Μηχανική Μάθηση: Μαθηματικό Υπόβαθρο

Κωνσταντίνος Καραμανής

The University of Texas at Austin & Archimedes/Athena RC

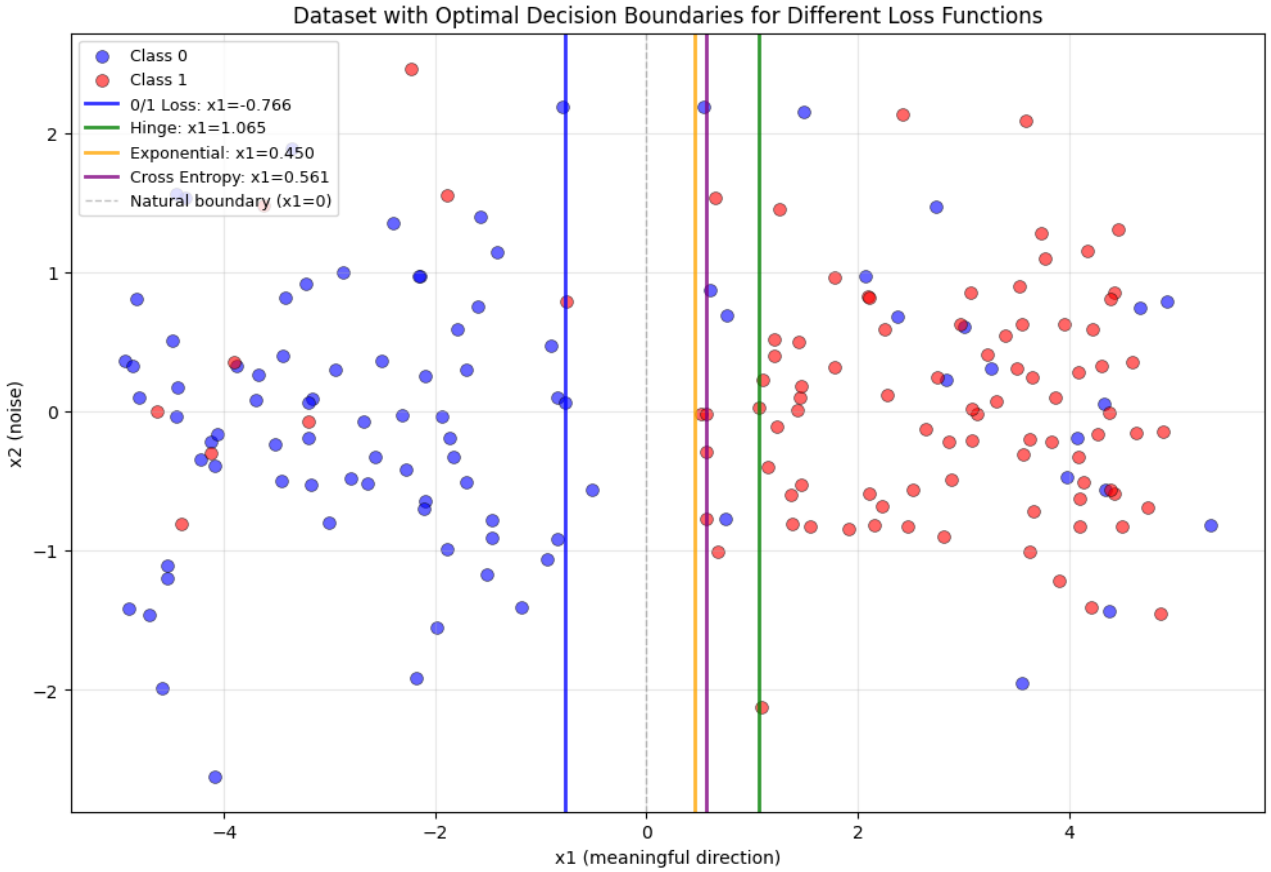
constantine@utexas.edu

https://caramanis.github.io/

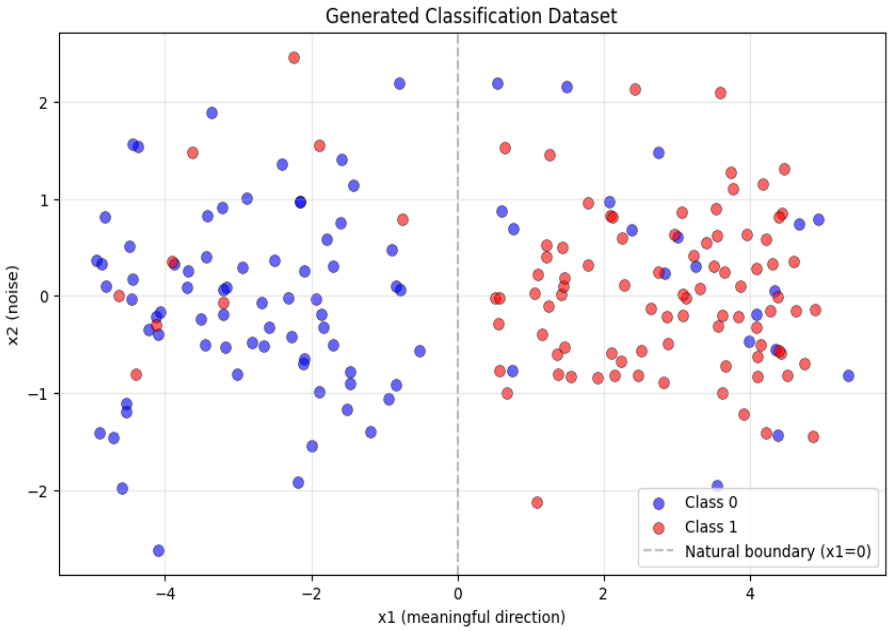
Ας θυμηθούμε τα προηγούμενα...

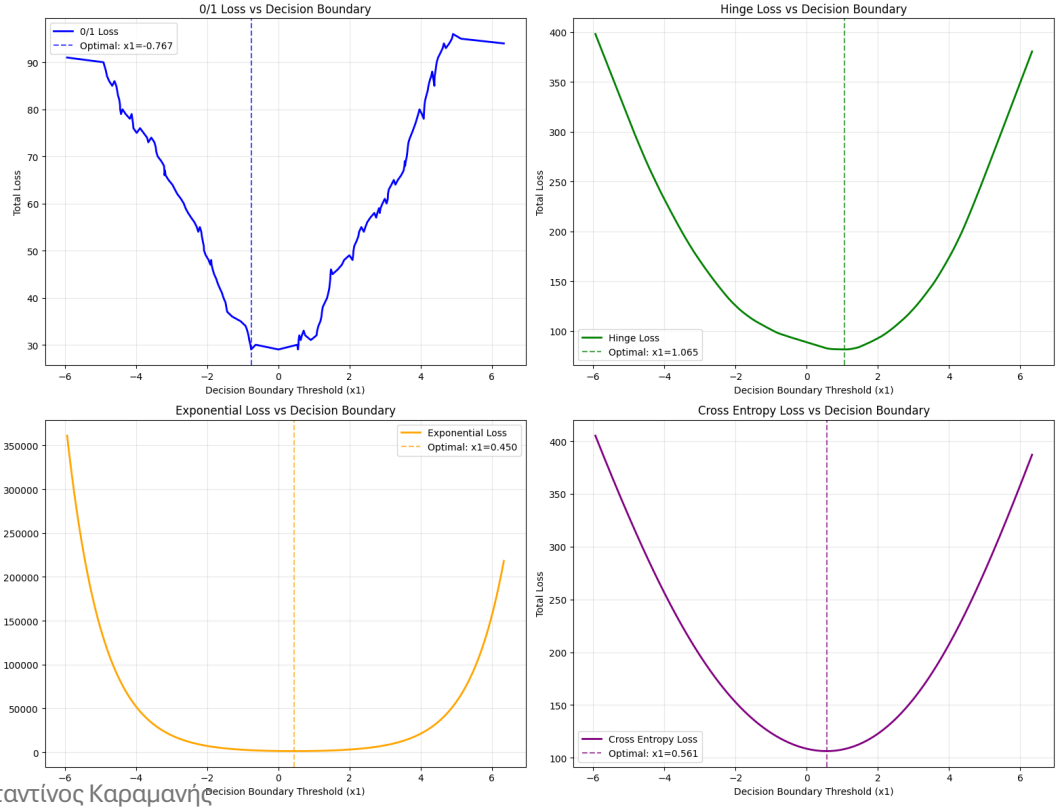
Πώς διαφέρουν οι 4 συναρτήσεις απώλειας

Όταν αλλάζουμε την συνάρτηση απώλειας, αλλάζουμε και την βέλτιστη λύση

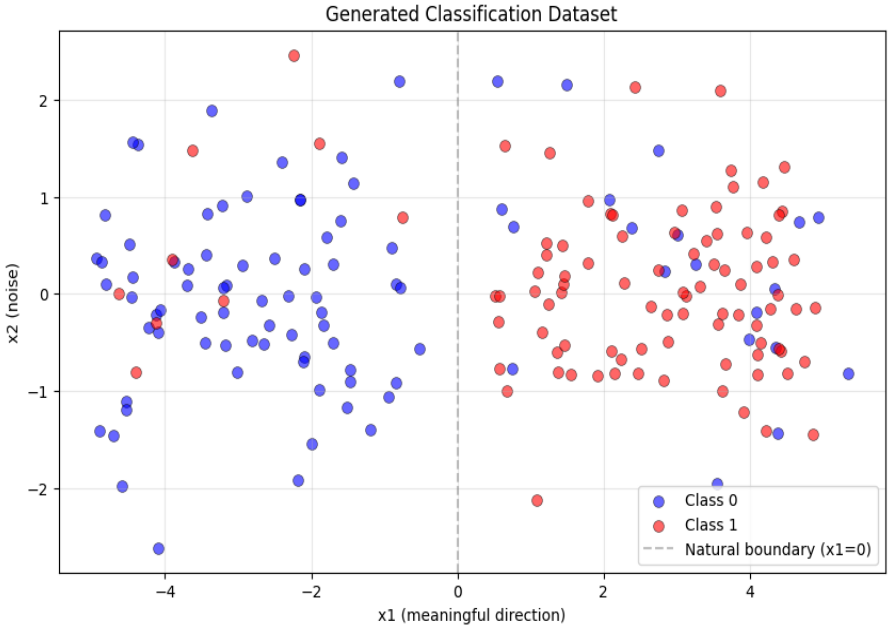


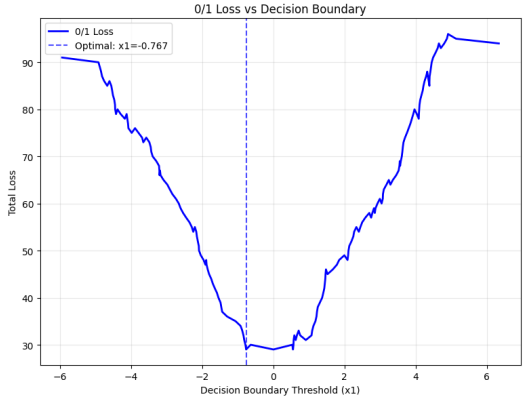
Πώς βρίσκουμε τη βέλτιστη λύση;

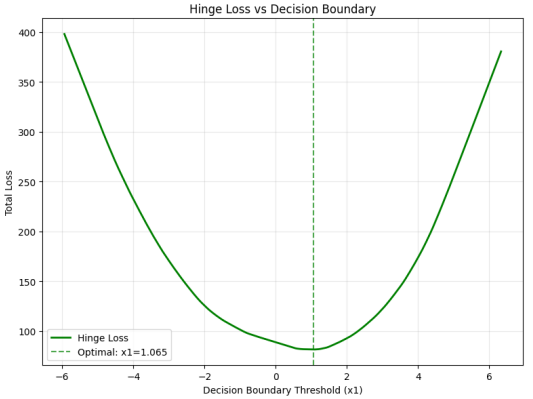


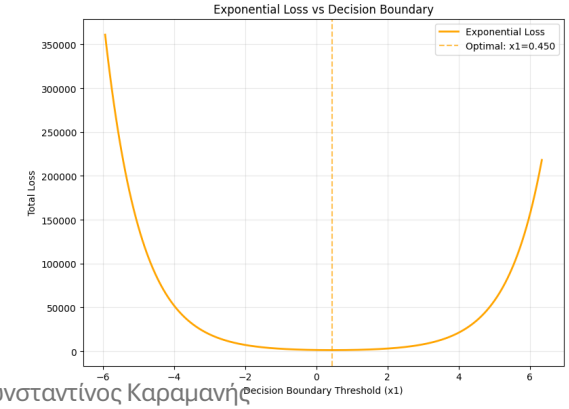


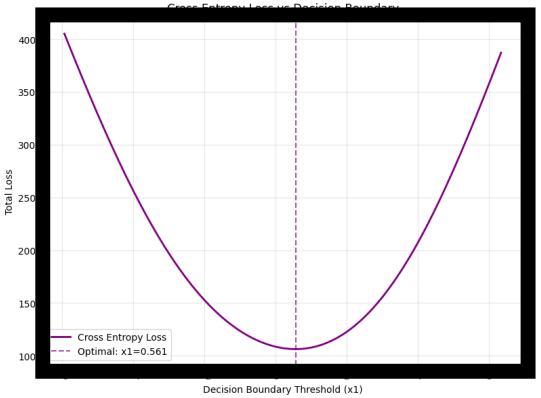
Πώς βρίσκουμε τη βέλτιστη λύση;



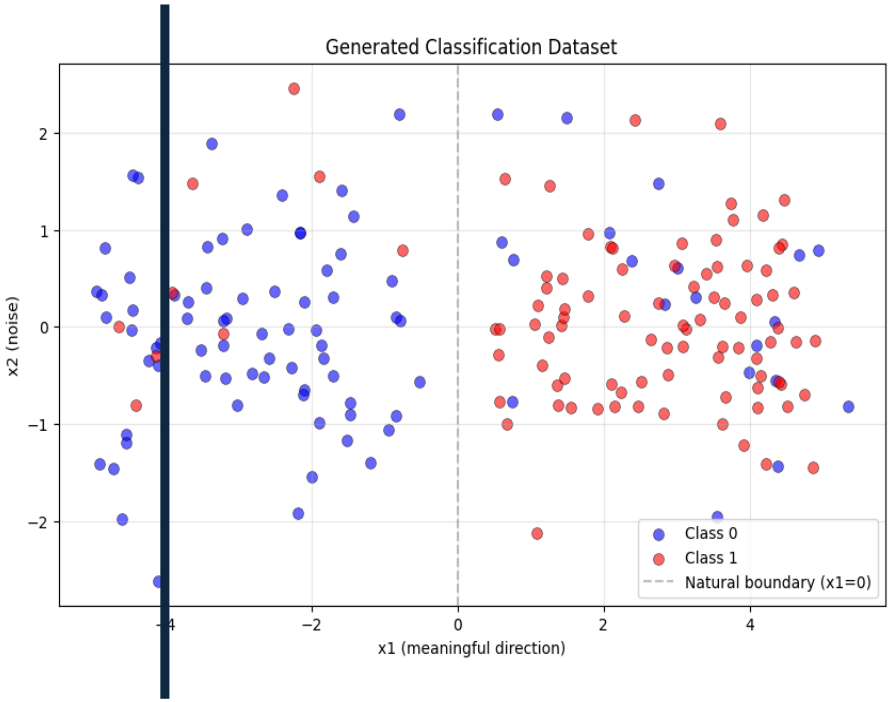


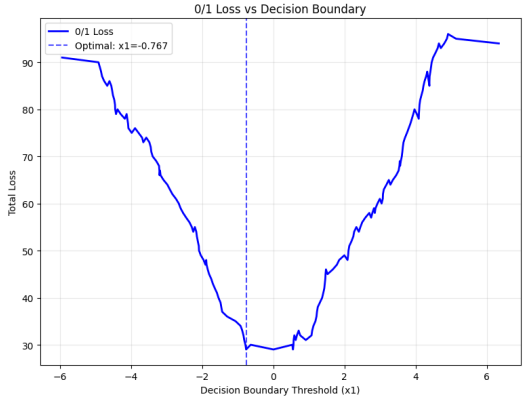


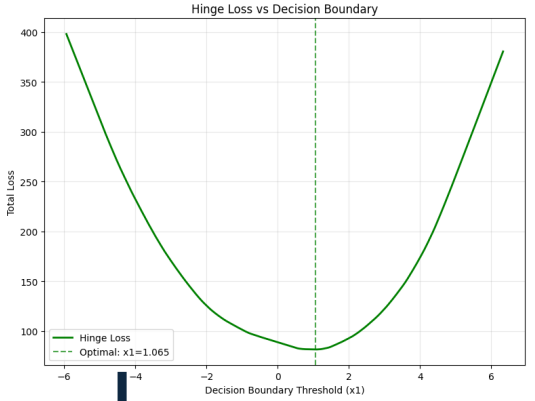


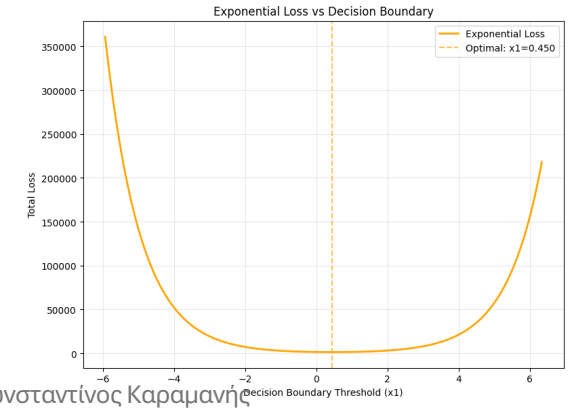


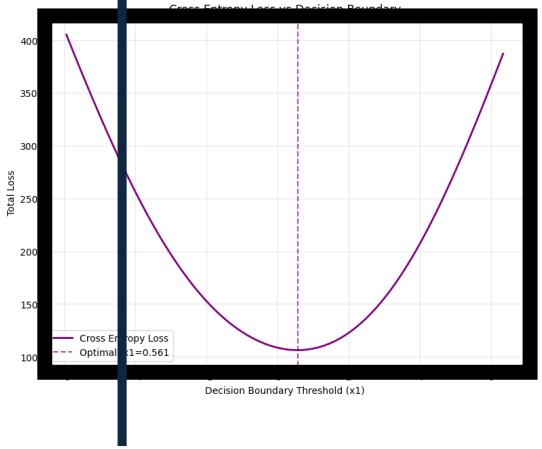
Πώς βρίσκουμε τη βέλτιστη λύση;



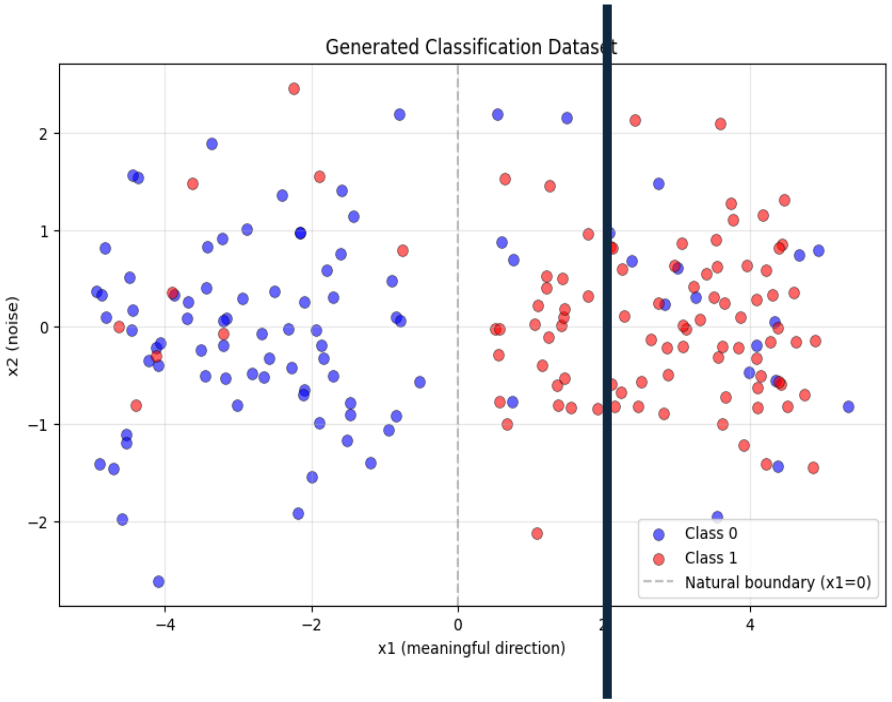


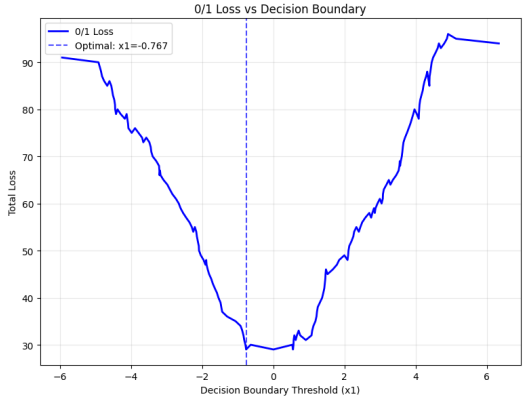


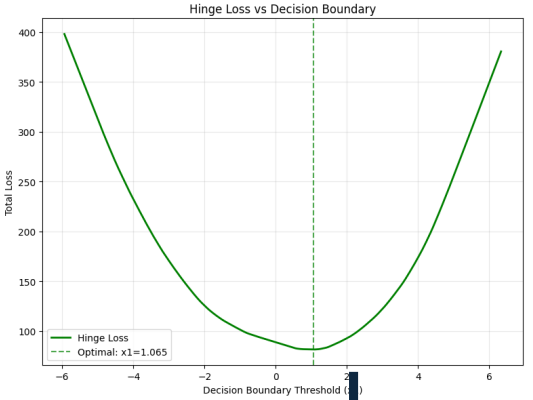


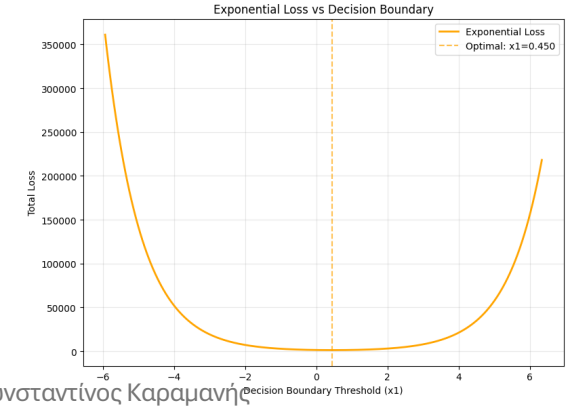


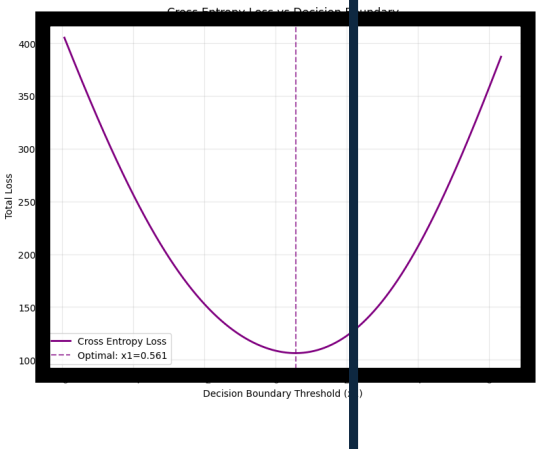
Πώς βρίσκουμε τη βέλτιστη λύση;



