

Εκπαίδευση ΤΝΔ με ελαχιστοποίηση του τετραγωνικού σφάλματος εκπαίδευσης

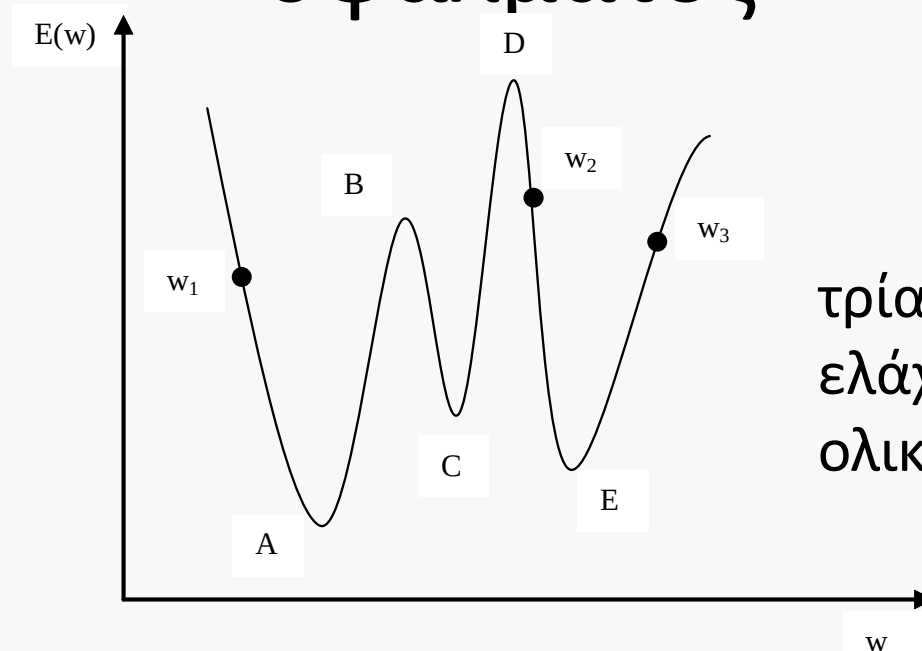
Ελαχιστοποίηση συνάρτησης σφάλματος

- Εκπαίδευση ΤΝΔ: μπορεί να διατυπωθεί ως **πρόβλημα ελαχιστοποίησης μιας συνάρτησης σφάλματος $E(w)$ ως προς το διάνυσμα $w=(w_1,...w_L)$ 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

- Εστω συνάρτηση σφάλματος $E(w)$ την οποία θέλουμε να ελαχιστοποιήσουμε ως προς w :
να βρούμε το **σημείο ελαχίστου w^*** στο οποίο η συνάρτηση $E(w^*)$ γίνεται ελάχιστη.
- Τα **ακρότατα** μιας συνάρτησης ικανοποιούν τη συνθήκη ότι η **$\partial E / \partial w_i = 0$** για κάθε $i=1, \dots, L$.
- 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 $\partial E / \partial w_i = 0, i=1, \dots, L$. 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:

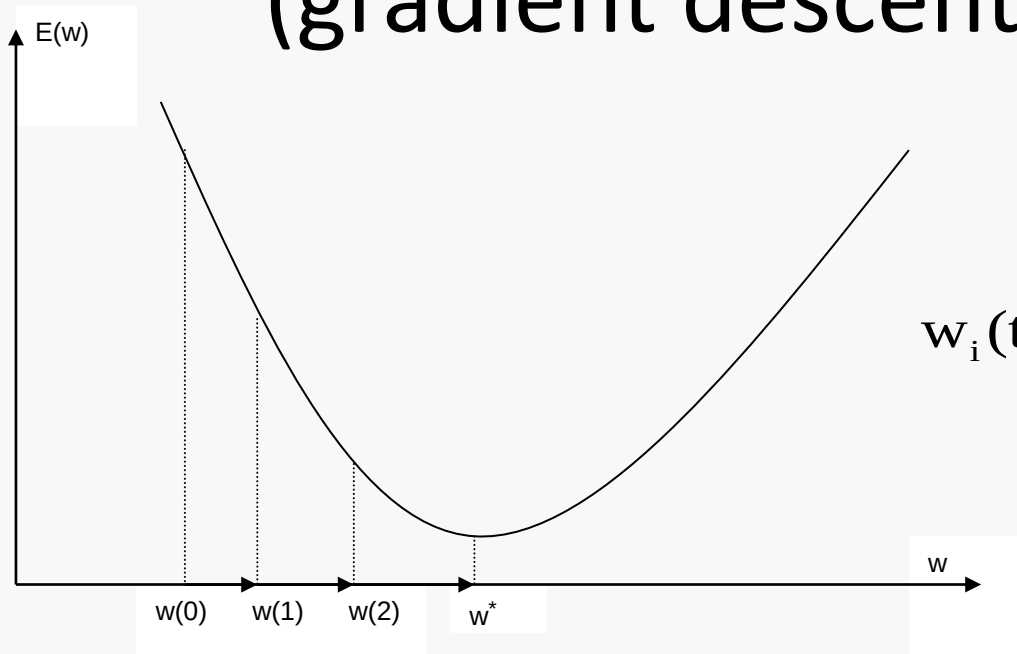
$$E(w(t+1)) \leq E(w(t)).$$

- **Οι αλγόριθμοι βελτιστοποίησης διαφοροποιούνται στον τρόπο με τον οποίο υπολογίζεται η μεταβολή $\Delta w(t)$.**
- Συνήθως χρησιμοποιείται πληροφορία σχετική με την κλίση της συνάρτησης.
- Η επαναληπτική διαδικασία συγκλίνει σε **ένα τοπικό ελάχιστο 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))



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

- 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, αλλιώς $t:=t+1$ 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 $D=\{(x^n, t^n)\}$, $n=1, \dots, N$
- $x^n=(x_{n1}, \dots, x_{nd})^T$ και t^n αριθμός
- Εκπαίδευση απλού νευρώνα με βάρη $w=(w_0, w_1, \dots, w_d)^T$ και συναρτ. ενεργοποίησης $g(u)$.
- Για είσοδο το x^n : $u(x^n; w)=\sum_i w_i x_i + w_0$, $o(x^n; w)=g(u(x^n; w))$
- Στην περίπτωση που για κάποιο διάνυσμα βαρών η εκπαίδευση είναι τέλεια θα ισχύει:
$$o(x^n; w)=t^n \text{ για κάθε } n=1, \dots, N$$
- 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**:

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

- 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

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

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

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

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

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

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

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

Εκπαίδευση του απλού νευρώνα με gradient descent (ομαδική ενημέρωση)

1. Αρχικοποίηση: Θέτουμε $t=0$, αρχικές τιμές βαρών $w(0)$ και ορίζουμε την τιμή του ρυθμού μάθησης η .
2. Σε κάθε επανάληψη t , έστω $w(t)$ το διάνυσμα των βαρών.
 - Αρχικοποιούμε: $\frac{\partial E}{\partial w_i} = 0, i=0, \dots, L$
 - Για $n=1, \dots, N$:
 - 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: $\delta^n = (t^n - o(x^n; w))$.
 - $\frac{\partial E}{\partial w_i} := \frac{\partial E}{\partial w_i} - \delta^n g'(u(x^n; w)) x_{ni}, i=0, \dots, d, x_{n0} = 1$
 - Ενημερώνουμε τις τιμές των βαρών: $w_i(t+1) = w_i(t) - \eta \frac{\partial E}{\partial w_i}, i=1, \dots, L$
3. Ελέγχουμε για τερματισμό της μεθόδου. Αν ναι, τερματίζουμε.
4. $t:=t+1$, 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**.
- Η ομαδική ενημέρωση αντιστοιχεί στην μαθηματικά αυστηρή υλοποίηση της μεθόδου gradient descent για την ελαχιστοποίηση του σφάλματος $E(w)$: $w_i(t+1) = w_i(t) - \eta \frac{\partial E}{\partial w_i}$, $i=1, \dots, L$
- 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 $E^n(w)$.
- Εναλλακτική προσέγγιση για την ελαχιστοποίηση του $E(w)$.
- Σε κάθε επανάληψη τ (χρησιμοποιούμε διαφορετικό συμβολισμό σε σχέση το συμβολισμό t στην ομαδική ενημέρωση) εφαρμόζουμε τον κανόνα ενημέρωσης gradient descent για την ελαχιστοποίηση **κάποιου από τα επιμέρους σφάλματα $E^n(w)$** :

