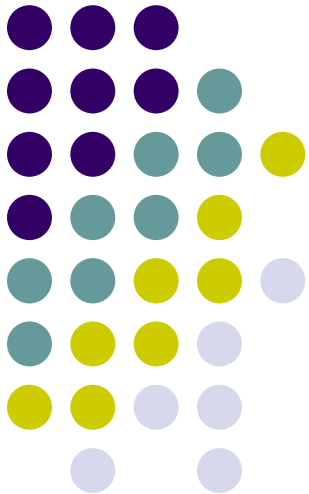
Θεωρία Αποφάσεων

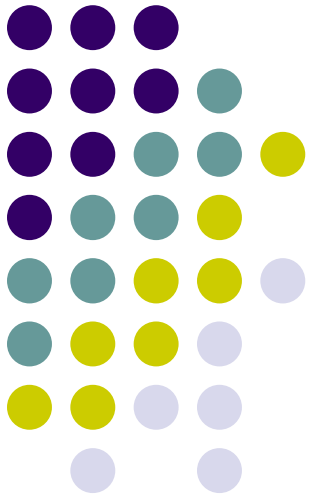


Σ. Λυκοθανάσης, Καθηγητής,
Δ. Κοσμόπουλος, Αν. Καθηγητής

Τμήμα Μηχανικών Η/Υ & Πληροφορικής - Εργαστήριο Αναγνώρισης Προτύπων
Διευθυντής: Σ. Λυκοθανάσης, Καθηγητής

Κεφάλαιο 5

Γραμμικές Διακρίνουσες Συναρτήσεις



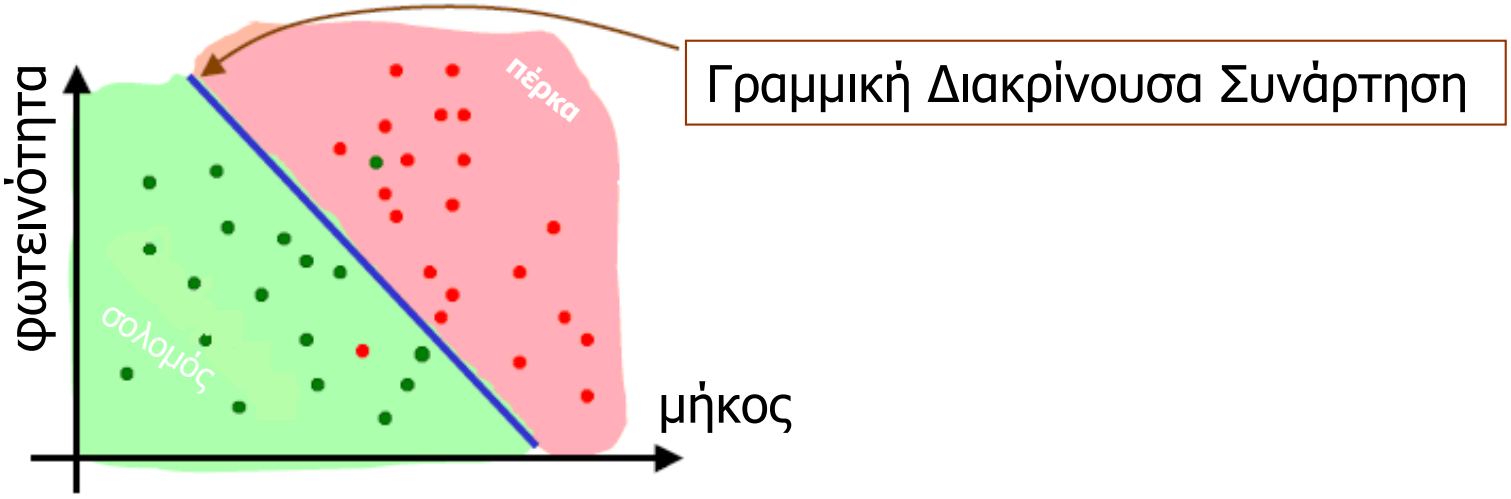
Γραμμικές Διακρίνουσες Συναρτήσεις

Δεν είναι γνωστή η σ.π.π.

Έχουμε δεδομένα (με ετικέτα)

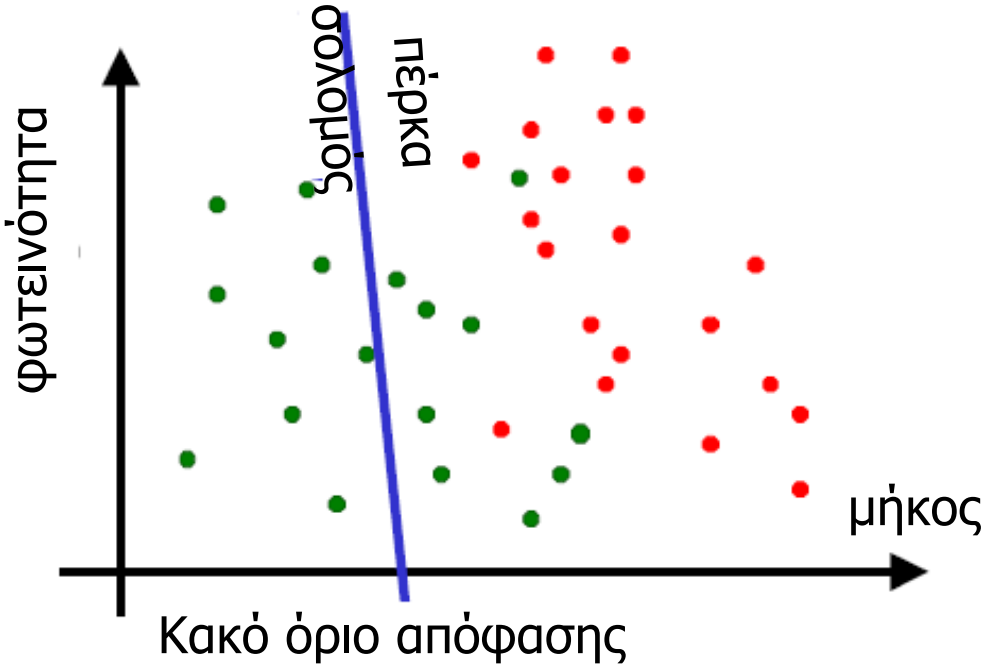
Σολομός Πέρκα Σολομός Σολομός

Γνωρίζουμε την μορφή της διακρίνουσας συνάρτησης



Πρέπει να εκτιμήσουμε τις παραμέτρους της διακρίνουσας συνάρτησης
(Σε αυτήν την περίπτωση, τις παραμέτρους της ευθείας)

Βασική Ιδέα

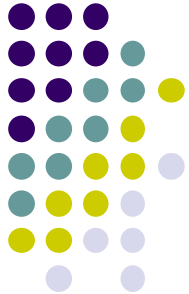


- Έχουμε δείγματα από δυό κλάσεις
- Υποθέτουμε ότι μπορούν να διαχωριστούν από γραμμικό όριο απόφασης l(θ) , όπου θ άγνωστοι παράμετροι
- Αναζητάμε το «καλύτερο» όριο για τα συγκεκριμένα δεδομένα βελτιστοποιώντας το θ

Τι σημαίνει «καλύτερο»?

Παραμετρικές Μέθοδοι

Διακρίνουσες Συναρτήσεις



Υποθέτουμε ότι είναι γνωστή η

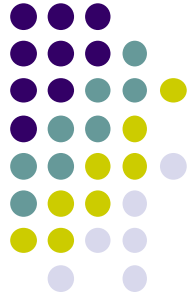
?

Θεωρούμε ότι είναι γνωστή η μορφή

?

Παραμετρικές Μέθοδοι

Διακρίνουσες Συναρτήσεις



Υποθέτουμε ότι είναι γνωστή η μορφή της σ.π.π των κλάσεων $p_1(\boldsymbol{x}|\boldsymbol{\theta}_1)$, $p_2(\boldsymbol{x}|\theta_2)$

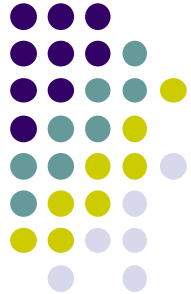
Εκτιμούμε ? από τα δεδομένα

Θεωρούμε ότι είναι γνωστή η μορφή των διακρινουσών συναρτήσεων $I(\theta_1)$, $I(\theta_2)$ με παραμέτρους $\theta_1, \theta_2, \ldots$

Εκτιμούμε ? από τα δεδομένα

Παραμετρικές Μέθοδοι

Διακρίνουσες Συναρτήσεις



Υποθέτουμε ότι είναι γνωστή η μορφή της σ.π.π των κλάσεων $p_1(\boldsymbol{x}|\boldsymbol{\theta}_1)$, $p_2(\boldsymbol{x}|\theta_2)$

Εκτιμούμε $\theta_1, \theta_2 \ldots$ από τα δεδομένα

Χρησιμοποιούμε Ε[illegible] ταξινομητή για να βρούμε τις περιοχές απόφασης

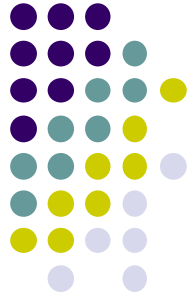
Θεωρούμε ότι είναι γνωστή η μορφή των διακρινουσών συναρτήσεων $I(\theta_1)$, $I(\theta_2)$ με παραμέτρους $\theta_1, \theta_2 \ldots$

Εκτιμούμε $\theta_1, \theta_2 \ldots$ από τα δεδομένα

Χρησιμοποιούμε τ[illegible]

Παραμετρικές Μέθοδοι

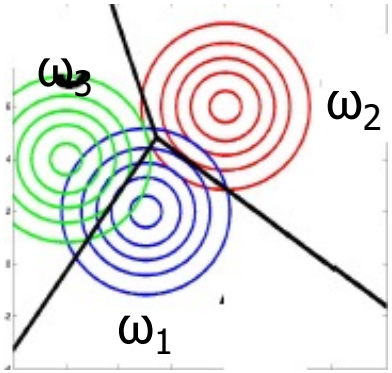
Διακρίνουσες Συναρτήσεις



Υποθέτουμε ότι είναι γνωστή η μορφή της σ.π.π των κλάσεων $p_1(\boldsymbol{x}|\boldsymbol{\theta}_1), p_2(\boldsymbol{x}|\theta_2)$

Εκτιμούμε $\theta_1, \theta_2 \ldots$ από τα δεδομένα

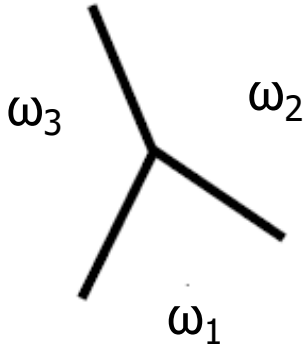
Χρησιμοποιούμε Bayesian ταξινομητή για να βρούμε τις περιοχές απόφασης



Θεωρούμε ότι είναι γνωστή η μορφή των διακρινουσών συναρτήσεων $l(\theta_1), l(\theta_2)$ με παραμέτρους $\theta_1, \theta_2 \ldots$

Εκτιμούμε $\theta_1, \theta_2 \ldots$ από τα δεδομένα

Χρησιμοποιούμε τις διακρίνουσες συναρτήσεις για ταξινόμηση

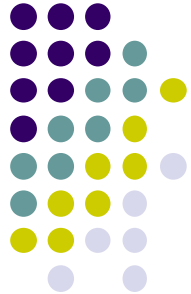


Θεωρητικά, ο Bayesian ταξινομητης ελαχιστοποιεί το ρίσκο

- Στη πράξη όμως ?

Παραμετρικές Μέθοδοι

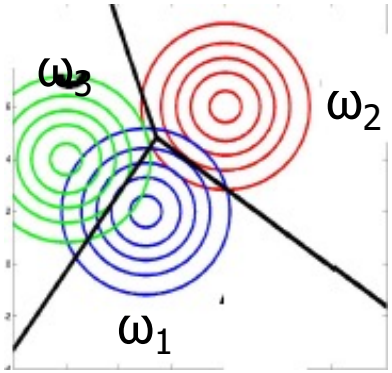
Διακρίνουσες Συναρτήσεις



Υποθέτουμε ότι είναι γνωστή η μορφή της σ.π.π των κλάσεων $p_1(\boldsymbol{x}|\boldsymbol{\theta}_1), p_2(\boldsymbol{x}|\theta_2)$

Εκτιμούμε $\theta_1, \theta_2 \ldots$ από τα δεδομένα

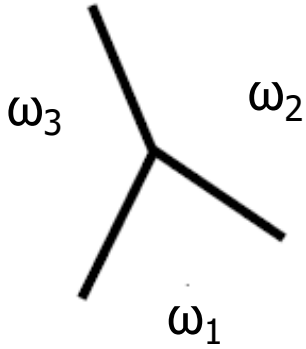
Χρησιμοποιούμε Bayesian ταξινομητή για να βρούμε τις περιοχές απόφασης



Θεωρούμε ότι είναι γνωστή η μορφή των διακρινουσών συναρτήσεων $l(\theta_1), l(\theta_2)$ με παραμέτρους $\theta_1, \theta_2 \ldots$

Εκτιμούμε $\theta_1, \theta_2 \ldots$ από τα δεδομένα

Χρησιμοποιούμε τις διακρίνουσες συναρτήσεις για ταξινόμηση

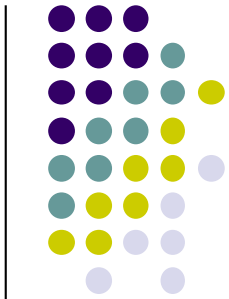


Θεωρητικά, ο Bayesian ταξινομητης ελαχιστοποιεί το ρίσκο

- Στη πράξη όμως δεν είμαστε σίγουροι αν η μορφή των μοντέλων που υποθέτουμε είναι η σωστή
- Στην πράξη ίσως να μην μας χρειάζονται οι πραγματικές σ.π.π. ...

Η ακριβής εκτίμηση των σ.π.π. είναι πιο δύσκολη από την ακριβή εκτίμηση των διακρινουσών συνάρτήσεων

Γραμμικές Διακρίνουσες Συναρτήσεις



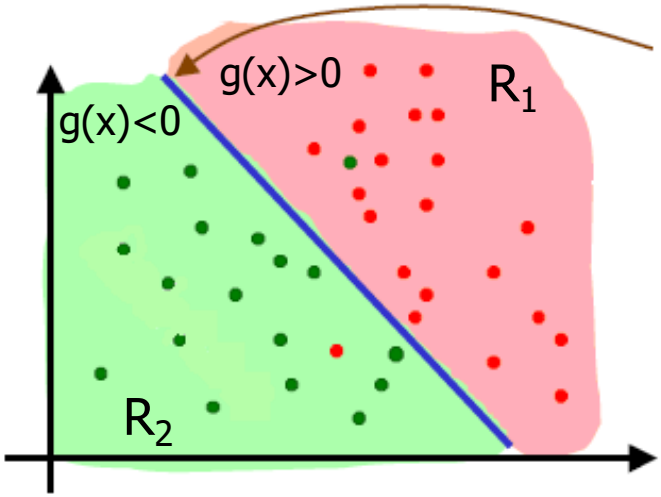
- Θα ασχοληθούμε κυρίως με διακρίνουσες συναρτήσεις οι οποίες είναι είτε γραμμικές ως προς τα στοιχεία του $\boldsymbol{x}$ είτε γραμμικές ως προς ένα δεδομένο σύνολο συναρτήσεων του $\boldsymbol{x}$.
- Οι γραμμικές διακρίνουσες συναρτήσεις έχουν πολλές ελκυστικές αναλυτικές ιδιότητες.
- Μπορεί να είναι βέλτιστες εάν οι αντίστοιχες κατανομές είναι συνεργατικές, όπως για παράδειγμα Gaussian με ίδιες διασπορές.
- Τέτοιου είδους κατανομές μπορούν να προκύψουν μέσα από εύστοχη επιλογή των ανιχνευτών χαρακτηριστικών.
- Ακόμα όμως και στην περίπτωση όπου δεν είναι βέλτιστες, συνήθως επιλέγονται διότι η απλότητά τους αποτελεί πολύ σημαντικό πλεονέκτημα.

2-Κλάσεις

Μια διακρίνουσα συνάρτηση είναι γραμμική αν μπορεί να γραφεί ως:

$$g(x) = w^t x + w_0$$

όπου το w ονομάζεται **διάνυσμα των βαρών** και το w_0 **βάρος κατωφλίου**

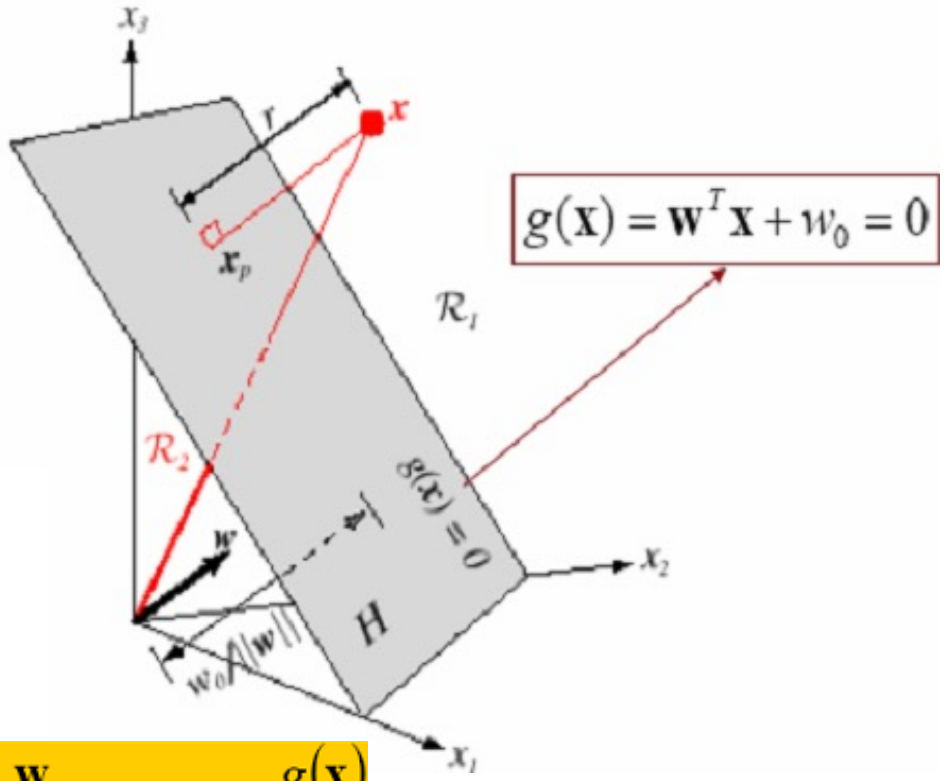


Όριο απόφασης g(x)=0

$g(x) > 0 \Rightarrow x \in$ κλάση 1
$g(x) < 0 \Rightarrow x \in$ κλάση 2
$g(x) = 0 \Rightarrow$ οποιαδήποτε κλάση

2-Κλάσεις

Το διάνυσμα βαρών, **w**, καθορίζει τον προσανατολισμό του υπερεπιπέδου απόφασης και το βάρος κατωφλίου, w_0, καθορίζει τη σχετική θέση του ως προς την αρχή των αξόνων.



$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0$$

- Το όριο απόφασης

  $g(x) = w^t x + w_0 = 0$

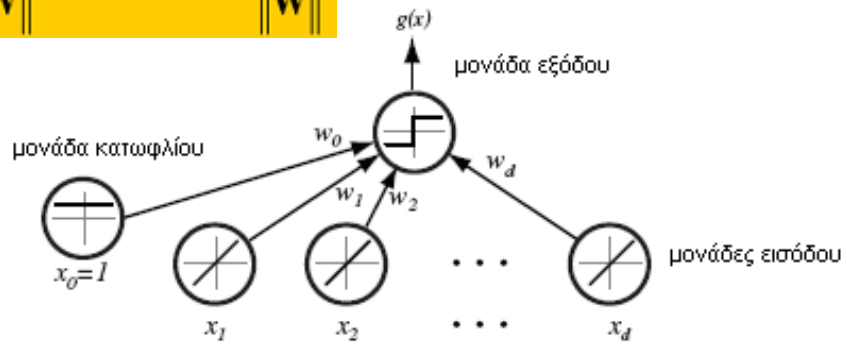
  είναι ένα υπερεπίπεδο

Ένα σύνολο διανυσμάτων x, τα οποία για κάποιες αριθμητικές (scalar) τιμές $\alpha_0, \ldots, \alpha_d$ ικανοποιούν την εξίσωση:

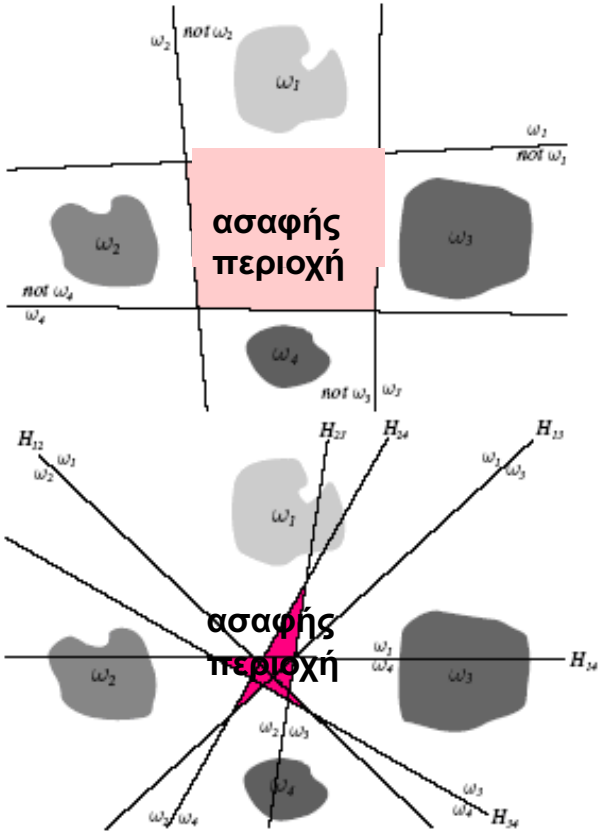
$$\alpha_0 + \alpha_1 x^{(1)} + \ldots + \alpha_d x^{(d)} = 0$$

$$\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}, \text{ όπου } r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}$$

Ένα υπερεπίπεδο είναι:
- Ένα σημείο σε 1-D
- Μια ευθεία σε 2-D
- Ενα επίπεδο σε 3-D



Η Περίπτωση Πολλών Κλάσεων



ω_i / όχι ω_i διχοτομήσεις

ω_i / ω_j διχοτομήσεις με τα αντίστοιχα H_{ij} όρια απόφασης

Η Περίπτωση Πολλών Κλάσεων

$$g_i(\boldsymbol{x}) = \boldsymbol{w}_i^t \boldsymbol{x} + w_{i0} \quad i=1,\ldots m$$

Γραμμική Μηχανή (linear machine): $\boldsymbol{x}$ in ω_i αν $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x})$ $\forall j \neq i$

Σύνορα Απόφασης: H_{ij}: $g_i(\boldsymbol{x}) = g_j(\boldsymbol{x}) \rightarrow (\boldsymbol{w}_i - \boldsymbol{w}_j)^t \boldsymbol{x} + (w_{i0} - w_{j0}) = 0$

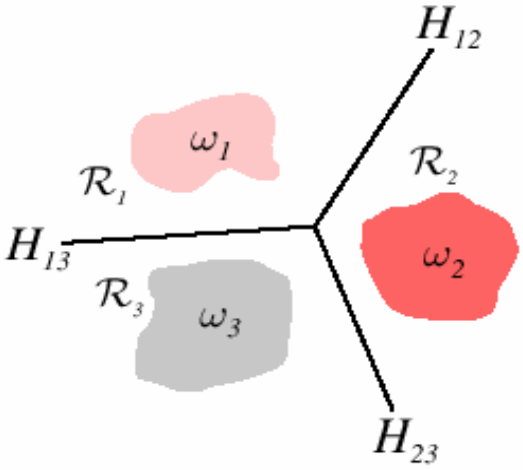
- τμήμα υπερεπιπέδου κάθετο στο διάνυσμα $\boldsymbol{w}_i - \boldsymbol{w}_j$.