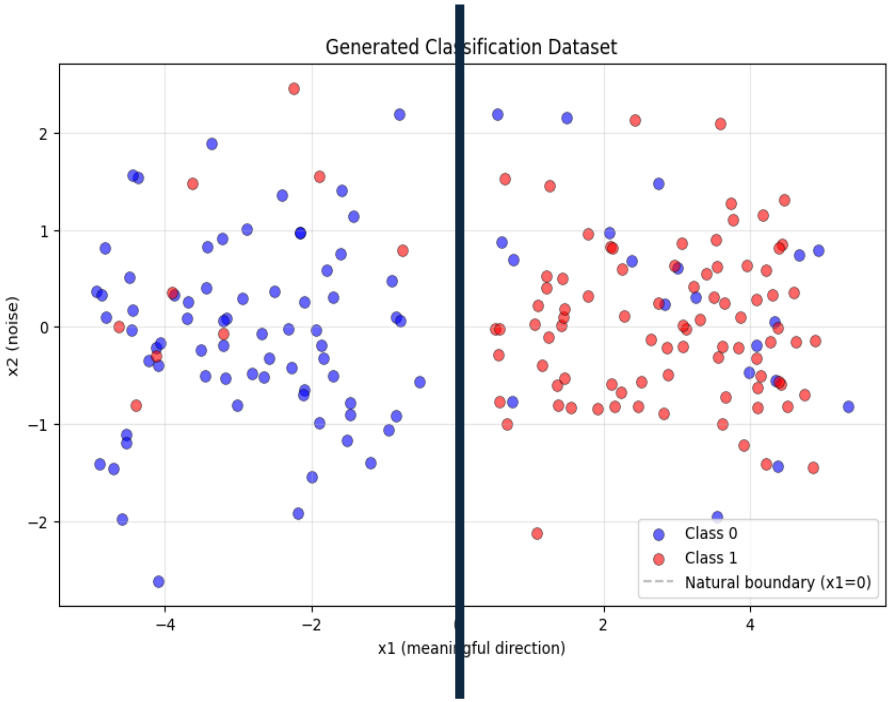


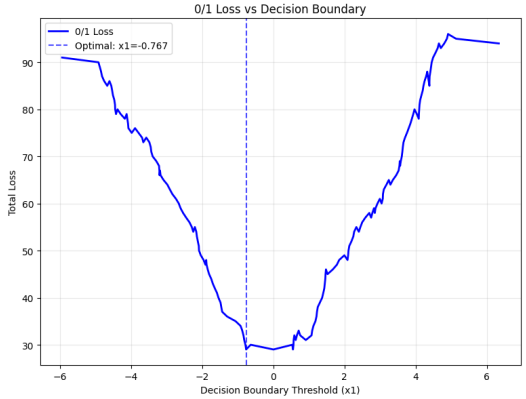


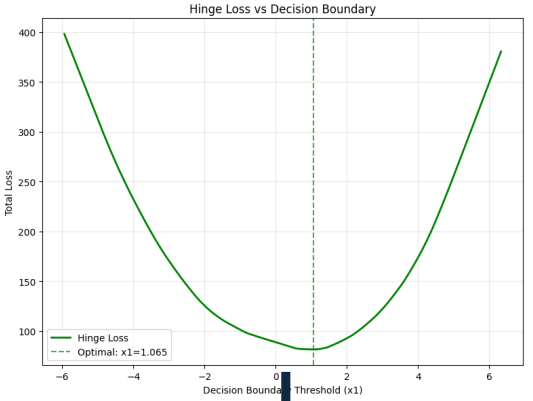


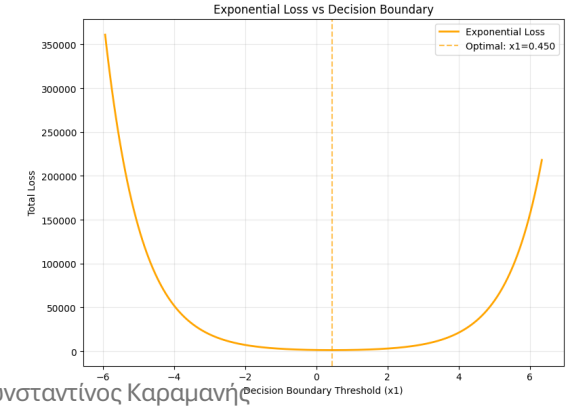


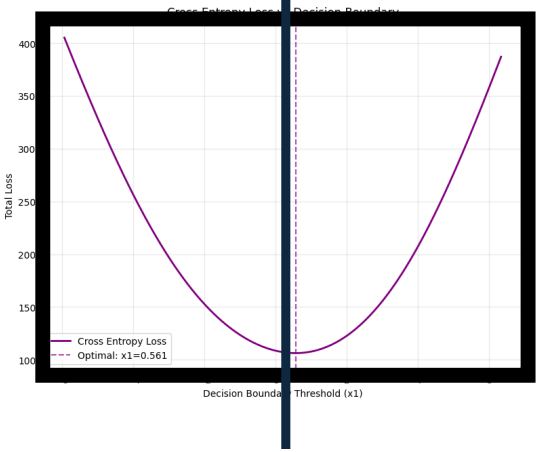
Πώς βρίσκουμε τη βέλτιστη λύση;



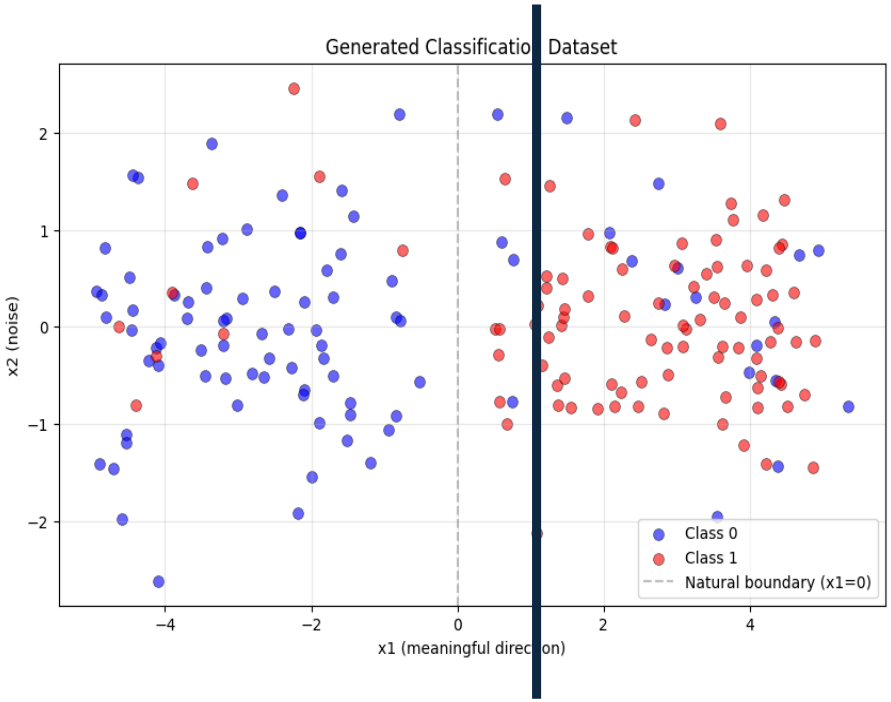


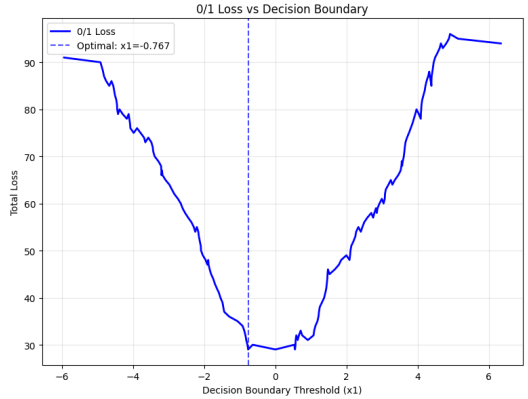


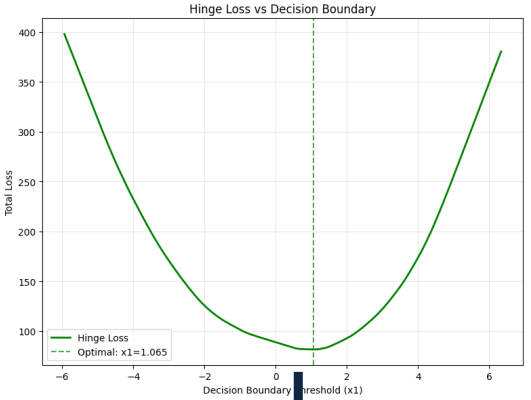


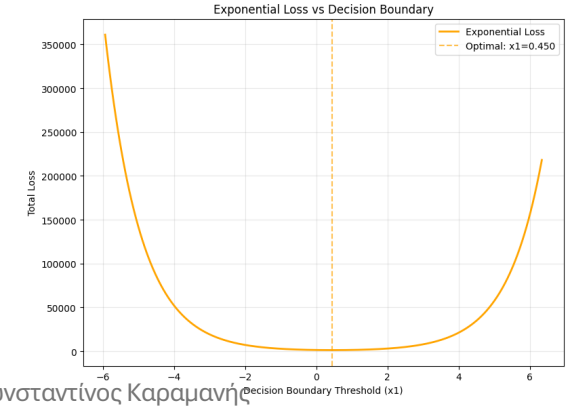


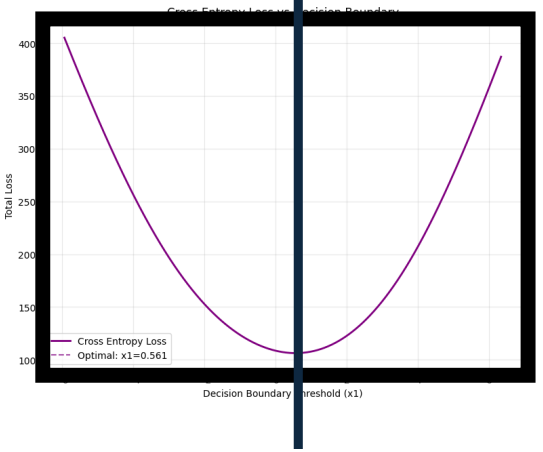
Πώς βρίσκουμε τη βέλτιστη λύση;



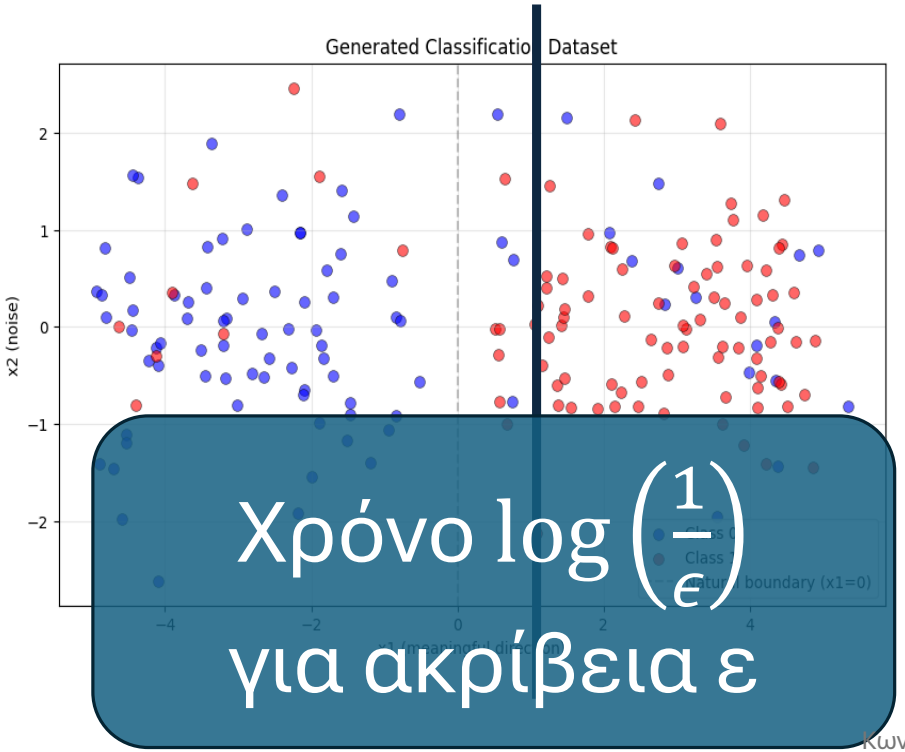


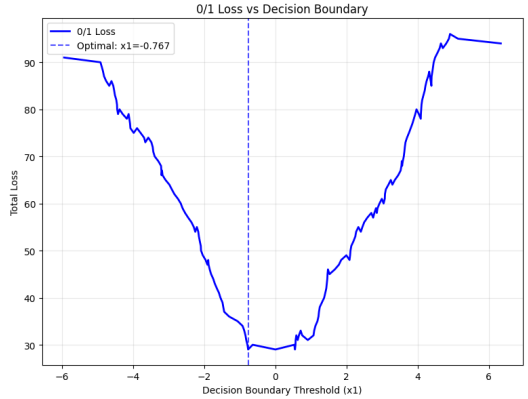


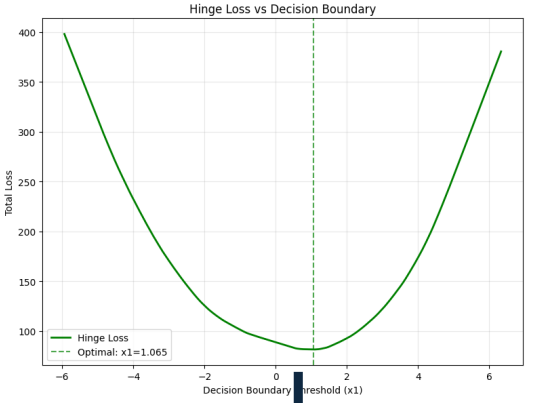


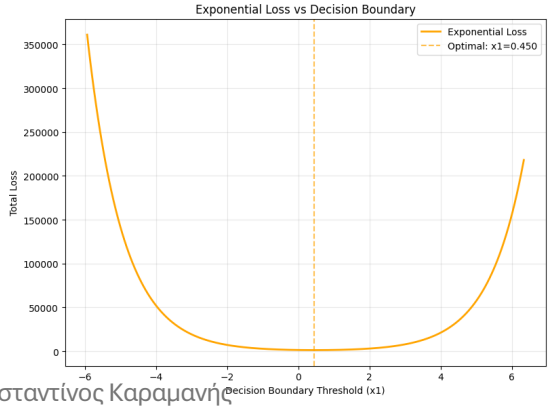


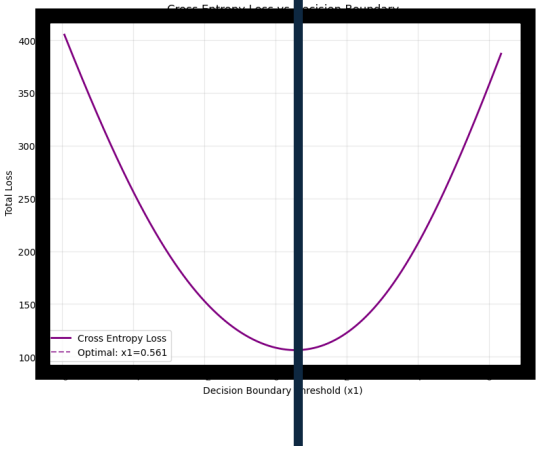
Πώς βρίσκουμε τη βέλτιστη λύση;





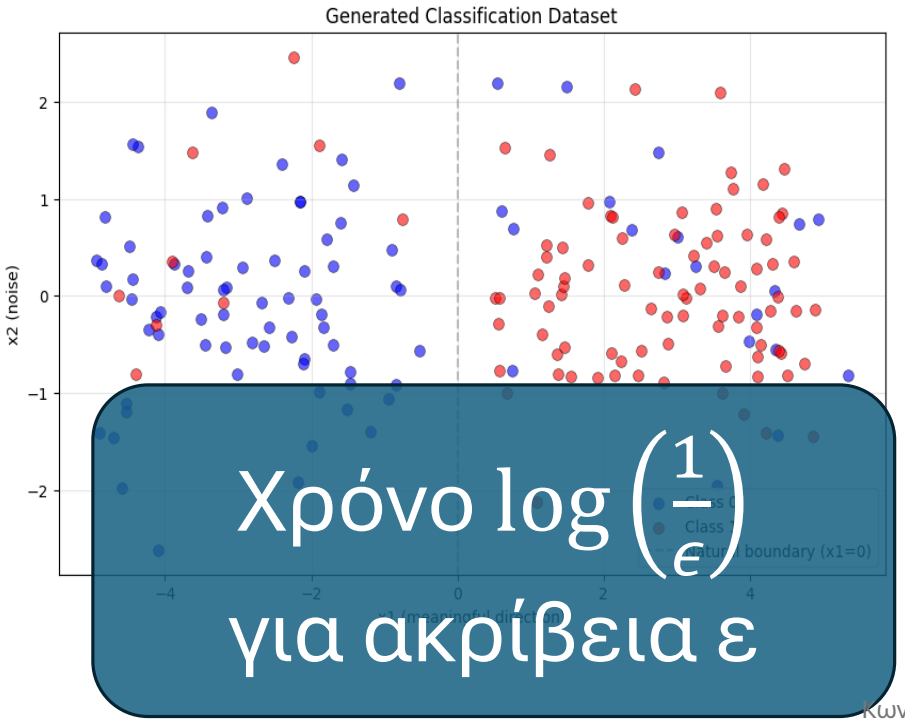


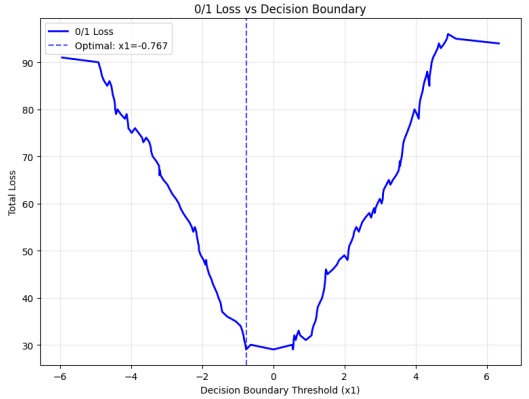


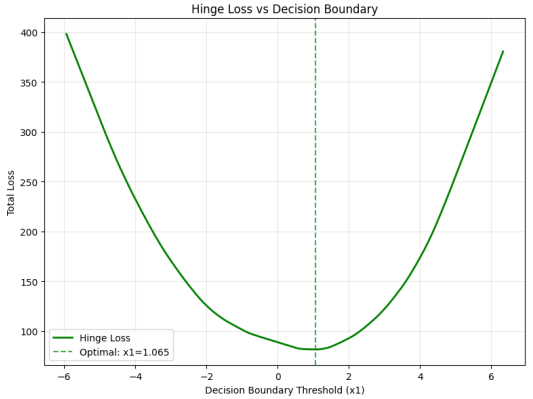


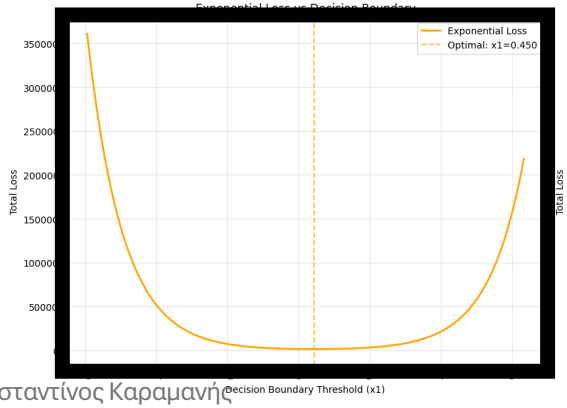
Χρόνο $\log\left(\frac{1}{\epsilon}\right)$
για ακρίβεια ε

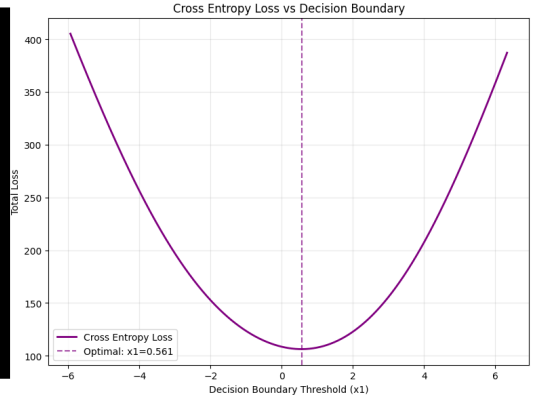
Πώς βρίσκουμε τη βέλτιστη λύση;





