσμα βαρών: $w=(w_1, w_2, \ldots, w_d)^T$
- ✓ Εκτεταμένο (extended) διάνυσμα βαρών:

$$w_e=(w_0, w_1, w_2, \ldots, w_d)^T$$

- ✓ Εκτεταμένο (extended) διάνυσμα εισόδου:

$$x_e=(1, x_1, x_2, \ldots, x_d)^T$$

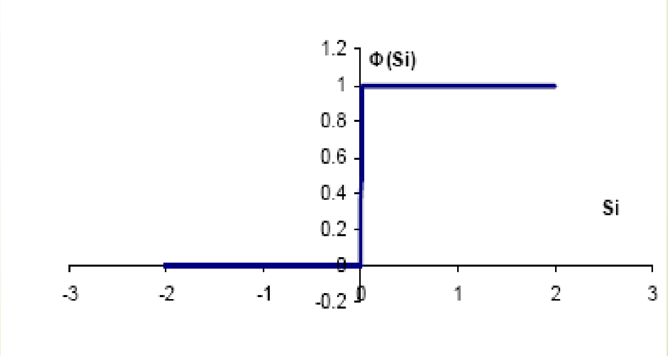
$$\mathbf{u(x)=w_e^T x_e \quad \leftrightarrow \quad u(x)=w^T x + w_0}$$

Συναρτήσεις ενεργοποίησης

Βηματική Συνάρτηση (ή συνάρτηση κατωφλίου):

- Η συνάρτηση ενεργοποίησης στο βιολογικό νευρώνα
- Χαρακτηρίζεται **από δύο τιμές a και b**.
- Αν **$x<0$ τότε $g(x)=a$** και εάν **$x>0$ τότε $g(x)=b$**. Συνήθως χρησιμοποιούνται οι τιμές a=0 και b=1 είτε a=-1 και b=1.
- **Η βηματική συνάρτηση έχει το μειονέκτημα ότι η παράγωγός της είναι μηδέν.**

Δεδομένου ότι μάθηση στα ΤΝΔ είναι η μεταβολή των τιμών των βαρών και μεταβολή σχετίζεται με την παράγωγο, η βηματική συνάρτηση δεν θεωρείται βολική ως συνάρτηση ενεργοποίησης των νευρώνων στα ΤΝΔ



Συναρτήσεις ενεργοποίησης

Σιγμοειδείς συναρτήσεις

- ✓ Έχουν **μορφή τελικού σίγμα**
- ✓ Αποτελούν **συνεχείς και παραγωγίσιμες** προσεγγίσεις της βηματικής.
- ✓ Στο όριο που η κλίση γίνεται πολύ μεγάλη, η σιγμοειδής γίνεται βηματική.
- ✓ **Δύο βασικοί τύποι**:

1) **Λογιστική:**

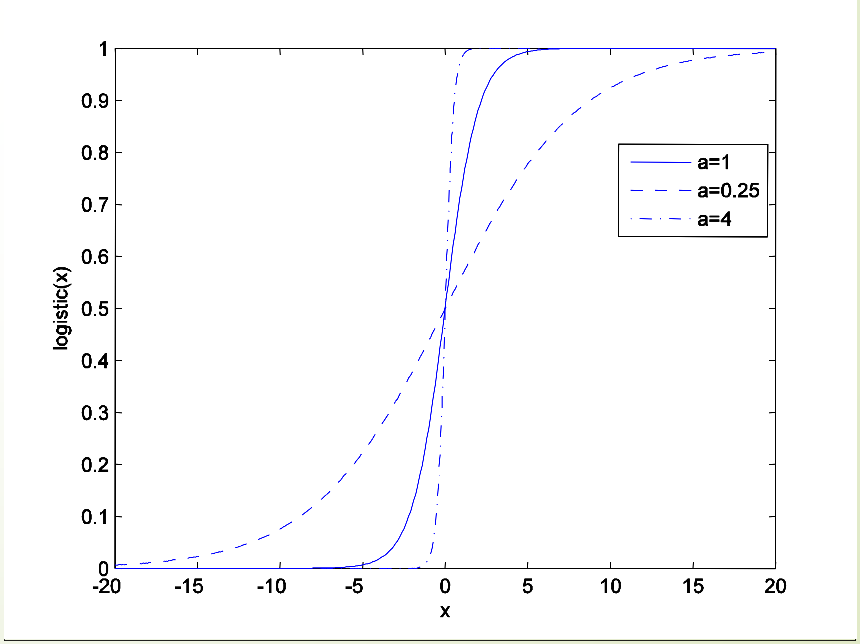
σ(x)=1/(1+exp(-ax))

(a: κλίση, συνήθως a=1)

δίνει τιμές στο (0,1)

σ'(x)=σ(x)(1-σ(x)) (για a=1)

σ''(x)=σ(x)(1-σ(x))(1-2σ(x))



Μπορούμε να υπολογίσουμε την παράγωγο σ'(x) ξέροντας μόνο το σ(x) χωρίς να χρειάζεται η τιμή του x.