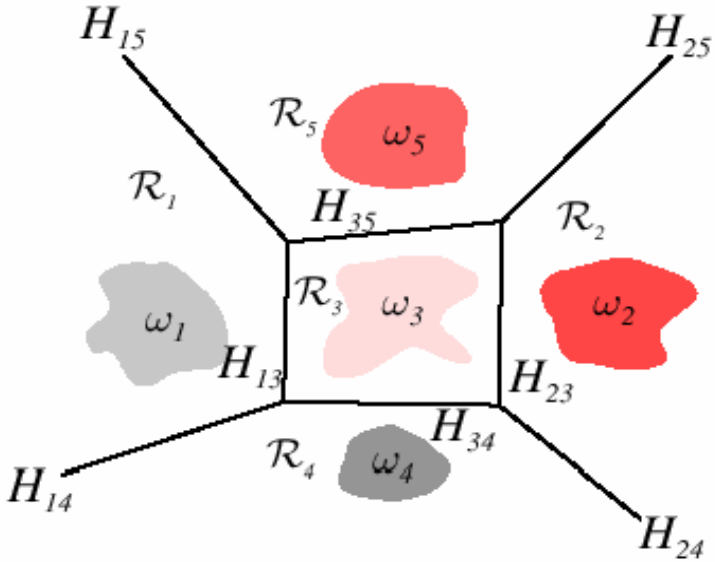
Απόσταση του $\boldsymbol{x}$ από το H_{ij}: $(g_i(\boldsymbol{x}) - g_j(\boldsymbol{x})) / \|\boldsymbol{w}_i - \boldsymbol{w}_j\|$

- σημαντικές είναι οι διαφορές των διανυσμάτων βαρών.

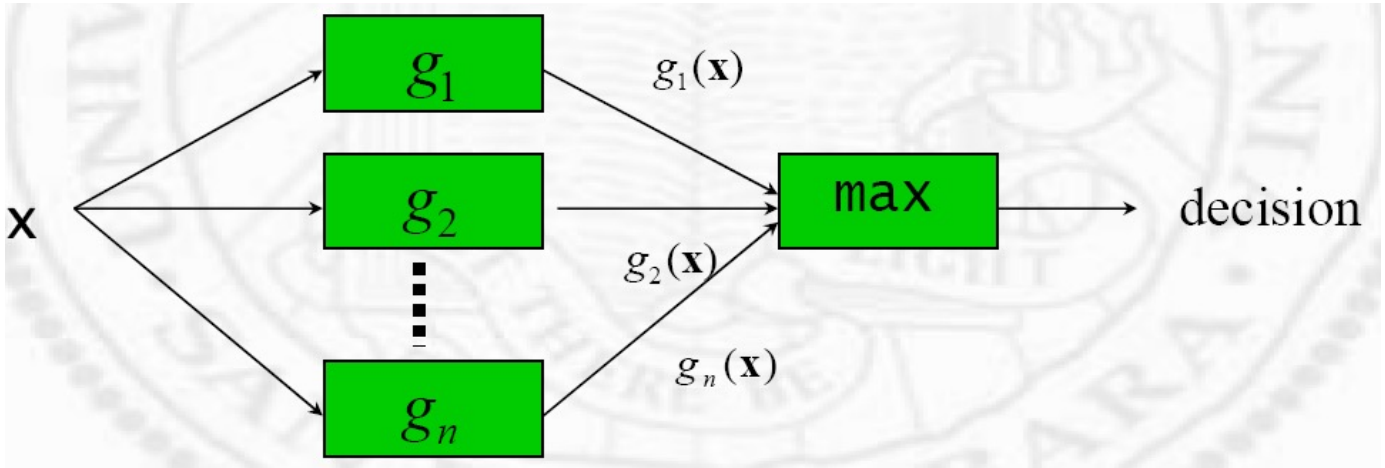
Για δυο όμορες περιοχές R_i , R_j

Κυρτές περιοχές απόφασης.





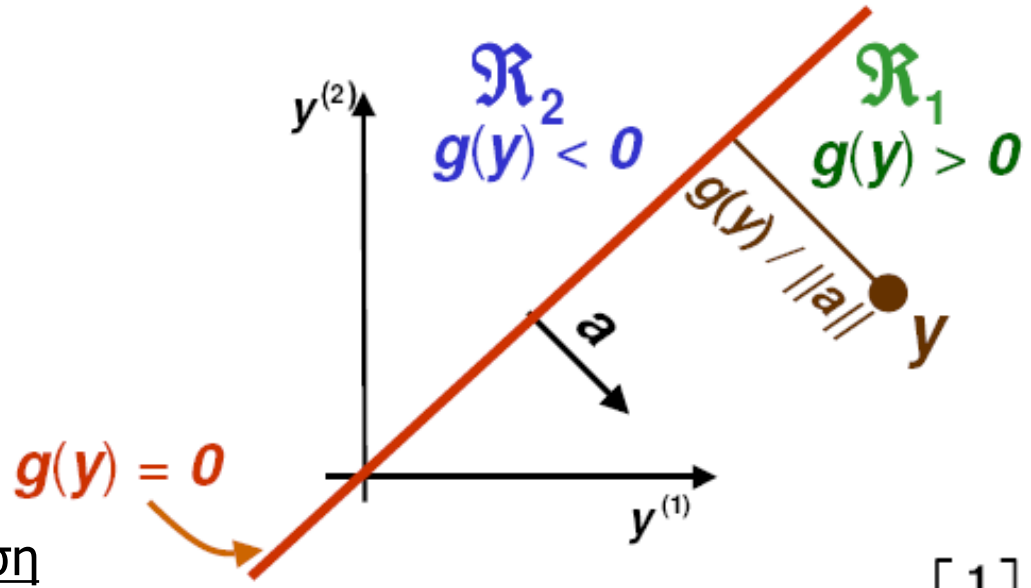
Η Περίπτωση Πολλών Κλάσεων



Επαυξημένα Διανύσματα χαρακτηριστικών και βαρών

$$g(x) = w^t x + w_0$$

$$g(x) = \underbrace{\begin{bmatrix} w_0 & w^t \end{bmatrix}}_{\text{νέο διάνυσμα a}} \underbrace{\begin{bmatrix} 1 \\ x \end{bmatrix}}_{\text{νέο διάνυσμα y}} = a^t y = g(y)$$



Παλιά διατύπωση

$$g(x) = w^t x + w_0$$

$$\begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

Νέα διατύπωση

$$g(y) = a^t y$$

$$\begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$y_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

Μηδενικό λάθος ταξινόμησης

Όταν $g(y_i) > 0 \Rightarrow y_i$ ταξινομείται στην ω_1
$g(y_i) < 0 \Rightarrow y_i$ ταξινομείται στην ω_2

Άρα μηδενικό λάθος ταξινόμησης όταν:

$$\begin{cases} g(y_i) > 0 & \forall y_i \in \omega_1 \\ g(y_i) < 0 & \forall y_i \in \omega_2 \end{cases} \Rightarrow \begin{cases} a^t y_i > 0 & \forall y_i \in \omega_1 \\ a^t y_i < 0 & \forall y_i \in \omega_2 \end{cases}$$

Αντίστοιχα μηδενικό λάθος ταξινόμησης όταν:

$$\begin{cases} a^t y_i > 0 & \forall y_i \in \omega_1 \\ a^t (-y_i) > 0 & \forall y_i \in \omega_2 \end{cases}$$

Κανονικοποίηση

$$y_i \rightarrow -y_i \qquad \forall y_i \in \omega_2$$

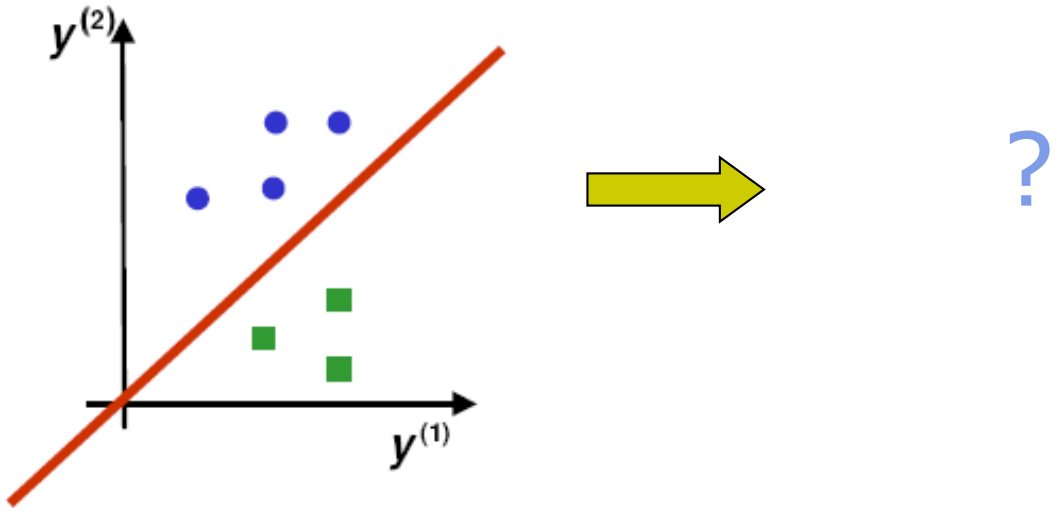
Ώστε:

$$a^t y_i > 0 \qquad \forall y_i$$

Κανονικοποίηση: αντικαθιστώντας όλα τα δείγματα εκπαίδευσης της κλάσης ω_2 με τα αρνητικά τους, ζητούμε τα διανύσματα διαχωρισμού (separating vectors) που πρέπει να ικανοποιούν την σχέση: $a^t y_i > 0 \qquad \forall y_i$

Κανονικοποίηση

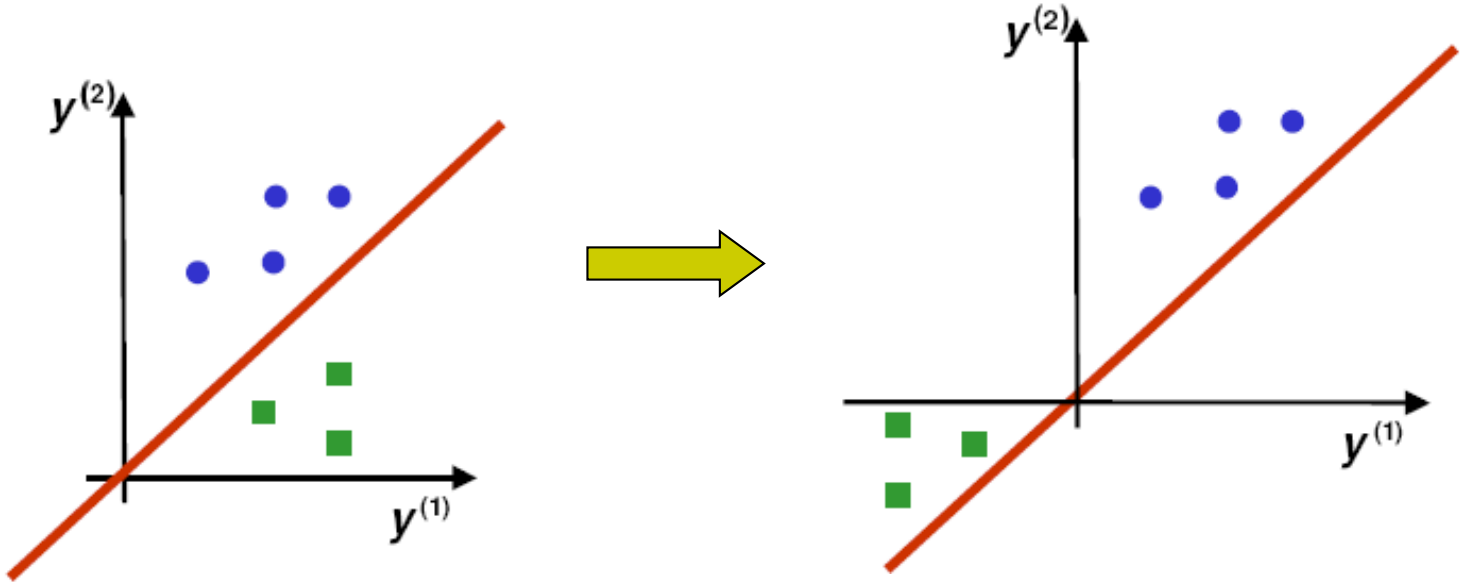
Κίνητρο: παύουμε να ασχολούμαστε με τις ετικέτες ω_1, ω_2



Αναζήτηση υπερεπιπέδου που διαχωρίζει τα δείγματα (πρότυπα) των δυο κλάσεων

Κανονικοποίηση

Κίνητρο: παύουμε να ασχολούμαστε με τις ετικέτες ω_1, ω_2

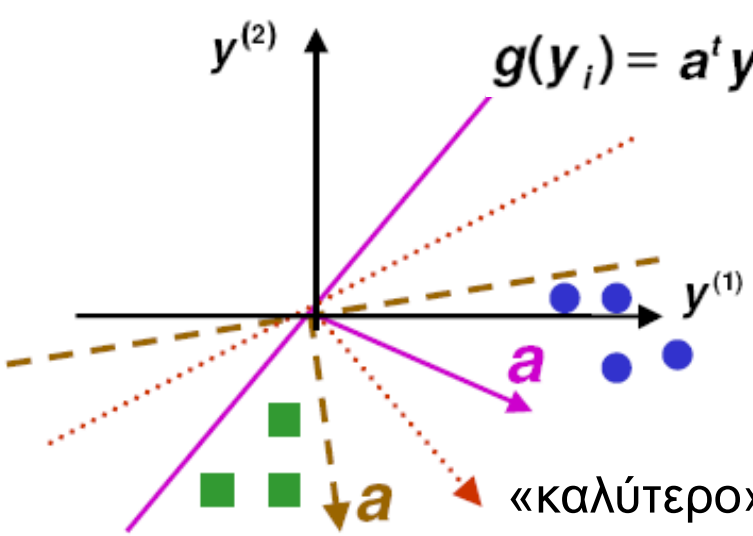


Αναζήτηση υπερεπιπέδου που διαχωρίζει τα δείγματα (πρότυπα) των δυο κλάσεων

Αναζήτηση υπερεπιπέδου που τοποθετεί τα δείγματα (πρότυπα) των δυο κλάσεις στην ίδια (θετική) πλευρά

Περιοχή Λύσης

Αναζήτηση διανύσματος βαρών $\boldsymbol{a}$,ώστε για όλα τα δείγματα $y_1, \ldots, y_n$

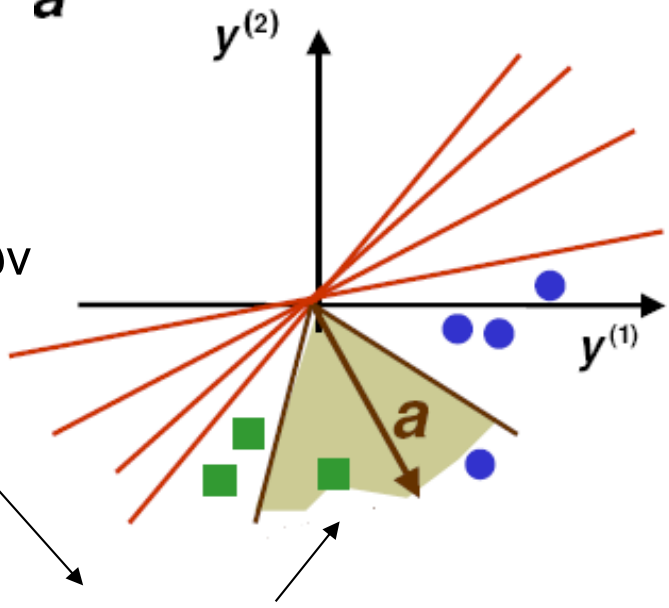


$$g(y_i) = a^t y_i = \sum_{k=0}^{d} a_k y_i^{(k)} > 0$$

«καλύτερο» $\boldsymbol{a}$

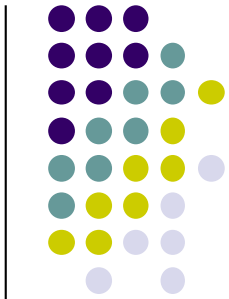
Στη γενική περίπτωση υπάρχουν πολλές λύσεις για το $\boldsymbol{a}$

Το σύνολο όλων των δυνατών λύσεων



Περιοχή λύσης για το $\boldsymbol{a}$

Διαδικασίες Βελτιστοποίησης



- **Πρόβλημα: Εύρεση του a που ικανοποιεί το σύνολο των γραμμικών ανισοτήτων $a^t y_i > 0$ για κάθε i=1,…,n.**

$$a^t y_i = \sum_{k=0}^{d} a_k y_i^{(k)} > 0$$

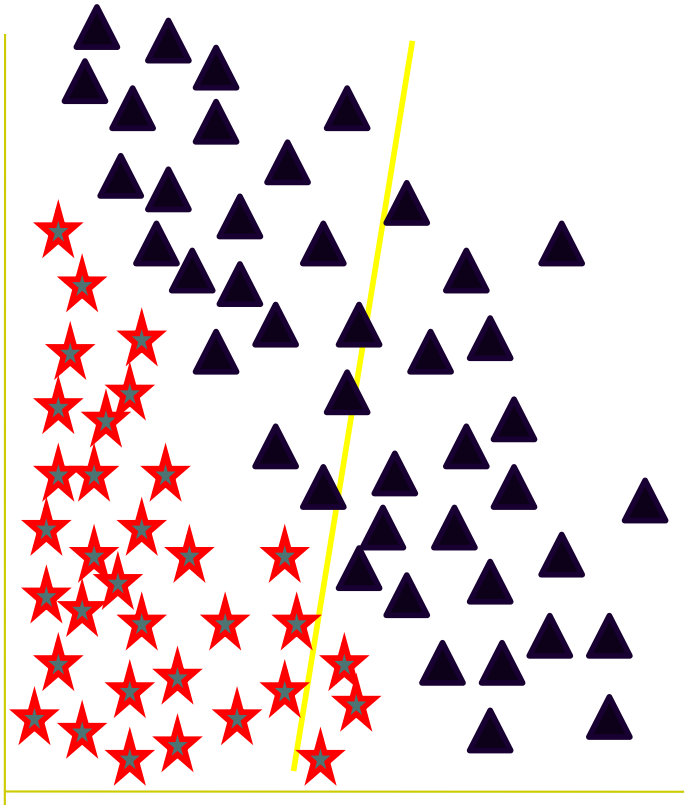
- **Πώς βρίσκουμε την κατάλληλη λύση;**
  - Ορίζουμε μια συνάρτηση κριτηρίου, *J*(**a**), η οποία όταν ελαχιστοποιείται το **a** είναι **ένα διάνυσμα λύσης.**
  - Μια πρώτη σκέψη θα ήταν... ο αριθμός των λάθος ταξινομημένων δειγμάτων εκπαίδευσης

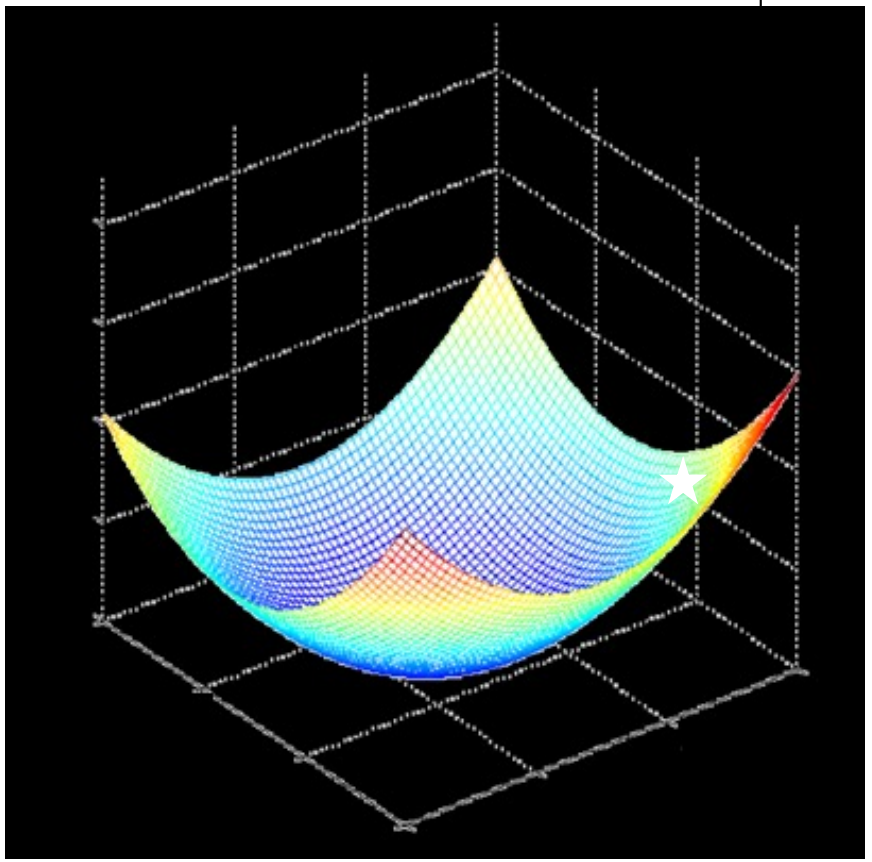
$$Y_M(a) = \{\text{δείγμα } y_i \text{ με } a^t y_i < 0\}$$

$$J(a) = |Y_M(a)|$$

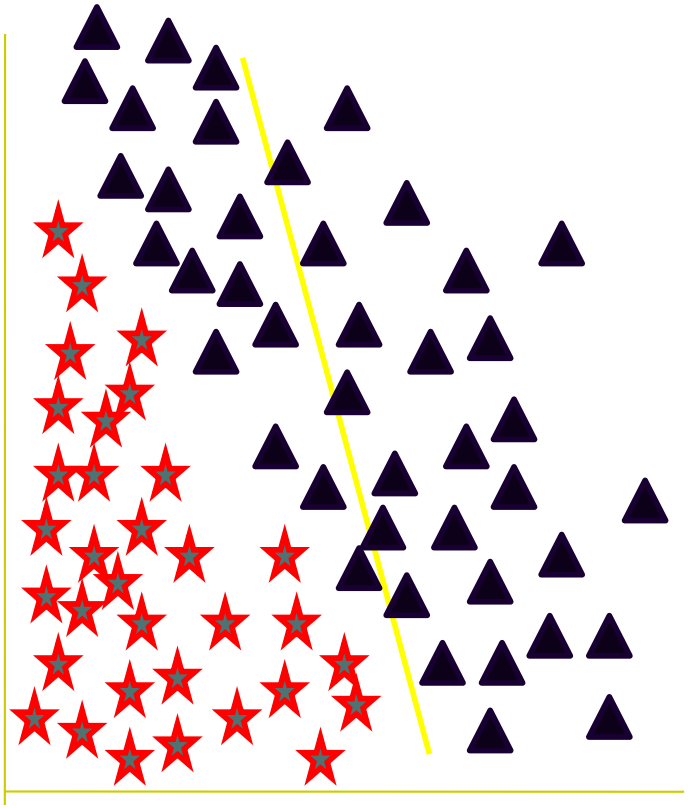
  - Με αυτό τον τρόπο μετασχηματίζουμε το πρόβλημα της εξαντλητικής αναζήτησης σε πρόβλημα ελαχιστοποίησης μιας βαθμωτής συνάρτησης.