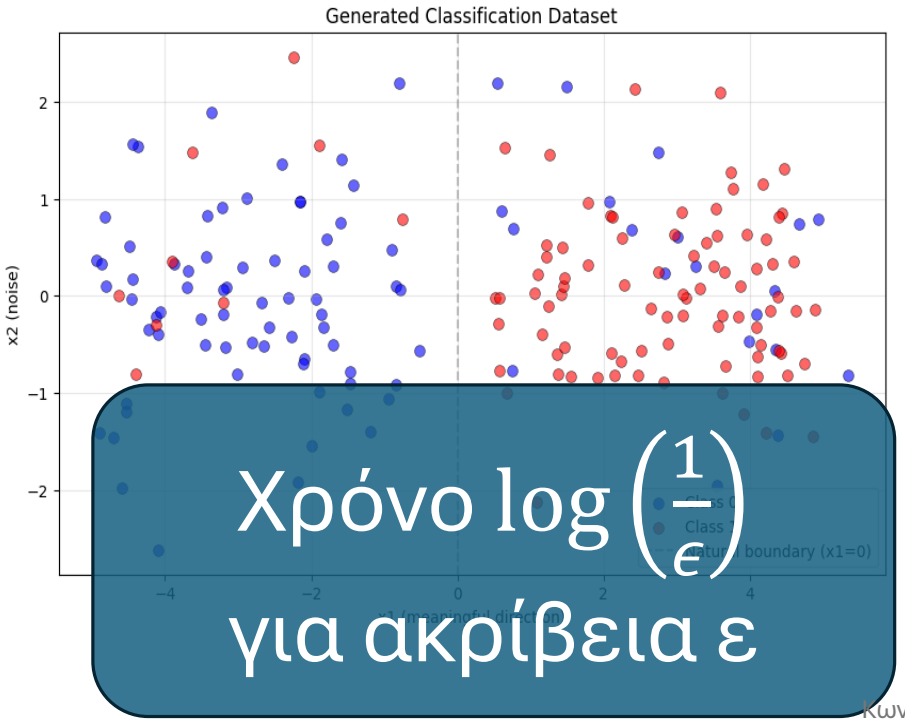


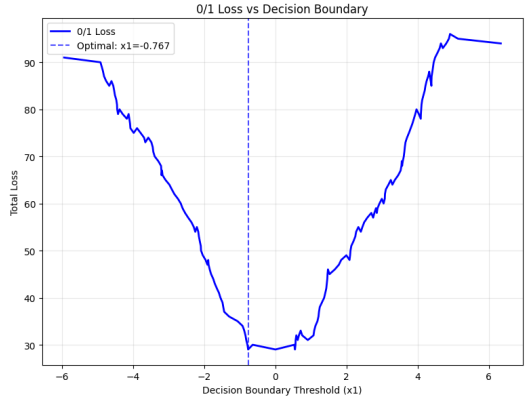


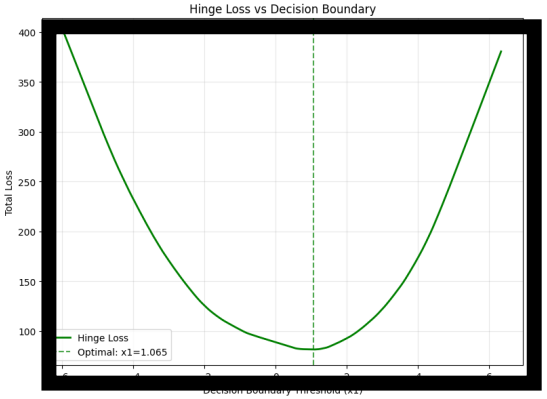


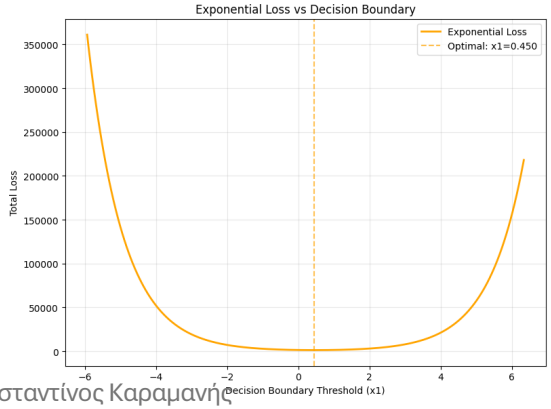
Χρόνο $\log\left(\frac{1}{\epsilon}\right)$
για ακρίβεια ε

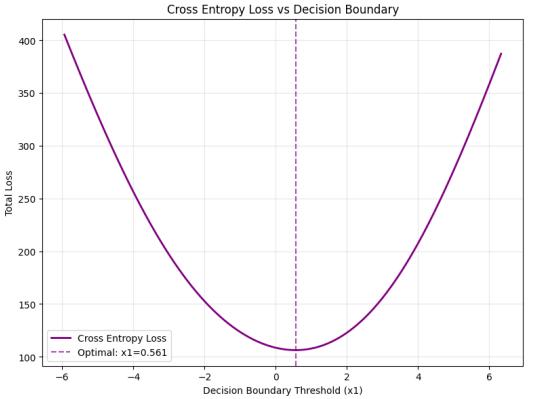
Πώς βρίσκουμε τη βέλτιστη λύση;





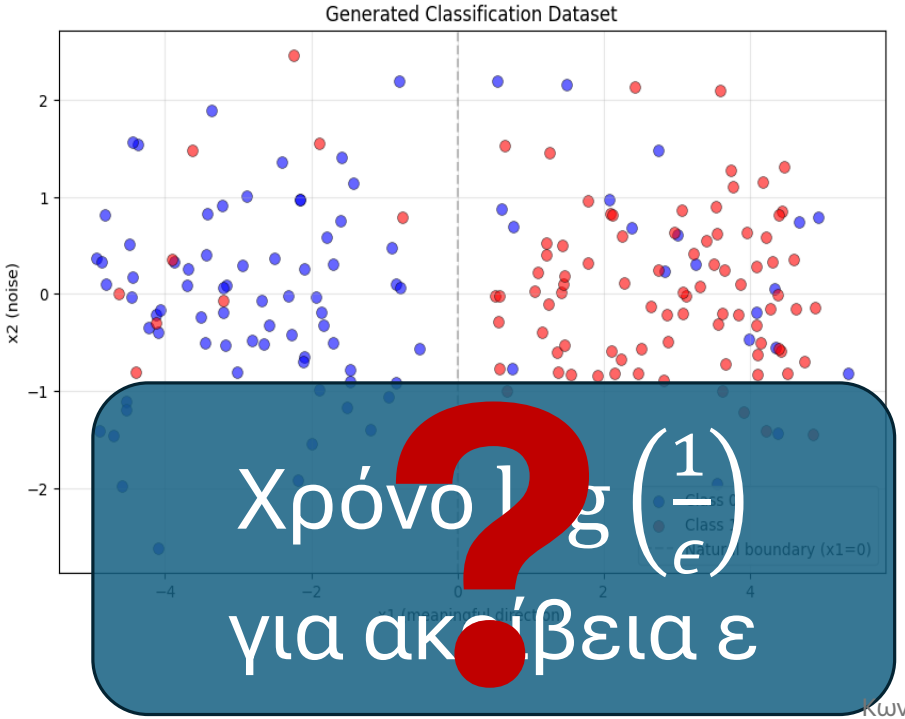




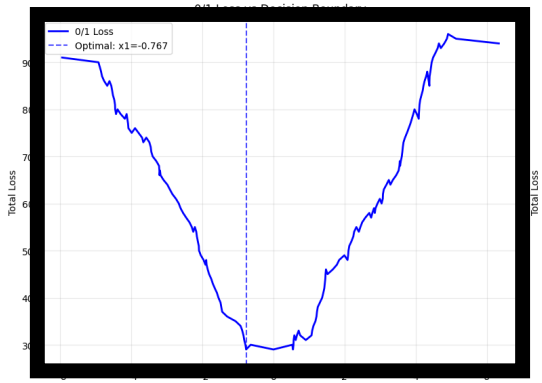


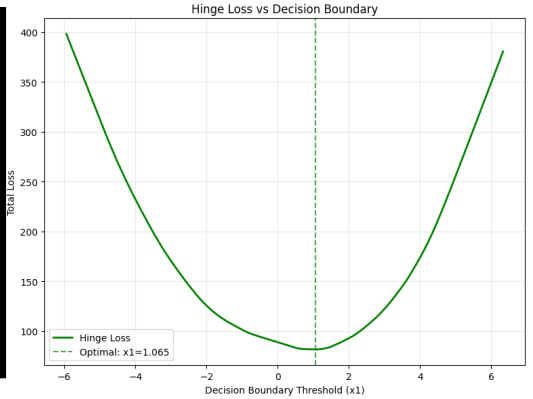
Χρόνο $\log\left(\frac{1}{\epsilon}\right)$ για ακρίβεια ε

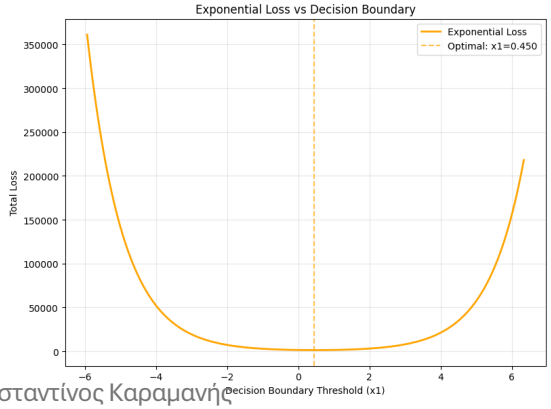
Πώς βρίσκουμε τη βέλτιστη λύση;

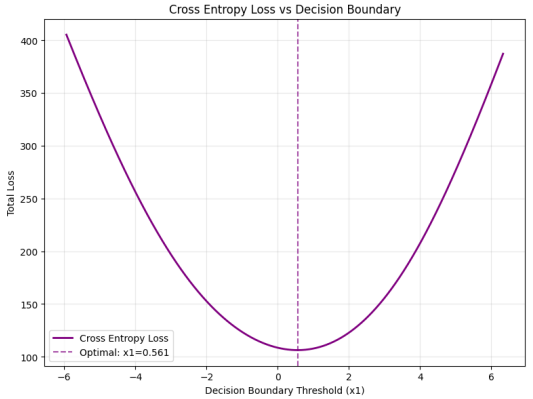


Χρόνο $\log\left(\frac{1}{\epsilon}\right)$ για ακρίβεια ε

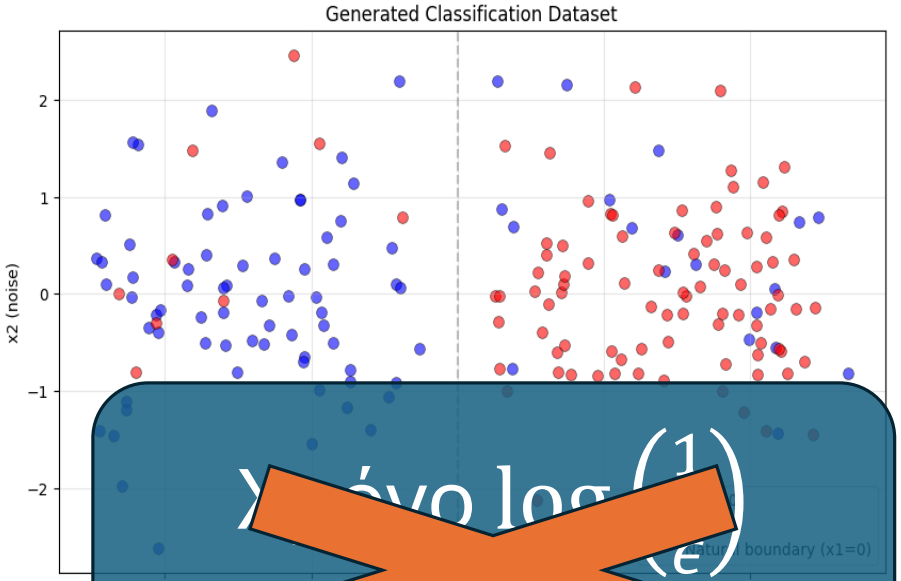


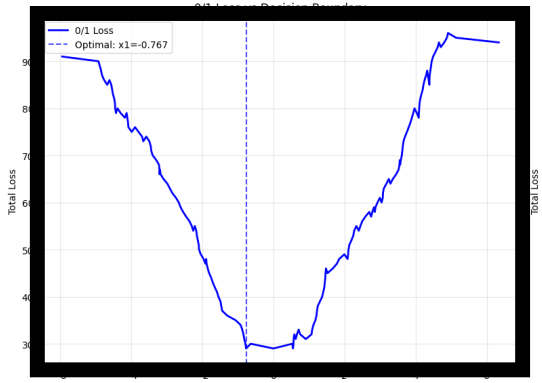


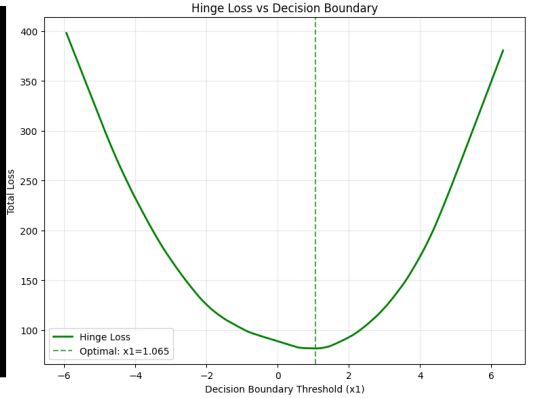


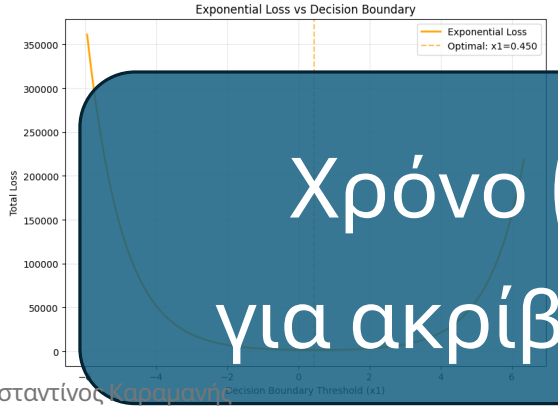


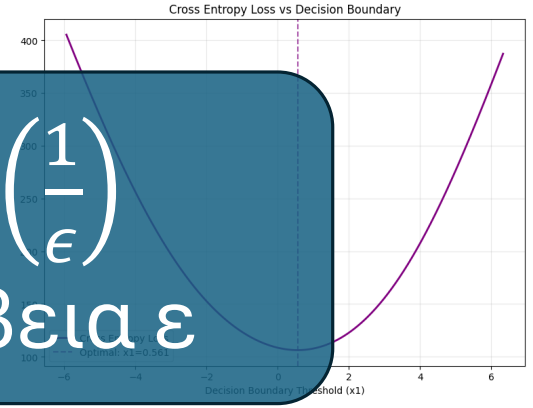
Πώς βρίσκουμε τη βέλτιστη λύση;







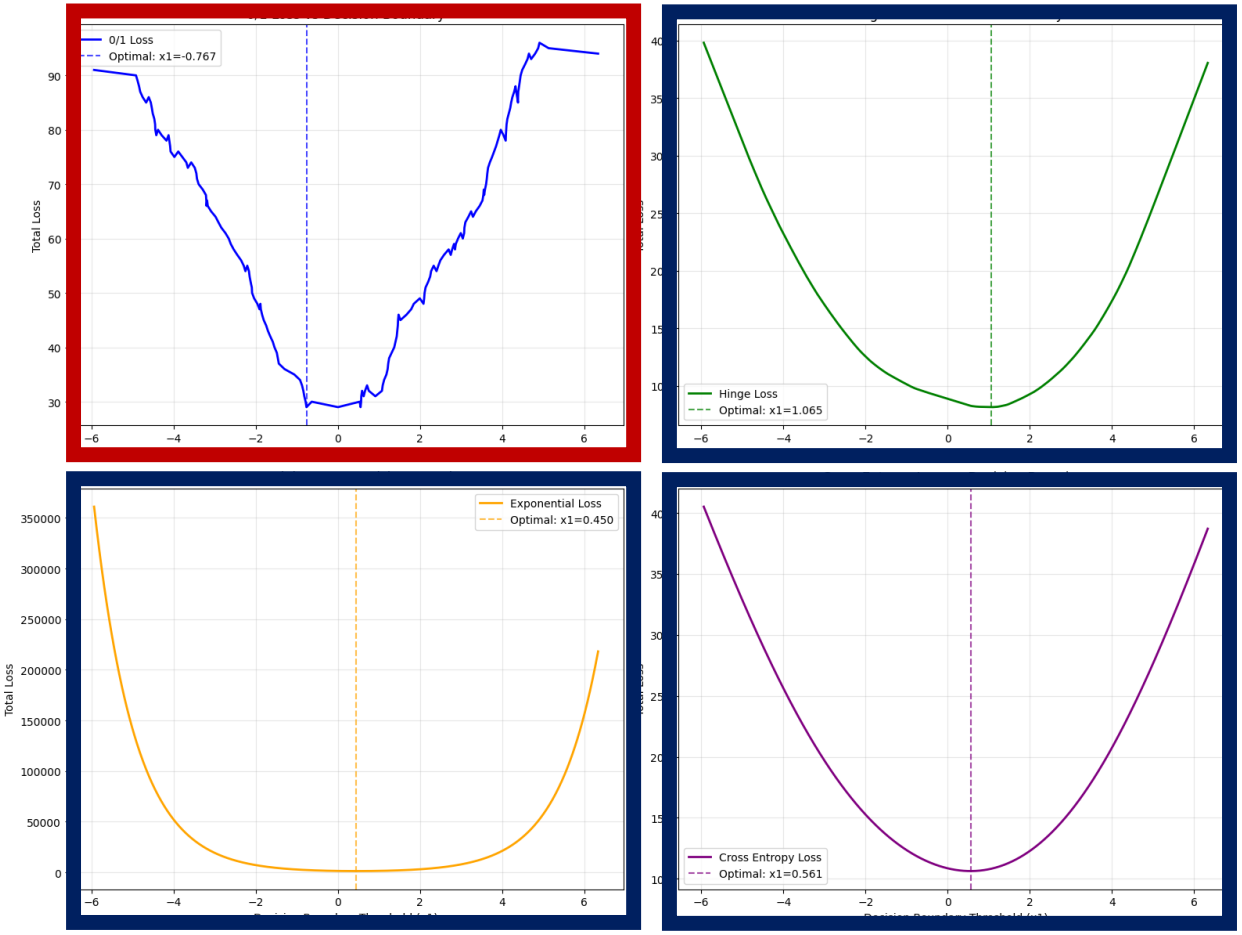




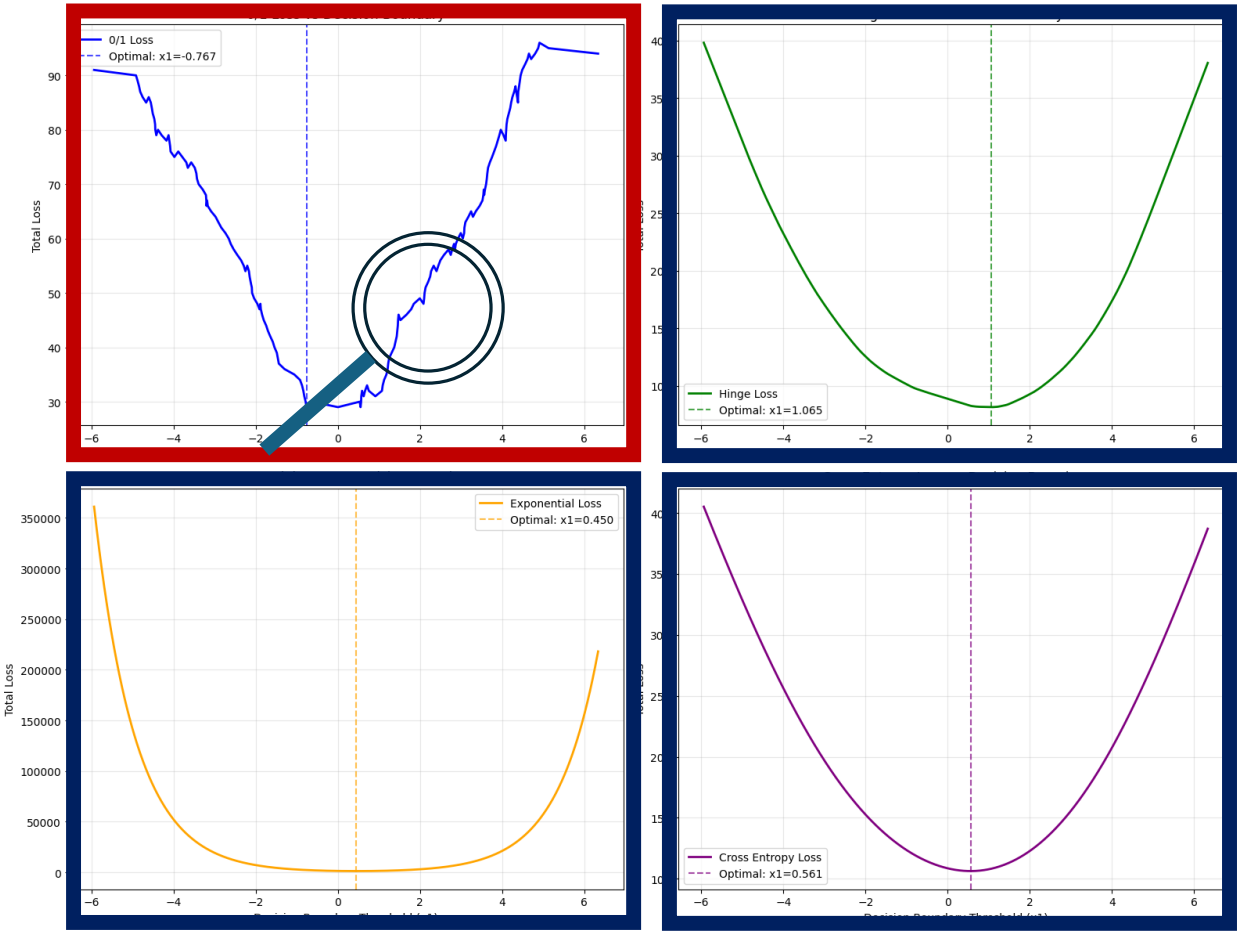
Χρόνο $\log\left(\frac{1}{\epsilon}\right)$
για ακρίβεια ε