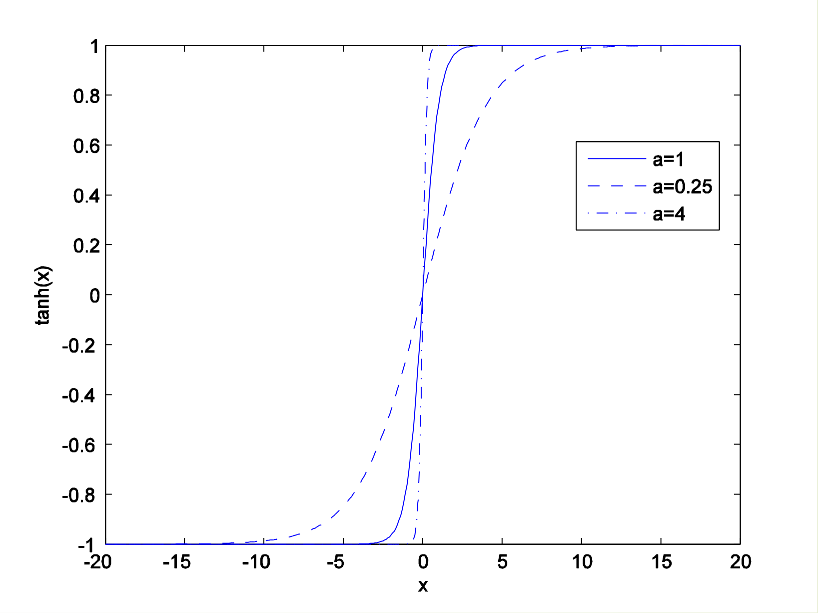
Συναρτήσεις ενεργοποίησης

2) Υπερβολική εφαπτομένη:

$$\tanh(x)=\frac{e^{ax}-e^{-ax}}{e^{ax}+e^{-ax}}$$

(a: κλίση, συνήθως a=1)
δίνει τιμές στο (-1,1)

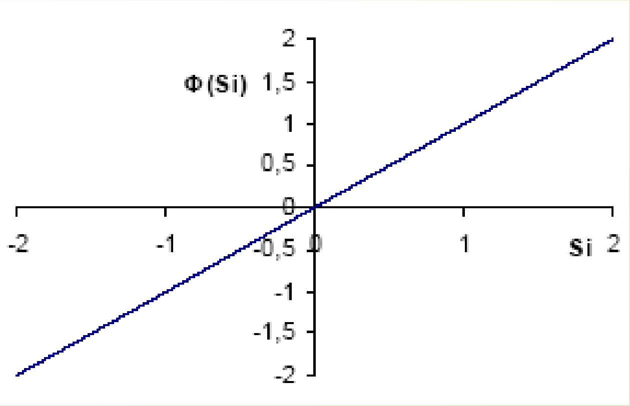
$\tanh'(x)=1-\tanh^2(x)$ (για a=1)



Γραμμική συνάρτηση

$$g(x)=x, \qquad g'(x)=1$$

δίνει τιμές στο R

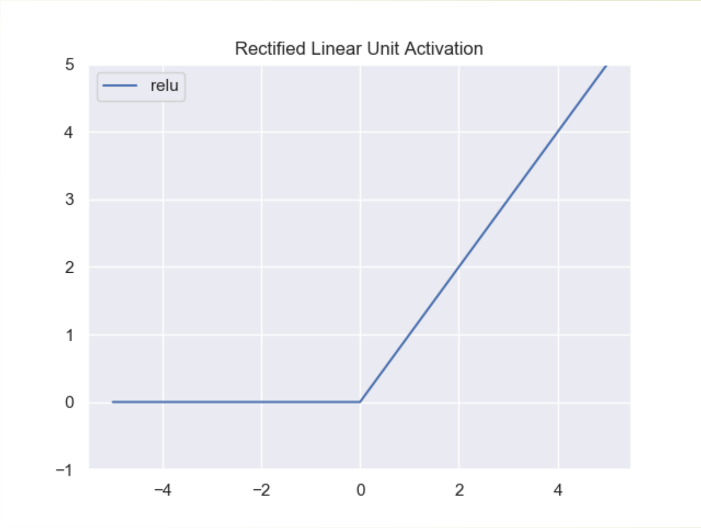


Συναρτήσεις ενεργοποίησης

Rectified linear unit (Relu)

$$g(x) = \max(0,x)$$

- Χαρακτηριστικά
  - ✓ Μη γραμμική
  - ✓ $(g(-1) + g(1) = 1,\ g(-1+1) = 0)$
  - ✓ Αραιότερες ενεργοποιήσεις-υπολογιστικά ευνοϊκή
  - ✓ Ιδανική για βαθιά δίκτυα