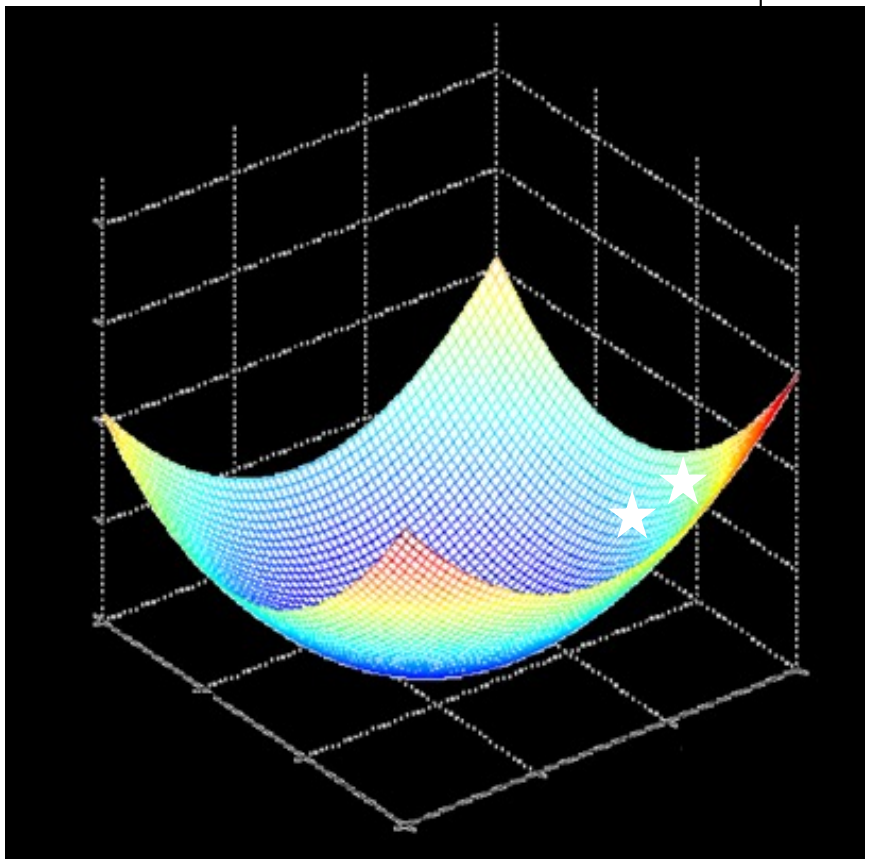
Πώς βρίσκουμε την κατάλληλη λύση;



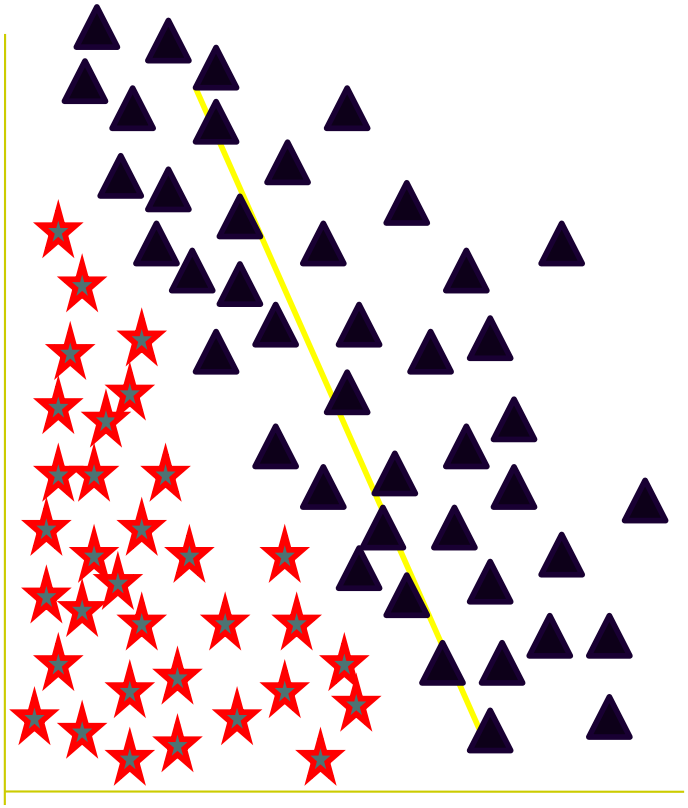


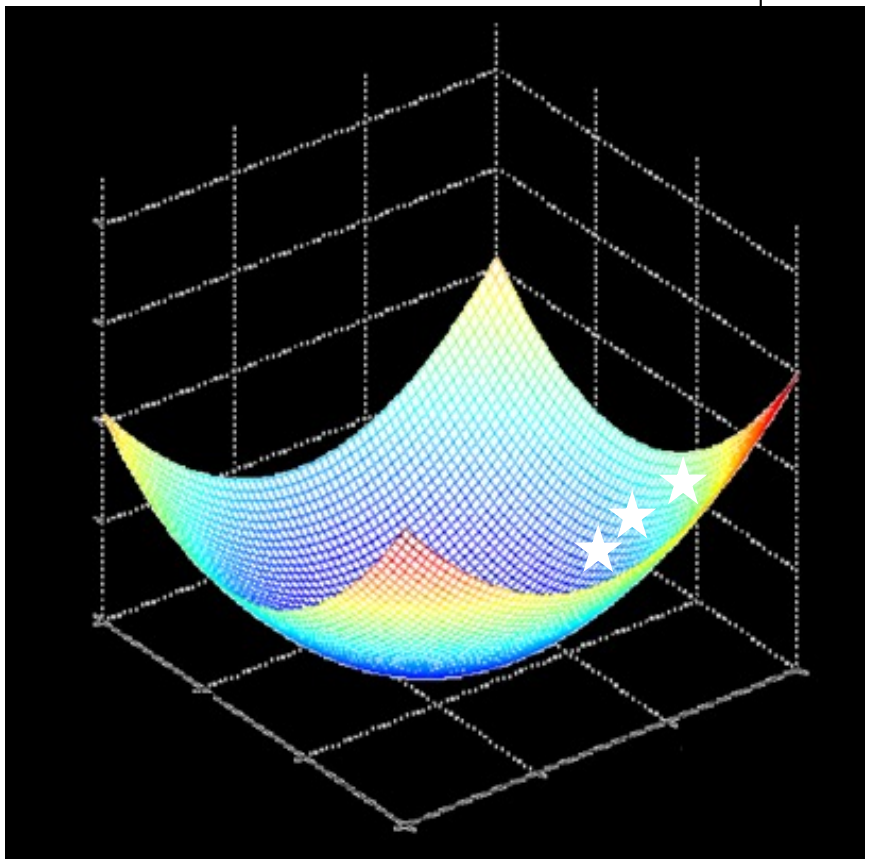
Πώς βρίσκουμε την κατάλληλη λύση;



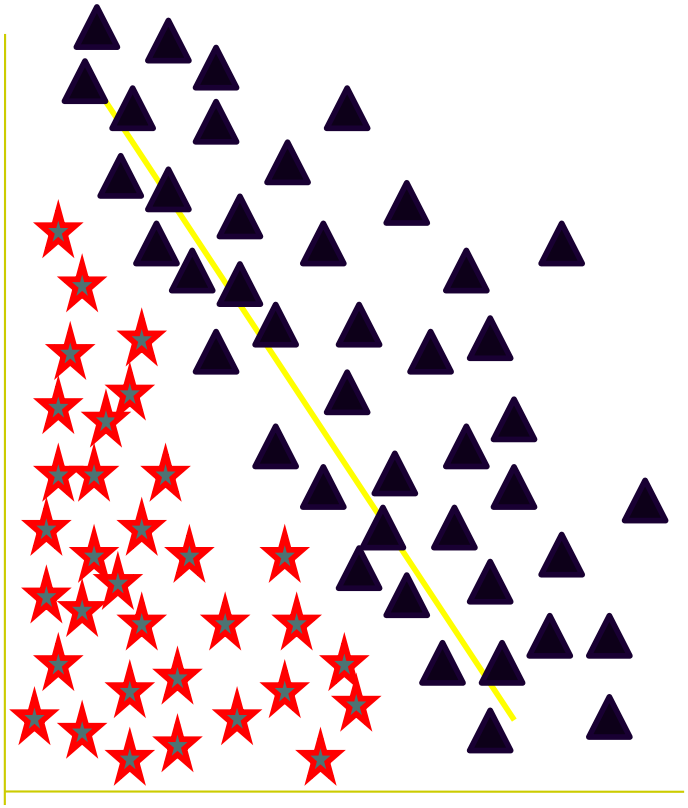


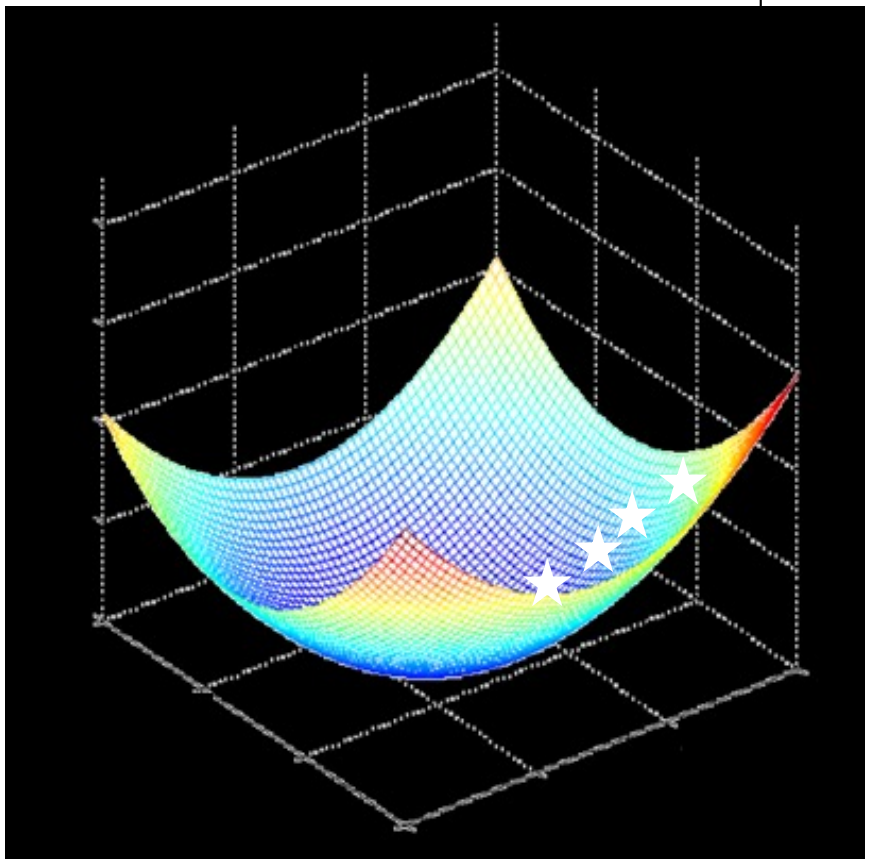
Πώς βρίσκουμε την κατάλληλη λύση;



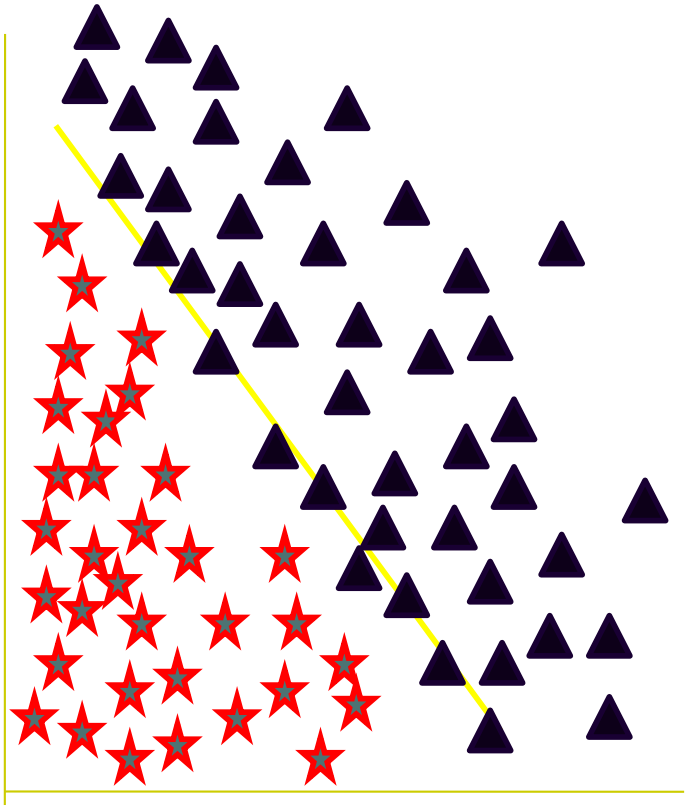


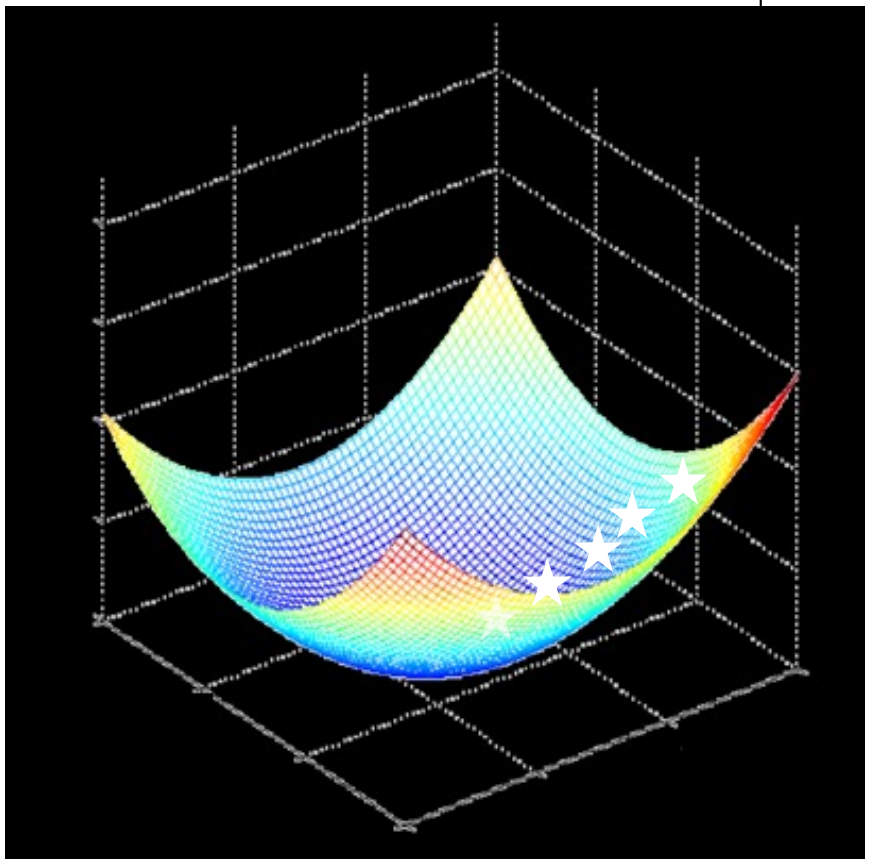
Πώς βρίσκουμε την κατάλληλη λύση;



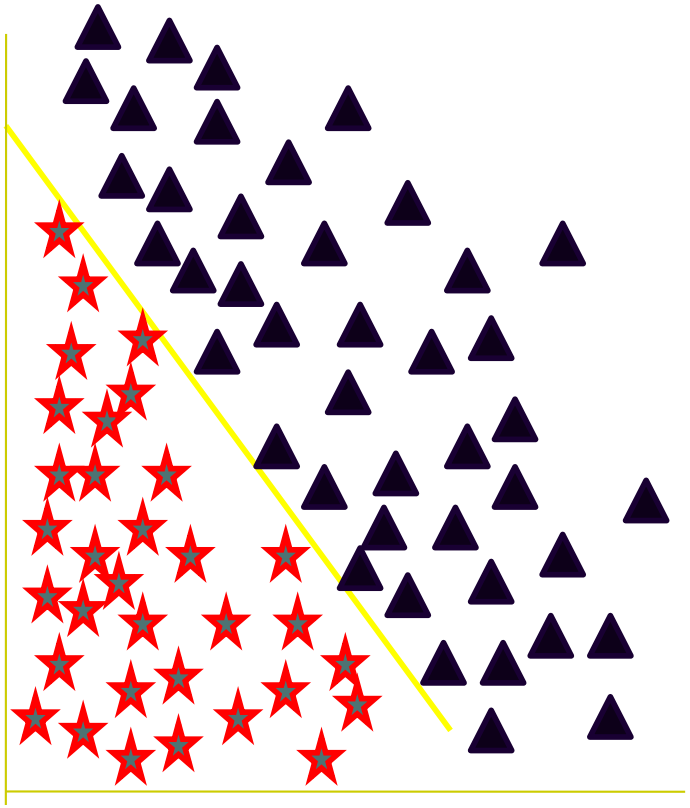


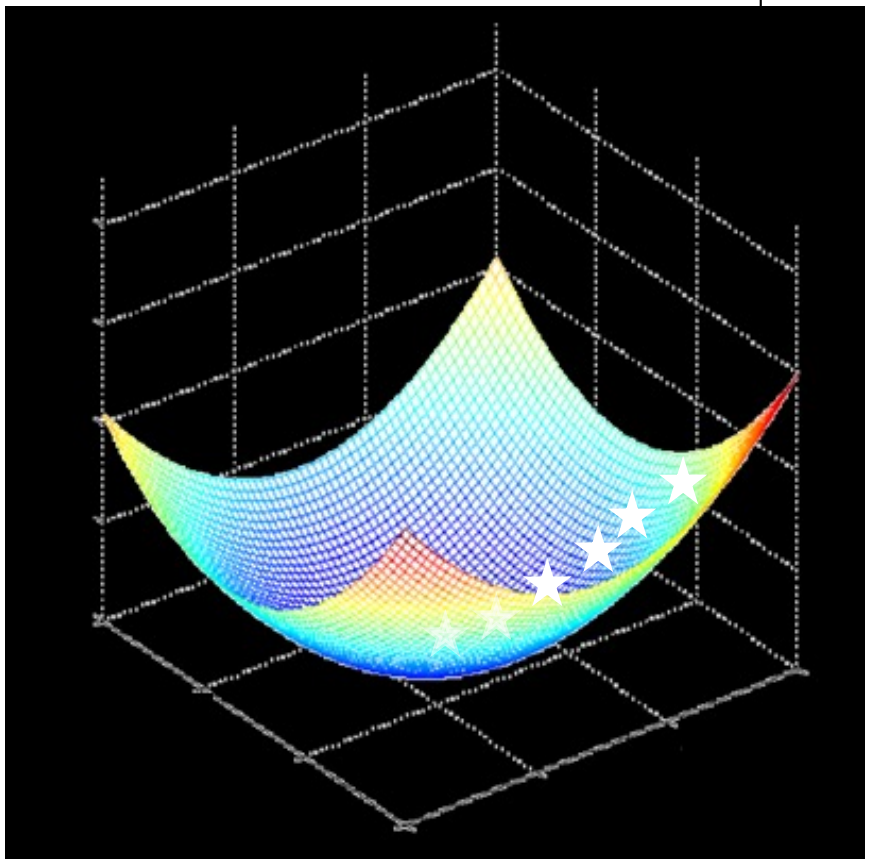
Πώς βρίσκουμε την κατάλληλη λύση;





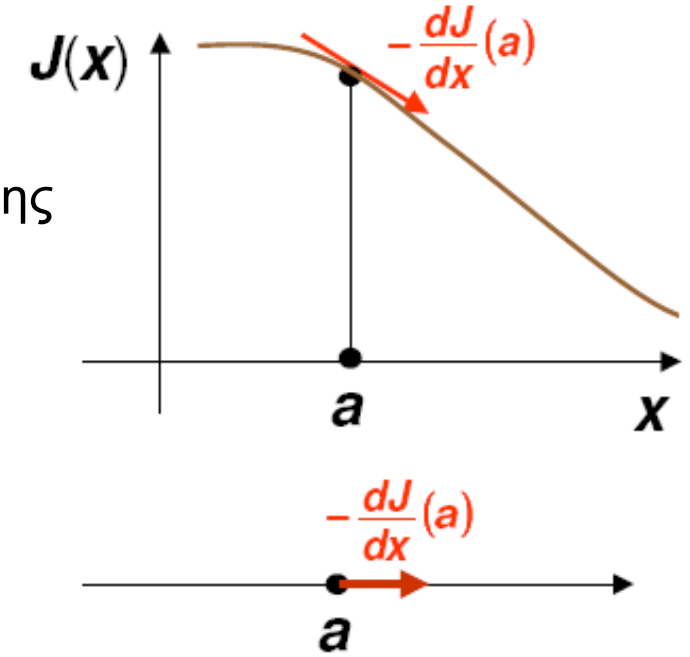
Πώς βρίσκουμε την κατάλληλη λύση;