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

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**)

- Εάν όμως γίνουν πολλές τέτοιες επαναλήψεις και ο ρυθμός μάθησης μειώνεται (ξεκινώντας από μια αρχική τιμή η_0) ως $\eta(k)=\eta_0/k$ αποδεικνύεται ότι με πιθανότητα 1, ο αλγόριθμος θα συγκλίνει σε τοπικό ελάχιστο του συνολικού σφάλματος $E(w)$.
- 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.

Εκπαίδευση του απλού νευρώνα με gradient descent (σειριακή ενημέρωση)

- 1 Αρχικοποίηση των βαρών $w(0)$, και καθορισμός του ρυθμού μάθησης η .
Αρχικοποίηση του μετρητή επαναλήψεων ($\tau:=0$) και του μετρητή εποχών ($t:=0$).
- 2 Σε κάθε εποχή t , έστω $w(t)$ το διάνυσμα βαρών του νευρώνα.
 - Έναρξη εποχής t , αποθήκευση του τρέχοντος διανύσματος βαρών $w_{old}=w(\tau)$.
 - Για $n=1\dots N$
 - Επιλογή του παραδείγματος εκπαίδευσης (x^n, t^n) και εφαρμογή του στο νευρώνα και υπολογισμός των $u(x^n; w)$ και $o(x^n; w)$.
 - Ενημέρωση των βαρών
$$w_i(\tau+1)=w_i(\tau)+\eta(t^n-o(x^n; w))g'(u(x^n; w))x_{ni}, \quad i=0,\dots,d, \quad x_{n0}=1$$
 - $\tau:=\tau+1$.
 - Τέλος εποχής k , έλεγχος τερματισμού ($|w(\tau)-w_{old}|<\epsilon$). Εάν ναι, τερματισμός.
 - $k:=k+1$, goto 2.

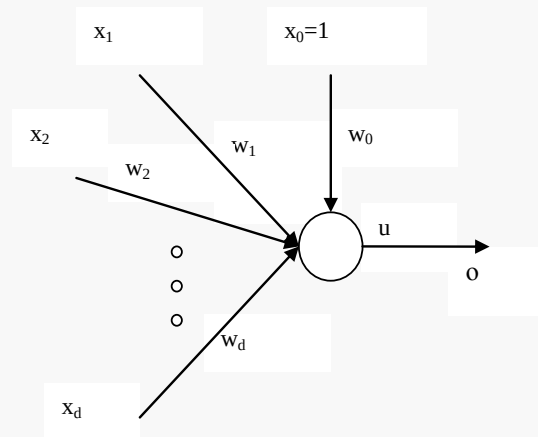
Εκπαίδευση του απλού νευρώνα με gradient descent (**σειριακή ενημέρωση**)