Μειονεκτήματα:

- ✓ Δεν είναι άνω φραγμένη
- ✓ Για αρνητικές παραμέτρους, η τοπική παράγωγος είναι 0, επομένως δεν ανταποκρίνονται κατά την εκπαίδευση (dying Relu problem)

Νευρώνας Perceptron

- Το Perceptron είναι η απλούστερη μορφή Νευρωνικού Δικτύου, το οποίο χρησιμοποιείται για την **ταξινόμηση γραμμικά διαχωριζόμενων** προτύπων, που ακολουθεί το μοντέλο McCulloch – Pitts.
- Χρησιμοποιεί ως συνάρτηση ενεργοποίησης τη **συνάρτηση προσήμου** (sign function):

$$\operatorname{sgn}(u) = \begin{cases} +1 \; \alpha\nu \; u \geq 0 \\ -1 \; \alpha\nu \; u < 0 \end{cases}$$

Αλγόριθμος εκπαίδευσης Perceptron

1. Αρχικοποίηση

Αρχικοποιούμε τα βάρη και το κατώφλι με τιμές στην περιοχή [-0.5, 0.5]

2. Ενεργοποίηση-Υπολογισμός εξόδου

$$\boldsymbol{y(n) = g(u(n))} \qquad u(n) = \sum_{i=0}^{m} w_i(n)\, x_i(n) \qquad ή \quad \boldsymbol{u(n) = W(n)^T \times X(n)}$$

3. Προσαρμογή βαρών

$$\boldsymbol{w_i(n+1)} = w_i(n) + \Delta w_i(n) \qquad \boldsymbol{\Delta w_i(n)} = \eta\, x_i(n)\, e(n) \qquad \boldsymbol{e(n)} = d(n) - y(n)$$

$$\boldsymbol{w_i(n+1) = w_i(n) + \eta\, [d(n) - y(n)]\, x_i(n)}$$

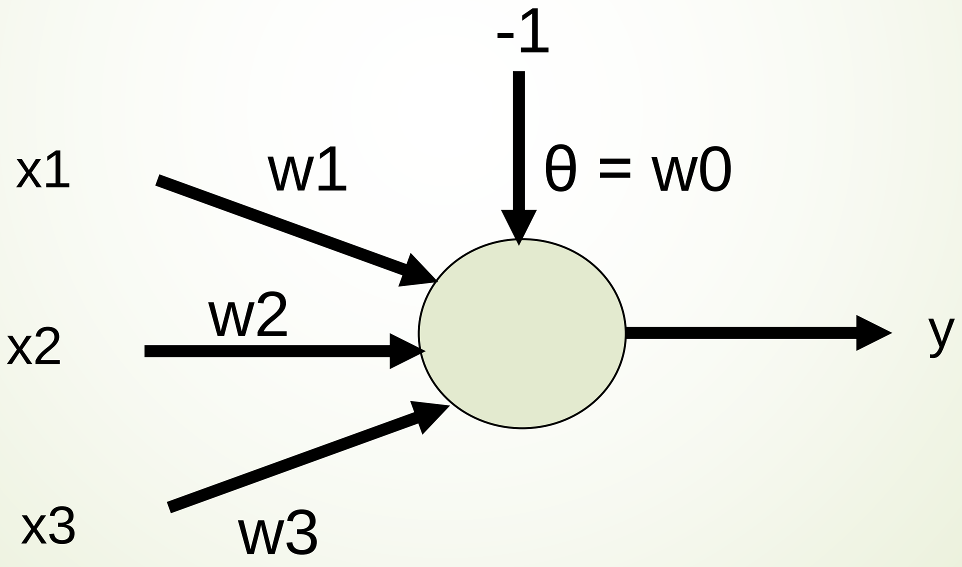
4. Έλεγχος-Επανάληψη (από βήμα 2)

$$e(n) < \varepsilon \quad (\varepsilon = 10^{-1} - 10^{-4})$$

Παράδειγμα εκπαίδευσης Perceptron

Έχουμε τον παρακάτω νευρώνα τριών εισόδων. Χρησιμοποιούμε τον αλγόριθμο του Perceptron για να επιλύσουμε το παρακάτω απλό πρόβλημα ταξινόμησης προτύπων:

$$X1 = [1, -1, 1]^T \rightarrow d1 = 0 \text{ και } X2 = [1, 1, -1]^T \rightarrow d2 = 1$$



Θεωρήστε ότι τα βάρη w1,w2,w3 έχουν αρχικές τιμές: [0.5, -1, -0.5] και το κατώφλι θ=w0=-0.5. Επίσης, η = 1.