Χρόνο $\left(\frac{1}{\epsilon}\right)$
για ακρίβεια ε

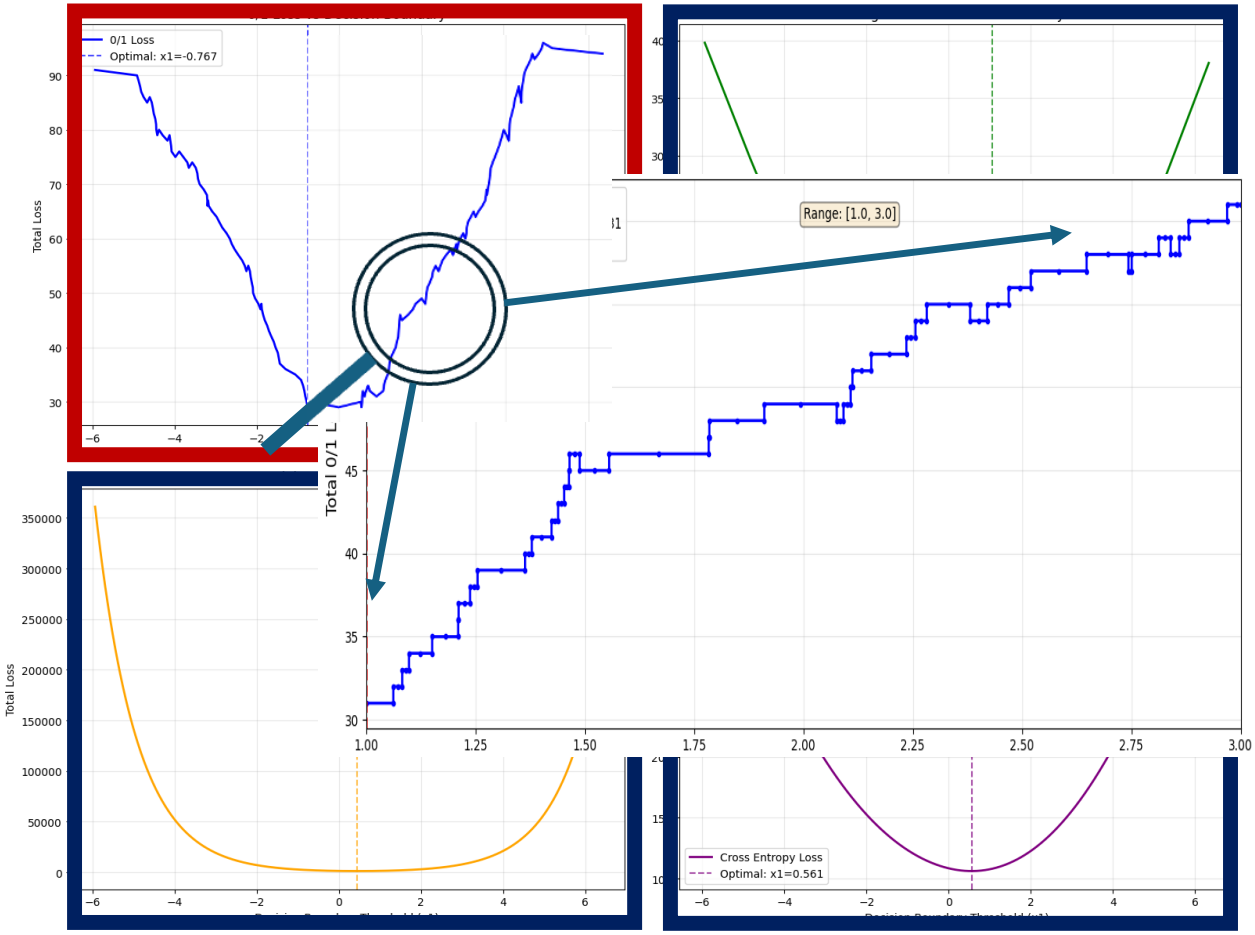
Πώς διαφέρει η πρώτη συνάρτηση απώλειας από τις άλλες τρείς;



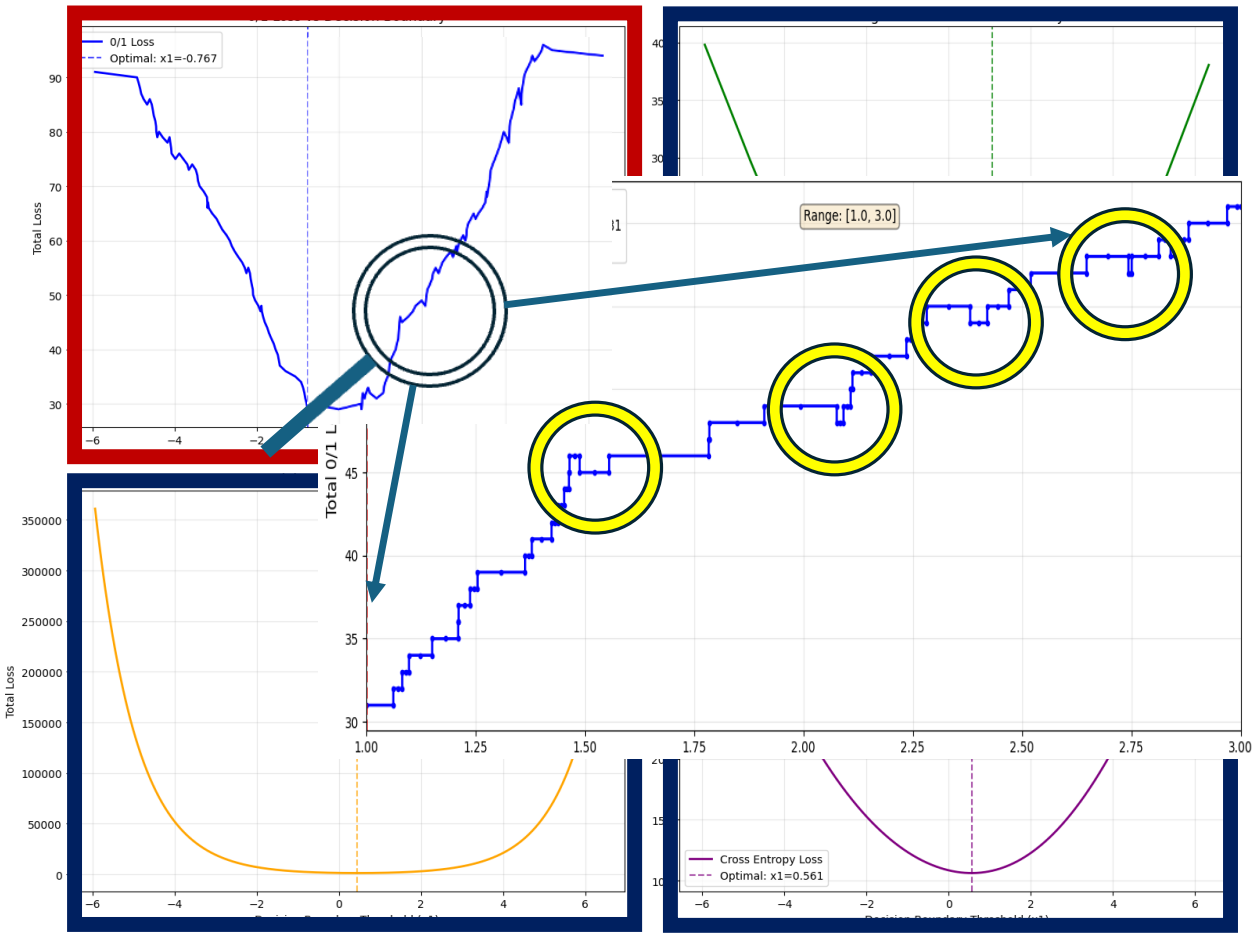
Πώς διαφέρει η πρώτη συνάρτηση απώλειας από τις άλλες τρείς;



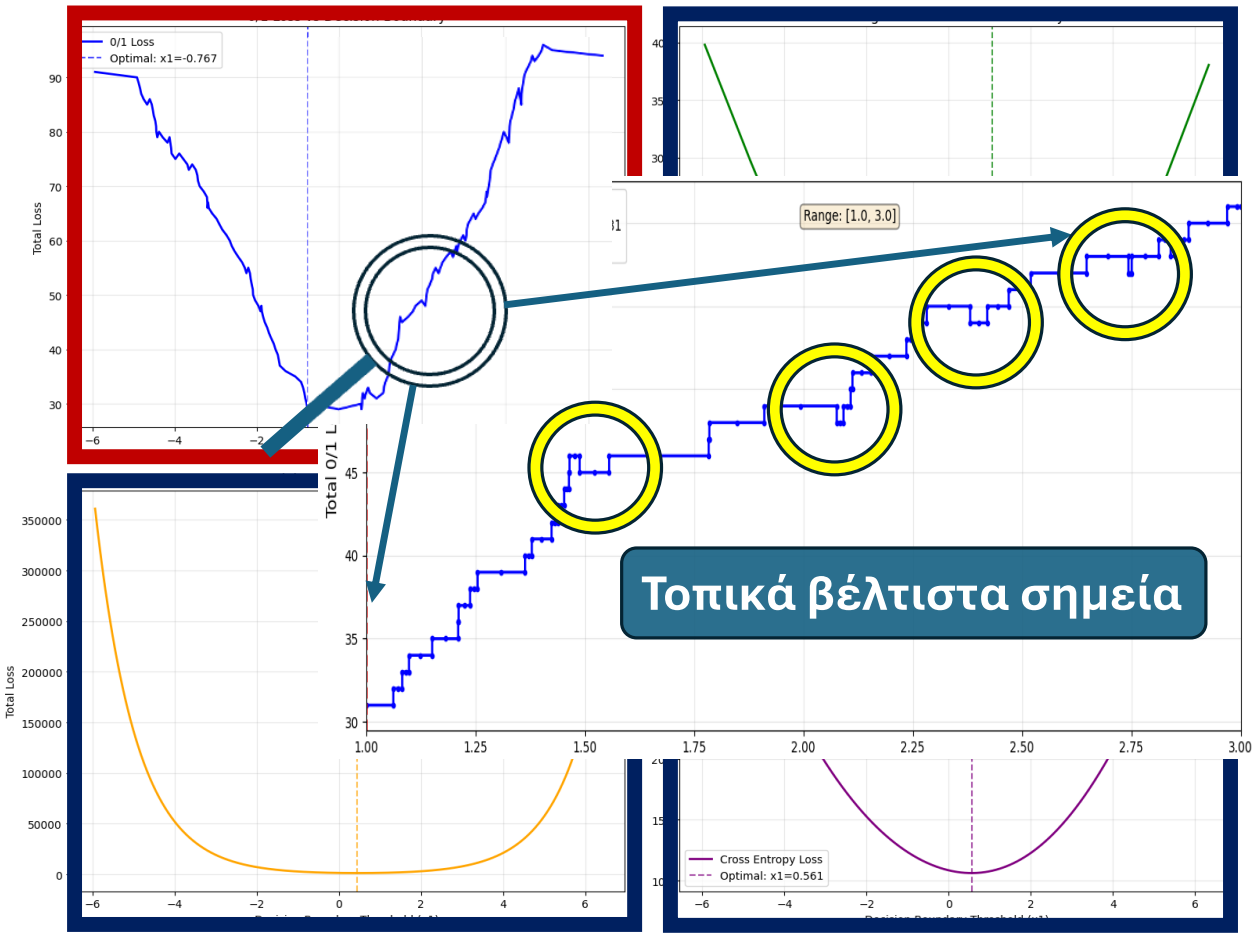
Πώς διαφέρει η πρώτη συνάρτηση απώλειας από τις άλλες τρεις;



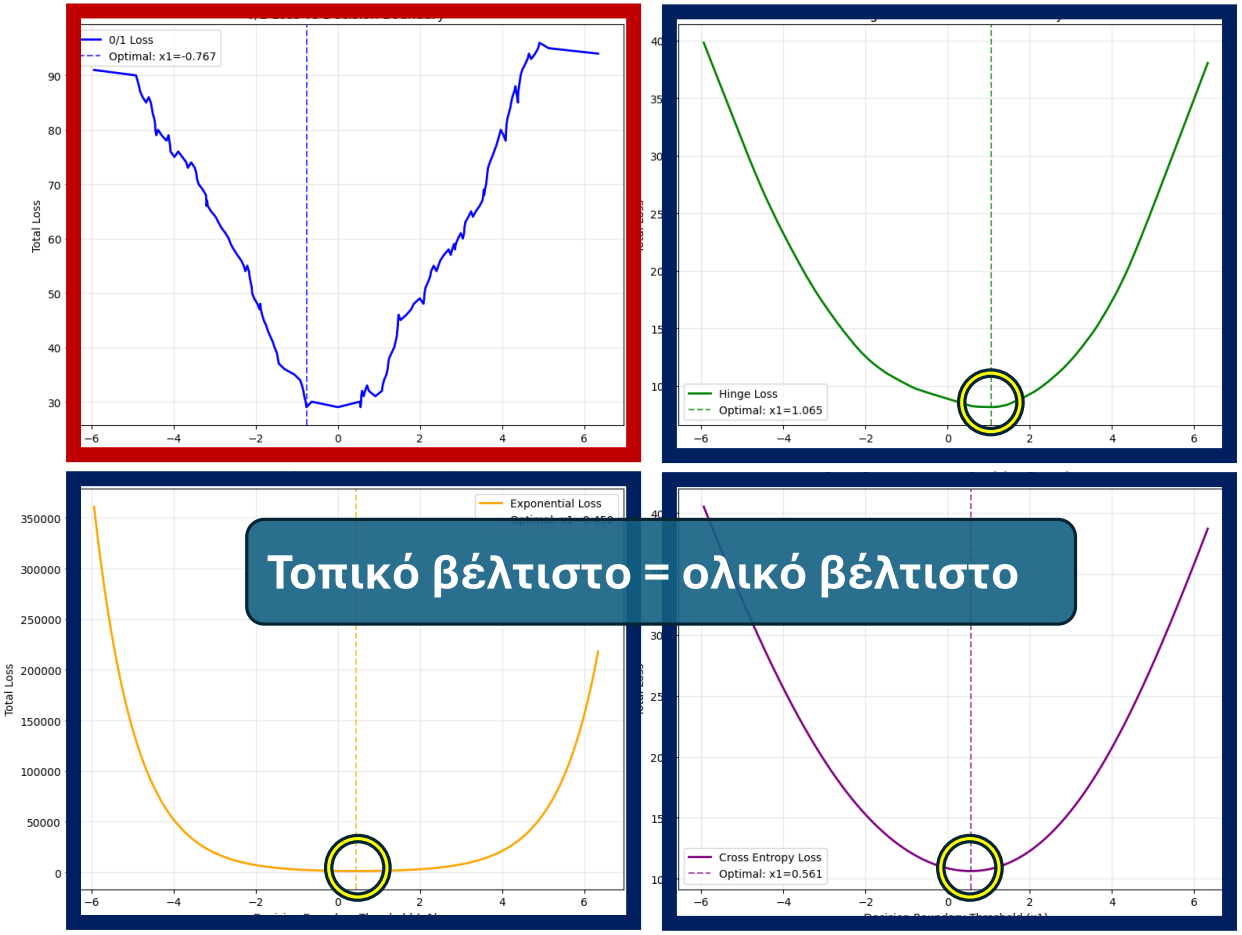
Πώς διαφέρει η πρώτη συνάρτηση απώλειας από τις άλλες τρεις;



Πώς διαφέρει η πρώτη συνάρτηση απώλειας από τις άλλες τρεις;

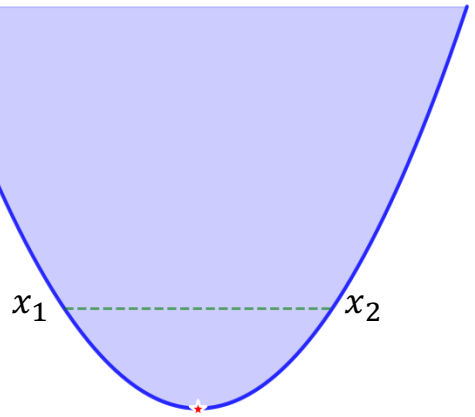


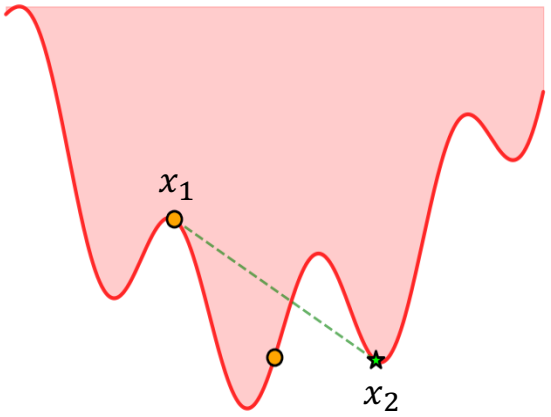
Πώς διαφέρει η πρώτη συνάρτηση απώλειας από τις άλλες τρεις;



Τοπικό βέλτιστο = ολικό βέλτιστο

Κυρτές συναρτήσεις (convex functions)





Μια συνάρτηση λέγεται **κυρτή**, εάν η χορδή μεταξύ οποιωνδήποτε δύο σημείων της γραφικής της παράστασης βρίσκεται πάνω από την καμπύλη.

Αναλυτικός ορισμός: για οποιαδήποτε $\lambda \in [0,1], x_1, x_2$

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

Κυρτές συναρτήσεις (convex functions)

Εάν η συνάρτηση είναι κυρτή:

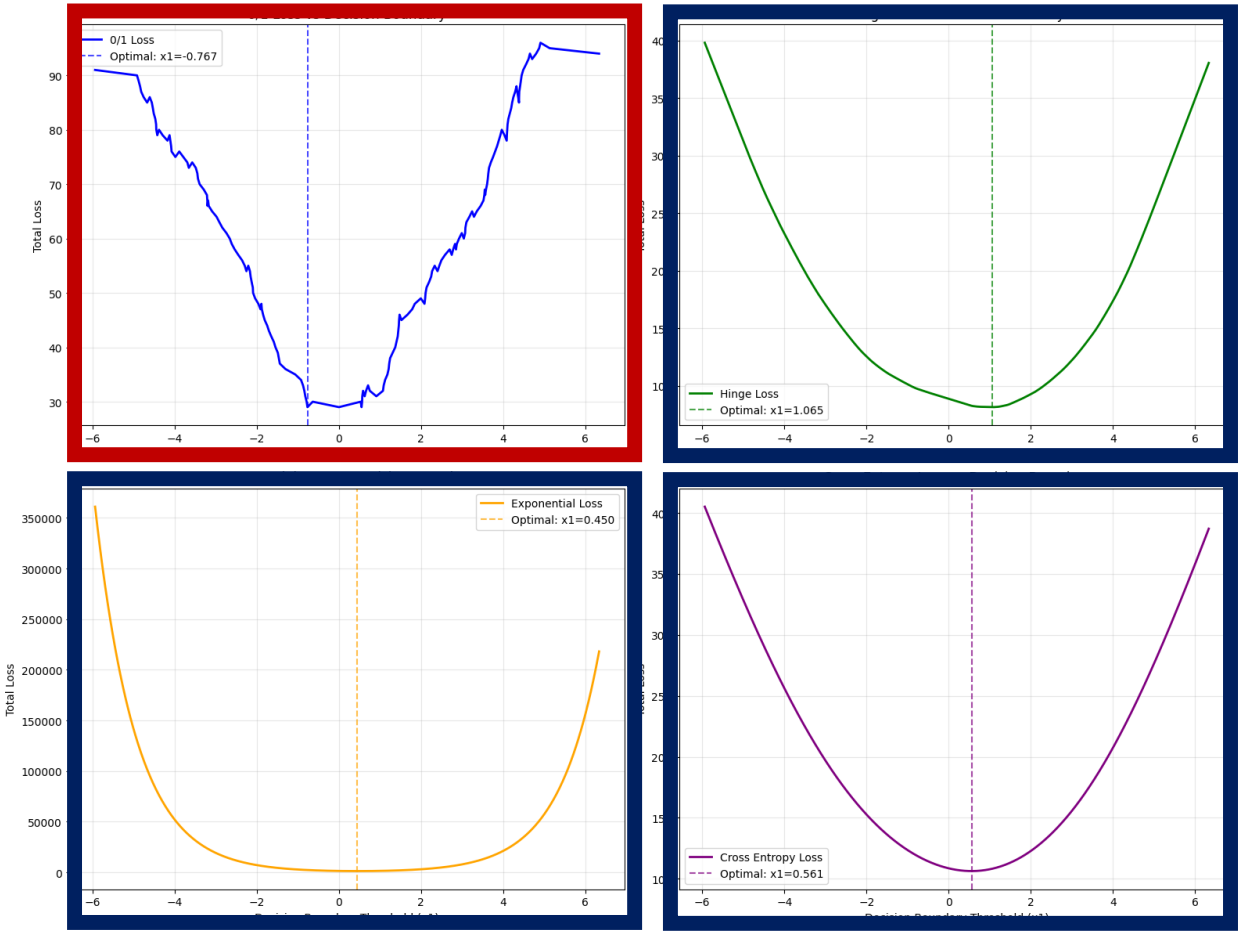
Τοπικό βέλτιστο = ολικό βέλτιστο

Μία συνάρτηση λέγεται ***κυρτή*** εάν η χορδή
βρίσκεται επάνω από την καμπύλη

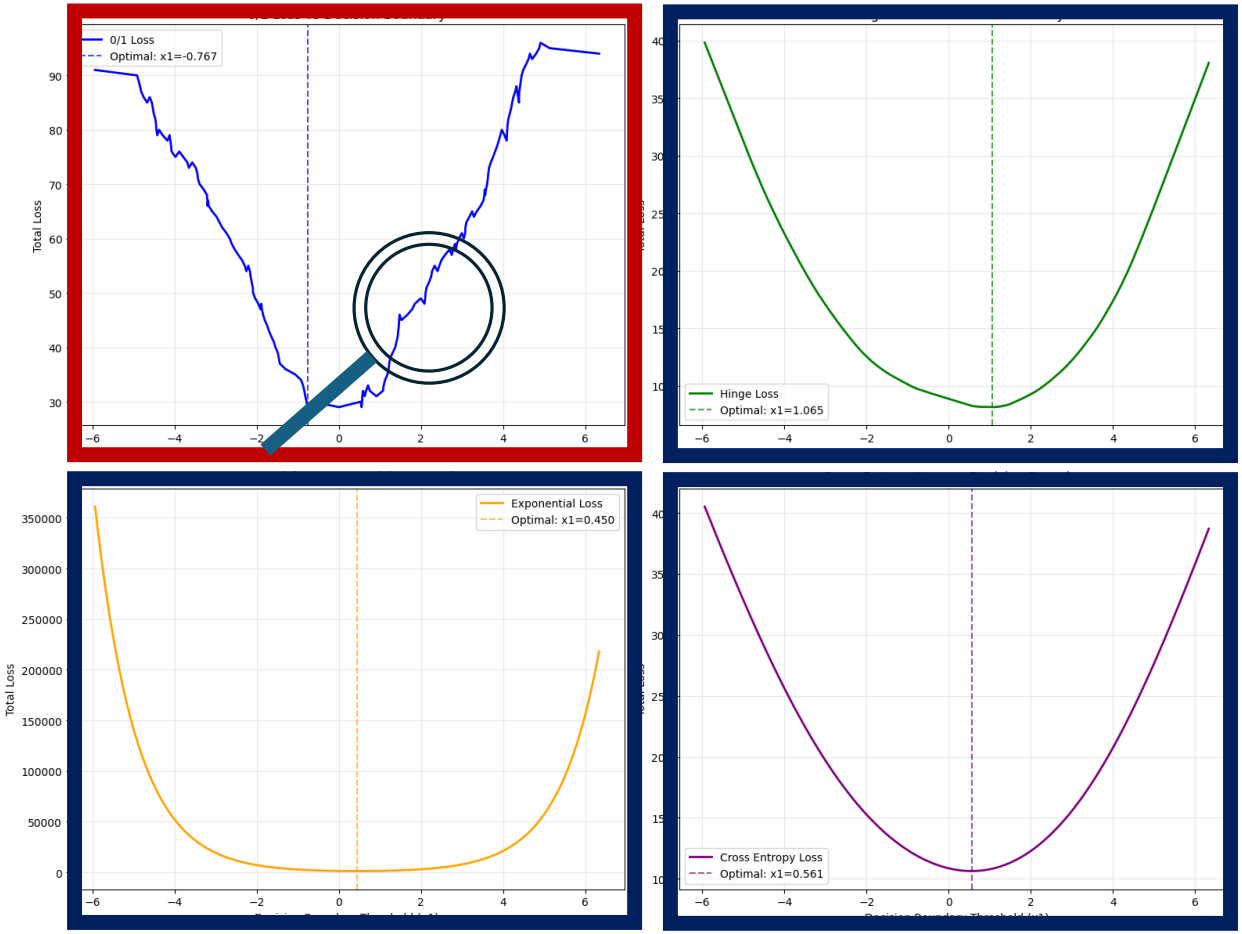
Αναλυτικός ορισμός: για οποιαδήποτε $\lambda \in [0,1], x_1, x_2$