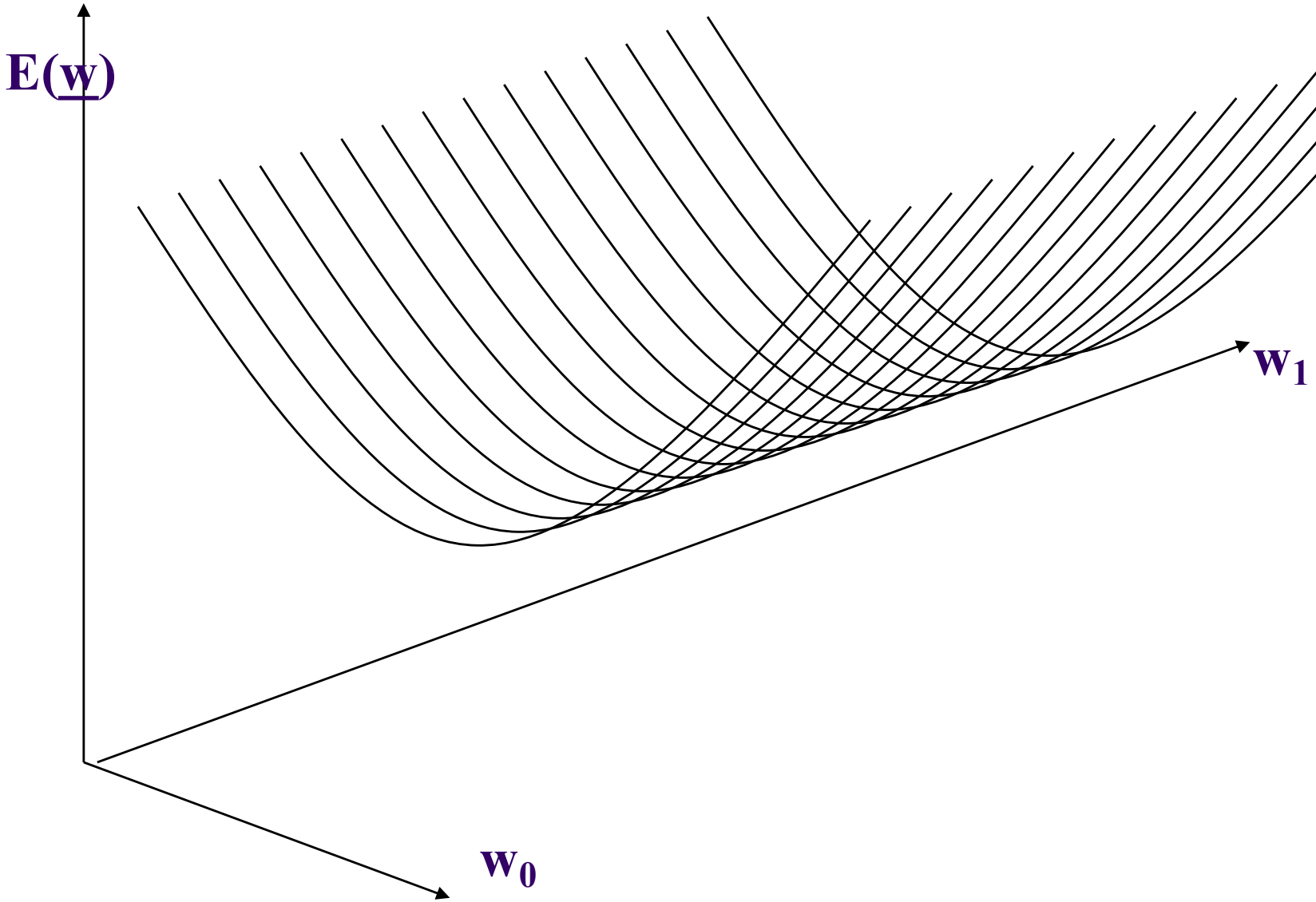


Αλγόριθμος της Πιο Απότομης Καθόδου (Steepest Descent)

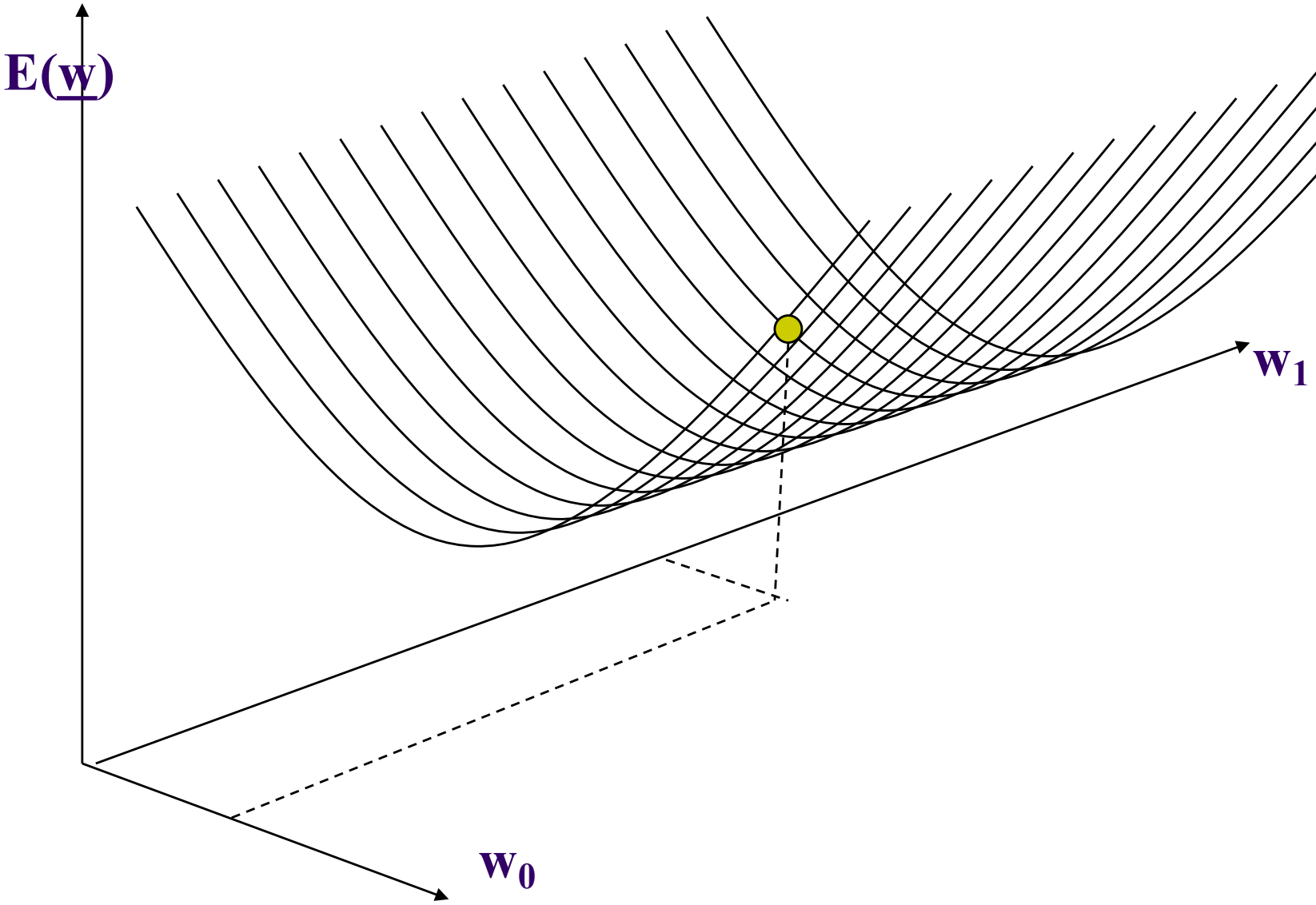
- **Πώς ελαχιστοποιούμε την $J(\mathbf{a})$;**
  - Επιλέγουμε κάποιο αρχικό σημείο $\mathbf{a}_1$ και υπολογίζουμε την τιμή $J(\mathbf{a}_1)$.
  - Υπολογίζουμε την κλίση $\nabla J(\mathbf{a}_1)$ στο $J(\mathbf{a}_1)$:
  - Παίρνουμε το επόμενο σημείο $\mathbf{a}_2$ κινούμενοι στην κατεύθυνση αρνητικής κλίσης (steepest descent), κατά μια ποσότητα $\eta(k)$, τον λεγόμενο ρυθμό μάθησης (learning rate) ή το βήμα (stepsize).



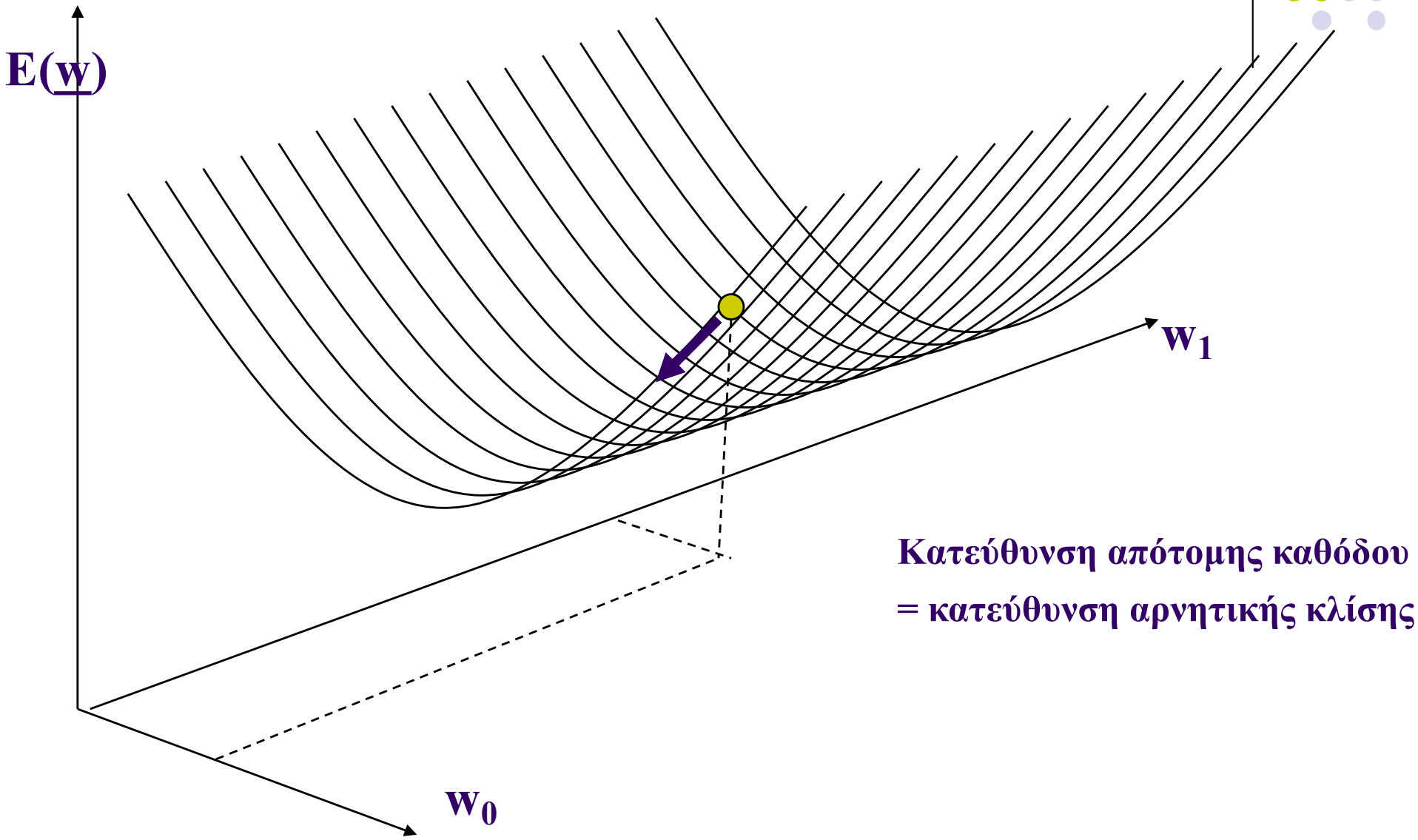
Αλγόριθμος της Πιο Απότομης Καθόδου



Αλγόριθμος της Πιο Απότομης Καθόδου

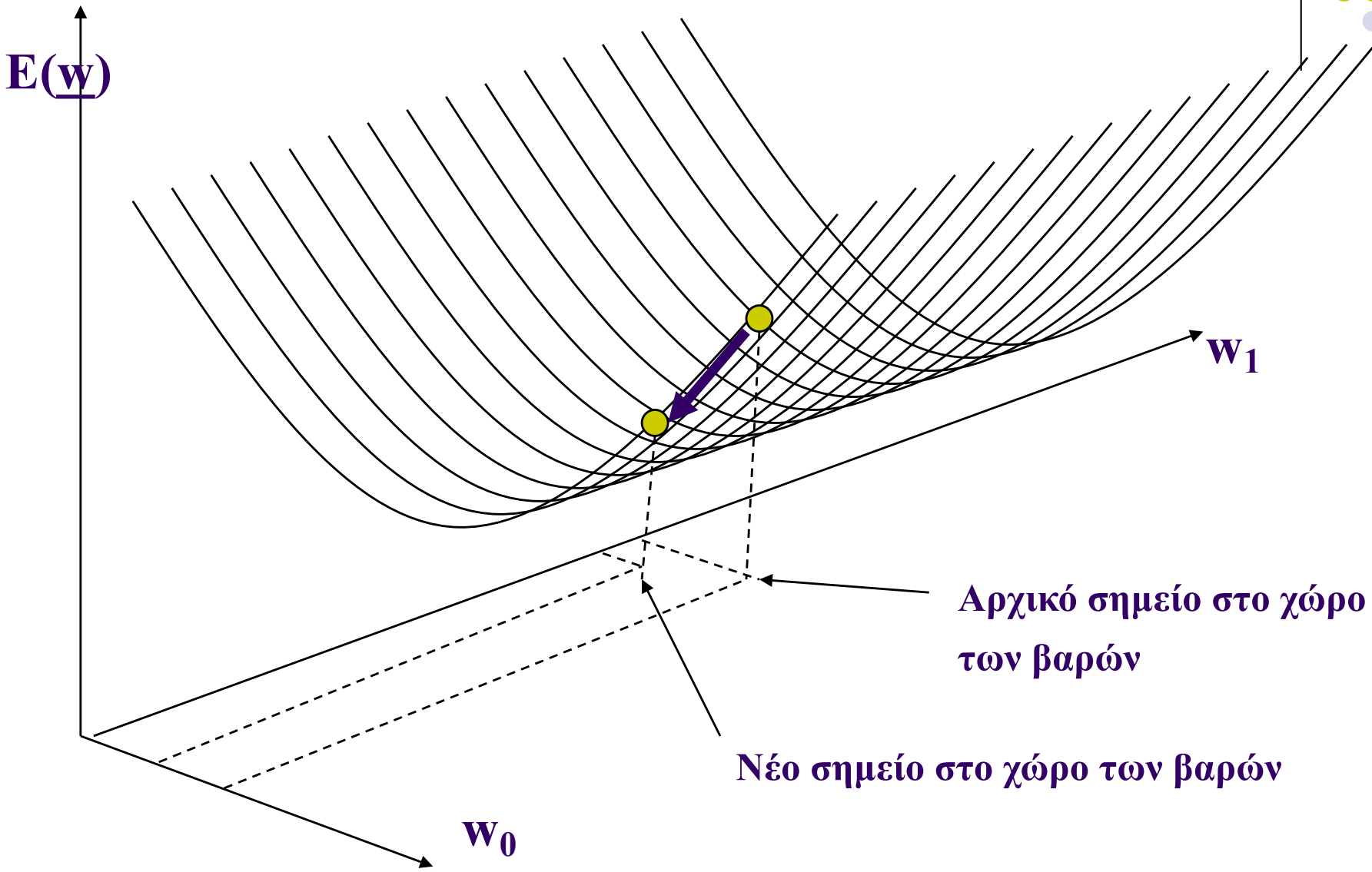


Αλγόριθμος της Πιο Απότομης Καθόδου

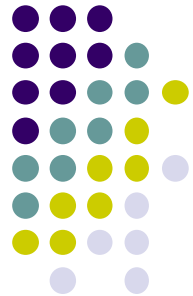


Κατεύθυνση απότομης καθόδου
= κατεύθυνση αρνητικής κλίσης

Αλγόριθμος της Πιο Απότομης Καθόδου



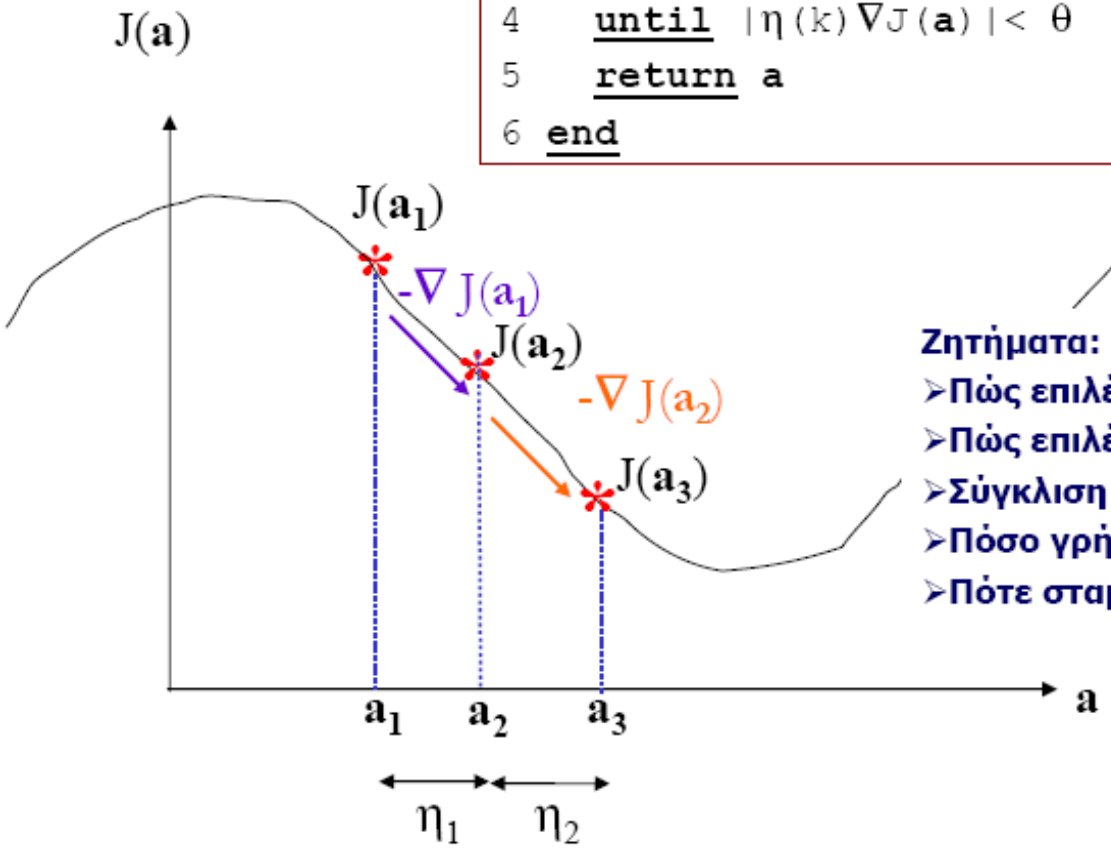
Αλγόριθμος της Πιο Απότομης Καθόδου (Steepest Descent)



Αλγόριθμος 1. Πιο Απότομη Κάθοδος (Steepest Descent)

```
1 begin initialize a, threshold θ, η(0)>0, k=0
2   do k ← k+1
3   a ← a-η(k)∇J(a)
4   until |η(k)∇J(a)|< θ
5   return a
6 end
```

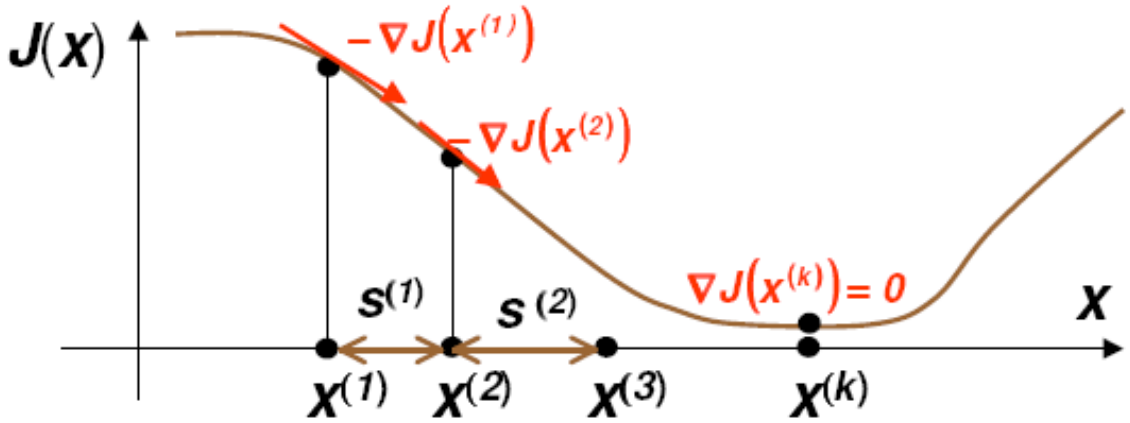
$$\mathbf{a}_{k+1} = \mathbf{a}_k - \eta_k \nabla J(\mathbf{a}_k)$$



Ζητήματα:

- **Πώς επιλέγουμε την συνάρτηση κριτηρίου;**
- **Πώς επιλέγουμε τον ρυθμό μάθησης *η(k)*;**
- **Σύγκλιση σε τοπικό/ολικό ελάχιστο;**
- **Πόσο γρήγορα συγκλίνουμε, πόσο ομαλά;**
- **Πότε σταματάμε;**

Αλγόριθμος της Πιο Απότομης Καθόδου (Steepest Descent)



$$s^{(k+1)} = x^{(k+1)} - x^{(k)} = \eta^{(k)}\left(-\nabla J\left(x^{(k)}\right)\right)$$

Αλγόριθμος Πιο Απότομης Καθόδου για την ελαχιστοποίηση της $J(x)$

set $k = 1$ and $x^{(1)}$ αρχική εκτίμηση για το x

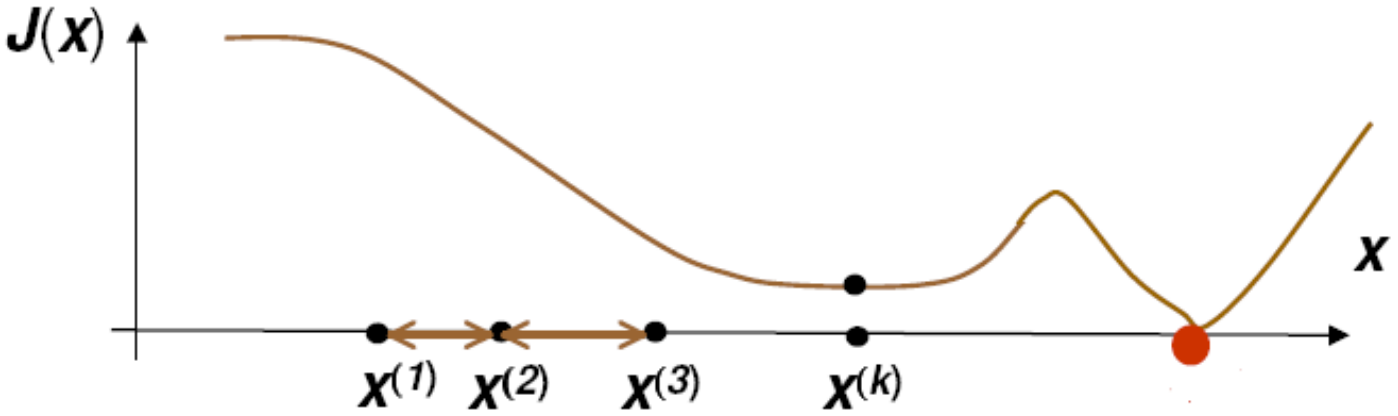
while $\eta^{(k)}\left|\nabla J\left(x^{(k)}\right)\right| > \varepsilon$