Παράδειγμα εκπαίδευσης Perceptron

- Θεωρούμε ότι βρισκόμαστε στο βήμα k, θέτουμε ως είσοδο το διάνυσμα X1 και υπολογίζουμε τη συνολική είσοδο

$$u(k+1) = \mathbf{w}^T(k) * X1$$

όπου w = [-0.5, 0.5, -1, -0.5] και X1 = [-1, 1, -1, 1]

Χρησιμοποιούμε εκτεταμένα διανύσματα, όπου η πρώτες τιμές αντιστοιχούν στην είσοδο κατωφλίου. Οπότε

$$u(k+1) = \begin{bmatrix} -0.5 \\ 0.5 \\ -1 \\ -0.5 \end{bmatrix} * [-1, 1, -1, 1] = 1.5, \text{ και } y(k+1) = \mathrm{sgn}(1.5) = +1$$

- Ανανεώνουμε τα βάρη, αφού υπολογίσουμε το σφάλμα

$$e = d1 - y(k+1) = 0-1 = -1$$

Παράδειγμα εκπαίδευσης Perceptron

$$\boldsymbol{w}(k+1) = \boldsymbol{w}\,(k) + \eta \cdot e \cdot X_1$$
$$= [-0.5, 0.5, -1, -0.5] + 1(-1)[-1, 1, -1, 1] = [0.5, -0.5, 0, -1.5]$$

- Προχωρούμε στο επόμενο βήμα k+2, θέτοντας ως είσοδο το διάνυσμα X2 και υπολογίζουμε την έξοδο

$$u(k+2) = \boldsymbol{w}^T(k+1) \cdot X_2 = \begin{bmatrix} 0.5 \\ -0.5 \\ 0 \\ -1.5 \end{bmatrix} \cdot [-1, 1, 1, -1] = 0.5 > 0$$

και y(k+2) = sgn(0.5) = +1

- Ανανεώνουμε τα βάρη, αφού υπολογίσουμε το σφάλμα

e = d2-y(k+2) = 1-1 = 0

Αφού e=0 τα βάρη δεν ανανεώνονται

Παράδειγμα εκπαίδευσης Perceptron

- Προχωρούμε στο επόμενο βήμα k+3, θέτοντας ως είσοδο το διάνυσμα X1 και υπολογίζουμε την έξοδο

$$u(k+3) = \boldsymbol{w}^T(k+2) \cdot X_1 = \begin{bmatrix} 0.5 \\ -0.5 \\ 0 \\ -1.5 \end{bmatrix} \cdot [-1,1,-1,1] = -2.5$$

και y(k+3) = sgn(-2.5) = -1

- Ανανεώνουμε τα βάρη, αφού υπολογίσουμε το σφάλμα

e = d1-y(k+3) = 0-(-1) = 1

$$\boldsymbol{w}(k+3) = \boldsymbol{w}\,(k+2) + \eta \cdot e \cdot X_1$$
$$= [0.5, -0.5, 0, -1.5] + 1 \cdot 1[-1,1,-1,1] = [-0.5, 0.5, -1, -0.5]$$

κ.ο.κ.

Εκπαίδευση ΤΝΔ με ελαχιστοποίηση του τετραγωνικού σφάλματος εκπαίδευσης

(Διασκευή διαφανειών από ΕΑΠ-ΠΛΗ31)

Διδάσκων:

Ι. ΧΑΤΖΗΛΥΓΕΡΟΥΔΗΣ

Πανεπιστήμιο Πατρών, Τμήμα Μηχ/κών Η/Υ και Πληροφορικής

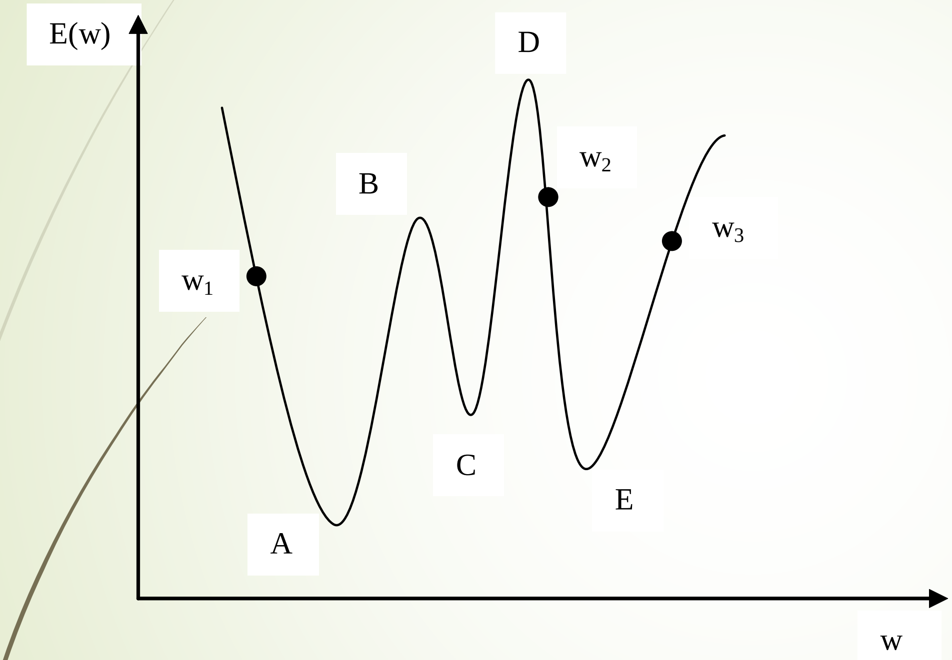
Ελαχιστοποίηση συνάρτησης σφάλματος

- Εκπαίδευση ΤΝΔ: μπορεί να διατυπωθεί ως **πρόβλημα ελαχιστοποίησης μιας συνάρτησης σφάλματος E(w) ως προς το διάνυσμα $\mathbf{w=(w_1,\ldots w_L)}$** των παραμέτρων του ΤΝΔ (βάρη και πολώσεις).
- Συνήθως αυτό που χρειάζεται είναι ο υπολογισμός των **μερικών παραγώγων $\partial E/\partial w_i$** του σφάλματος ως προς τις παραμέτρους του ΤΝΔ
- Πολλές αποδοτικές **μέθοδοι αριθμητικής ελαχιστοποίησης** βασίζονται στις μερικές παραγώγους
- Πιο δημοφιλής μέθοδος για τα ΤΝΔ: **gradient descent** (κάθοδος βασισμένη στην κλίση)
- Είναι και η απλούστερη

Ελαχιστοποίηση συνάρτησης σφάλματος

- Έστω συνάρτηση σφάλματος E(w) την οποία θέλουμε να ελαχιστοποιήσουμε ως προς w:
  - ✓ να βρούμε το **σημείο ελαχίστου** $\mathbf{w^*}$ στο οποίο η συνάρτηση $E(w^*)$ γίνεται ελάχιστη.
- Τα **ακρότατα** μιας συνάρτησης ικανοποιούν τη συνθήκη ότι η $\boldsymbol{\partial E/\partial w_i=0}$ για κάθε i=1,…,L.
- Μια συνάρτηση μπορεί να έχει περισσότερα του ενός ελάχιστα που ονομάζονται **τοπικά ελάχιστα**.
- Το καλύτερο (αυτό με την μικρότερη τιμή) από τα τοπικά ελάχιστα ονομάζεται **ολικό ελάχιστο**.

Ελαχιστοποίηση συνάρτησης σφάλματος



τρία τοπικά ελάχιστα: (A, C, E)
ολικό ελάχιστο: A

•**Αναλυτική** εύρεση ελαχίστων: λύση του συστήματος εξισώσεων **$\partial E/\partial w_i=0$, i=1,…,L. Δυνατή μόνο όταν η E(w) είναι τετραγωνική.**

•Καταφεύγουμε σε **μεθόδους αριθμητικής ανάλυσης (επαναληπτικές)**

Ελαχιστοποίηση συνάρτησης σφάλματος

Επαναληπτικές μέθοδοι:

- ξεκινούν από μια αρχική τιμή (συνήθως τυχαία) $w^{(0)}$.
- σε κάθε επανάληψη t το διάνυσμα των βαρών τροποποιείται κατά $\Delta w(t)$:

$$w(t+1) = w(t) + \Delta w(t)$$

ώστε η συνάρτηση να μειώνεται:

$$E(w(t+1)) \leq E(w(t)).$$

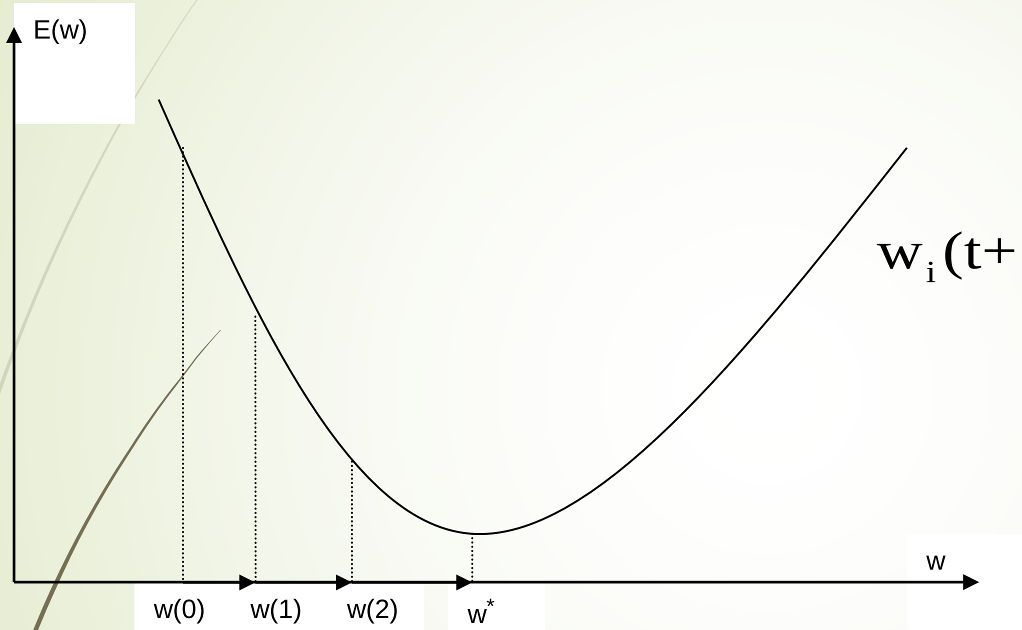
Οι αλγόριθμοι βελτιστοποίησης διαφοροποιούνται στον τρόπο με τον οποίο υπολογίζεται η μεταβολή Δw(t).

- Συνήθως χρησιμοποιείται πληροφορία σχετική με την κλίση της συνάρτησης.
- Η επαναληπτική διαδικασία συγκλίνει σε **ένα τοπικό ελάχιστο** $\mathbf{w}^*$ της συνάρτησης E(w).

Ελαχιστοποίηση συνάρτησης σφάλματος

- Οι μέθοδοι υλοποιούν **τοπική ελαχιστοποίηση**.
- **Αν η συνάρτηση E(w) έχει πολλά τοπικά ελάχιστα, το ελάχιστο στο οποίο θα καταλήξει η μέθοδος εξαρτάται από την αρχική τιμή του διανύσματος $w^{(0)}$ (η οποία συνήθως επιλέγεται τυχαία).**
- Υπάρχει πιθανότητα ‘εγκλωβισμού’ σε ανεπιθύμητα (με υψηλή τιμή) τοπικά ελάχιστα της συνάρτησης σφάλματος.
- Μια απλή λύση: πολλές εκτελέσεις από διαφορετικές αρχικές τιμές. Κρατάμε την καλύτερη από τις λύσεις που βρίσκουμε.

Κάθοδος με βάση την κλίση (gradient descent-GD)



$$w_i(t+1)=w_i(t)-\eta\frac{\partial E}{\partial w_i},\ \ i=1,\ldots,L$$

- Ξεκινάμε από μια αρχική τιμή των βαρών $w_i(0)$ (συνήθως τυχαία).
- Σε κάθε επανάληψη t:
  - ✓ Υπολογισμός της κλίσης και **ενημέρωση των w_i**
  - ✓ Ελέγχουμε για τερματισμό της μεθόδου
  - ✓ Αν ναι, τερματίζουμε, αλλοιώς t:=t+1 και συνεχίζουμε.

Ρυθμός μάθησης

η : ονομάζεται **βήμα καθόδου**

- Στην περίπτωση της εκπαίδευσης των ΤΝΔ ονομάζεται **ρυθμός μάθησης (learning rate).**
- Καθορίζει εάν θα μετακινηθούμε στην κατεύθυνση μείωσης της συνάρτησης με μικρά ή μεγάλα βήματα.
- Μικρός ρυθμός μάθησης συνεπάγεται ομαλή κάθοδο προς το τοπικό ελάχιστο, αλλά απαιτούνται περισσότερες επαναλήψεις.
- Μεγάλος ρυθμός μάθησης συνεπάγεται ταχύτερη κάθοδο (μεγαλύτερα βήματα, λιγότερες επαναλήψεις), αλλά και αυξημένη πιθανότητα εμφάνισης **ταλαντώσεων** γύρω από το σημείο ελαχίστου.

Εκπαίδευση του απλού νευρώνα με ελαχιστοποίηση σφάλματος

- Σύνολο παραδειγμάτων εκπαίδευσης $D=\{(x^n,t^n)\}$, $n=1,\ldots,N$

  $x^n=(x_{n1,\ldots,}x_{nd})^T$ και t^n αριθμός
- Εκπαίδευση απλού νευρώνα με βάρη $w=(w_{0,}w_{1,}\ldots,w_d)^T$ και συναρτηση ενεργοποίησης $g(u)$.
- Για είσοδο το x^n : $u(x^n;\, w)=\Sigma_i\, w_i\, x_i\ + w_0$, $o(x^n;\, w)=g(u(x^n;\, w))$
- Στην περίπτωση που για κάποιο διάνυσμα βαρών η εκπαίδευση είναι τέλεια θα ισχύει:

$$o(x^n;\, w)=t^n \text{ για κάθε } n=1,\ldots,N$$

δηλαδή η έξοδος του νευρώνα για είσοδο x^n θα είναι ίση με την επιθυμητή t^n.

