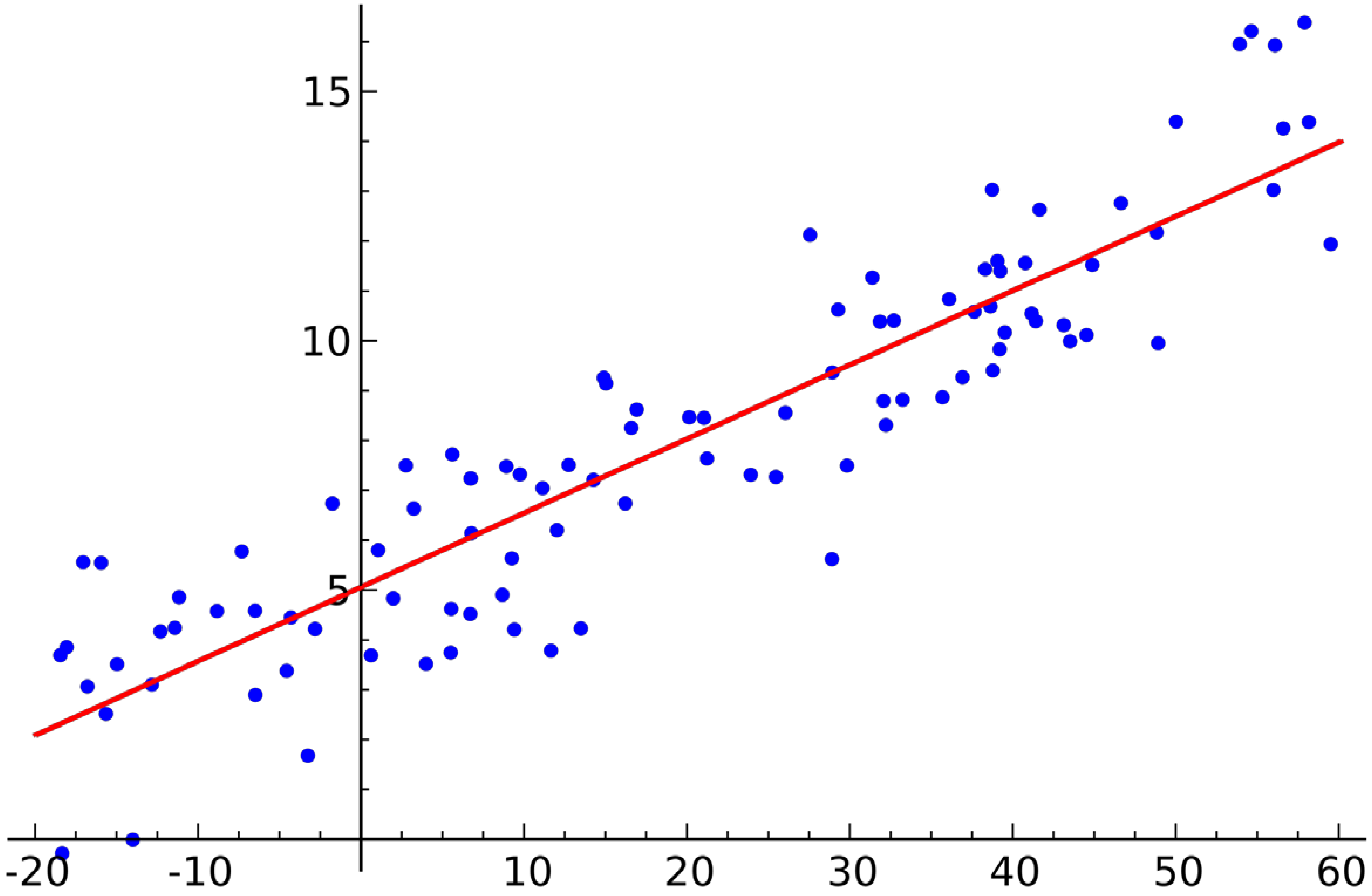