επίλεξε $\eta^{(k)}$

$x^{(k+1)} = x^{(k)} - \eta^{(k)} \nabla J(x)$

$k = k + 1$

Αλγόριθμος της Πιο Απότομης Καθόδου



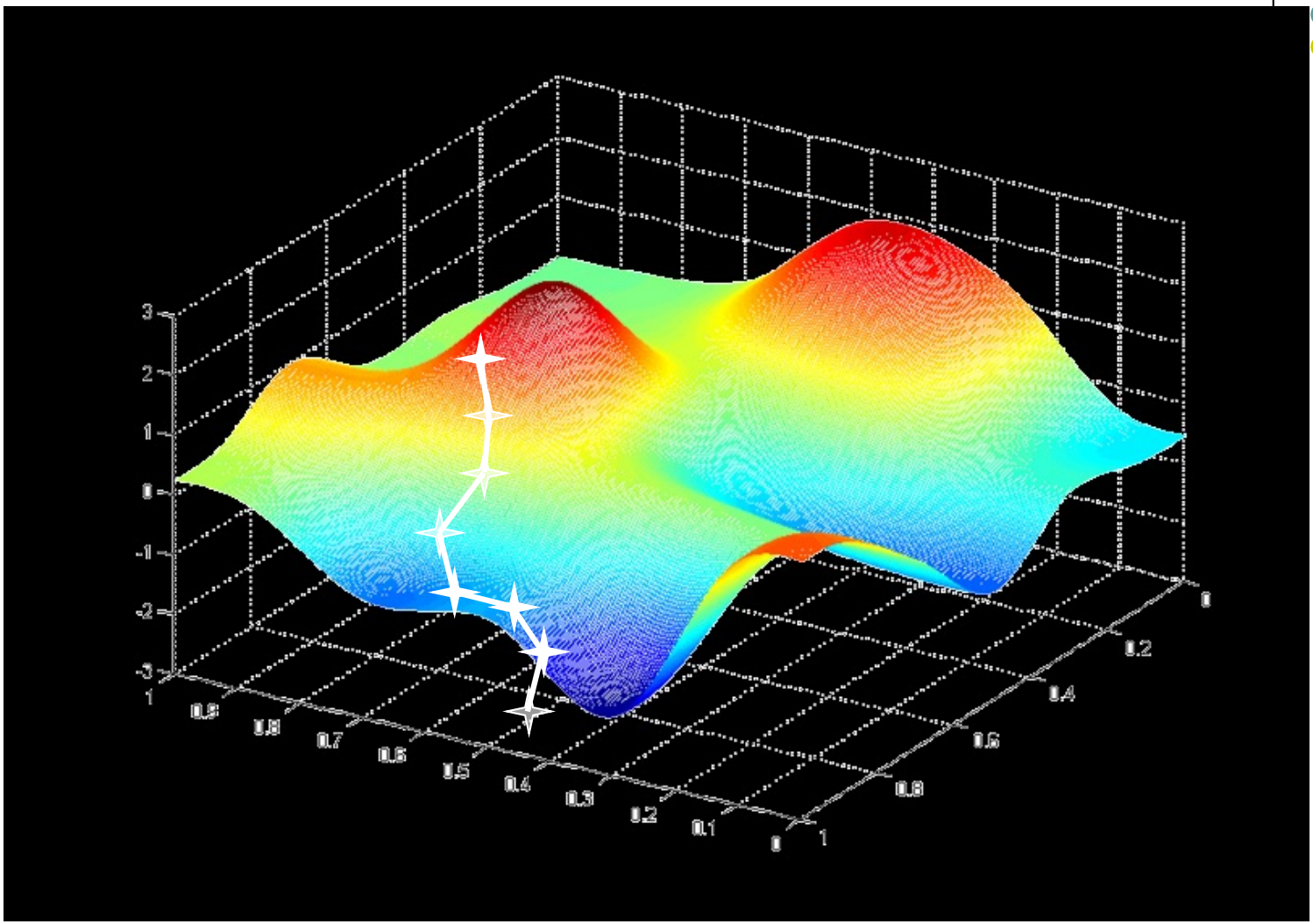
Συνολικό ελάχιστο

- Ο αλγόριθμος εγγυάται την εύρεση τοπικού ελάχιστου μόνο!

- Παρ'αυτά χρησιμοποιείται συχνά επειδή είναι απλός και μπορεί να εφαρμοστεί σε σχεδόν οποιαδήποτε συνάρτηση.

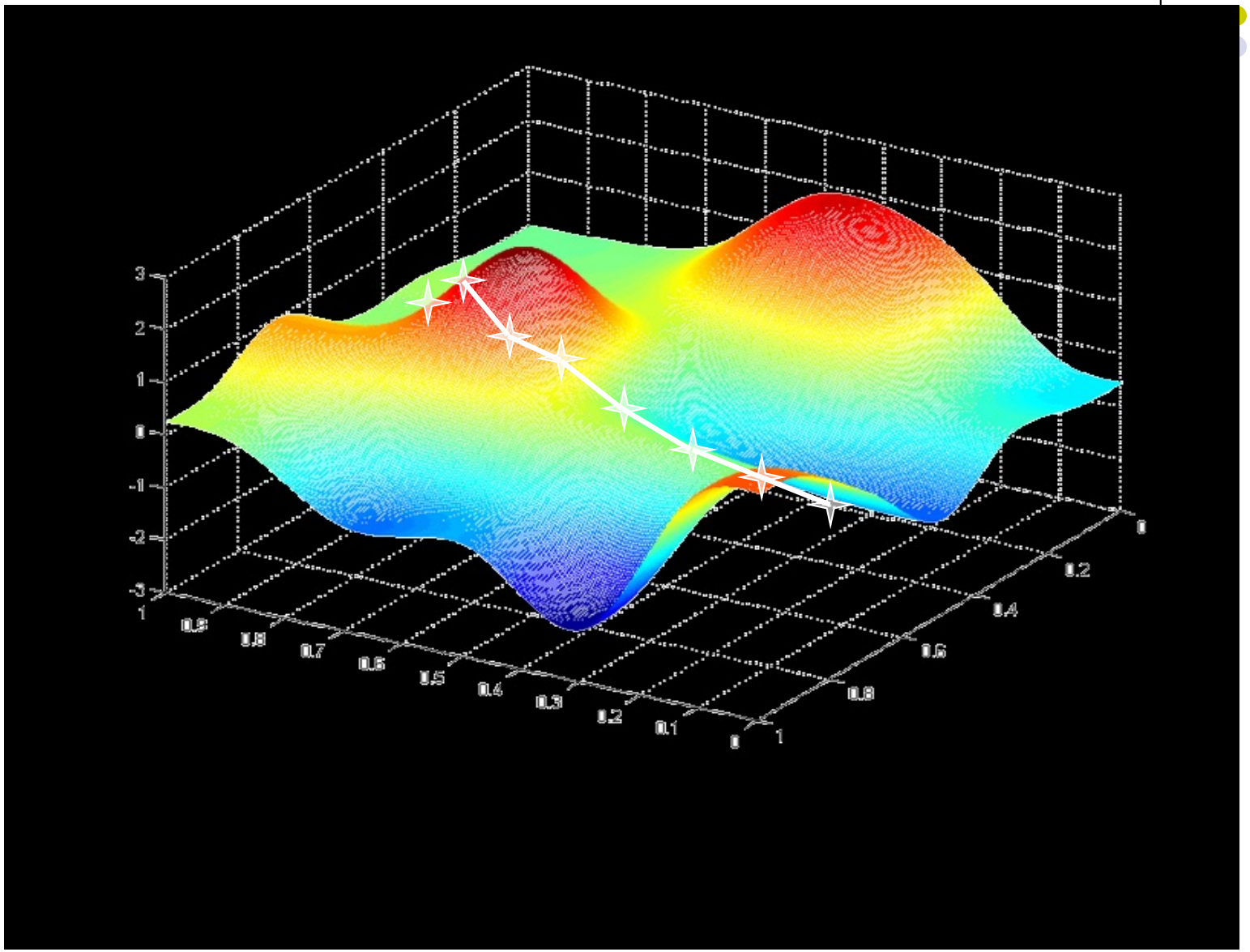
Τοπικό Βέλτιστο -1

$J(\theta_0,\theta_1)$



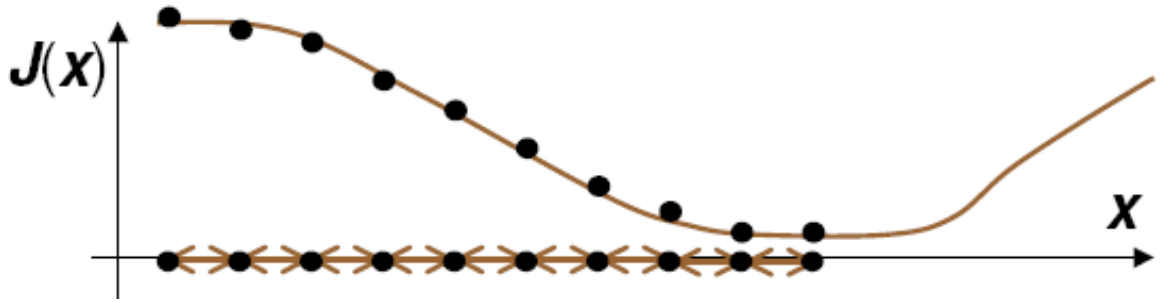
Τοπικό Βέλτιστο -1

$J(\theta_0,\theta_1)$

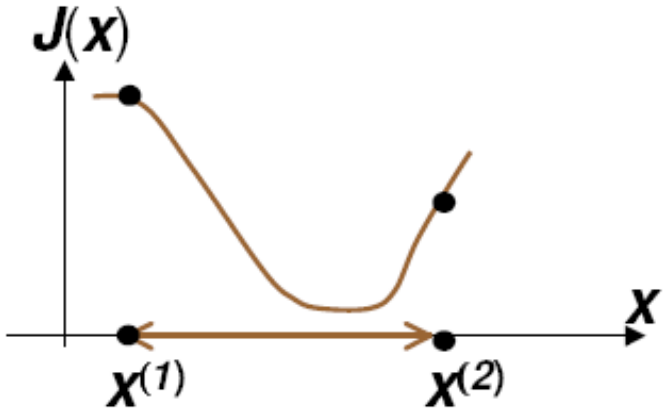


Επιλογή παραμέτρου μάθησης

Αν το η είναι πολύ μικρό, χρειάζονται πολλές επαναλήψεις



Αν το η είναι πολύ μεγάλο μπορεί να προσπεράσει το ελάχιστο και να μην το ξαναεντοπίσει (αν συνεχίσουμε να το προσπερνάμε)



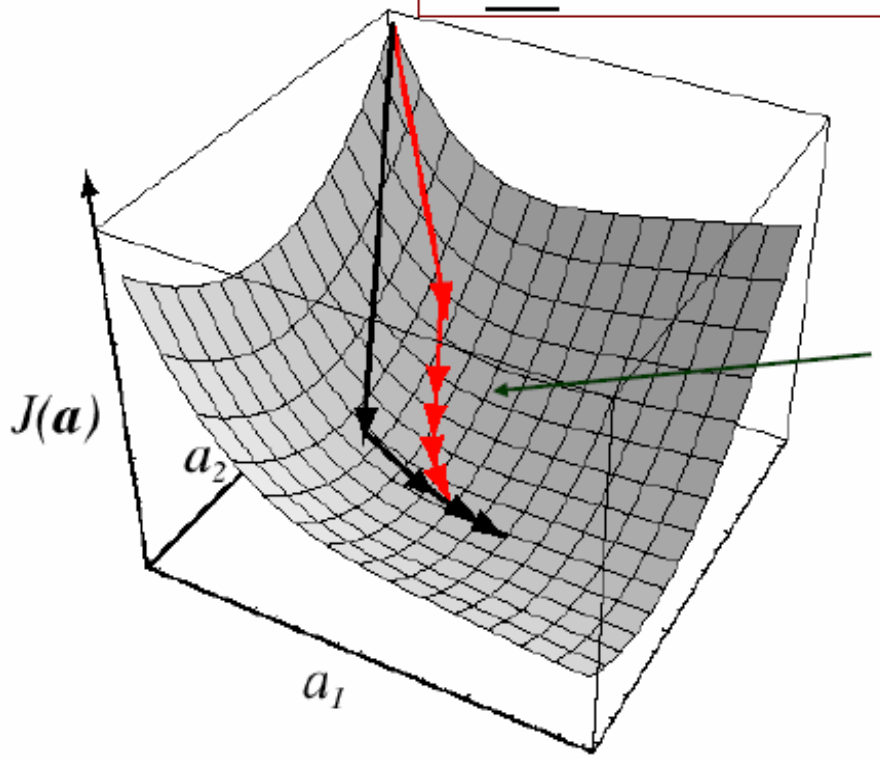
Αλγόριθμος Καθόδου Newton (Newton Descent)

Αλγόριθμος 2. Κάθοδος Newton (Newton Descent)

```
1 begin initialize a, threshold θ
2   a ← a-H^-1∇J(a)
3   until |H^-1∇J(a)|< θ
4   return a
5 end
```

$$\mathbf{a}_{k+1} = \mathbf{a}_k - \mathbf{H}_k^{-1}\nabla J(\mathbf{a}_k)$$

$$\mathbf{H}_k = \left[\partial^2 J(\mathbf{a})/\partial a_i \partial a_j\right]_{\mathbf{a}=\mathbf{a}_k}$$



Κόκκινο: Steepest Descent
Μαύρο: Newton Descent

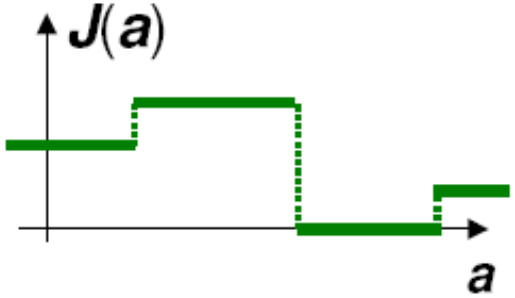
Newton: μεγαλύτερη βελτίωση σε κάθε βήμα πληρώνοντας το υπολογιστικό κόστος της αντιστροφής του Hessian πίνακα *H*.

Αλγόριθμος Καθόδου Newton (Newton Descent)

- Μεγάλο το μέτρο της $2^{ης}$ παραγώγου -> μεγάλη κλίση -> πρέπει να επιβραδύνουμε
- Μικρό το μέτρο της $2^{ης}$ παραγώγου -> μικρή κλίση -> πρέπει να επιταχύνουμε
- Παίρνουμε επομένως την αντίστροφη Hessian
- Θετικές τιμές -> φτάνουμε σε ελάχιστο (κοίλα προς τα πάνω)
- Αρνητικές τιμές -> φτάνουμε σε μέγιστο (κοίλα προς τα κάτω – αντιστροφή κατεύθυνσης)

Το Κριτήριο Perceptron

Πλήθος των λάθος ταξινομημένων δειγμάτων εκπαίδευσης.



Αλλά: Αυτή η συνάρτηση είναι ασυνεχής οπότε δεν είναι διαφόρισιμη.

Μια καλύτερη επιλογή: Η συνάρτηση κριτηρίου perceptron:

$$J_p(a) = \sum_{y \in Y_M} \left(-a^t y\right)$$

όπου Y_M είναι το σύνολο των δειγμάτων που δεν έχουν ταξινομηθεί σωστά.

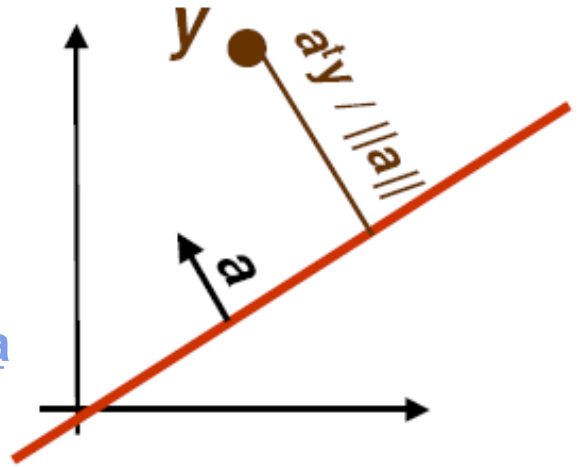
Κριτήριο Perceptron

$$J_p(a) = \sum_{y \in Y_M} (-a^t y)$$

Αν το y δεν ταξινομηθεί σωστά: $a^t y \leq 0$

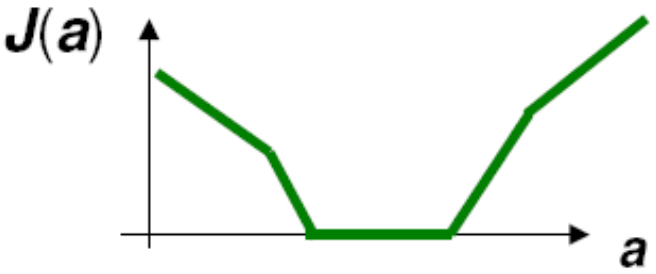
Οπότε $J_p(a) \geq 0$

Αν το Y_M είναι κενό, τότε $J_p(\mathbf{a})=0$.



Η $J_p(\mathbf{a})$ δεν είναι ποτέ αρνητική και μηδενίζεται όταν το **a** είναι διάνυσμα λύσης.

Γεωμετρικά, η $J_p(\mathbf{a})$ είναι ανάλογη (με συντελεστή $-\|\boldsymbol{a}\|$) του αθροίσματος των αποστάσεων των λάθος ταξινομημένων δειγμάτων από το σύνορο απόφασης.



Η $J_p(a)$ είναι κατά τμήματα (piecewise) γραμμική και άρα προσφέρεται για την εφαρμογή του αλγορίθμου gradient decent

Σωρηδόν Κανόνας Perceptron

$$J_p(a) = \sum_{y \in Y_M} (-a^t y)$$

Η κλίση του $J_p(a)$ είναι $\nabla J_p(a) = \sum_{y \in Y_M} (-y)$

- Y_M δείγματα ταξινομούνται λάθος από το διάνυσμα $a^{(k)}$
- Δεν είναι δυνατόν να λυθεί αναλυτικά η εξίσωση $\nabla J_p(a) = 0$ λόγω του Y_M

Κανόνας ενημέρωσης για τον αλγόριθμο απότομης καθόδου:

$$x^{(k+1)} = x^{(k)} - \eta^{(k)} \nabla J(x)$$

Οπότε ο κανόνας ενημέρωσης για την $J_p(a)$ είναι: ?

Σωρηδόν Κανόνας Perceptron

$$J_p(a) = \sum_{y \in Y_M} (-a^t y)$$

Η κλίση του $J_p(a)$ είναι $\nabla J_p(a) = \sum_{y \in Y_M} (-y)$

- Y_M δείγματα ταξινομούνται λάθος από το διάνυσμα $a^{(k)}$
- Δεν είναι δυνατόν να λυθεί αναλυτική η εξίσωση $\nabla J_p(a) = 0$ λόγω του Y_M

Κανόνας ενημέρωσης για τον αλγόριθμο απότομης καθόδου:

$$x^{(k+1)} = x^{(k)} - \eta^{(k)} \nabla J(x)$$

Οπότε ο κανόνας ενημέρωσης για την $J_p(a)$ είναι:

$$a^{(k+1)} = a^{(k)} + \eta^{(k)} \sum_{y \in Y_M} y$$

Καλείται σωρηδόν κανόνας batch (batch rule) επειδή βασίζεται σε όλα τα λαθος ταξινομημένα δείγματα

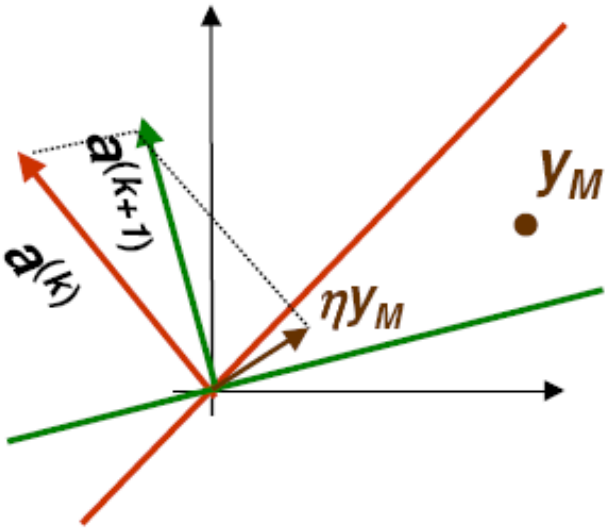
Κανόνας Single-Sample Perceptron

Οπότε ο κανόνας Single-Sample Perceptron για την $J_p(a)$ είναι:

$$a^{(k+1)} = a^{(k)} + \eta^{(k)} y_M$$

•Προσέξτε ότι το y_M είναι ένα διάνυσμα το οποίο ταξινομείται λάθος από το $a^{(k)}$

Γεωμετρική Ερμηνεία:



Κανόνας Single-Sample Perceptron

Οπότε ο κανόνας Single-Sample Perceptron για την $J_p(a)$ είναι:

$$a^{(k+1)} = a^{(k)} + \eta^{(k)} y_M$$

•Προσέξτε ότι το y_M είναι ένα διάνυσμα το οποίο ταξινομείται λάθος από το $a^{(k)}$

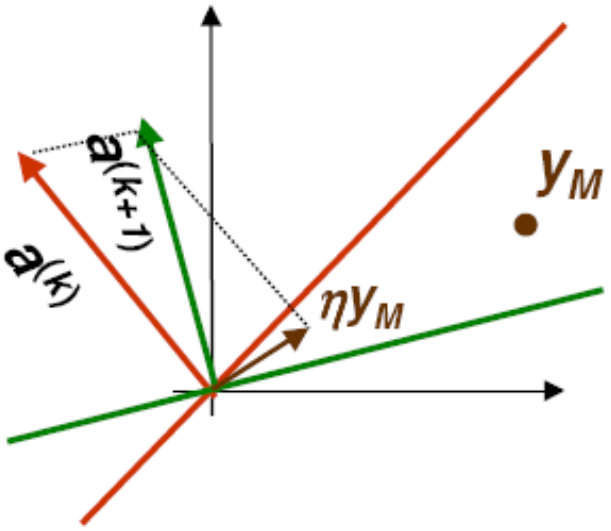
Γεωμετρική Ερμηνεία:

• Το y_M ταξινομείται λάθος από το $a^{(k)}$

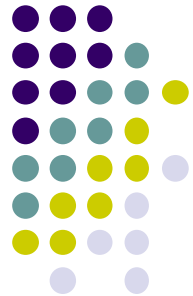
$$\left(a^{(k)}\right)^t y_M \leq 0$$

•Το y_M βρίσκεται στην λάθος πλευρά του υπερεπιπέδου

•Προσθέτοντας το ηy_M το διάνυσμα a μετακινεί το νέο υπερεπίπεδο προς την σωστή κατεύθυνση όσον αφορά στο y_M



Fixed-Increment Single-Sample Perceptron



- Αντί να δοκιμάζουμε το διάνυσμα βαρών *a(k)* σε όλα τα δείγματα και να το διορθώνουμε βάσει του συνόλου Y_k των λάθος ταξινομημένων δειγμάτων, χρησιμοποιούμε τα δείγματα ένα κάθε φορά και ανάλογα με την ταξινόμηση του ανανεώνουμε ή όχι το διάνυσμα βαρών.
- Αν επιπλέον, χρησιμοποιήσουμε ένα σταθερό βήμα *η(k)=1*, τότε προκύπτει ο αλγόριθμος:

Αλγόριθμος 4. Fixed-Increment Single-Sample Perceptron

```
1 begin initialize a, k=0
2   do k ← (k+1) mod n
3     If y^k is misclassified by a, then a ← a + y^k
4   until all patterns properly classified
5   return a
6 end
```

Κυκλική Σειρά Δεδομένων (με πράσινο υποδηλώνονται τα λάθος ταξινομημένα δείγματα):

$y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_1 \quad y_2 \quad y_3 \quad y_4$

$\rightarrow \quad y^1 \; y^2 \; y^3 \; y^0 \; y^1 \ldots = y_2 \quad y_1 \quad y_3 \quad y_2 \quad y_4$

Variable-Increment Perceptron with Margin

- Χρησιμοποιούμε τα δείγματα ένα κάθε φορά και διορθώνουμε το διάνυσμα βαρών *a(k)* όταν το εσωτερικό γινόμενό του με το δείγμα y^k είναι μικρότερο από κάποιο προκαθορισμένο θετικό όριο b: $a(k)y^k < b$.
- Αν επιπλέον, χρησιμοποιήσουμε ένα μεταβαλλόμενο βήμα $\eta(k)$, τότε προκύπτει ο αλγόριθμος:

Αλγόριθμος 5. Variable-Increment Perceptron w. Margin

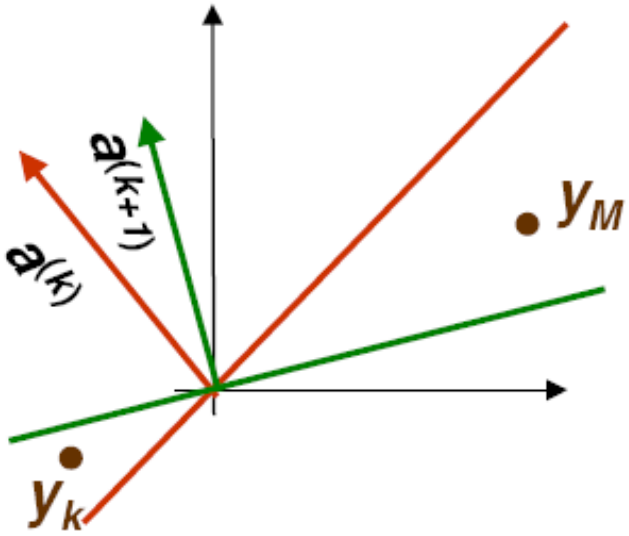
```
1 begin initialize a, threshold θ, margin b, η(0), k=0
2   do  k ← (k+1) mod n
3     if a^t y^k<b, then a ← a + η(k) y^k
4   until a^t y^k>b for all k
5   return a
6 end
```

Συνθήκες Σύγκλισης:

$$\eta(k) \geq 0, \quad \lim_{m\to\infty} \sum_{k=1}^{m} \eta(k) = \infty, \quad \lim_{m\to\infty} \frac{\sum_{k=1}^{m} \eta^2(k)}{\left(\sum_{k=1}^{m} \eta(k)\right)^2} = 0$$

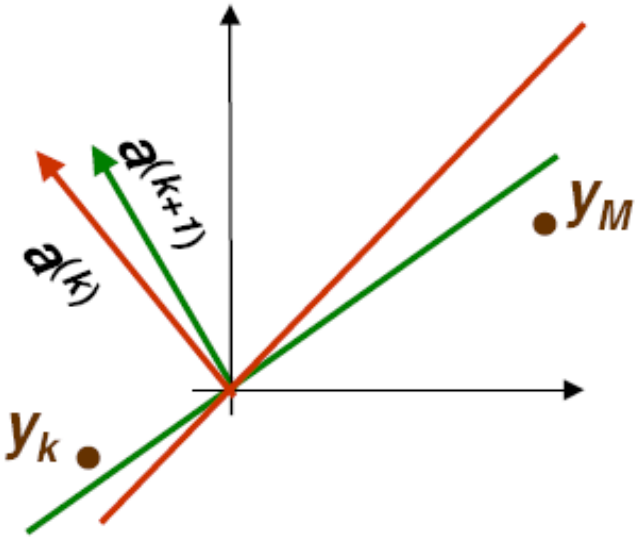
Κανόνας Single-Sample Perceptron

$$a^{(k+1)} = a^{(k)} + \eta^{(k)} y_M$$



Αν η παράμετρος η είναι πολύ μεγάλη

?

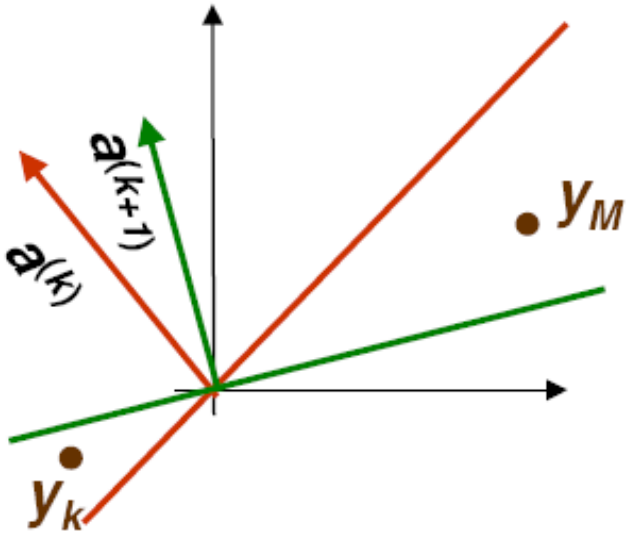


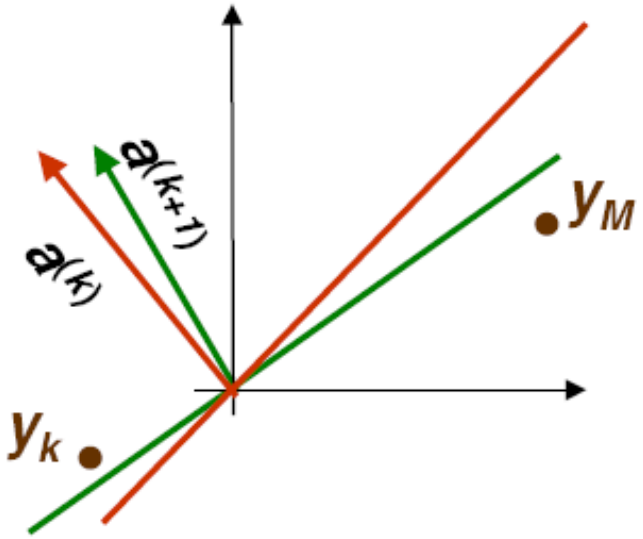
Αν η παράμετρος η είναι πολύ μικρή

?

Κανόνας Single-Sample Perceptron

$$a^{(k+1)} = a^{(k)} + \eta^{(k)} y_M$$

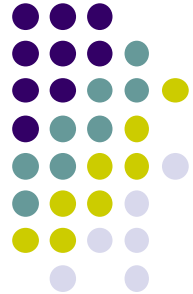




Αν η παράμετρος η είναι πολύ μεγάλη το πρώην σωστά ταξινομημένο δείγμα y_k τώρα ταξινομείται λάθος

Αν η παράμετρος η είναι πολύ μικρή το y_M παραμένει ταξινομημένο λάθος

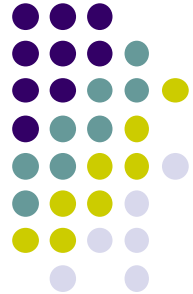
Παράδειγμα



	features				grade
name	***good attendance?***	***tall?***	***sleeps in class?***	***chews gum?***	
Jane	*yes (1)*	*yes (1)*	*no (-1)*	*no (-1)*	*A*
Steve	*yes (1)*	*yes (1)*	*yes (1)*	*yes (1)*	*F*
Mary	*no (-1)*	*no (-1)*	*no (-1)*	*yes (1)*	*F*
Peter	*yes (1)*	*no (-1)*	*no (-1)*	*yes (1)*	*A*

- ***class 1***: Φοιτητές που παίρνουν βαθμό Α
- ***class 2***: Φοιτητές που παίρνουν βαθμο F

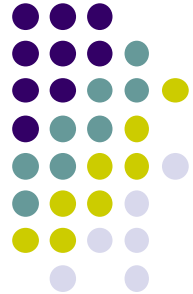
Παράδειγμα



	features					grade
name	***extra***	***good attendance?***	***tall?***	***sleeps in class?***	***chews gum?***	
Jane	?	*yes (1)*	*yes (1)*	*no (-1)*	*no (-1)*	*A*
Steve		*yes (1)*	*yes (1)*	*yes (1)*	*yes (1)*	*F*
Mary		*no (-1)*	*no (-1)*	*no (-1)*	*yes (1)*	*F*
Peter		*yes (1)*	*no (-1)*	*no (-1)*	*yes (1)*	*A*

Μετατροπή των δειγμάτων $x_1, \ldots, x_n$ σε επαυξημένα δείγματα $y_1, \ldots, y_n$ προσθέτοντας μια νέα διάσταση με τιμή 1

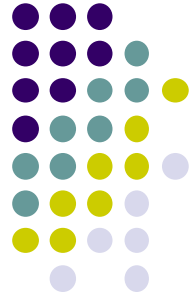
Παράδειγμα



	features					grade
name	***extra***	***good attendance?***	***tall?***	***sleeps in class?***	***chews gum?***	
Jane	*1*	*yes (1)*	*yes (1)*	*no (-1)*	*no (-1)*	*A*
Steve	*1*	*yes (1)*	*yes (1)*	*yes (1)*	*yes (1)*	*F*
Mary	*1*	*no (-1)*	*no (-1)*	*no (-1)*	*yes (1)*	*F*
Peter	*1*	*yes (1)*	*no (-1)*	*no (-1)*	*yes (1)*	*A*

Μετατροπή των δειγμάτων $x_1, \ldots, x_n$ σε επαυξημένα δείγματα $y_1, \ldots, y_n$ προσθέτοντας μια νέα διάσταση με τιμή 1

Παράδειγμα



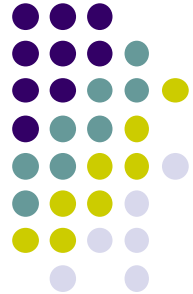
	features					grade
name	*extra*	*good attendance?*	*tall?*	*sleeps in class?*	*chews gum?*	
Jane	*1*	*yes (1)*	*yes (1)*	*no (-1)*	*no (-1)*	*A*
Steve	*1*	*yes (1)*	*yes (1)*	*yes (1)*	*yes (1)*	*F*
Mary	*1*	*no (-1)*	*no (-1)*	*no (-1)*	*yes (1)*	*F*
Peter	*1*	*yes (1)*	*no (-1)*	*no (-1)*	*yes (1)*	*A*

- Αντικατάσταση όλων των δειγμάτων της κλάσης c_2 από τα αρνητικά τους

$$\boldsymbol{y}_i \rightarrow -\boldsymbol{y}_i \qquad \forall \boldsymbol{y}_i \in \boldsymbol{c}_2$$

Αναζήτηση διανύσματος βαρών ***a,*** ώστε $\boldsymbol{a}^t \boldsymbol{y}_i > 0 \qquad \forall \boldsymbol{y}_i$

Παράδειγμα



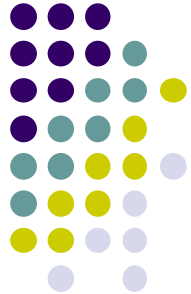
	features					grade
name	*extra*	*good attendance?*	*tall?*	*sleeps in class?*	*chews gum?*	
Jane	*1*	*yes (1)*	*yes (1)*	*no (-1)*	*no (-1)*	*A*
Steve	*-1*	*yes (-1)*	*yes (-1)*	*yes (-1)*	*yes (-1)*	*F*
Mary	*-1*	*no (1)*	*no (1)*	*no (1)*	*yes (-1)*	*F*
Peter	*1*	*yes (1)*	*no (-1)*	*no (-1)*	*yes (1)*	*A*

- Αντικατάσταση όλων των δειγμάτων της κλάσης c_2 από τα αρνητικά τους

$$\boldsymbol{y}_i \rightarrow -\boldsymbol{y}_i \qquad \forall \boldsymbol{y}_i \in \boldsymbol{c}_2$$

Αναζήτηση διανύσματος βαρών ***a,*** ώστε $\boldsymbol{a}^t \boldsymbol{y}_i > 0 \qquad \forall \boldsymbol{y}_i$

Παράδειγμα

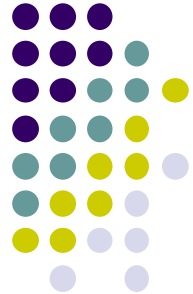


	features					grade
name	***extra***	***good attendance?***	***tall?***	***sleeps in class?***	***chews gum?***	
Jane	*1*	*yes (1)*	*yes (1)*	*no (-1)*	*no (-1)*	*A*
Steve	*-1*	*yes (-1)*	*yes (-1)*	*yes (-1)*	*yes (-1)*	*F*
Mary	*-1*	*no (1)*	*no (1)*	*no (1)*	*yes (-1)*	*F*
Peter	*1*	*yes (1)*	*no (-1)*	*no (-1)*	*yes (1)*	*A*

- Το δείγμα ταξινομείται λάθος όταν $a^t y_i = \sum_{k=0}^{4} a_k y_i^{(k)} < 0$
- Κανόνας ενημέρωσης single sample $a^{(k+1)} = a^{(k)} + \eta^{(k)} y_M$

Θέτουμε σταθερή παράμετρο μαθήσης $\eta^{(k)} = 1$: $\boxed{a^{(k+1)} = a^{(k)} + y_M}$

Παράδειγμα



Θέτουμε ίσα αρχικά βάρη $\boldsymbol{a}^{(1)}$ = [0.25. 0.25, 0.25, 0.25, 0.25]

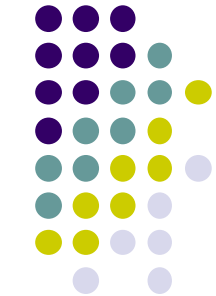
Επισκεπτόμαστε διαδοχικά όλα τα δείγματα, τροποποιώντας τα βάρη μετά από την εύρεση λάθος ταξινομημένου δείγματος

name	$\boldsymbol{a}^t\boldsymbol{y}$	*misclassified?*
Jane	*0.25*1+0.25*1+0.25*1+0.25*(-1)+0.25*(-1) >0*	*no*
Steve	*0.25*(-1)+0.25*(-1)+0.25*(-1)+0.25*(-1)+0.25*(-1)<0*	*yes*

Νέα βάρη:

$\boldsymbol{a}^{(2)}$ = ?

Παράδειγμα



Θέτουμε ίσα αρχικά βάρη $\boldsymbol{a}^{(1)}=[0.25, 0.25, 0.25, 0.25]$

Επισκεπτόμαστε διαδοχικά όλα τα δείγματα, τροποποιώντας τα βάρη μετά από την εύρεση λάθος ταξινομημένου δείγματος

name	$\boldsymbol{a}^t\boldsymbol{y}$	*misclassified?*
Jane	*0.25*1+0.25*1+0.25*1+0.25*(-1)+0.25*(-1) >0*	*no*
Steve	*0.25*(-1)+0.25*(-1)+0.25*(-1)+0.25*(-1)+0.25*(-1)<0*	*yes*

Νέα βάρη:

$$\boldsymbol{a}^{(2)} = \boldsymbol{a}^{(1)} + \boldsymbol{y}_M = [0.25\ \ 0.25\ \ 0.25\ \ 0.25\ \ 0.25] + [-1\ \ -1\ \ -1\ \ -1\ \ -1] = [-0.75\ \ -0.75\ \ -0.75\ \ -0.75\ \ -0.75]$$

Παράδειγμα

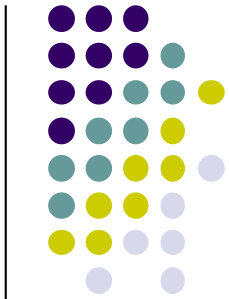
$$a^{(2)} = [-0.75 \ -0.75 \ -0.75 \ -0.75 \ -0.75]$$

name	$a^t y$	*misclassified?*
Mary	*-0.75*(-1)-0.75*1 -0.75 *1 -0.75 *1 -0.75*(-1) <0*	*yes*

Νέα βάρη:

$a^{(3)} =$?

Παράδειγμα



$$a^{(2)} = [-0.75 \ -0.75 \ -0.75 \ -0.75 \ -0.75]$$

name	$a^t y$	*misclassified?*
Mary	*-0.75*(-1)-0.75*1 -0.75 *1 -0.75 *1 -0.75*(-1) <0*	*yes*

Νέα βάρη:

$$a^{(3)} = a^{(2)} + y_M = [-0.75 \ -0.75 \ -0.75 \ -0.75 \ -0.75] +$$
$$+ [-1 \quad 1 \quad 1 \quad 1 \quad -1] =$$
$$= [-1.75 \quad 0.25 \quad 0.25 \quad 0.25 \quad -1.75]$$

Παράδειγμα

$$a^{(3)} = [-1.75 \quad 0.25 \quad 0.25 \quad 0.25 \quad -1.75]$$

name	$a^t y$	*misclassified?*
Peter	*-1.75 *1 +0.25* 1+0.25* (-1) +0.25 *(-1)-1.75*1 <0*	*yes*

Νέα βάρη:

$$a^{(4)} = a^{(3)} + y_M = [-1.75 \quad 0.25 \quad 0.25 \quad 0.25 \quad -1.75] +$$
$$+ [1 \quad 1 \quad -1 \quad -1 \quad 1] =$$
$$= [-0.75 \quad 1.25 \quad -0.75 \quad -0.75 \quad -0.75]$$

Παράδειγμα

$$a^{(4)} = [-0.75 \quad 1.25 \quad -0.75 \quad -0.75 \quad -0.75]$$

name	$a^t y$	*misclassified?*
Jane	*-0.75 *1 +1.25*1 -0.75*1 -0.75 *(-1) -0.75 *(-1)+0*	*no*
Steve	*-0.75*(-1)+1.25*(-1) -0.75*(-1) -0.75*(-1)-0.75*(-1)>0*	*no*
Mary	*-0.75 *(-1)+1.25*1-0.75*1 -0.75 *1 –0.75*(-1) >0*	*no*
Peter	*-0.75 *1+ 1.25*1-0.75* (-1)-0.75* (-1) -0.75 *1 >0*	*no*

Οπότε η διακρίνουσα συνάρτηση είναι:

Παράδειγμα

$$a^{(4)} = [-0.75 \quad 1.25 \quad -0.75 \quad -0.75 \quad -0.75]$$

name	$a^t y$	*misclassified?*
Jane	*-0.75 *1 +1.25*1 -0.75*1 -0.75 *(-1) -0.75 *(-1)+0*	*no*
Steve	*-0.75*(-1)+1.25*(-1) -0.75*(-1) -0.75*(-1)-0.75*(-1)>0*	*no*
Mary	*-0.75 *(-1)+1.25*1-0.75*1 -0.75 *1 –0.75*(-1) >0*	*no*
Peter	*-0.75 *1+ 1.25*1-0.75* (-1)-0.75* (-1) -0.75 *1 >0*	*no*

Οπότε η διακρίνουσα συνάρτηση είναι:

$$g(y) = -0.75 * y^{(0)} + 1.25 * y^{(1)} - 0.75 * y^{(2)} - 0.75 * y^{(3)} - 0.75 * y^{(4)}$$

Μετατρέποντας πάλι στα αρχικά χαρακτηριστικά ***x:***

Παράδειγμα

$$a^{(4)} = [-0.75 \;\; 1.25 \;\; -0.75 \;\; -0.75 \;\; -0.75]$$

name	$a^t y$	*misclassified?*
Jane	*-0.75 *1 +1.25*1 -0.75*1 -0.75 *(-1) -0.75 *(-1)+0*	*no*
Steve	*-0.75*(-1)+1.25*(-1) -0.75*(-1) -0.75*(-1)-0.75*(-1)>0*	*no*
Mary	*-0.75 *(-1)+1.25*1-0.75*1 -0.75 *1 –0.75*(-1) >0*	*no*
Peter	*-0.75 *1+ 1.25*1-0.75* (-1)-0.75* (-1) -0.75 *1 >0*	*no*

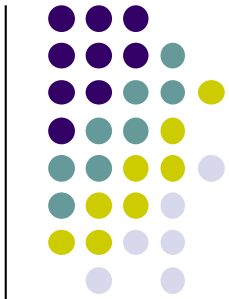
Οπότε η διακρίνουσα συνάρτηση είναι:

$$g(y) = -0.75 * y^{(0)} + 1.25 * y^{(1)} - 0.75 * y^{(2)} - 0.75 * y^{(3)} - 0.75 * y^{(4)}$$

Μετατρέποντας πάλι στα αρχικά χαρακτηριστικά ***x:***

$$g(x) = 1.25 * x^{(1)} - 0.75 * x^{(2)} - 0.75 * x^{(3)} - 0.75 * x^{(4)} - 0.75$$

Παράδειγμα



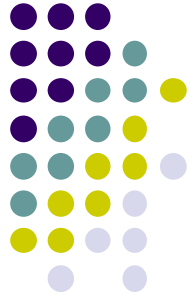
- Μετατρέποντας πάλι στα αρχικά χαρακτηριστικά $x^{(i)}$:

$$1.25 * x^{(1)} - 0.75 * x^{(2)} - 0.75 * x^{(3)} - 0.75 * x^{(4)} > 0.75 \Rightarrow$$
$$1.25 * x^{(1)} - 0.75 * x^{(2)} - 0.75 * x^{(3)} - 0.75 * x^{(4)} < 0.75 \Rightarrow$$

?

good attendance — tall — sleeps in class — chews gum

Παράδειγμα



- Μετατρέποντας πάλι στα αρχικά χαρακτηριστικά $x^{(i)}$

$$1.25 * x^{(1)} - 0.75 * x^{(2)} - 0.75 * x^{(3)} - 0.75 * x^{(4)} > 0.75 \Rightarrow$$ βαθμός A

$$1.25 * x^{(1)} - 0.75 * x^{(2)} - 0.75 * x^{(3)} - 0.75 * x^{(4)} < 0.75 \Rightarrow$$ βαθμός F

good attendance — tall — sleeps in class — chews gum

- Αυτό είναι μόνο ένα από τα πιθανά διανύσματα λύσης
- Αν ξεκινούσαμε με βάρη $a^{(1)}=[0, 0.5, 0.5, 0, 0]$,
  Η λύση θα ήταν $[-1, 1.5, -0.5, -1, -1]$

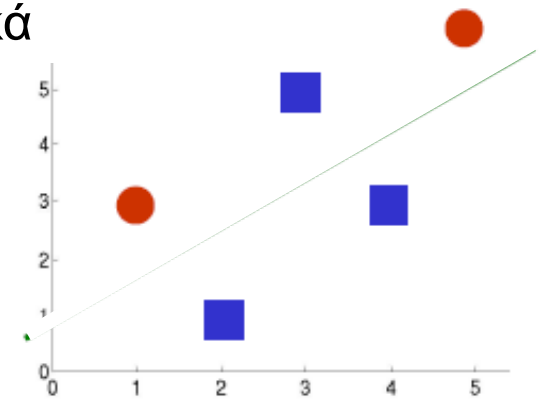
$$1.5 * x^{(1)} - 0.5 * x^{(2)} - x^{(3)} - x^{(4)} > 1 \Rightarrow$$ βαθμός A

$$1.5 * x^{(1)} - 0.5 * x^{(2)} - x^{(3)} - x^{(4)} < 1 \Rightarrow$$ βαθμός F

  - Σε αυτήν την λύση το χαρακτηριστικό «being tall» είναι το λιγότερο σημαντικό χαρακτηριστικό

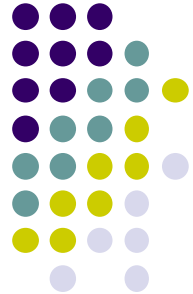
Παράδειγμα - Μη γραμμικά διαχωρίσιμα δεδομένα

- Υποθέστε ότι έχουμε 2 χαρακτηριστικά και τα δείγματα είναι:
  - Class 1: [2,1], [4,3], [3,5]
  - Class 2: [1,3] και [5,6]

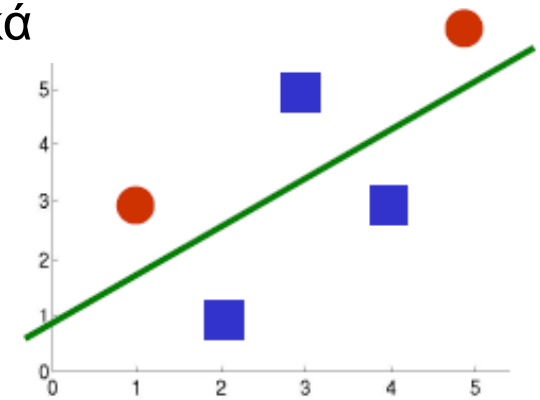


- Τα δείγματα αυτά δεν διαχωρίζονται με μια ευθεία
- Θέλουμε προσεγγιστική λύση μέσω μιας ευθείας - μια καλή επιλογή είναι ?