Εκπαίδευση του απλού νευρώνα με ελαχιστοποίηση σφάλματος

- Επομένως μπορούμε να ορίσουμε την **τετραγωνική συνάρτηση σφάλματος εκπαίδευσης:**

$$E(w)=\frac{1}{2}\sum_{n=1}^{N}(t^n-o(x^n;w))^2 \quad \Longleftrightarrow \quad E(w)=\sum_{n=1}^{N}E^n(w),\ E^n(w)=\frac{1}{2}(t^n-o(x^n;w))^2$$

- Ως άθροισμα τετραγώνων έχουμε κάτω φράγμα την τιμή μηδέν η οποία προκύπτει όταν έχουμε τέλεια εκπαίδευση.
- Η πιο σημαντική κατηγορία μεθόδων εκπαίδευσης ΤΝΔ για μάθησης με επίβλεψη προκύπτει από την **ενημέρωση του διανύσματος των βαρών w με σκοπό την ελαχιστοποίηση του τετραγωνικού σφάλματος E(w)**.
- Ευρύτερα χρησιμοποιούμενη μέθοδος ελαχιστοποίησης: κάθοδος με βάση την κλίση (gradient descent).

Μερική Παράγωγος του σφάλματος εκπαίδευσης

$$E(w)=\sum_{n=1}^{N} E^{n}(w),\ E^{n}(w)=\frac{1}{2}(t^{n}\text{-}o(x^{n};w))^{2} \qquad \frac{\partial E}{\partial w_i}=?$$

$$\frac{\partial E}{\partial w_i}=\sum_{n=1}^{N}\frac{\partial E^n}{\partial w_i} \qquad \frac{\partial E^n}{\partial w_i}=-(t^n\text{-}o(x^n;w))\frac{\partial o(x^n;w)}{\partial w_i}$$

$$\frac{\partial o(x^n;w)}{\partial w_i}=\frac{\partial g(u)}{\partial u}\frac{\partial u(x^n;w)}{\partial w_i}=g'(u)x_{ni},\ i=0,\ldots,d,\ x_{n0}=1$$

$$\frac{\partial E^n}{\partial w_i}=-(t^n\text{-}o(x^n;w))g'(u(x^n;w))x_{ni},\ i=0,\ldots,d,\ x_{n0}=1$$

$$\frac{\partial E}{\partial w_i}=-\sum_{n=1}^{N}(t^n\text{-}o(x^n;w))g'(u(x^n;w))x_{ni},\ i=0,\ldots,d,\ x_{n0}=1$$

Μερική Παράγωγος του σφάλματος εκπαίδευσης

- Υπολογισμός της μερικής παραγώγου που αντιστοιχεί στο σφάλμα για ένα παράδειγμα εκπαίδευσης (x^n, t^n):
  - εφαρμογή του x^n ως είσοδο στον νευρώνα και υπολογισμός της συνολικής εισόδου $u(x^n;w)$ και της εξόδου $o(x^n;w)$
  - υπολογισμός του **σφάλματος**: $\delta^n = (t^n - o(x^n;w))$
  - υπολογισμός των μερικών παραγώγων ως προς w_i

$$\frac{\partial E^n}{\partial w_i} = -(t^n - o(x^n;w))\, g'(u(x^n;w))\, x_{ni}, \quad i=0,\ldots,d, \; x_{n0}=1$$

Εκπαίδευση του απλού νευρώνα με gradient descent (ομαδική ενημέρωση)

1. Αρχικοποίηση: Θέτουμε k=0, αρχικές τιμές βαρών w(0) και ορίζουμε την τιμή του ρυθμού μάθησης *η*.
2. Σε κάθε επανάληψη k, έστω w(t) το διάνυσμα των βαρών.
   - Αρχικοποιούμε: $\frac{\partial E}{\partial w_i} = 0$, i=0,…,L
   - Για n=1,…,N:
     - ✓ εφαρμογή του x^n ως είσοδο στον νευρώνα και υπολογισμός της συνολικής εισόδου $u(x^n;w)$ και της εξόδου $o(x^n;w)$
     - ✓ υπολογισμός του *σφάλματος*: $\delta^n = (t^n - o(x^n;w))$.

     $$\frac{\partial E}{\partial w_i} := \frac{\partial E}{\partial w_i} - \delta^n g'(u(x^n;w))x_{ni}, \ \ i=0,...,d, \ x_{n0} = 1$$
   - Ενημερώνουμε τις τιμές των βαρών: $w_i(t+1) = w_i(t) - \eta \frac{\partial E}{\partial w_i}, \ \ i=1,...,L$
3. Ελέγχουμε για τερματισμό της μεθόδου. Αν ναι, τερματίζουμε.
4. k:=k+1, μετάβαση στο βήμα 2.