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

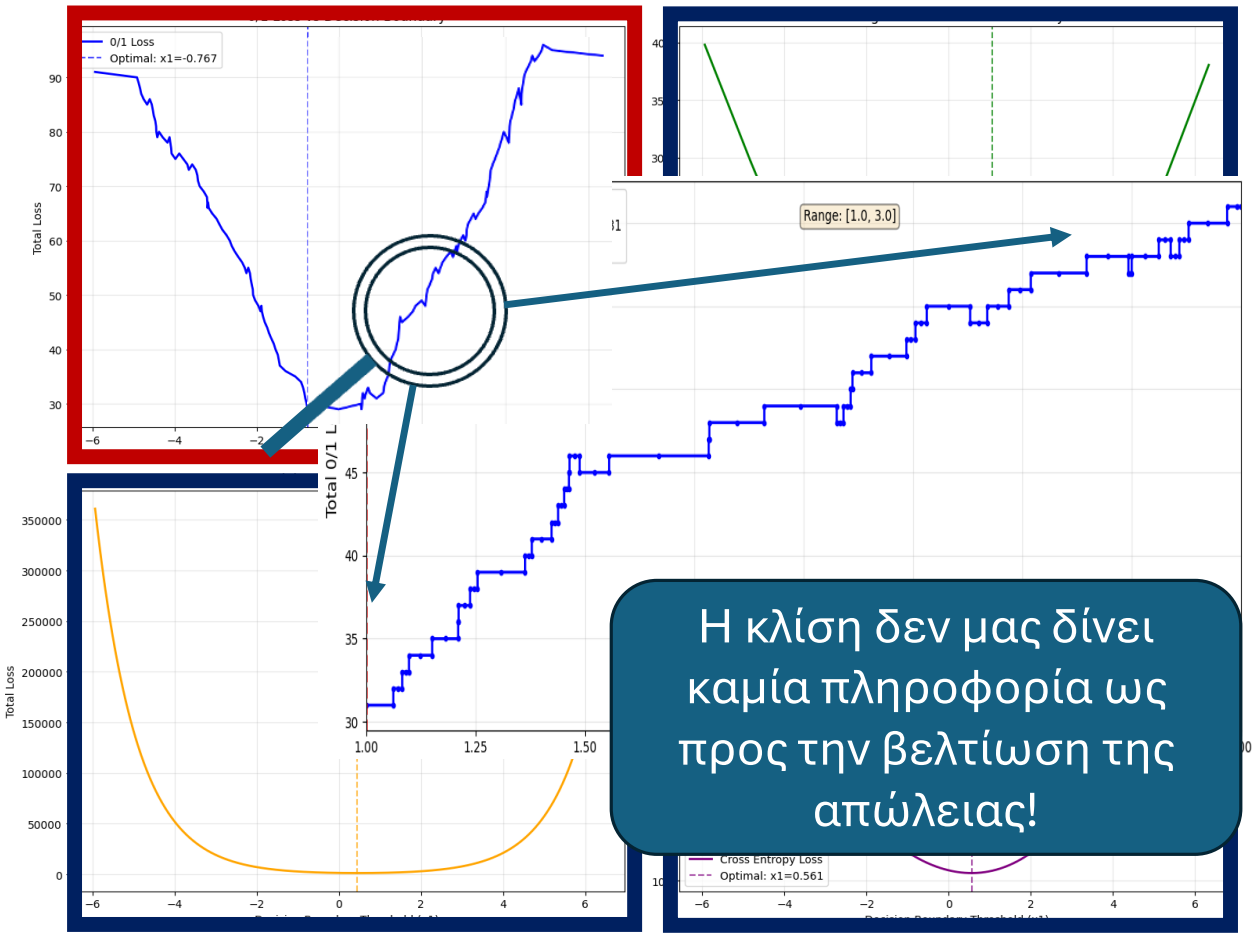
Πως διαφέρει η πρώτη συνάρτηση απώλειας από τις άλλες τρείς;



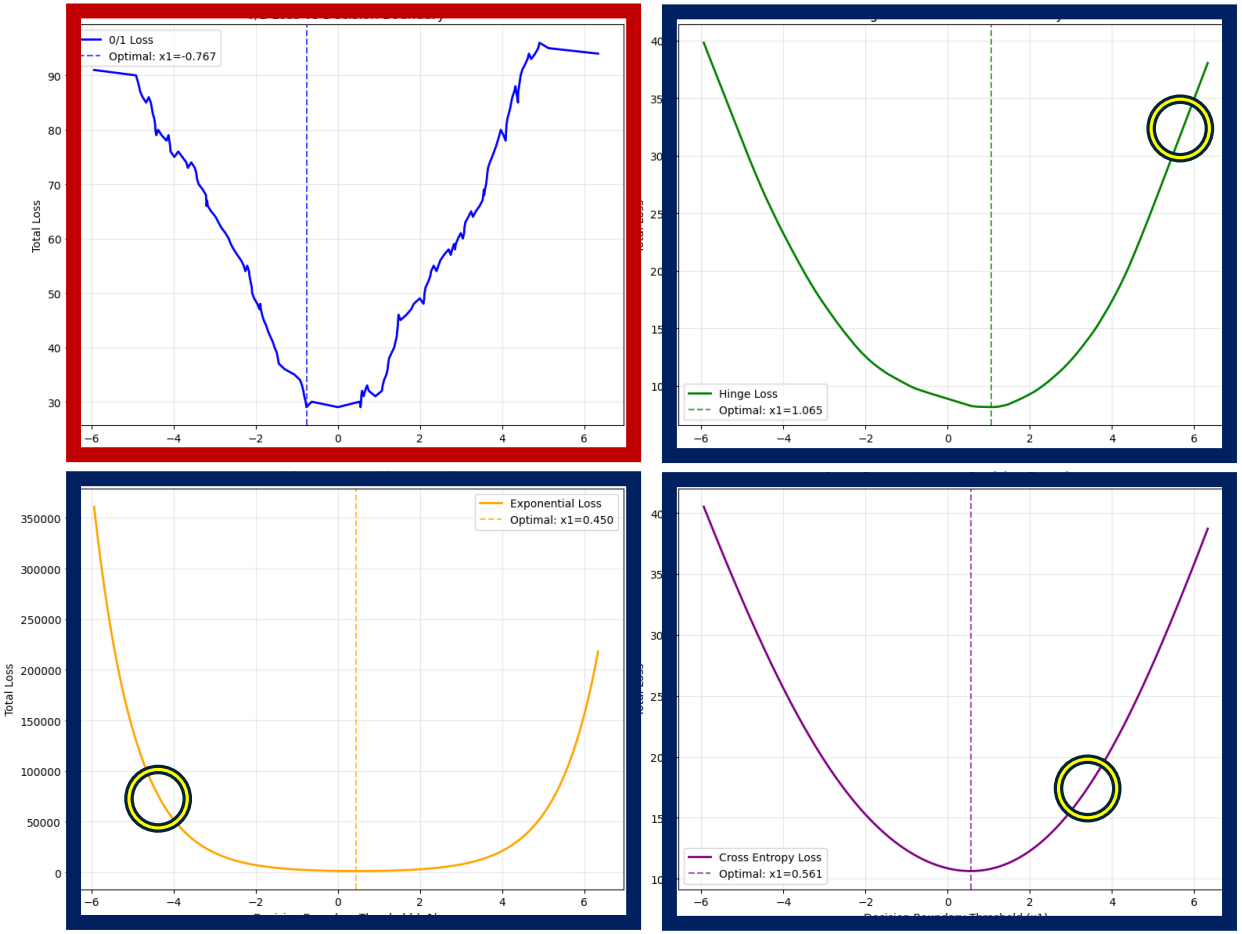
Πως διαφέρει η πρώτη συνάρτηση απώλειας από τις άλλες τρείς;



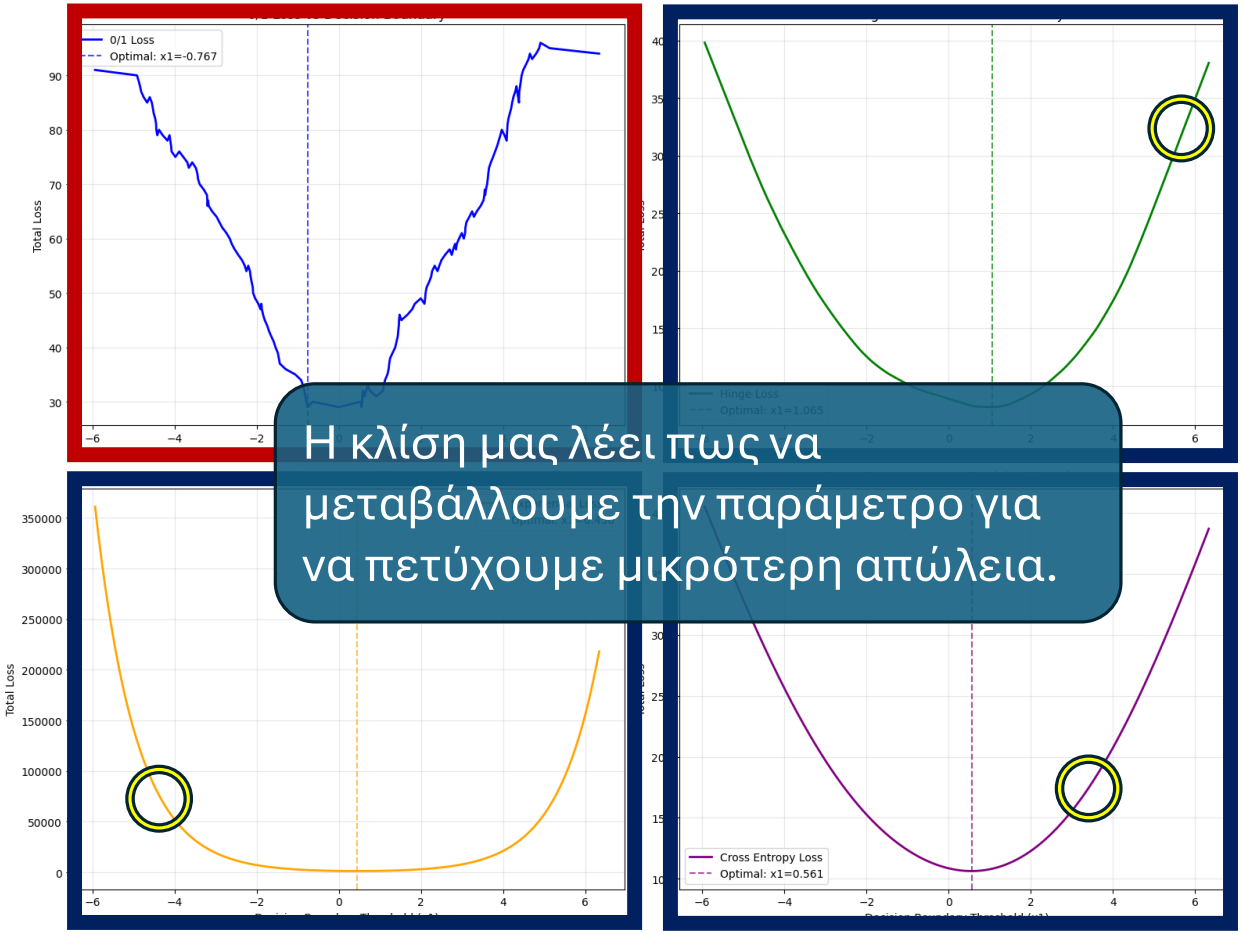
Πως διαφέρει η πρώτη συνάρτηση απώλειας από τις άλλες τρείς;



Πως διαφέρει η πρώτη συνάρτηση απώλειας από τις άλλες τρείς;

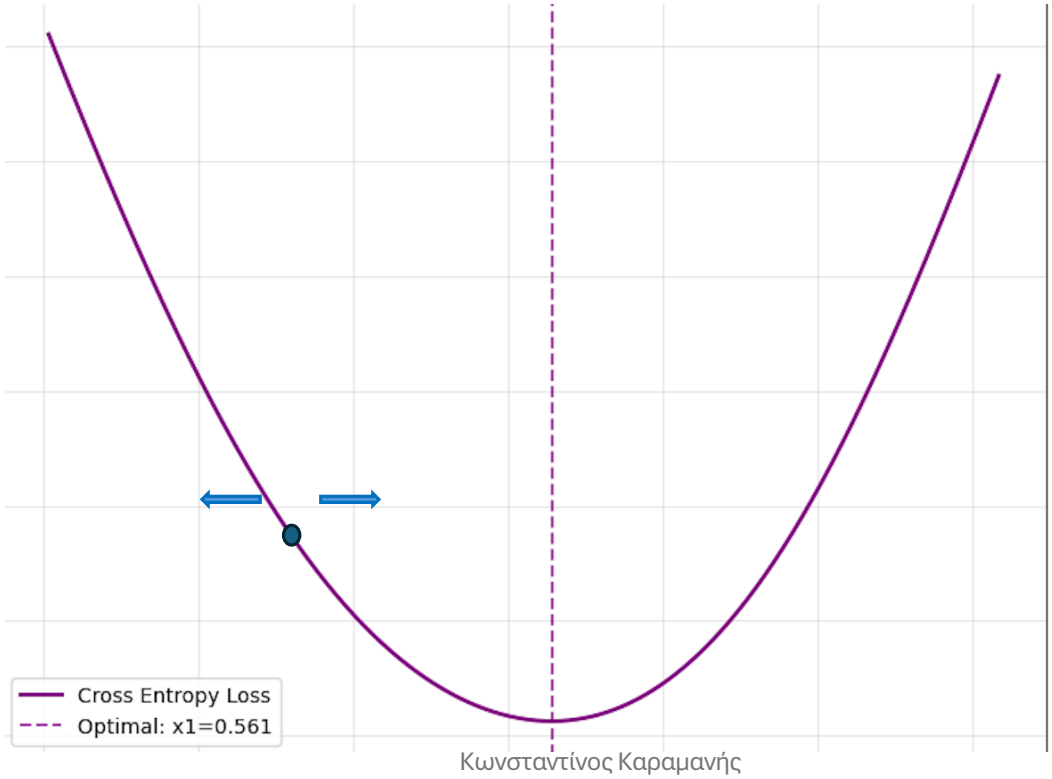


Πως διαφέρει η πρώτη συνάρτηση απώλειας από τις άλλες τρείς;

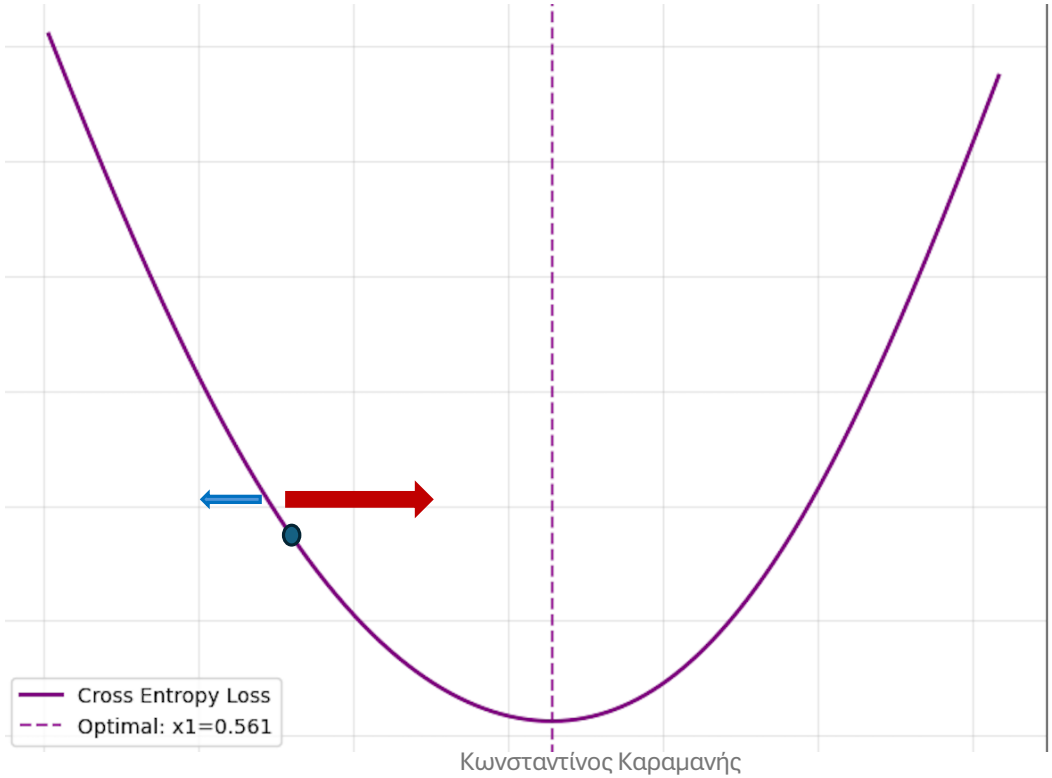


Η κλίση μας λέει πως να μεταβάλλουμε την παράμετρο για να πετύχουμε μικρότερη απώλεια.