Παράδειγμα - Μη γραμμικά διαχωρίσιμα δεδομένα

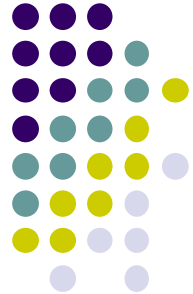


- Υποθέστε ότι έχουμε 2 χαρακτηριστικά και τα δείγματα είναι:
  - Class 1: [2,1], [4,3], [3,5]
  - Class 2: [1,3] και [5,6]
- Τα δείγματα αυτά δεν διαχωρίζονται με μια ευθεία

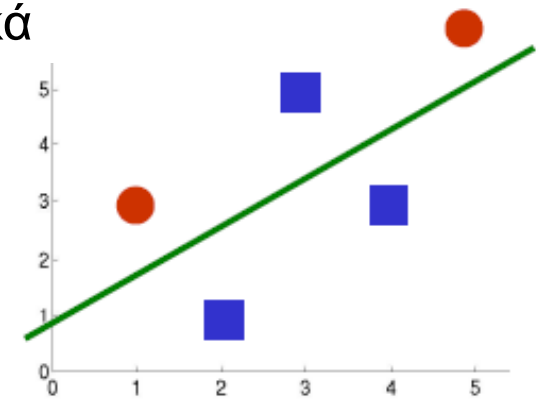


- Θέλουμε προσεγγιστική λύση μέσω μιας ευθείας - μια καλή επιλογή είναι η πράσινη ευθεία
  - Κάποια δείγματα περιέχουν θόρυβο, οπότε δεν είναι πρόβλημα να βρίσκονται στην λάθος μεριά της ευθείας

Παράδειγμα - Μη γραμμικά διαχωρίσιμα δεδομένα



- Υποθέστε ότι έχουμε 2 χαρακτηριστικά και τα δείγματα είναι:
  - Class 1: [2,1], [4,3], [3,5]
  - Class 2: [1,3] και [5,6]
- Τα δείγματα αυτά δεν διαχωρίζονται με μια ευθεία

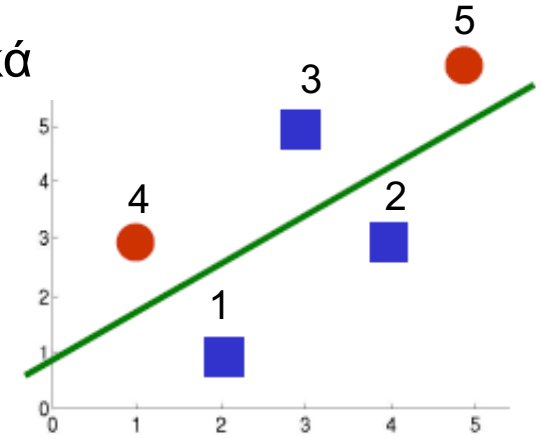


- Θέλουμε προσεγγιστική λύση μέσω μιας ευθείας - μια καλή επιλογή είναι η πράσινη ευθεία
  - Κάποια δείγματα περιέχουν θόρυβο, οπότε δεν είναι πρόβλημα να βρίσκονται στην λάθος μεριά της ευθείας

Βρίσκουμε y_1, y_2, y_3, y_4 προσθέτοντας ένα χαρακτηριστικό και κανονικοποιώντας

Παράδειγμα - Μη γραμμικά διαχωρίσιμα δεδομένα

- Υποθέστε ότι έχουμε 2 χαρακτηριστικά και τα δείγματα είναι:
  - Class 1: [2,1], [4,3], [3,5]
  - Class 2: [1,3] κ [5,6]
- Τα δείγματα αυτά δεν διαχωρίζονται με μια ευθεία



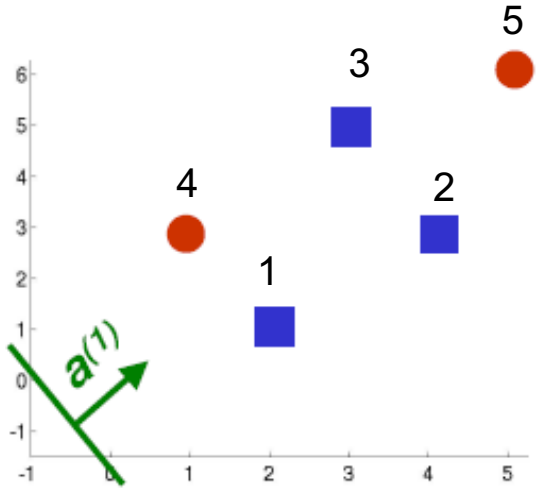
- Θέλουμε προσεγγιστική λύση μέσω μιας ευθείας - μια καλή επιλογή είναι η πράσινη ευθεία
  - Κάποια δείγματα περιέχουν θόρυβο, οπότε δεν είναι πρόβλημα να βρίσκονται στην λάθος μεριά της ευθείας

Βρίσκουμε y_1, y_2, y_3, y_4 προσθέτοντας ένα χαρακτηριστικό

και κανονικοποιώντας $y_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$ $y_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix}$ $y_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}$ $y_4 = \begin{bmatrix} -1 \\ -1 \\ -3 \end{bmatrix}$ $y_5 = \begin{bmatrix} -1 \\ -5 \\ -6 \end{bmatrix}$

Παράδειγμα - Μη γραμμικά διαχωρίσιμα δεδομένα

- Εφαρμόζουμε τον αλγόριθμο Perceptron Single Sample
- Αρχικά ίσα βάρη $a^{(1)} = [1\ 1\ 1]$
  - Η ευθεία $x^{(1)} + x^{(2)} + 1 = 0$
- στ. παραμετρος μάθ. $\eta = 1$



$$a^{(k+1)} = a^{(k)} + y_M$$

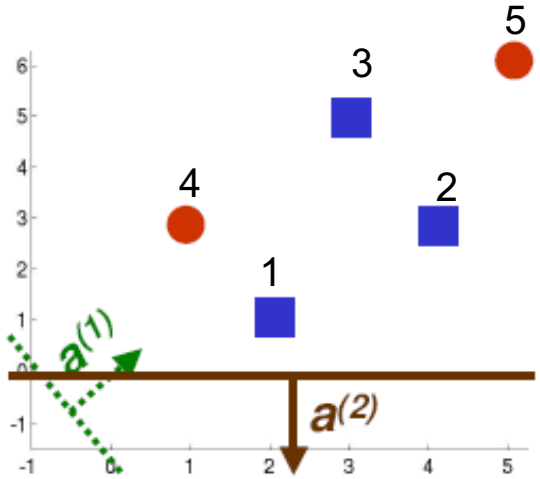
$$y_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad y_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} \quad y_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \quad y_4 = \begin{bmatrix} -1 \\ -1 \\ -3 \end{bmatrix} \quad y_5 = \begin{bmatrix} -1 \\ -5 \\ -6 \end{bmatrix}$$

- $y^t_1 a^{(1)} = [1\ 1\ 1]^*[1\ 2\ 1]^t > 0$ ✓
- $y^t_2 a^{(1)} = [1\ 1\ 1]^*[1\ 4\ 3]^t > 0$ ✓
- $y^t_3 a^{(1)} = [1\ 1\ 1]^*[1\ 3\ 5]^t > 0$ ✓

Παράδειγμα - Μη γραμμικά διαχωρίσιμα δεδομένα

$$a^{(1)} = [1\ 1\ 1] \qquad a^{(k+1)} = a^{(k)} + y_M$$

$$y_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad y_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} \quad y_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \quad y_4 = \begin{bmatrix} -1 \\ -1 \\ -3 \end{bmatrix} \quad y_5 = \begin{bmatrix} -1 \\ -5 \\ -6 \end{bmatrix}$$



- $y^t_4 a^{(1)} = [1\ 1\ 1]*[-1\ -1\ -3]^t = -5 < 0$

$$a^{(2)} = a^{(1)} + y_M = [1\ 1\ 1] + [-1\ -1\ -3] = [0\ 0\ -2]$$

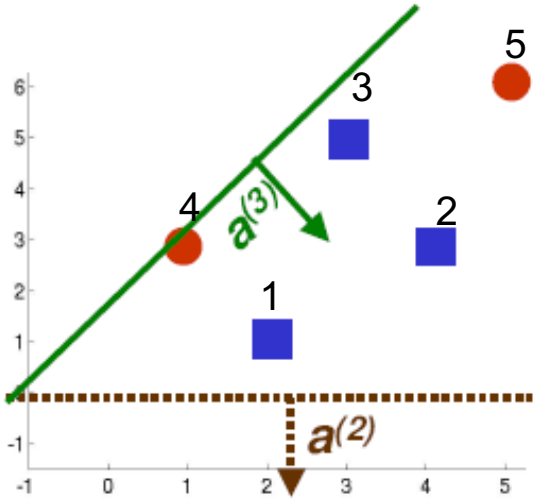
- $y^t_5\, a^{(2)} = [0\ 0\ -2]*[-1\ -5\ -6]^t = 12 > 0$ ✔
- $y^t_1\, a^{(2)} = [0\ 0\ -2]*[1\ 2\ 1]^t < 0$

$$a^{(3)} = a^{(2)} + y_M = [0\ 0\ -2] + [1\ 2\ 1] = [1\ 2\ -1]$$

Παράδειγμα - Μη γραμμικά διαχωρίσιμα δεδομένα

$$a^{(3)} = [1\ 2\ -1] \qquad a^{(k+1)} = a^{(k)} + y_M$$

$$y_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad y_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} \quad y_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \quad y_4 = \begin{bmatrix} -1 \\ -1 \\ -3 \end{bmatrix} \quad y_5 = \begin{bmatrix} -1 \\ -5 \\ -6 \end{bmatrix}$$



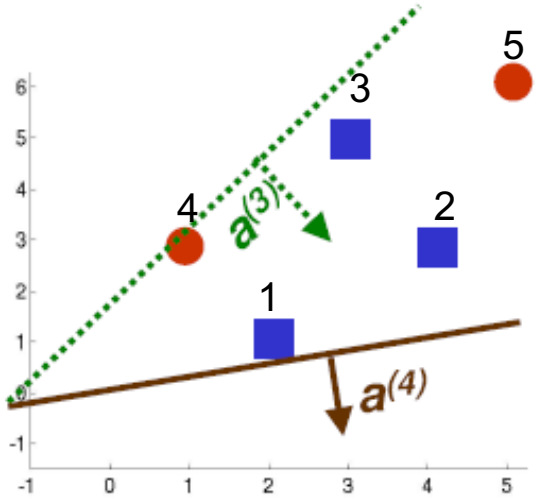
- $y^t_2\, a^{(3)} = [1\ 4\ 3]*[1\ 2\ -1]^t = 6 > 0$ ✔
- $y^t_3\, a^{(3)} = [1\ 3\ 5]*[1\ 2\ -1]^t > 0$ ✔
- $y^t_4\, a^{(3)} = [-1\ -1\ -3]*[1\ 2\ -1]^t = 0$

$$a^{(4)} = a^{(3)} + y_M = [1\ 2\ -1] + [-1\ -1\ -3] = [0\ 1\ -4]$$

Παράδειγμα - Μη γραμμικά διαχωρίσιμα δεδομένα

$$a^{(4)} = [0\ 1\ -4] \qquad a^{(k+1)} = a^{(k)} + y_M$$

$$y_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad y_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} \quad y_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \quad y_4 = \begin{bmatrix} -1 \\ -1 \\ -3 \end{bmatrix} \quad y_5 = \begin{bmatrix} -1 \\ -5 \\ -6 \end{bmatrix}$$



- $y^t_2\, a^{(3)} = [1\ 4\ 3]^*[1\ 2\ -1]^t = 6 > 0$ ✔
- $y^t_3\, a^{(3)} = [1\ 3\ 5]^*[1\ 2\ -1]^t > 0$ ✔
- $y^t_4\, a^{(3)} = [-1\ -1\ -3]^*[1\ 2\ -1]^t = 0$

$$a^{(4)} = a^{(3)} + y_M = [1\ 2\ -1] + [-1\ -1\ -3] = [0\ 1\ -4]$$

Παράδειγμα - Μη γραμμικά διαχωρίσιμα δεδομένα

- Μπορούμε να το συνεχίσουμε για πάντα

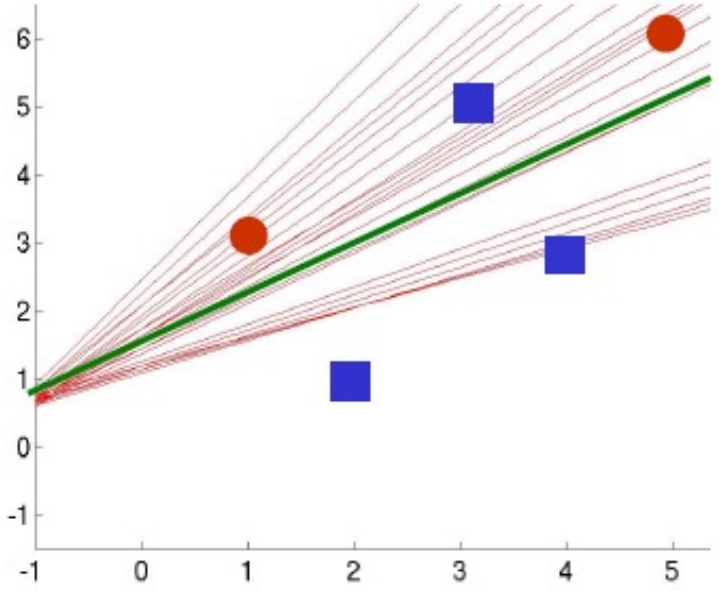
Δεν υπάρχει διάνύσμα λύσης a το οποίο να ικανοποιεί για κάθε i τη σχέση:

$$a^t y_i = \sum_{k=0}^{5} a_k y_i^{(k)} > 0$$

Πρέπει να σταματήσουμε αλλά σε «καλό» σημείο:

•Οι λύσεις για επαναλήψεις 900 ως 915:
Κάποιες είναι καλές, κάποιες όχι

•Πώς θα σταματήσουμε σε καλή λύση?



Σύγκλιση

- Αν οι κλάσεις είναι γραμμικά διαχωρίσιμες, χρησιμοποιείστε σταθερή παράμετρο μαθησης, $\eta^{(k)} = c$

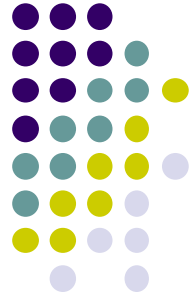
• Και ο σωρηδόν και ο single sample κανόνας συγκλίνουν στη σωστή λύση (η οποία μπορεί να βρίσκεται οπουδήποτε στην περιοχή λύσης)

- Αν οι κλάσεις δεν είναι γραμμικά διαχωρίσιμες:

•Ο αλγόριθμος δεν σταματάει, αλλα ψάχνει για λύση η οποία δεν υπάρχει
•Χρησιμοποιώντας κατάλληλη παράμετρο μάθησης, μπορεί να επιτευχθεί πάντα σύγκλιση: $\eta^{(k)} \to 0$ όσο $k \to \infty$

•Π.χ. Αντίστροφος ρυθμός μάθησης: $\eta^{(k)} = \frac{\eta^{(1)}}{k}$
•Για αντίστροφο γραμμικό ρυθμό μπορεί να αποδειχθεί η σύγκλιση για την γραμμική περίπτωση
•Δεν υπάρχει εγγύηση ότι ο αλγόριθμος σταμάτησε σε καλό σημείο, αλλά η επιλογή του αντίστροφου ρυθμού μάθησης δικαιολογείται για διάφορους λόγους

Κριτήριο Perceptron και αλγόριθμος Απότομης Καθόδου



- Γραμμικά διαχωρίσιμα δεδομένα

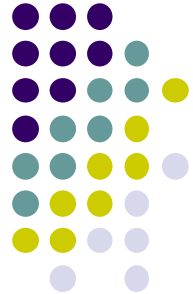
  Κριτήριο Perceptron με αλγόριθμο Απότομης Καθόδου δίνουν καλά αποτελέσματα

- Γραμμικά μη- διαχωρίσιμα δεδομένα

  Πρέπει να βρεθεί τρόπος να σταματήσει ο αλγόριθμος perceptron σε «καλό» σημείο, το οποίο μπορεί να είναι δύσκολο

Σωρηδόν αλγόριθμος	Single sample αλγόριθμος
Πιο ομαλή κλίση επειδή χρησιμοποιούνται όλα τα δείγματα	Ευκολότερη η ανάλυση του Επικεντρώνεται περισσότερο από ότι θα ήταν επιθυμητό σε μεμονομένα, πιθανώς με θόρυβο δείγματα

Παράδειγμα υλοποίησης

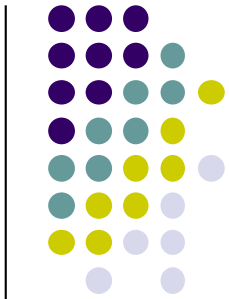


$f(x)=x^4-3x^3+2$ $f'(x)=4x^3-9x^2$

```
#include <stdio.h>
#include <stdlib.h>
#include <math.h>

int main ()
{
      // From calculation, we expect that the
local minimum occurs at x=9/4
      // The algorithm starts at x=6

      double xOld = 0;
      double xNew = 6;
      double eps = 0.01; // step size
      double precision = 0.00001;
      while (fabs(xNew - xOld) > precision)
      {
              xOld = xNew;
              xNew = xNew - eps*(4*xNew*xNew*xNew-
9*xNew*xNew);
      }

      printf ("Local minimum occurs at %lg\n",
xNew);

}
```

Βρίσκει τοπικό ελάχιστο
2.24996 σε 70 επαναλήψεις

Μέθοδοι Χαλάρωσης (Relaxation Procedures)



- Συνάρτηση κριτηρίου:

$$J_r(\mathbf{a}) = \frac{1}{2}\sum_{\mathbf{y}\in Y}\frac{\left(\mathbf{a}^t\mathbf{y}-b\right)^2}{\|\mathbf{y}\|^2}$$

- όπου *Y(a)*είναι το σύνολο των δειγμάτων για τα οποία $a^t y<b$.

- Αν το *Y(a)*είναι κενό, τότε $J_r(a)=0$.Η $J_r(a)$ δεν είναι ποτέ αρνητική και μηδενίζεται αν και μόνο αν $a^t y<b$ για όλα τα δείγματα εκπαίδευσης.

- Το διάνυσμα κλίσεων είναι:

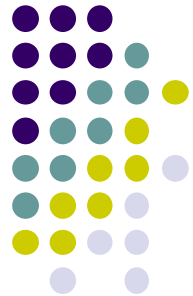
$$\nabla J_r(\mathbf{a}) = \left[\frac{\partial J_r}{\partial a_i}\right] = \sum_{y\in Y}\frac{\mathbf{a}^t\mathbf{y}-b}{\|\mathbf{y}\|^2}\mathbf{y}$$

- Αναδρομική σχέση:

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \eta(k)\sum_{y\in Y_k}\frac{b-\mathbf{a}^t\mathbf{y}}{\|\mathbf{y}\|^2}\mathbf{y}$$

Προηγούμενη προσέγγιση: αργή σύγκλιση κοντά στη λύση και επηρεάζεται από **y** με μεγάλο μέτρο

Batch Relaxation with Margin



Αλγόριθμος 6. Batch Relaxation with Margin

1 **begin initialize** **a**, margin b, $\eta(0)$, $k=0$

2 **do** $k \leftarrow (k+1) \bmod n$

3 $\mathcal{Y}_k = \{\}$

4 $j=0$

5 **do** $j \leftarrow j+1$

6 **if** $\mathbf{a}^t\mathbf{y}^k < b$, **then** append $\mathbf{y}^j$ to $\mathcal{Y}_k$

7 **until** $j=n$

8 $\mathbf{a} \leftarrow \mathbf{a} + \eta(k) \sum_{y \in \mathcal{Y}} \frac{b - \mathbf{a}^t\mathbf{y}}{\|\mathbf{y}\|^2} \mathbf{y}$

9 **until** $\mathcal{Y}_k = \{\}$

10 **return** **a**

11 **end**

Single-Sample Relaxation with Margin

Αλγόριθμος 7. Single-Sample Relaxation with Margin

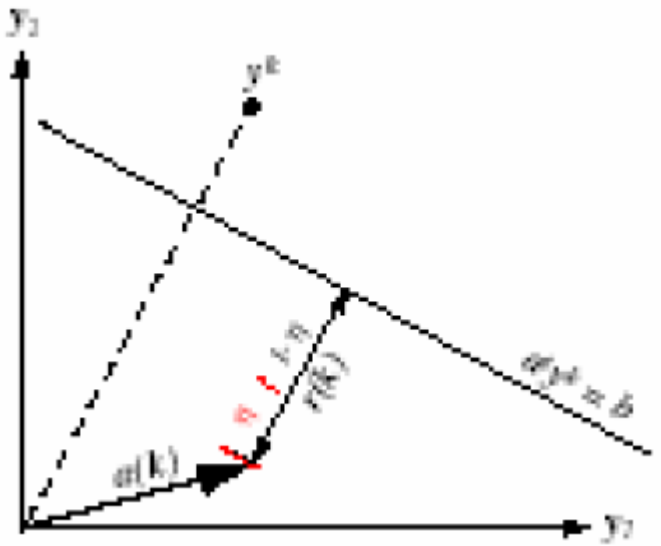
```
1 begin initialize a, margin b, η(0), k=0
2   do  k ← (k+1) mod n
3       if a^t y^k < b, then  a ← a + η(k) (b − a^t y^k) / ||y^k||^2  y^k
4   until a^t y^k > b for all y^k
5   return a
6 end
```

Σε κάθε βήμα, το διάνυσμα βαρών a(k), μετατοπίζεται προς το υπερεπίπεδο $a^t y^k=b$ κατά ένα ποσοστό, η(k),της απόστασής του, r(k),από αυτό.