Εκπαίδευση του απλού νευρώνα με gradient descent (ομαδική ενημέρωση)

- **Ομαδική ενημέρωση:** η ενημέρωση των βαρών πραγματοποιείται **μια φορά** στο τέλος κάθε εποχής με βάση την μερική παράγωγο του συνολικού σφάλματος, αθροίζοντας δηλαδή τις μερικές παραγώγους των επιμέρους σφαλμάτων.
- Ο μετρητής επαναλήψεων t μετράει τις εποχές. Μια **εποχή** θεωρείται το πέρασμα όλων των παραδειγμάτων του συνόλου εκπαίδευσης.
- Η ομαδική ενημέρωση αντιστοιχεί στην μαθηματικά αυστηρή υλοποίηση της μεθόδου gradient descent για την ελαχιστοποίηση του σφάλματος E(w):

$$w_i(t+1)=w_i(t)-\eta\frac{\partial E}{\partial w_i},\ i=1,...,L$$

- Σε κάθε εποχή t το σφάλμα E(w) θα πρέπει να μειώνεται (εάν ο ρυθμός μάθησης είναι επαρκώς μικρός)

Εκπαίδευση του απλού νευρώνα με gradient descent (σειριακή ενημέρωση)

- Η συνάρτηση **E(w)** που θέλουμε να ελαχιστοποιήσουμε έχει την εξής χρήσιμη ιδιότητα: **εκφράζεται ως το άθροισμα των επιμέρους σφαλμάτων E^n(w)**.
- Μια εναλλακτική προσέγγιση για την ελαχιστοποίηση του E(w):
  - ✓ Σε κάθε επανάληψη τ (δηλ. μετά από το πέρασμα κάθε παραδείγματος) εφαρμόζουμε τον κανόνα ενημέρωσης gradient descent για την ελαχιστοποίηση **κάποιου από τα επιμέρους σφάλματα E^n(w):**

$$w_i(\tau+1)=w_i(\tau)+n(t^n-o(x^n;w))g'(u(x^n;w))x_{ni},\ \ i=0,...,d,\ x_{n0}=1$$

Εκπαίδευση του απλού νευρώνα με gradient descent (σειριακή ενημέρωση)

- Αποδεικνύεται ότι εάν όλοι οι όροι $E^n(w)$ επιλέγονται το ίδιο συχνά, τότε το τελικό αποτέλεσμα της μεθόδου είναι η ελαχιστοποίηση του συνολικού σφάλματος E(w)
- δηλαδή λειτουργώντας σε κάθε βήμα στην κατεύθυνση μείωσης ενός όρου, επιτυγχάνουμε στο τέλος τη μείωση του αθροίσματος των όρων.
- Αυτό το γεγονός δεν πρέπει να θεωρηθεί ως κάτι προφανές δεδομένου ότι σε κάθε βήμα η αλλαγή των βαρών για την μείωση του όρου $E^n(w)$ δεν μειώνει απαραίτητα και το συνολικό σφάλμα E(w), διότι μπορεί να υπάρχουν άλλοι όροι $E^m(w)$ που να αυξάνουν με την αλλαγή των βαρών.

Εκπαίδευση του απλού νευρώνα με gradient descent (σειριακή ενημέρωση)

- Η παραπάνω διαδικασία ονομάζεται στοχαστική (stochastic) gradient descent ή on-line gradient descent ή σειριακή (sequential) gradient descent.
- Θα την ονομάζουμε μέθοδο gradient descent *με* **σειριακή ενημέρωση** των βαρών.
- Ενώ στην ομαδική ενημέρωση έχουμε μία ενημέρωση των βαρών ανά εποχή (κύκλος εκπαίδευσης), στην σειριακή ενημέρωση έχουμε N ενημερώσεις.

Εκπαίδευση του γραμμικού νευρώνα

Ο γραμμικός νευρώνας έχει συνάρτηση ενεργοποίησης g(u)=u, επομένως g'(u)=1.

- **Ομαδική ενημέρωση**: $w_i(k+1)=w_i(k)+n\sum_{n=1}^{N}(t^n\text{-}o(x^n;w))x_{ni},\ i=0,...,d,\ x_{n0}=1$
- **Σειριακή ενημέρωση**: $w_i(\kappa+1)=w_i(\kappa)+n(t^n\text{-}o(x^n;w))x_{ni},\ i=0,...,d,\ x_{n0}=1$
- Αν $\delta^n=t^n\text{-}o(x^n;w)$: $w_i(\kappa+1)=w_i(\kappa)+n\ \delta^n x_{ni}$

κανόνας δέλτα (delta rule)