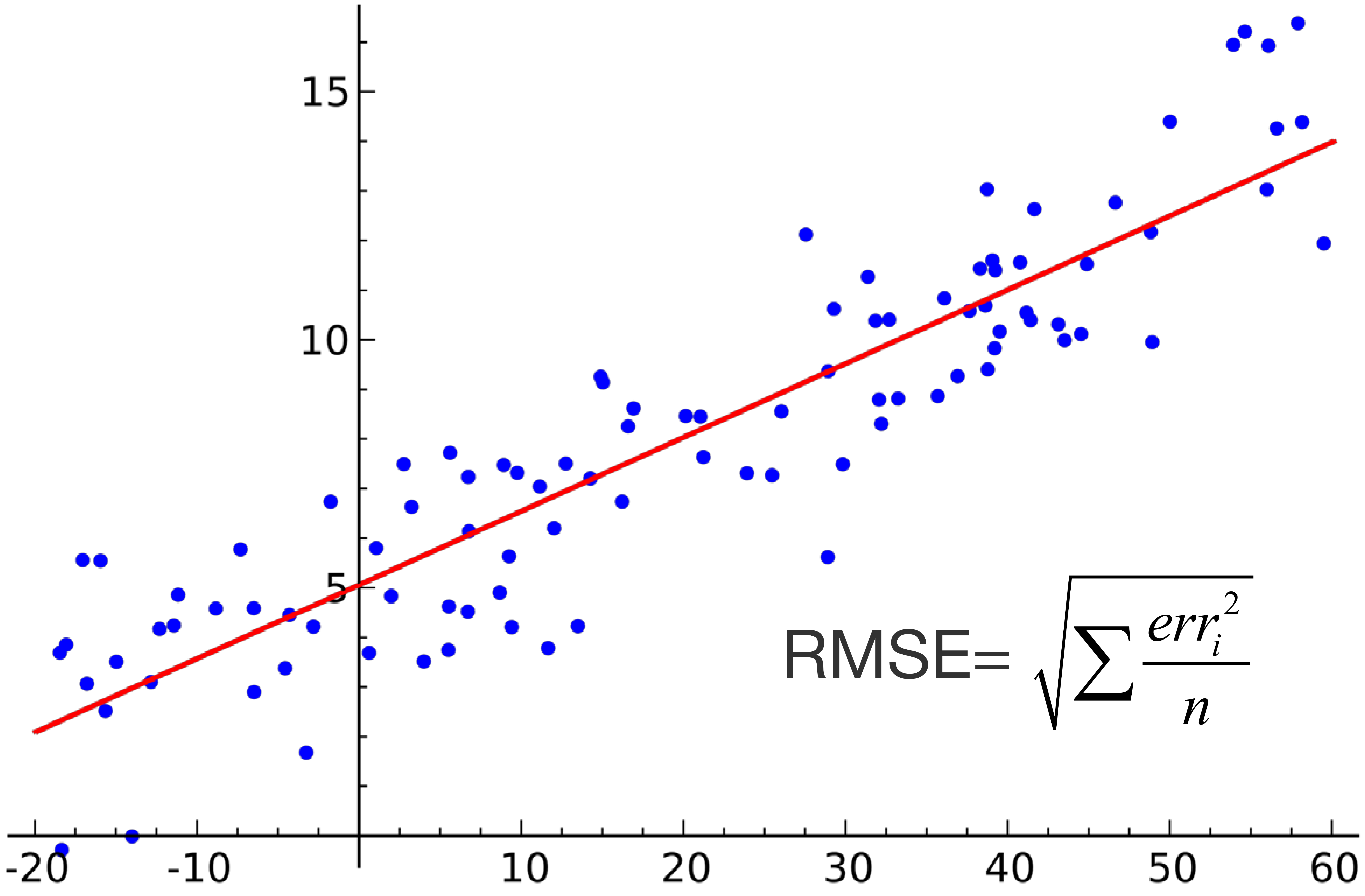
Training models on big data