η(k)<1⇨underrelaxation

η(k)>1⇨overrelaxation

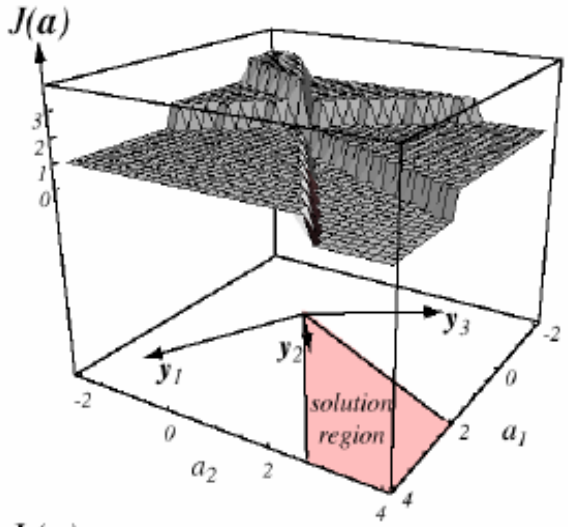
0<η(k)<2για σύγκλιση

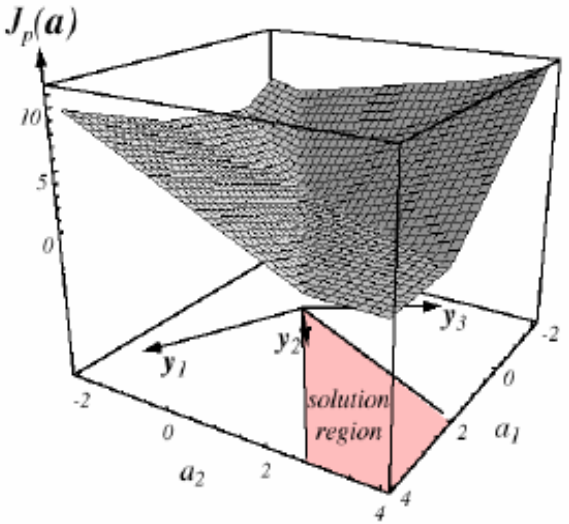


Τέσσερις Συναρτήσεις Κόστους

Πλήθος εσφαλμένων ταξινομήσεων

Bad





Κριτήριο Perceptron

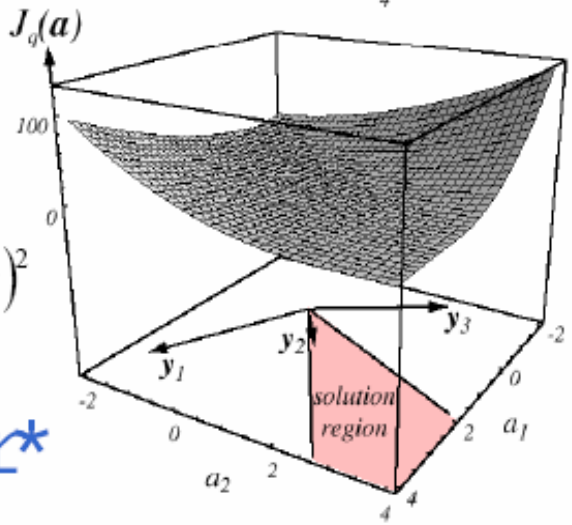
$$J_p(\mathbf{a}) = \sum_{\mathbf{y}\in\mathcal{Y}} \left(-\mathbf{a}^t\mathbf{y}\right)$$

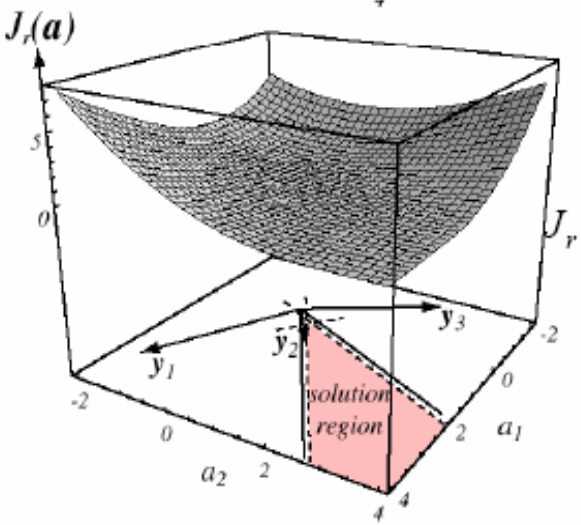
Good!

Συνολικό τετραγωνικό λάθος - Total square error (TSE)

$$J_q(\mathbf{a}) = \sum_{\mathbf{y}\in\mathcal{Y}} \left(\mathbf{a}^t\mathbf{y}\right)^2$$

Better*





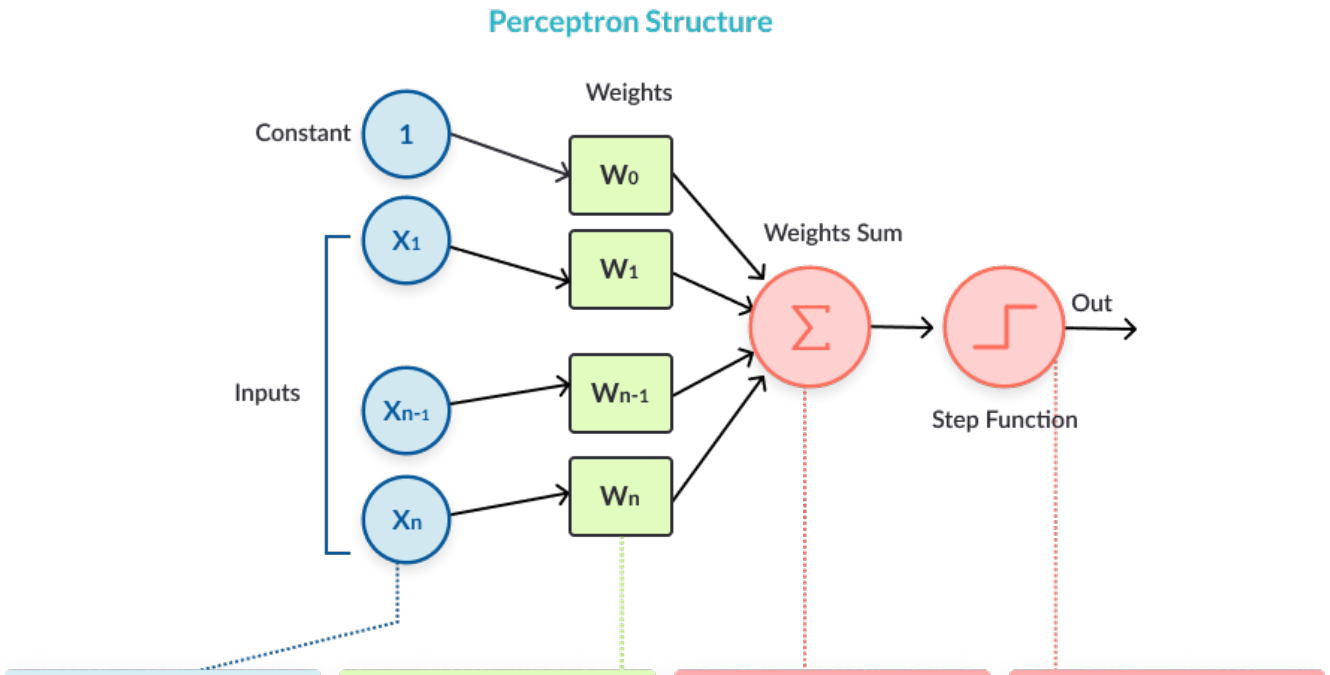
TSE with margin

$$J_r(\mathbf{a}) = \frac{1}{2}\sum_{\mathbf{y}\in\mathcal{Y}} \frac{\left(\mathbf{a}^t\mathbf{y}-b\right)^2}{\|\mathbf{y}\|}$$

Best*

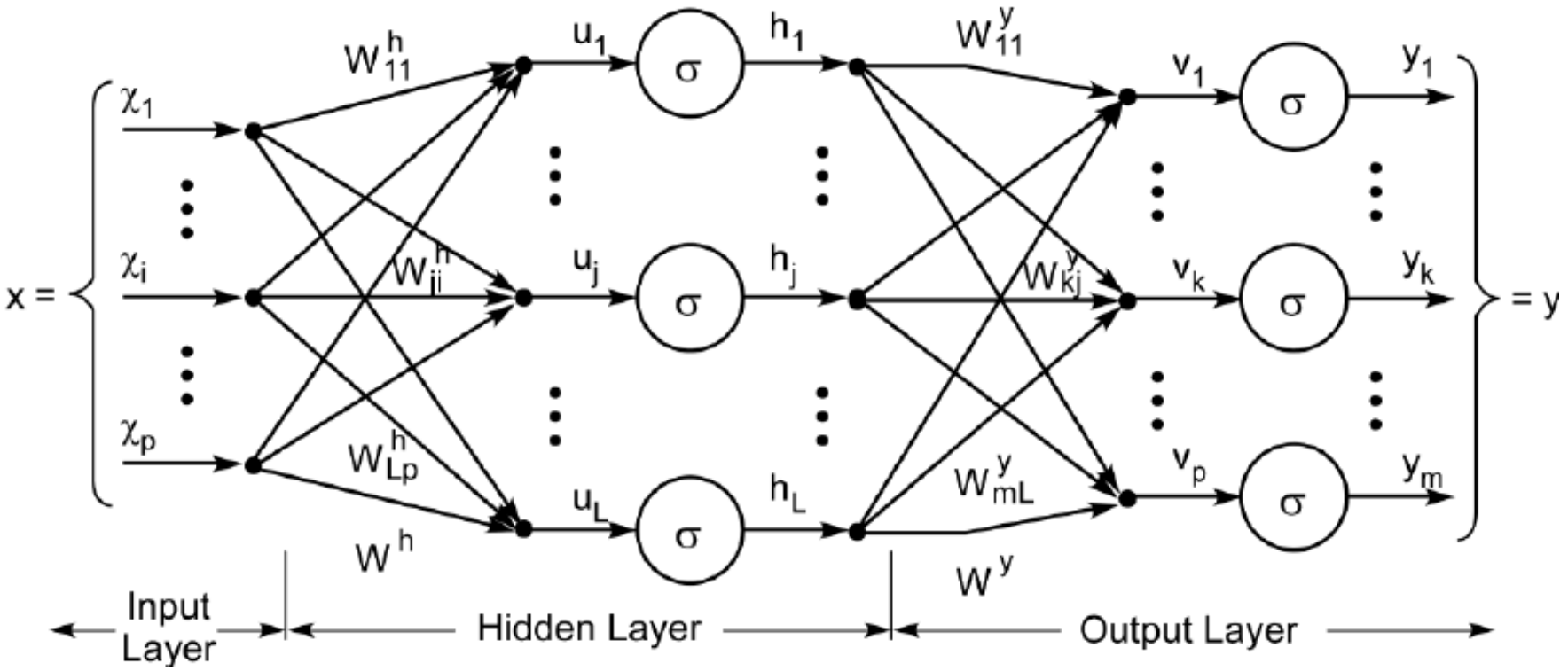
* Αλλά θα μπορούσε να έχει μεγάλο υπολογιστικό κόστος

Perceptron



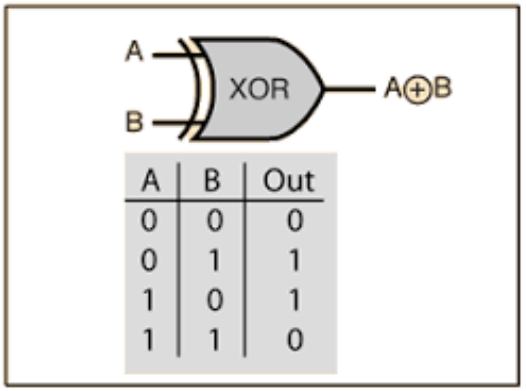
INPUT VALUES	WEIGHTS AND BIAS	WEIGHTED SUM	ACTIVATION FUNCTION
The perceptron takes real values as its inputs. For example, if the perceptron is tasked with classifying Iris flowers (an open deep learning data set), two inputs could be the length and width of the flower petals.	The weights represent the relative importance of each of the weights to the classification decision. A "bias weight" is added, and multiplied by a constant equal to 1.	The input values are multiplied by the weights and summed up, to create one aggregate value which is fed into the activation function.	The activation function generates a classification decision. For example, the Iris is classified as "Setosa" or "Versicolor".

Multilayer Perceptron

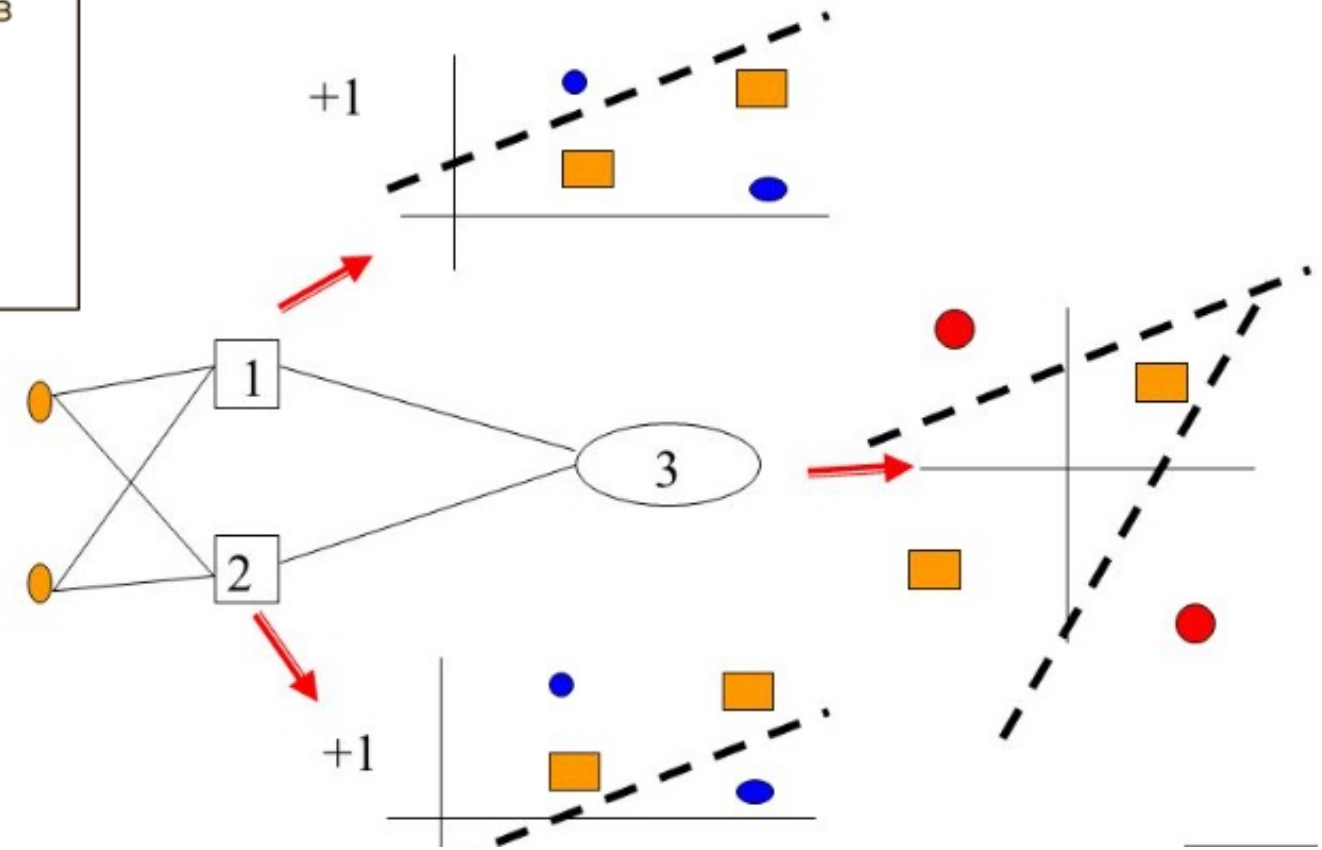


Multilayer Perceptron

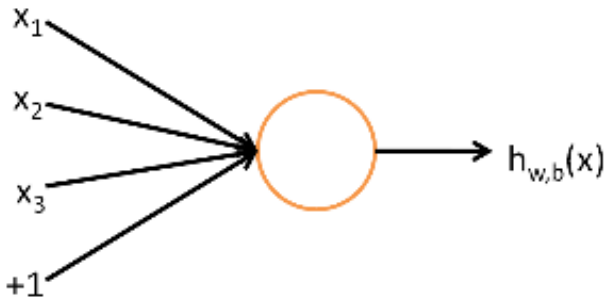
Λύση στο πρόβλημα XOR



A	B	Out
0	0	0
0	1	1
1	0	1
1	1	0



Νευρώνας



$$h_{W,b}(x) = f(W^T x) = f(\sum_{i=1}^{3} W_i x_i + b)$$

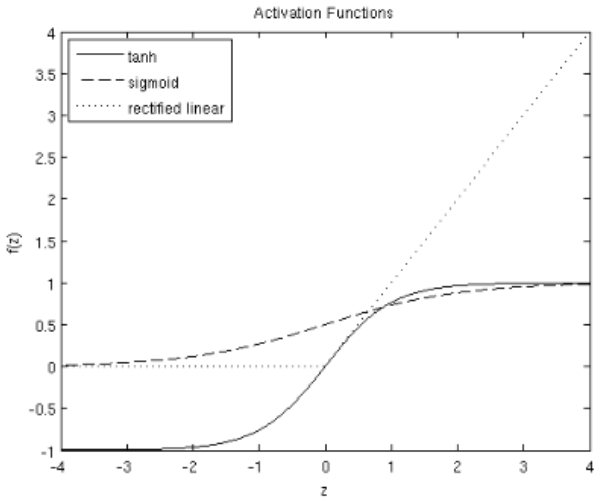
- Παραδοσιακές συναρτήσεις ενεργοποίησης (s-shaped):

$$f(z) = \frac{1}{1+\exp(-z)}$$

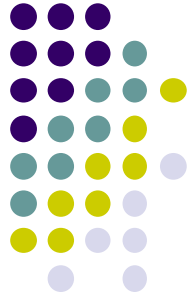
$$f(z) = \tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

- Πρόσφατα (rectified linear):

$$f(z) = \max(0, x)$$



ReLU

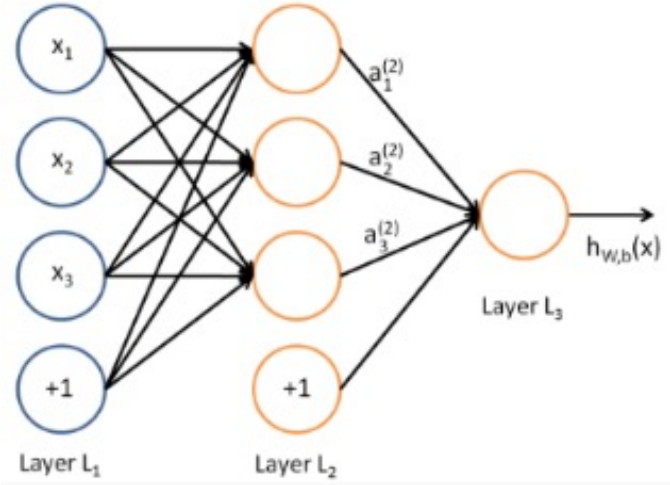


- Βιολογικά δόκιμο: μονόπλευρο, αντί για αντισυμμετρικό όπως η tanh.
- Αραιή ενεργοποίηση: για τυχαία ενεργοποίηση μόνο το 50% έχει μη μηδενική έξοδο.
- Διάδοση της παραγώγου: όχι vanishing / exploding gradient.
- Εύκολος υπολογισμός: σύγκριση πρόσθεση πολλαπλασιασμός.
- Scale-invariant: max(0,ax)=a * max(0,x)

- Μη συνεχής στο μηδέν

Νευρωνικό δίκτυο

- $a_i^{(l)}$ Η έξοδος του κόμβου i στο επίπεδο l. Με z συμβολίζουμε την αντίστοιχη είσοδο



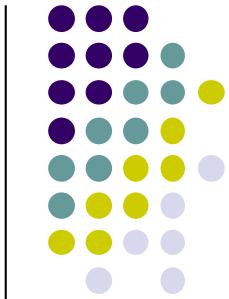
είσοδος

$$
\begin{aligned}
a_1^{(2)} &= f(W_{11}^{(1)}x_1 + W_{12}^{(1)}x_2 + W_{13}^{(1)}x_3 + b_1^{(1)}) \\
a_2^{(2)} &= f(W_{21}^{(1)}x_1 + W_{22}^{(1)}x_2 + W_{23}^{(1)}x_3 + b_2^{(1)}) \\
a_3^{(2)} &= f(W_{31}^{(1)}x_1 + W_{32}^{(1)}x_2 + W_{33}^{(1)}x_3 + b_3^{(1)}) \\
h_{W,b}(x) &= a_1^{(3)} = f(W_{11}^{(2)}a_1^{(2)} + W_{12}^{(2)}a_2^{(2)} + W_{13}^{(2)}a_3^{(2)} + b_1^{(2)})
\end{aligned}
$$

έξοδος

$$
\begin{aligned}
z^{(2)} &= W^{(1)}x + b^{(1)} \\
a^{(2)} &= f(z^{(2)}) \\
z^{(3)} &= W^{(2)}a^{(2)} + b^{(2)} \\
h_{W,b}(x) &= a^{(3)} = f(z^{(3)})
\end{aligned}
$$

Γραμμικές Διακρίνουσες Συναρτήσεις



- Στόχος:
  Η σχεδίαση γραμμικών ως προς το διάνυσμα χαρακτηριστικών x συναρτήσεων διάκρισης που ορίζουν υπερεπίπεδα ως επιφάνειες απόφασης.

- Γιατί;

Απλή μορφή, εύκολη υλοποίηση, βέλτιστες για Γκαουσσιανές σ.π.π.

- Πώς;

Διατυπώνοντας το πρόβλημα εύρεσης των παραμέτρων (βαρών) ως πρόβλημα βελτιστοποίησης μιας συνάρτησης κριτηρίου (κόστους).

- Τι είναι η συνάρτηση κριτηρίου;

Μια βαθμωτή συνάρτηση των βαρών που θα πρέπει να ελαχιστοποιηθεί, π.χ. Η πιθανότητα λάθος ταξινόμησης κατά την εκπαίδευση.

- Είναι δύσκολο να επιτευχθεί;

Ναι, γενικώς είναι δύσκολη η σχεδίαση ενός γραμμικού ταξινομητή που να ελαχιστοποιεί το ρίσκο

- Επομένως;

Χρησιμοποιούμε εναλλακτικά κριτήρια (απλές συναρτήσεις των βαρών) και επαναληπτικές μεθόδους βελτιστοποίησης (καθόδου κατά την κλίση του κριτηρίου – gradient descent).