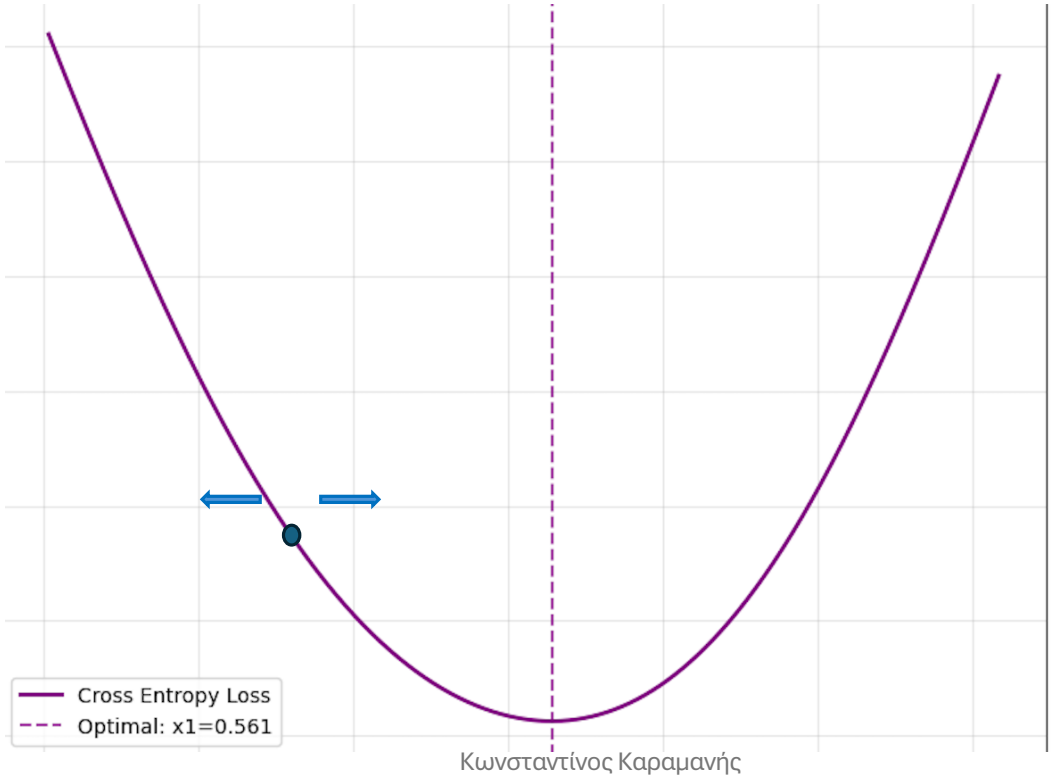
Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)



Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

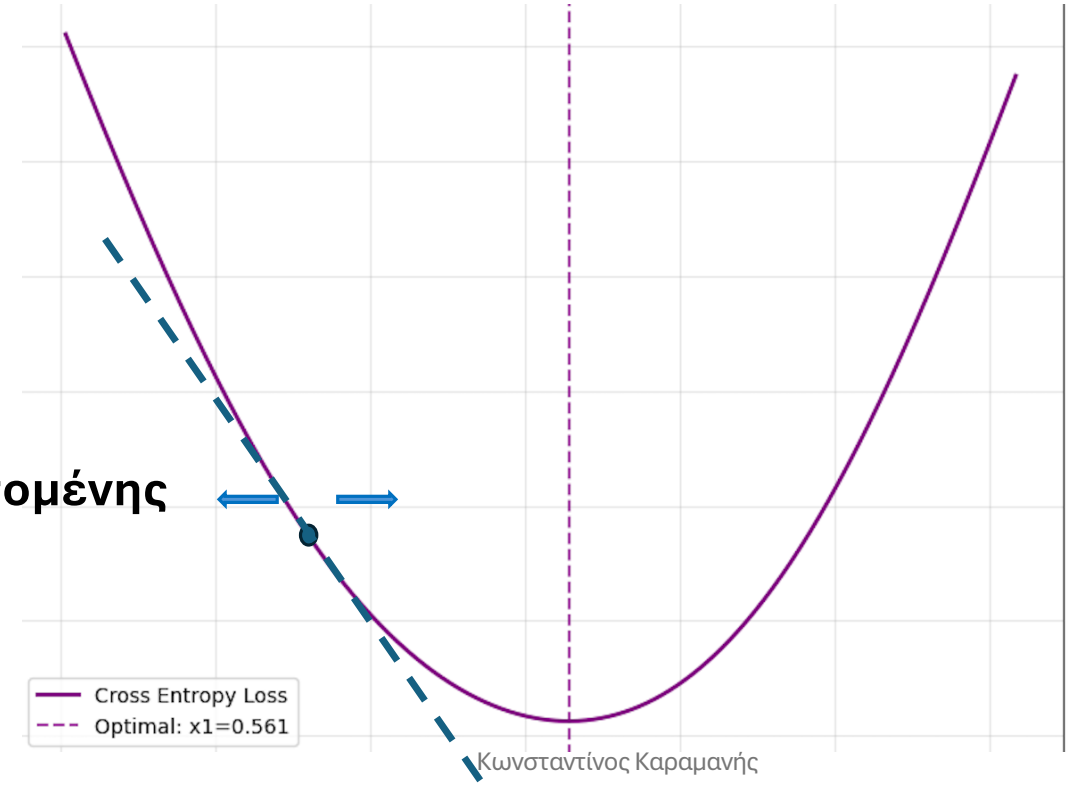


Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)



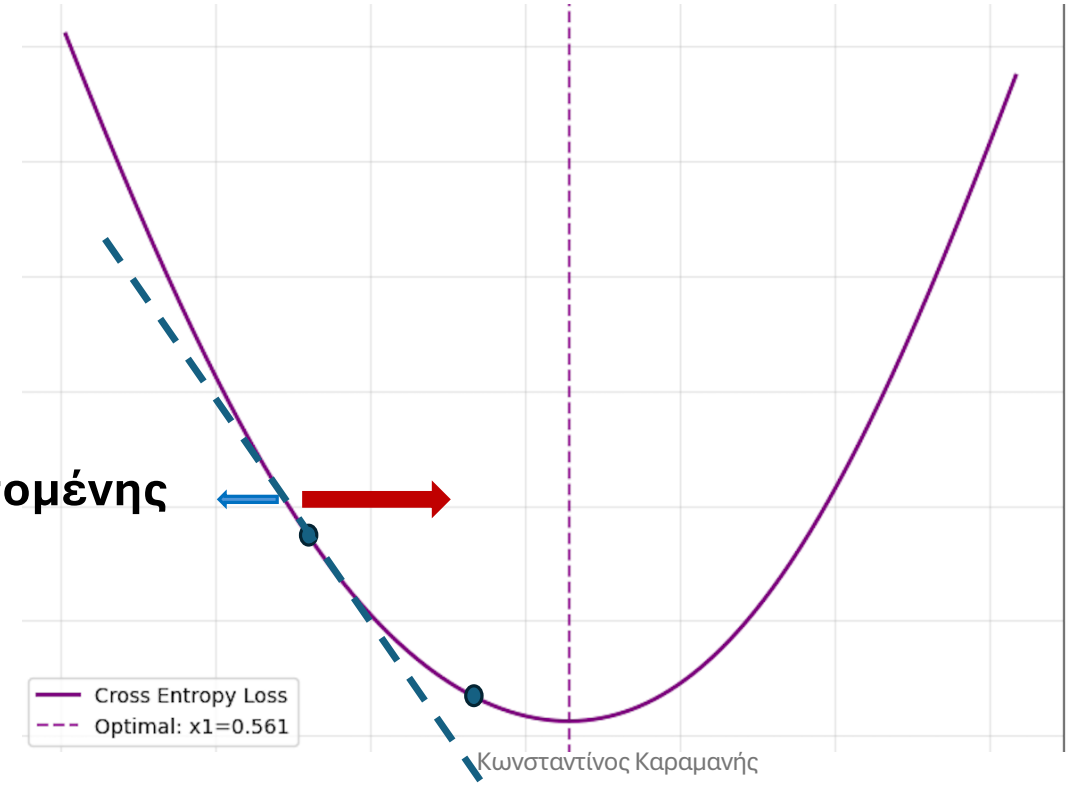
Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

κλίση της εφαπτομένης
= η **παράγωγος**



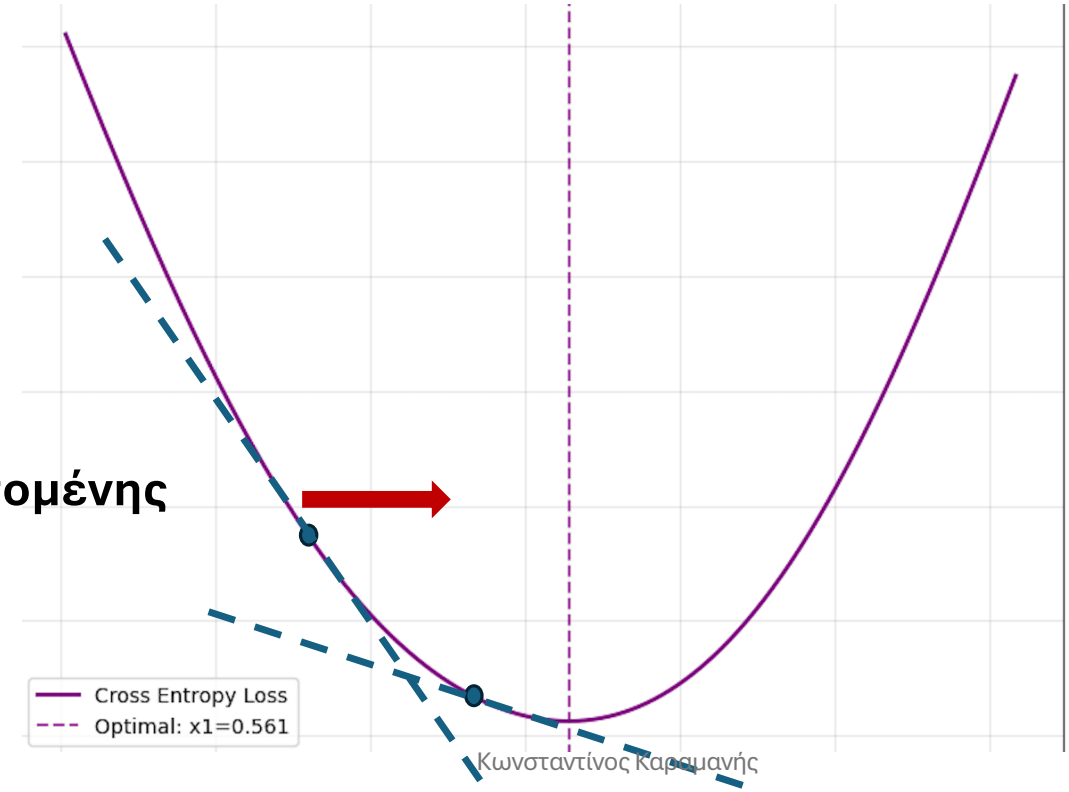
Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

κλίση της εφαπτομένης
= η **παράγωγος**



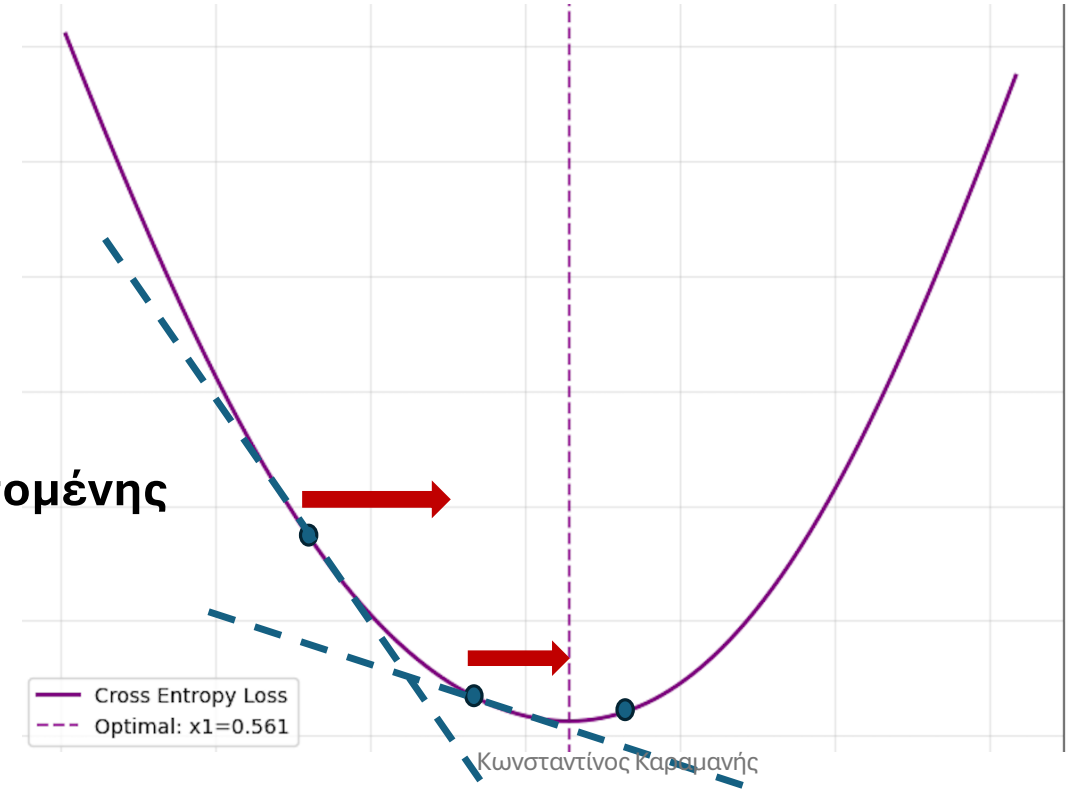
Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

κλίση της εφαπτομένης
= η **παράγωγος**



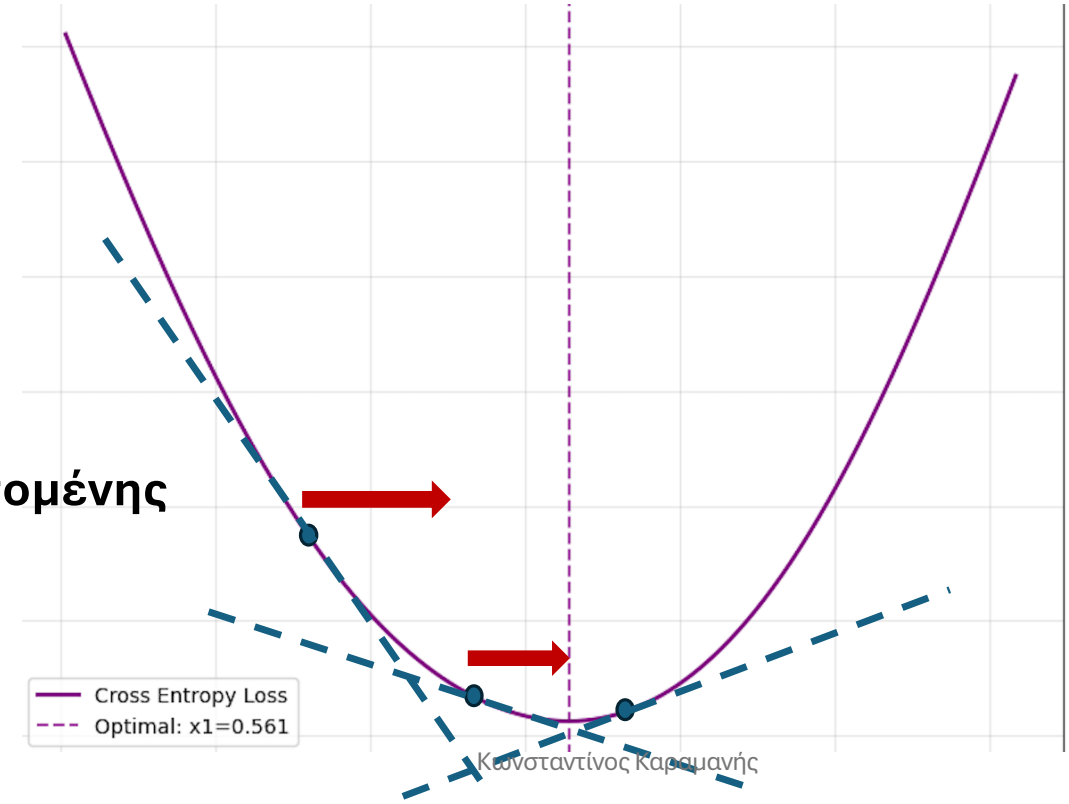
Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

κλίση της εφαπτομένης
= η **παράγωγος**



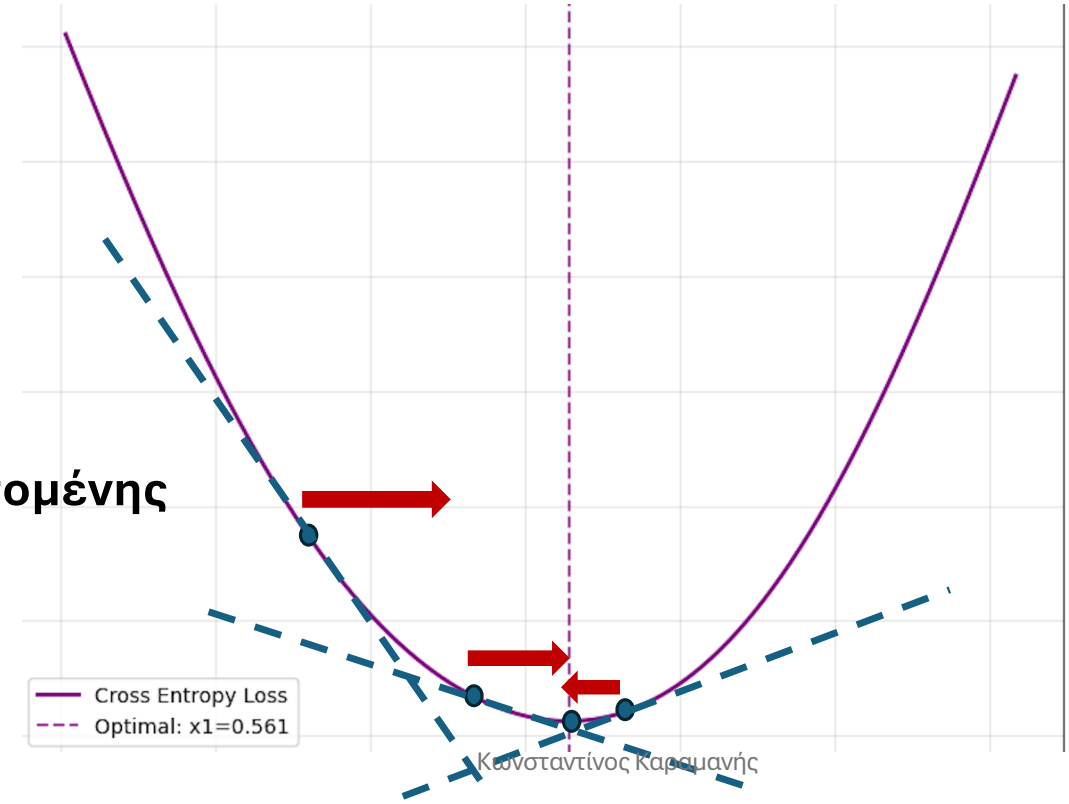
Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

κλίση της εφαπτομένης
= η **παράγωγος**



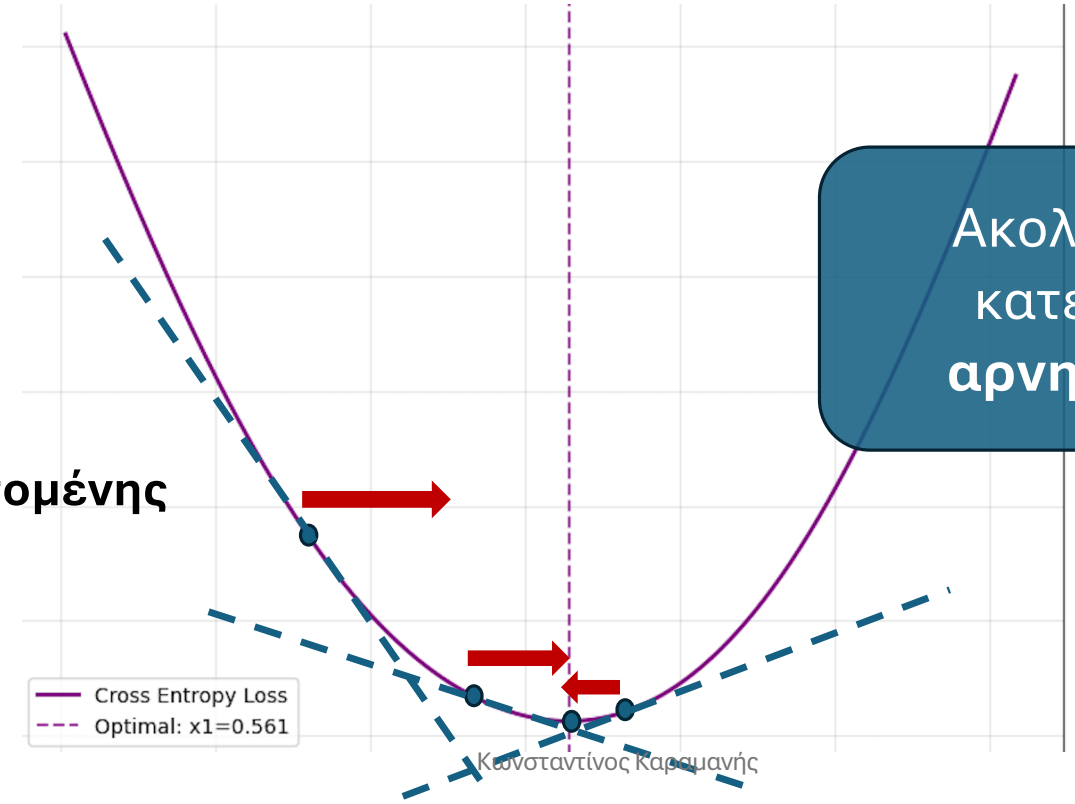
Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

κλίση της εφαπτομένης
= η **παράγωγος**



Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

κλίση της εφαπτομένης
= η **παράγωγος**



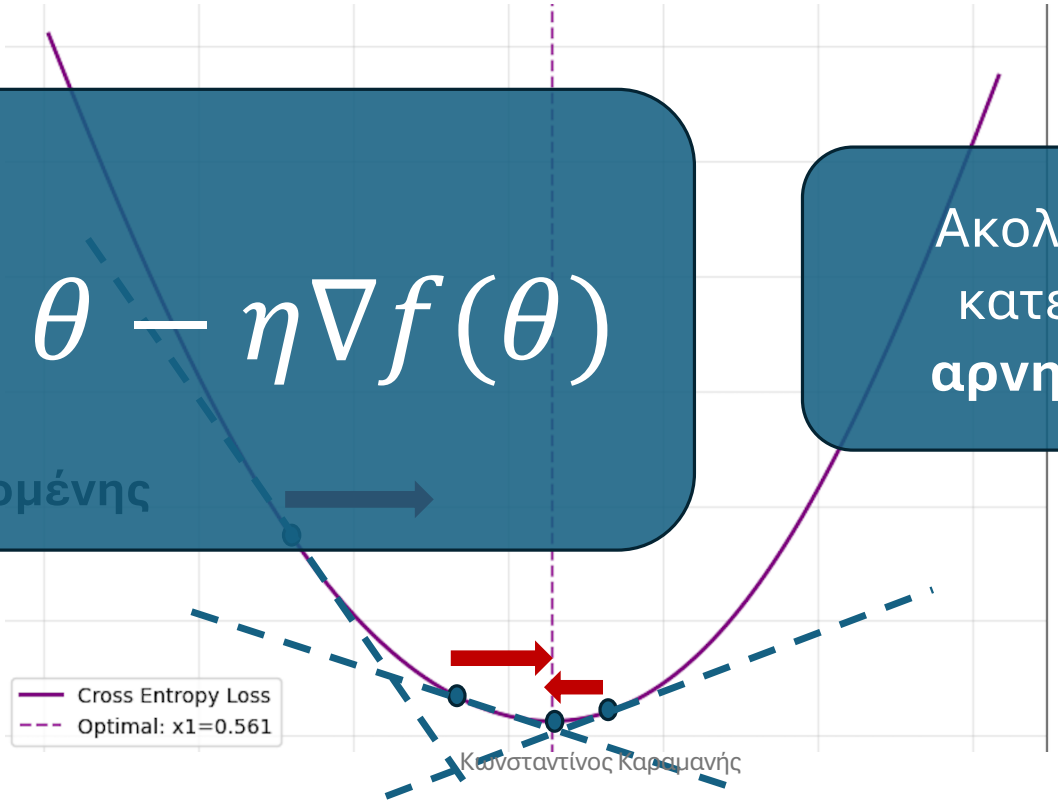
Ακολουθούμε την κατεύθυνση της **αρνητικής κλίσης**

Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

$$\theta_+ = \theta - \eta \nabla f(\theta)$$

κλίση της εφαπτομένης
= η **παράγωγος**

Ακολουθούμε την κατεύθυνση της **αρνητικής κλίσης**



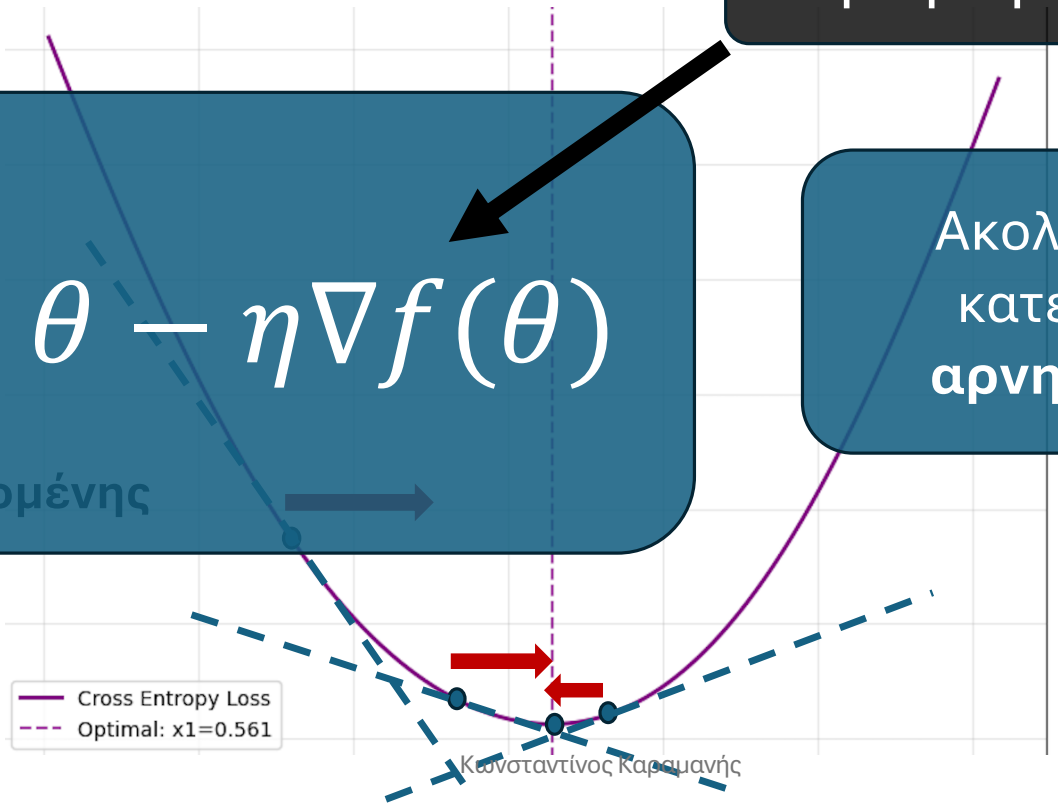
Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

Παράγωγος

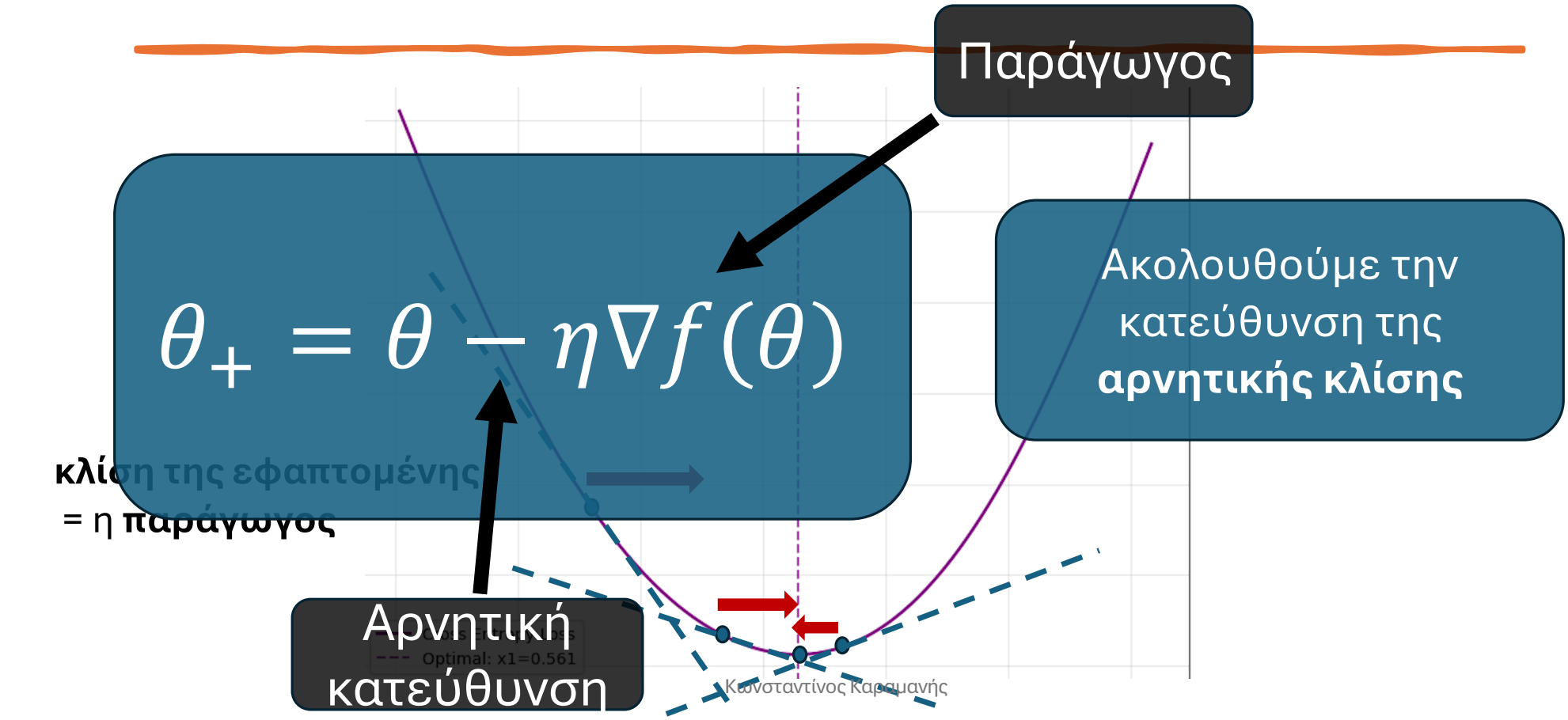
$$\theta_+ = \theta - \eta \nabla f(\theta)$$

Ακολουθούμε την κατεύθυνση της **αρνητικής κλίσης**

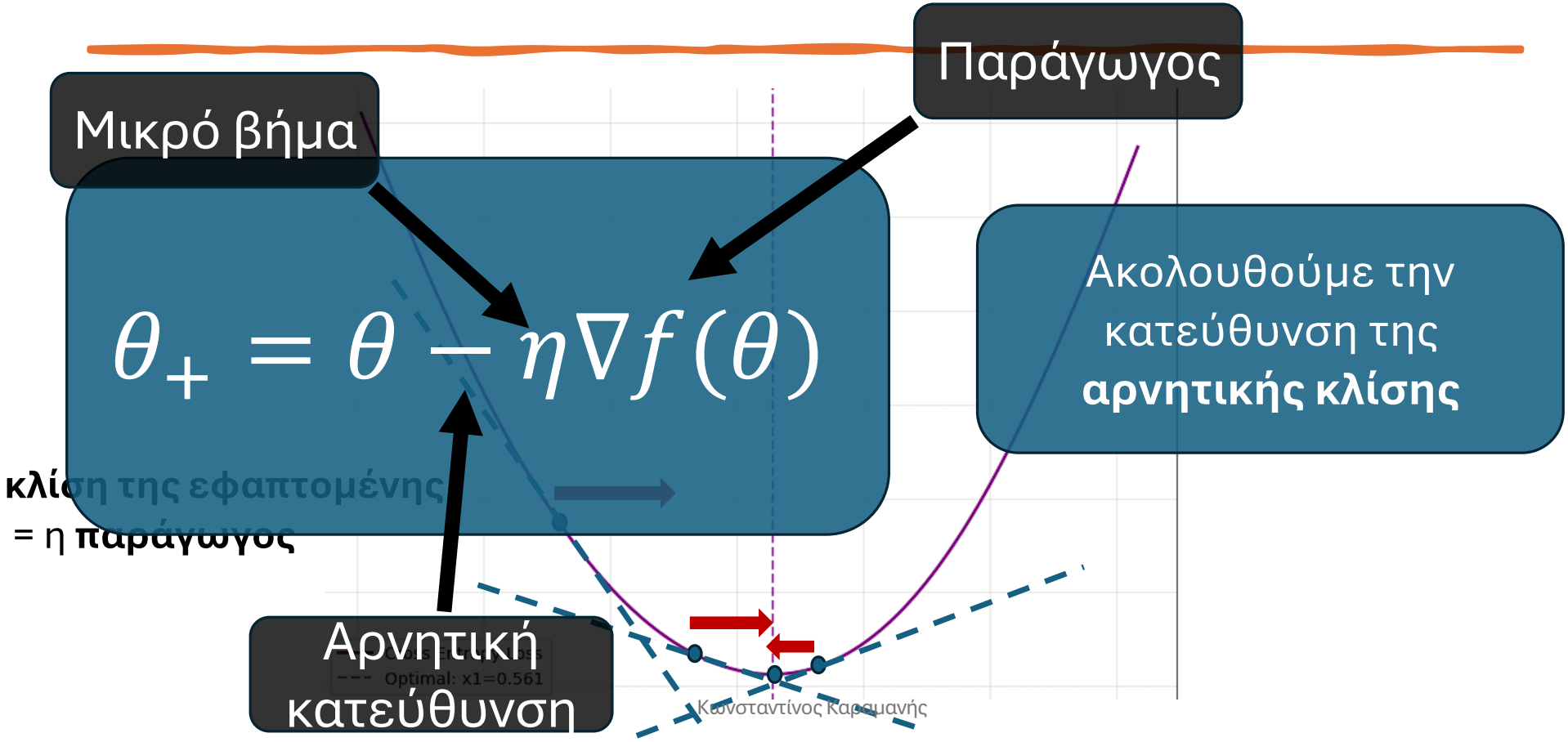
κλίση της εφαπτομένης
= η **παράγωγος**



Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

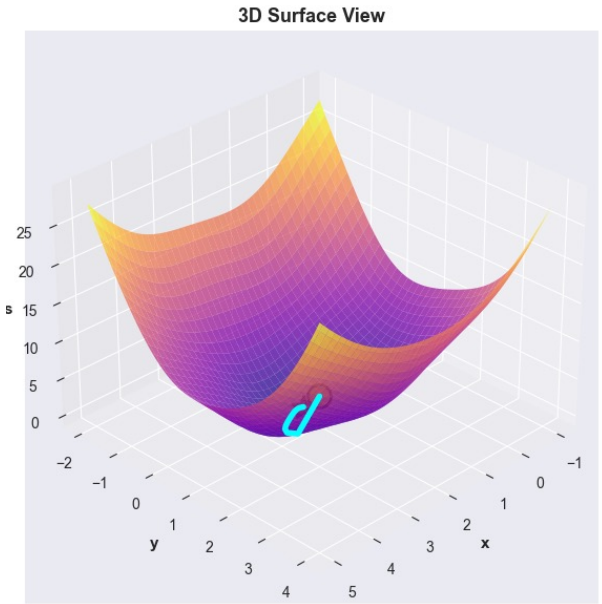


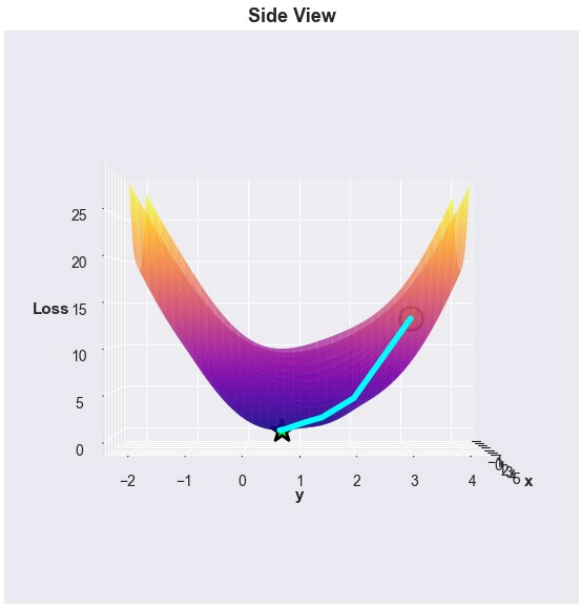
Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

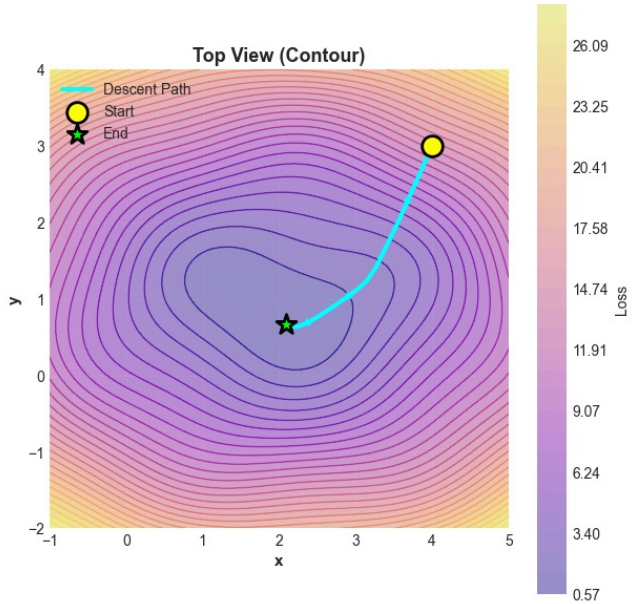


Μέθοδος: Gradient Descent (Κάθοδος Κλίσης)

Ακολουθούμε την κατεύθυνση της αρνητικής κλίσης







Gradient Descent – Παραδείγματα (Α)

$f(x) = 2x^2 - 6x + 3$, με $x = 1$

Ποια αξία του x μειώνει την τιμή $f(x)$;

(α) $x = 1$: είναι ήδη η βέλτιστη τιμή

(β) $x = 1.1$: δίνει μικρότερη αξία. $f(1.1) < f(1)$

(γ) $x = 0.9$: δίνει μικρότερη αξία. $f(0.9) < f(1)$

Gradient Descent – Παραδείγματα (A)

$f(x) = 2x^2 - 6x + 3$, με $x = 1$

Ποια αξία του x μειώνει την τιμή $f(x)$;

(α) $x = 1$: είναι ήδη η βέλτιστη τιμή

(β) $x = 1.1$: δίνει μικρότερη αξία. $f(1.1) < f(1)$

(γ) $x = 0.9$: δίνει μικρότερη αξία. $f(0.9) < f(1)$

$$f(x) = 2x^2 - 6x + 3$$

$$\nabla f(x) = 4x - 6 = -2$$

Gradient Descent – Παραδείγματα (Α)

$f(x) = 2x^2 - 6x + 3$, με $x = 1$

Ποια αξία του x μειώνει την τιμή $f(x)$;

(α) $x = 1$: είναι ήδη η βέλτιστη τιμή

(β) $x = 1.1$: δίνει μικρότερη αξία. $f(1.1) < f(1)$

(γ) $x = 0.9$: δίνει μικρότερη αξία. $f(0.9) < f(1)$

$f(x) = 2x^2 - 6x + 3$

$\nabla f(x) = 4x - 6 = -2$

Οπότε: $x_+ = x \ - \eta(-2)$

$f(1) = -1$

$f(1.1) = -1.18$

$f(0.9) = -0.78$

Gradient Descent – Παραδείγματα (Β)

$f(x) = 4x^2 - x - 5$, με $x = 1$

Ποια αξία του x μειώνει την τιμή $f(x)$;

(α) $x = 1$: είναι ήδη η βέλτιστη τιμή

(β) $x = 1.1$: δίνει μικρότερη αξία. $f(1.1) < f(1)$

(γ) $x = 0.9$: δίνει μικρότερη αξία. $f(0.9) < f(1)$

Gradient Descent – Παραδείγματα (Β)

$f(x) = 4x^2 - x - 5$, με $x = 1$

Ποια αξία του x μειώνει την τιμή $f(x)$;

(α) $x = 1$: είναι ήδη η βέλτιστη τιμή

(β) $x = 1.1$: δίνει μικρότερη αξία. $f(1.1) < f(1)$

(γ) $x = 0.9$: δίνει μικρότερη αξία. $f(0.9) < f(1)$

$f(x) = 4x^2 - x - 5$

$\nabla f(x) = 8x - 1 = 7$

Οπότε: $x_+ = x\ - \eta(7)$

$f(1) = -2$

$f(1.1) = -1.26$

$f(0.9) = -2.66$

Gradient Descent – Παραδείγματα (Γ)

$f(x) = 3x^2 - 6x + 1$, με $x = 1$

Ποια αξία του x μειώνει την τιμή $f(x)$;

(α) $x = 1$: είναι ήδη η βέλτιστη τιμή

(β) $x = 1.1$: δίνει μικρότερη αξία. $f(1.1) < f(1)$

(γ) $x = 0.9$: δίνει μικρότερη αξία. $f(0.9) < f(1)$

Gradient Descent – Παραδείγματα (Γ)

$f(x) = 3x^2 - 6x + 1$, με $x = 1$

Ποια αξία του x μειώνει την τιμή $f(x)$;

(α) $x = 1$: είναι ήδη η βέλτιστη τιμή

(β) $x = 1.1$: δίνει μικρότερη αξία. $f(1.1) < f(1)$

(γ) $x = 0.9$: δίνει μικρότερη αξία. $f(0.9) < f(1)$

$f(x) = 3x^2 - 6x + 1$

$\nabla f(x) = 6x - 6 = 0$

Οπότε: $x_+ = x\ - \eta(0)$

$f(1) = -2$

$f(1.1) = -1.97$

$f(0.9) = -1.97$

Gradient Descent – Παραδείγματα (Δ)

$f(x, y) = x^2 + 3y^2 - 2x + y$, με $x = 1, y = 1$

Ποια αξία του (x, y) μειώνει την τιμή $f(x, y)$;

(α) $(x, y) = (1,1)$: είναι ήδη η βέλτιστη τιμή

(β) $(x, y) = (0.9,1.1)$: δίνει μικρότερη αξία.

(γ) $(x, y) = (1,0.9)$: δίνει μικρότερη αξία.

Gradient Descent – Παραδείγματα (Δ)

$f(x,y) = x^2 + 3y^2 - 2x + y$, με $x = 1, y = 1$

Ποια αξία του (x,y) μειώνει την τιμή $f(x,y)$;

(α) $(x,y) = (1,1)$: είναι ήδη η βέλτιστη τιμή

(β) $(x,y) = (0.9,1.1)$: δίνει μικρότερη αξία.

(γ) $(x,y) = (1,0.9)$: δίνει μικρότερη αξία.

$f(x,y) = x^2 + 3y^2 - 2x + y$

$\nabla f(x) = \begin{pmatrix} 2x-2 \\ 6y+1 \end{pmatrix} = \begin{pmatrix} 0 \\ 7 \end{pmatrix}$

Οπότε: $\begin{pmatrix} x_+ \\ y_+ \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} - \eta \begin{pmatrix} 0 \\ 7 \end{pmatrix}$

Gradient Descent – Παραδείγματα (Δ)

$f(x,y) = x^2 + 3y^2 - 2x + y$, με $x = 1, y = 1$

Ποια αξία του (x,y) μειώνει την τιμή $f(x,y)$;

(α) $(x,y) = (1,1)$: είναι ήδη η βέλτιστη τιμή

(β) $(x,y) = (0.9,1.1)$: δίνει μικρότερη αξία.

(γ) $(x,y) = (1,0.9)$: δίνει μικρότερη αξία.