- Ενώ στην ομαδική ενημέρωση έχουμε μία ενημέρωση των βαρών ανά εποχή, στην σειριακή ενημέρωση έχουμε N ενημερώσεις ($\tau = Nt$).
- **Πλεονεκτήματα** της σειριακής ενημέρωσης σε σχέση με την ομαδική ενημέρωση:
 - Τα παραδείγματα μπορεί να μην είναι γνωστά εκ των προτέρων: προβλήματα μάθησης κατά τη λειτουργία (on-line learning).
 - Συνήθως είναι ταχύτερη λόγω του ότι η ενημέρωση των βαρών γίνεται άμεσα μετά από κάθε παράδειγμα εκπαίδευσης.
 - Έχει παρατηρηθεί ότι μπορεί να αποφεύγει ρηχά τοπικά ελάχιστα.

Εκπαίδευση του απλού νευρώνα με gradient descent (**σειριακή ενημέρωση**)

- **Μειονεκτήματα** της σειριακής σε σχέση με την ομαδική ενημέρωση:
 - Η σειριακή ενημέρωση δεν εγγυάται τη μείωση της συνάρτησης σφάλματος.
 - Ο έλεγχος τερματισμού είναι δυσκολότερος διότι δεν υπολογίζεται η κλίση της συνολικής συνάρτησης σφάλματος παρά μόνο η κλίση των επιμέρους σφαλμάτων ανά παράδειγμα εκπαίδευσης.
- Δεν υπάρχουν ευρέως αποδεκτές μελέτες που να αποδεικνύουν την υπεροχή της μιας μεθοδολογίας έναντι της άλλης.
- Η τεχνική της σειριακής ενημέρωσης έχει χρησιμοποιηθεί περισσότερο στο πεδίο της εκπαίδευσης των ΤΝΔ.
- Υβριδική προσέγγιση: **on-line updates in batches**

Εκπαίδευση του γραμμικού νευρώνα

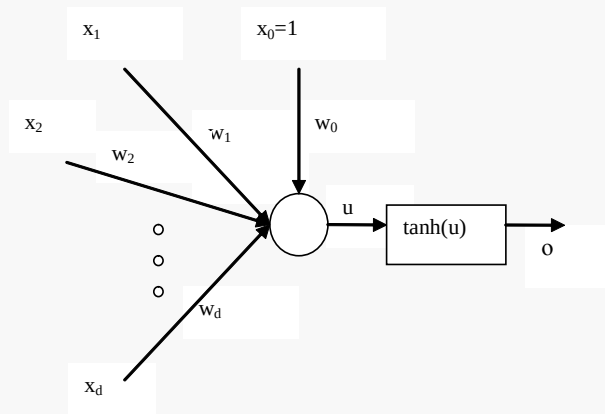


- Ο γραμμικός νευρώνας έχει συνάρτηση ενεργοποίησης $g(u)=u$, επομένως $g'(u)=1$.

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

κανόνας δέλτα (delta rule)

Εκπαίδευση του νευρώνα με συνάρτηση ενεργοποίησης την υπερβ. εφαιτομένη



- $g(u)=\tanh(u)$
- τιμές στο $(-1,1)$,
- $g'(u)=1-\tanh^2(u)$
- προσέγγιση του perceptron

- Ομαδική ενημέρωση: $w_i(k+1)=w_i(k)+n \sum_{n=1}^N (t^n - o(x^n; w))(1 - o^2(x^n; w))x_{ni}$, $i=0, \dots, d$, $x_{n0} = 1$
- Σειριακή ενημέρωση: $w_i(k+1)=w_i(k)+n (t^n - o(x^n; w))(1 - o^2(x^n; w))x_{ni}$, $i=0, \dots, d$, $x_{n0} = 1$
- **Η μέθοδος συγκλίνει πάντα** σε ελάχιστο της συνάρτησης σφάλματος. Αν το σύνολο εκπαίδευσης είναι γραμμικά διαχωρίσιμο θα βρούμε λύση χωρίς σφάλματα (ολικό ελάχιστο), αλλιώς καταλήγουμε σε λύση με σφάλματα.