$f(x,y) = x^2 + 3y^2 - 2x + y$

$\nabla f(x) = \begin{pmatrix} 2x-2 \\ 6y+1 \end{pmatrix} = \begin{pmatrix} 0 \\ 7 \end{pmatrix}$

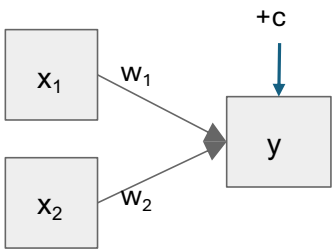
Οπότε: $\begin{pmatrix} x_+ \\ y_+ \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} - \eta \begin{pmatrix} 0 \\ 7 \end{pmatrix}$

$f(1,1) = 3$

$f(0.9,1.1) = 3.74$

$f(1,0.9) = 2.33$

Παράμετροι, Συνάρτηση Απώλειας + Gradient Descent



Οικογένεια μοντέλων: ένα γραμμικό νευρωνικό δίκτυο με παραμέτρους:

$$\theta = (w_1, w_2, c)$$

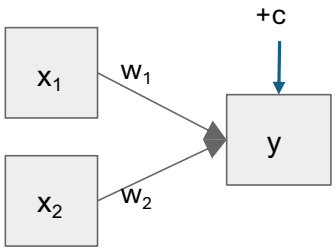
Συνάρτηση απώλειας: $L(y, \hat{y}) = (y - \hat{y})^2$

Απώλεια σαν συνάρτηση των παραμέτρων:

$$\hat{y} = w_1 x_1 + w_2 x_2 + c$$

$$\boldsymbol{L(\theta) = L(y, \hat{y}) = (y - \hat{y})^2 = (y - (w_1 x_1 + w_2 x_2 + c))^2}$$

Παράμετροι, Συνάρτηση Απώλειας + Gradient Descent



X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

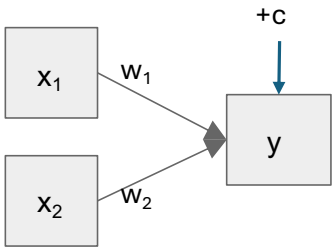
$$L(y,\hat{y}) = (y - \hat{y})^2 = (y - (w_1x_1 + w_2x_2 + c))^2$$

Συνολική απώλεια

$$\sum_i L(y_i,\hat{y}_i) = \sum_i (y_i - (w_1x_{i1} + w_2xi_2 + c))^2$$

$$\boldsymbol{L(\theta) = (0 - c)^2 + \left(1 - (w_2 + c)\right)^2 + \left(1 - (w_1 + c)\right)^2}$$
$$\boldsymbol{+ \left(0 - (w_1 + w_2 + c)\right)^2}$$

Παράμετροι, Συνάρτηση Απώλειας + Gradient Descent



$$L(y, \hat{y}) = (y - \hat{y})^2 = (y - (w_1 x_1 + w_2 x_2 + c))^2$$

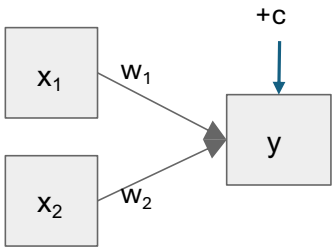
Συνολική απώλεια

$$\sum_i L(y_i, \hat{y}_i) = \sum_i (y_i - (w_1 x_{i1} + w_2 x i_2 + c))^2$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

$$\boldsymbol{L(\theta) = (0 - c)^2 + (1 - (w_2 + c))^2 + (1 - (w_1 + c))^2 + (0 - (w_1 + w_2 + c))^2}$$

Παράμετροι, Συνάρτηση Απώλειας + Gradient Descent



$$L(y, \hat{y}) = (y - \hat{y})^2 = (y - (w_1 x_1 + w_2 x_2 + c))^2$$

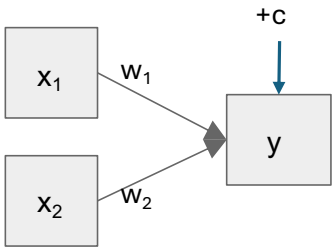
Συνολική απώλεια

$$\sum_i L(y_i, \hat{y}_i) = \sum_i (y_i - (w_1 x_{i1} + w_2 x i_2 + c))^2$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

$$\boldsymbol{L(\theta) = (0 - c)^2 + \left(1 - (w_2 + c)\right)^2 + \left(1 - (w_1 + c)\right)^2 + \left(0 - (w_1 + w_2 + c)\right)^2}$$

Παράμετροι, Συνάρτηση Απώλειας + Gradient Descent



$$L(y, \hat{y}) = (y - \hat{y})^2 = (y - (w_1 x_1 + w_2 x_2 + c))^2$$

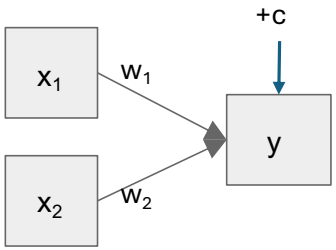
Συνολική απώλεια

$$\sum_i L(y_i, \hat{y}_i) = \sum_i (y_i - (w_1 x_{i1} + w_2 x i_2 + c))^2$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

$$\boldsymbol{L(\theta) = (0 - c)^2 + \left(1 - (w_2 + c)\right)^2 + \left(1 - (w_1 + c)\right)^2 + \left(0 - (w_1 + w_2 + c)\right)^2}$$

Παράμετροι, Συνάρτηση Απώλειας + Gradient Descent



X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

$$L(y,\hat{y}) = (y - \hat{y})^2 = (y - (w_1x_1 + w_2x_2 + c))^2$$

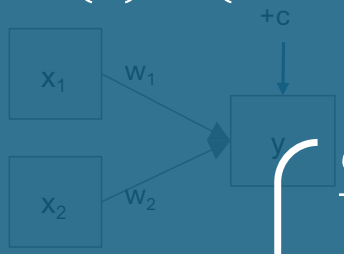
Συνολική απώλεια

$$\sum_i L(y_i,\hat{y}_i) = \sum_i (y_i - (w_1x_{i1} + w_2xi_2 + c))^2$$

$$\boldsymbol{L(\theta) = (0 - c)^2 + \left(1 - (w_2 + c)\right)^2 + \left(1 - (w_1 + c)\right)^2 + \left(0 - (w_1 + w_2 + c)\right)^2}$$

Απώλεια και η Παράγωγος της Απώλειας

$$L(\theta) = (0 - c)^2 + \big(1 - (w_2 + c)\big)^2 + \big(1 - (w_1 + c)\big)^2 + \big(0 - (w_1 + w_2 + c)\big)^2$$

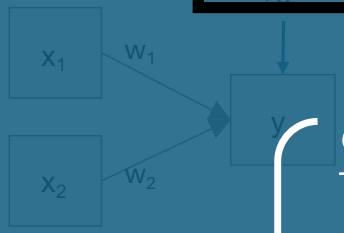


$$\nabla_\theta L(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \frac{\partial L(\theta)}{\partial\, w_2} \\ \frac{\partial L(\theta)}{\partial c} \end{pmatrix} = \begin{pmatrix} 4w_1 + 2w_2 + 4c - 2 \\ 2w_1 + 4w_2 + 4c - 2 \\ 4w_1 + 4w_2 + 8c - 4 \end{pmatrix}$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Απώλεια και η Παράγωγος της Απώλειας

$$L(\theta) = (0 - c)^2 + \big(1 - (w_2 + c)\big)^2 + \big(1 - (w_1 + c)\big)^2 + \big(0 - (w_1 + w_2 + c)\big)^2$$



$$\nabla_\theta L(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \frac{\partial L(\theta)}{\partial w_2} \\ \frac{\partial L(\theta)}{\partial c} \end{pmatrix} = \begin{pmatrix} 4w_1 + 2w_2 + 4c - 2 \\ 2w_1 + 4w_2 + 4c - 2 \\ 4w_1 + 4w_2 + 8c - 4 \end{pmatrix}$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Απώλεια και η Παράγωγος της Απώλειας

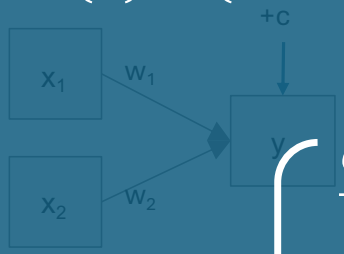
$$L(\theta) = (0 - c)^2 + \big(1 - (w_2 + c)\big)^2 + \big(1 - (w_1 + c)\big)^2 + \big(0 - (w_1 + w_2 + c)\big)^2$$

$$\nabla_\theta L(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \frac{\partial L(\theta)}{\partial w_2} \\ \frac{\partial L(\theta)}{\partial c} \end{pmatrix} = \begin{pmatrix} 4w_1 + 2w_2 + 4c - 2 \\ 2w_1 + 4w_2 + 4c - 2 \\ 4w_1 + 4w_2 + 8c - 4 \end{pmatrix}$$

0

Απώλεια και η Παράγωγος της Απώλειας

$$L(\theta) = (0 - c)^2 + \big(1 - (w_2 + c)\big)^2 + \big(1 - (w_1 + c)\big)^2 + \big(0 - (w_1 + w_2 + c)\big)^2$$



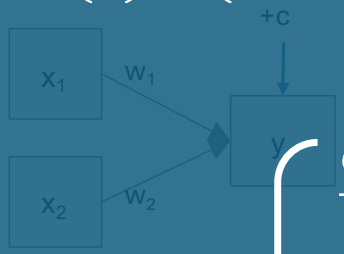
$$\nabla_\theta L(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \frac{\partial L(\theta)}{\partial w_2} \\ \frac{\partial L(\theta)}{\partial c} \end{pmatrix} = \begin{pmatrix} 4w_1 + 2w_2 + 4c - 2 \\ 2w_1 + 4w_2 + 4c - 2 \\ 4w_1 + 4w_2 + 8c - 4 \end{pmatrix}$$

$$0 + -2(1 - (w_2 + c))$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Απώλεια και η Παράγωγος της Απώλειας

$$L(\theta) = (0 - c)^2 + \big(1 - (w_2 + c)\big)^2 + \big(1 - (w_1 + c)\big)^2 + \big(0 - (w_1 + w_2 + c)\big)^2$$



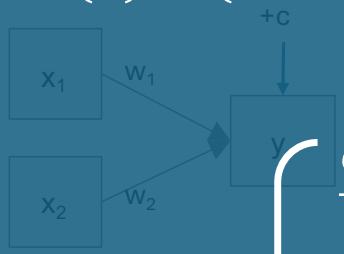
$$\nabla_\theta L(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \frac{\partial L(\theta)}{\partial w_2} \\ \frac{\partial L(\theta)}{\partial c} \end{pmatrix} = \begin{pmatrix} 4w_1 + 2w_2 + 4c - 2 \\ 2w_1 + 4w_2 + 4c - 2 \\ 4w_1 + 4w_2 + 8c - 4 \end{pmatrix}$$

$$0 + -2(1 - (w_2 + c)) + 0$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Απώλεια και η Παράγωγος της Απώλειας

$$L(\theta) = (0 - c)^2 + \big(1 - (w_2 + c)\big)^2 + \big(1 - (w_1 + c)\big)^2 + \big(0 - (w_1 + w_2 + c)\big)^2$$



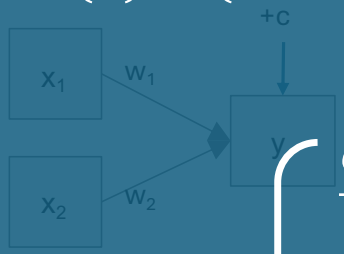
$$\nabla_\theta L(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \frac{\partial L(\theta)}{\partial w_2} \\ \frac{\partial L(\theta)}{\partial c} \end{pmatrix} = \begin{pmatrix} 4w_1 + 2w_2 + 4c - 2 \\ 2w_1 + 4w_2 + 4c - 2 \\ 4w_1 + 4w_2 + 8c - 4 \end{pmatrix}$$

$$\begin{matrix} 0 \\ + \\ -2(1 - (w_2 + c)) \\ + \\ 0 \\ + \\ -2(0 - (w_1 + w_2 + c)) \end{matrix}$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Απώλεια και η Παράγωγος της Απώλειας

$$L(\theta) = (0 - c)^2 + \big(1 - (w_2 + c)\big)^2 + \big(1 - (w_1 + c)\big)^2 + \big(0 - (w_1 + w_2 + c)\big)^2$$



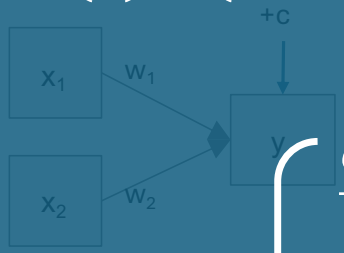
$$\nabla_\theta L(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \frac{\partial L(\theta)}{\partial\, w_2} \\ \frac{\partial L(\theta)}{\partial c} \end{pmatrix} = \begin{pmatrix} 4w_1 + 2w_2 + 4c - 2 \\ 2w_1 + 4w_2 + 4c - 2 \\ 4w_1 + 4w_2 + 8c - 4 \end{pmatrix}$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Απώλεια και η Παράγωγος της Απώλειας

Αρχική τιμή παραμέτρων: $\theta = (w_1, w_2, c) = (0,1,1)$

$$L(\theta) = (0 - c)^2 + \big(1 - (w_2 + c)\big)^2 + \big(1 - (w_1 + c)\big)^2 + \big(0 - (w_1 + w_2 + c)\big)^2$$



$$\nabla_\theta L(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \frac{\partial L(\theta)}{\partial w_2} \\ \frac{\partial L(\theta)}{\partial c} \end{pmatrix} \sum_i = \begin{pmatrix} 4w_1 + 2w_2 + 4c - 2 \\ 2w_1 + 4w_2 + 4c - 2 \\ 4w_1 + 4w_2 + 8c - 4 \end{pmatrix}$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Απώλεια και η Παράγωγος της Απώλειας

Αρχική τιμή παραμέτρων: $\theta = (w_1, w_2, c) = (0,1,1)$

$$L(\theta) = (0 - c)^2 + \big(1 - (w_2 + c)\big)^2 + \big(1 - (w_1 + c)\big)^2 + \big(0 - (w_1 + w_2 + c)\big)^2$$

$$= 1 + 1 + 0 + 4 = 6$$

$$\nabla_\theta L(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \frac{\partial L(\theta)}{\partial w_2} \\ \frac{\partial L(\theta)}{\partial c} \end{pmatrix} = \begin{pmatrix} 4w_1 + 2w_2 + 4c - 2 \\ 2w_1 + 4w_2 + 4c - 2 \\ 4w_1 + 4w_2 + 8c - 4 \end{pmatrix}$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Απώλεια και η Παράγωγος της Απώλειας

Αρχική τιμή παραμέτρων: $\theta = (w_1, w_2, c) = (0,1,1)$

$$L(\theta) = (0 - c)^2 + \big(1 - (w_2 + c)\big)^2 + \big(1 - (w_1 + c)\big)^2 + \big(0 - (w_1 + w_2 + c)\big)^2$$

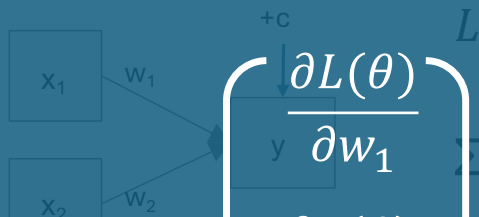
$$= 1 + 1 + 0 + 4 = 6$$

$$\nabla_\theta L(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \frac{\partial L(\theta)}{\partial w_2} \\ \frac{\partial L(\theta)}{\partial c} \end{pmatrix} = \sum_i \begin{pmatrix} 4w_1 + 2w_2 + 4c - 2 \\ 2w_1 + 4w_2 + 4c - 2 \\ 4w_1 + 4w_2 + 8c - 4 \end{pmatrix} = \begin{pmatrix} 4 \\ 6 \\ 8 \end{pmatrix}$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Μέθοδος Gradient Descent

Αρχική τιμή παραμέτρων: $\theta = (w_1, w_2, c) = (0,1,1)$



$$\nabla_\theta L(\theta) = \begin{pmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \frac{\partial L(\theta)}{\partial w_2} \\ \frac{\partial L(\theta)}{\partial c} \end{pmatrix} = \begin{pmatrix} 4 \\ 6 \\ 8 \end{pmatrix}$$

$$\theta^+ = \theta - \eta \nabla_\theta L(\theta) = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} - \eta \begin{pmatrix} 4 \\ 6 \\ 8 \end{pmatrix}$$

X1	X2	Y
0	0	0
0	1	1
1	0	1
1	1	0

Μέθοδος Gradient Descent

Αρχική τιμή παραμέτρων: $\theta = (w_1, w_2, c) = (0,1,1)$

Επιλέγοντας (π.χ.), $\eta = \frac{1}{10}$

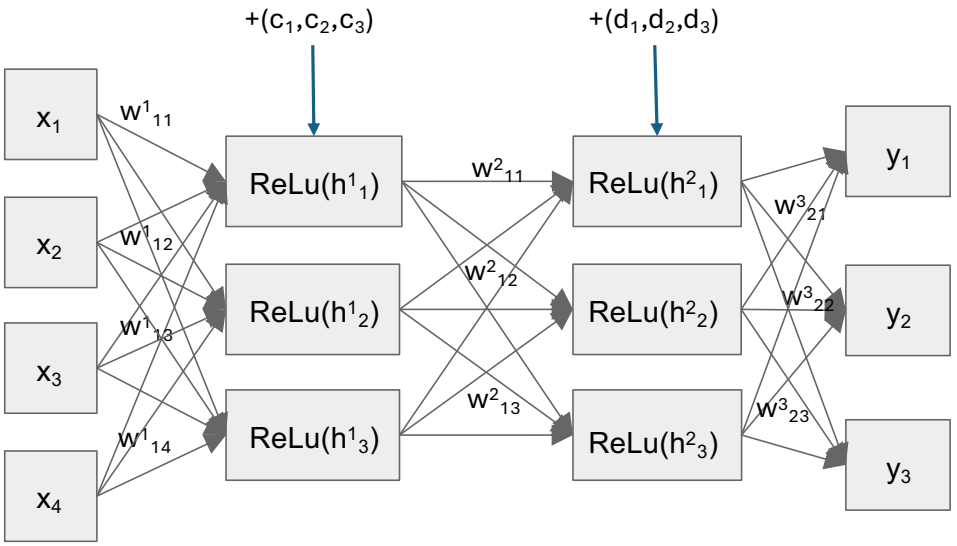
$$\theta^+ = \theta - \eta \nabla_\theta L(\theta) = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} - 0.1 \begin{bmatrix} 4 \\ 6 \\ 8 \end{bmatrix} = \begin{bmatrix} -.4 \\ 0.4 \\ 0.2 \end{bmatrix}$$

$$L(\theta^+) = 0.04 + .16 + 1.44 + .16$$

$$= 1.8 < 6$$

$$\theta^+ = \theta - \eta \nabla_\theta L(\theta) = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} - \eta \begin{bmatrix} 4 \\ 6 \\ 8 \end{bmatrix}$$

Gradient Descent + Νευρωνικά Δίκτυα

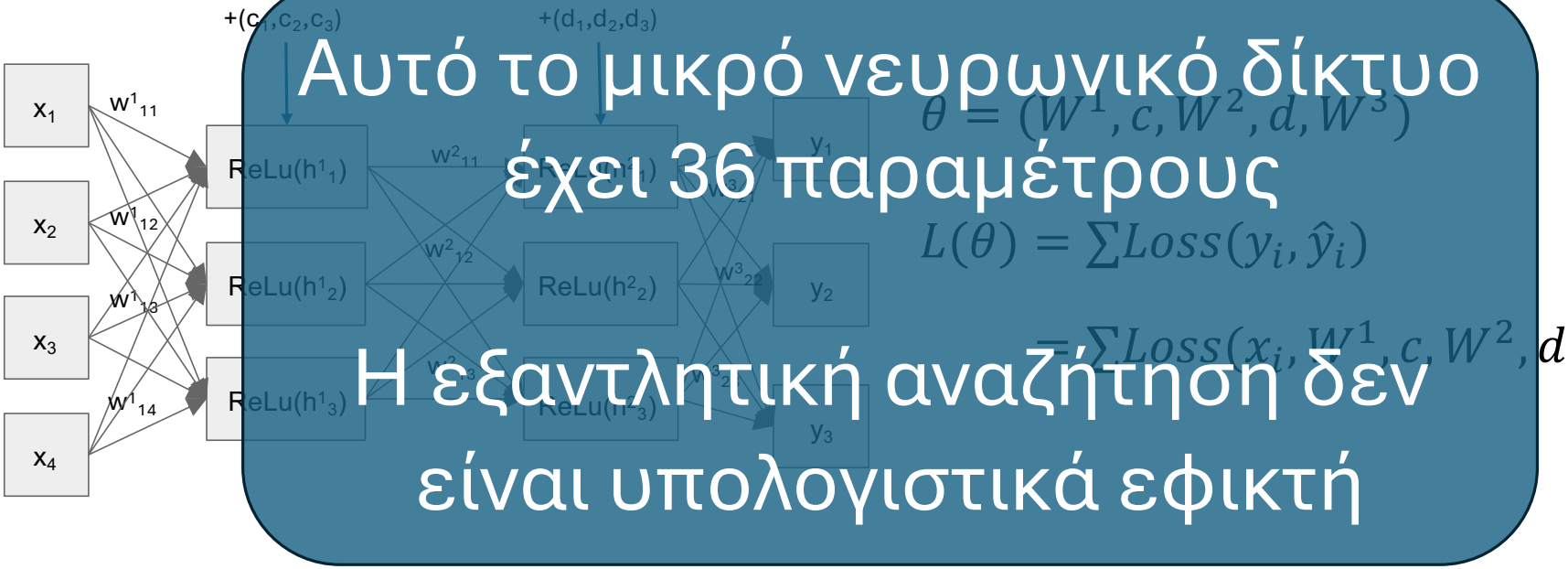


$$\theta = (W^1, c, W^2, d, W^3)$$

$$L(\theta) = \sum Loss(y_i, \hat{y}_i)$$

$$= \sum Loss(x_i, W^1, c, W^2, d, W^3)$$

Gradient Descent + Νευρωνικά Δίκτυα



Αυτό το μικρό νευρωνικό δίκτυο έχει 36 παραμέτρους

$$\theta = (W^1, c, W^2, d, W^3)$$

$$L(\theta) = \sum Loss(y_i, \hat{y}_i)$$

$$= \sum Loss(x_i, W^1, c, W^2, d, W^3)$$

Η εξαντλητική αναζήτηση δεν είναι υπολογιστικά εφικτή

Η Μέθοδος: Gradient Descent

1. Υπολογιστικά αποδοτική μέθοδος ακόμα και με πάρα πολλές παραμέτρους

2. Βρίσκει τοπικό βέλτιστο σημείο (λύση)

3. Εάν η συνάρτηση είναι κυρτή, τότε το τοπικό βέλτιστο σημείο είναι και ολικό βέλτιστο

Η Μέθοδος:
Gradient Descent

1. **Υπολογιστικά αποδοτική μέθοδος ακόμα και με πάρα πολλές παραμέτρους**

2. **Βρίσκει τοπικό βέλτιστο σημείο (λύση)**

3. Εάν η συνάρτηση είναι κυρτή, τότε το τοπικό βέλτιστο σημείο είναι και ολικό βέλτιστο