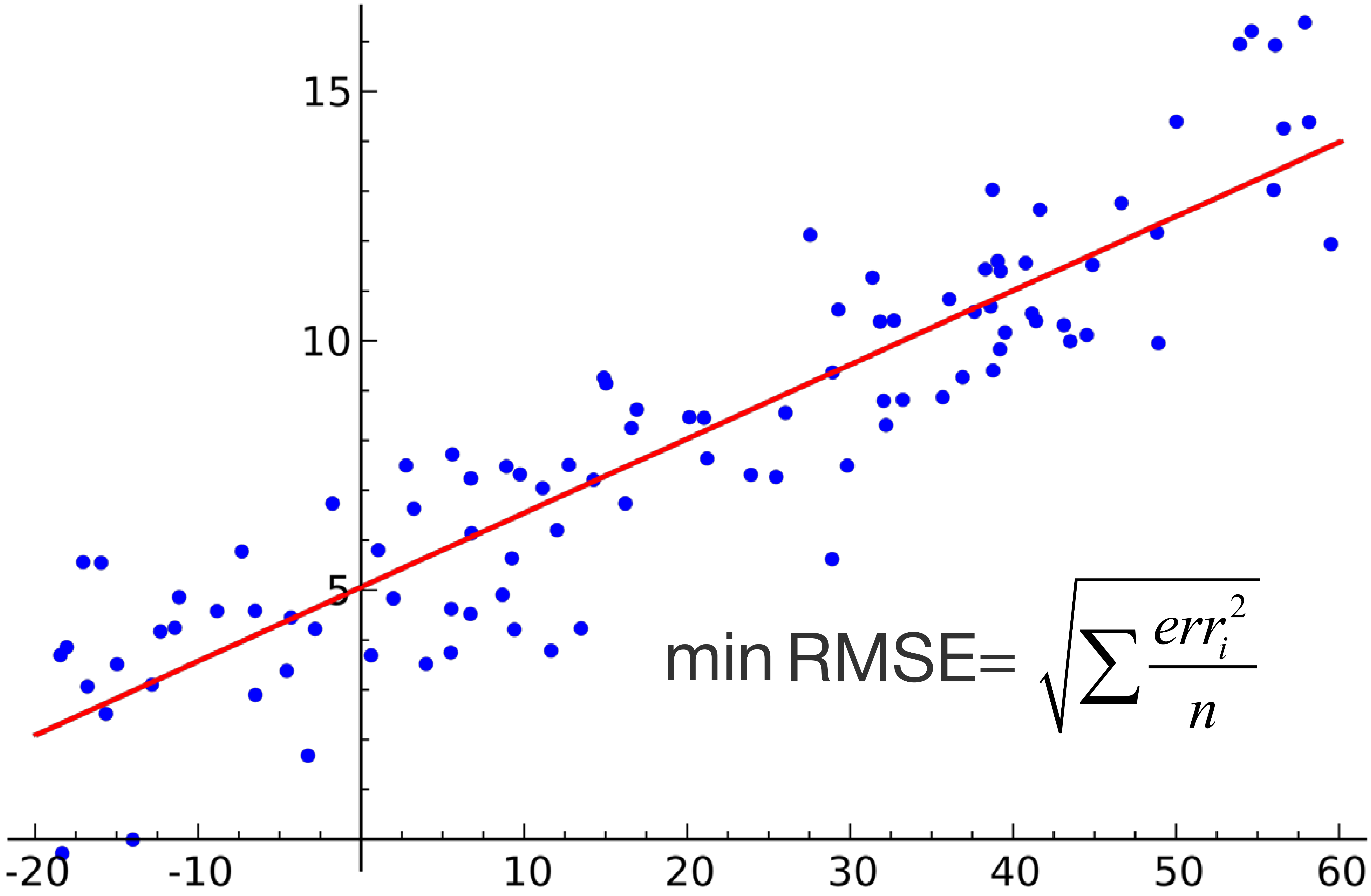
In this lesson you will learn:

- how to train algorithms on big data
- analytical solution
- gradient descend
- stochastic gradient descend


15
10
5
-20
-10
10
20
30
40
50
60



$$\text{RMSE} = \sqrt{\sum \frac{err_i^2}{n}}$$



$$\min \mathrm{RMSE} = \sqrt{\sum \frac{err_i^2}{n}}$$

features - x^i

features - x^i

labels - y^i

features - x^i

$$x^i = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_4 \end{pmatrix}$$

labels - y^i

$$y^i \in R$$

features - x^i

$$x^i = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_4 \end{pmatrix}$$

labels - y^i

$$y^i \in R$$

$$\hat{y} = w_0 + w_1 x_1 + \dots + w_n x_n$$

$w_0, \dots, w_n$ - parameters

$$x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix} \qquad w = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \ldots \\ w_n \end{pmatrix}$$

$$x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix} \qquad w = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \ldots \\ w_n \end{pmatrix}$$

$$\hat{y} = w^T x$$

$$L(w) = \sum_i (y^i - w^T x^i)^2$$

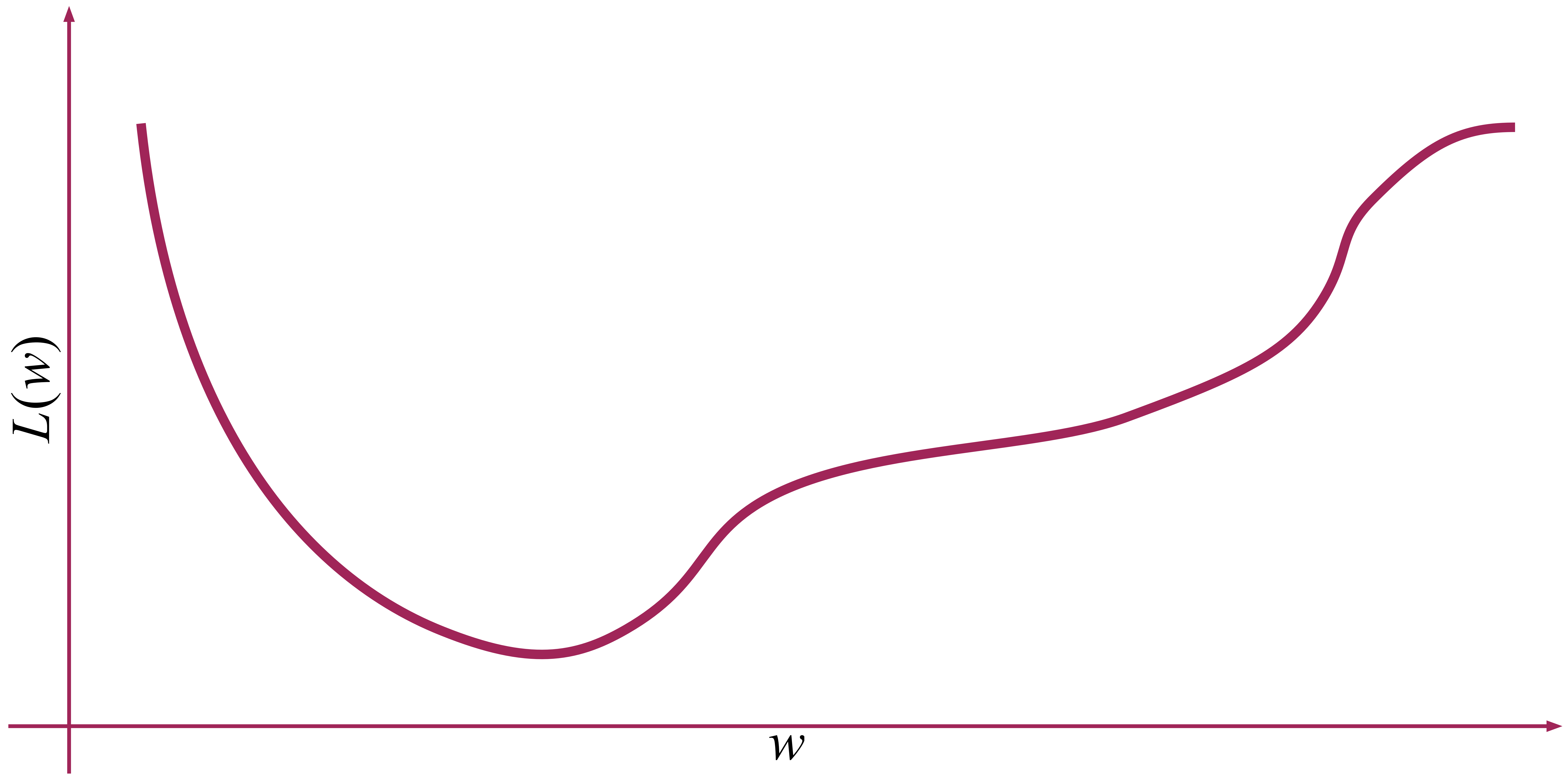
$$w = \arg\min L(w)$$

$$\text{RMSE} = \sqrt{\frac{\sum_i (y^i - w^T x^i)^2}{n}}$$

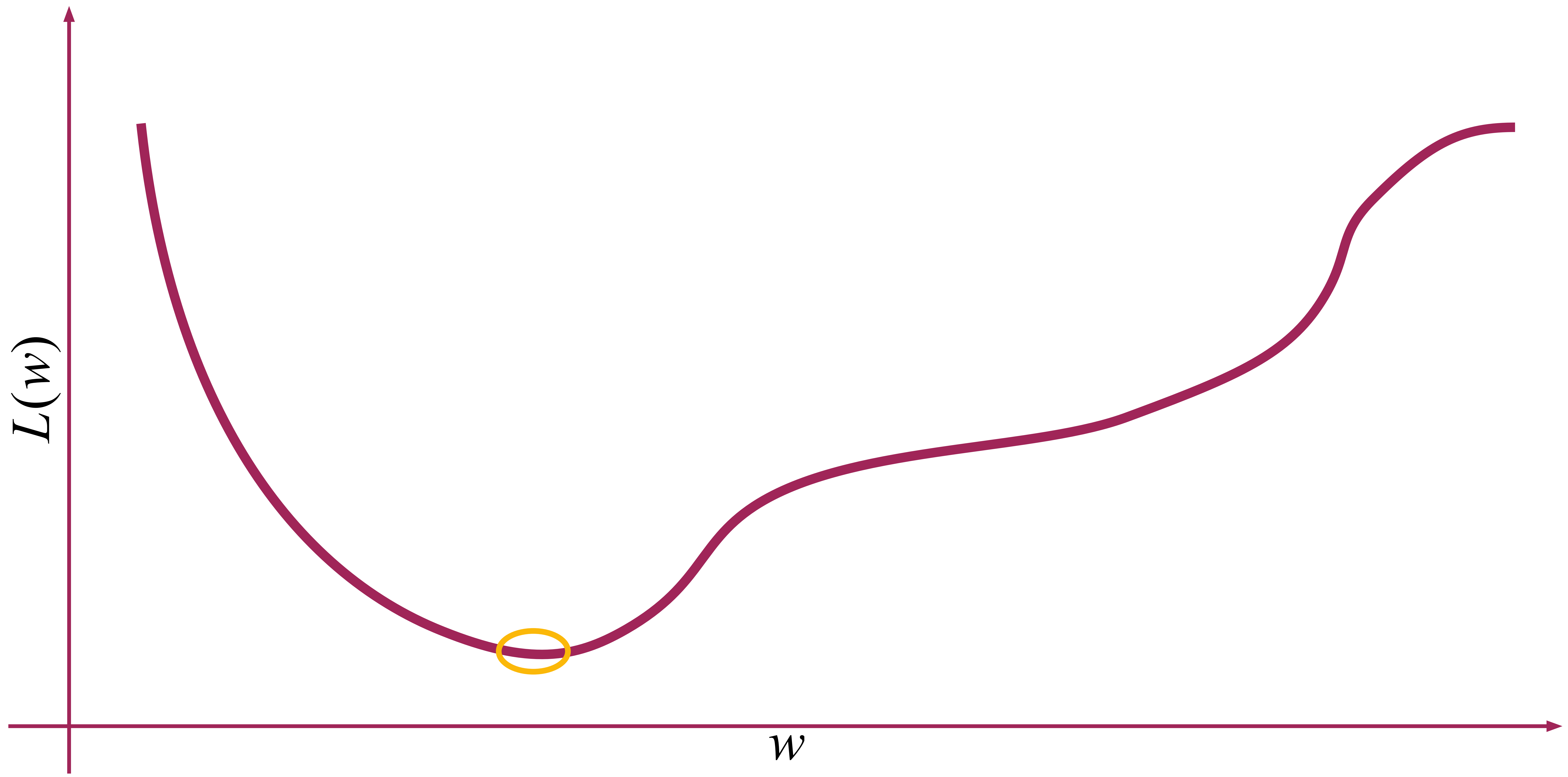
$$L(w) = \sum_i (y^i - w^T x^i)^2$$

$$w = \arg\min L(w)$$

Analytical solution



$L(w)$
w

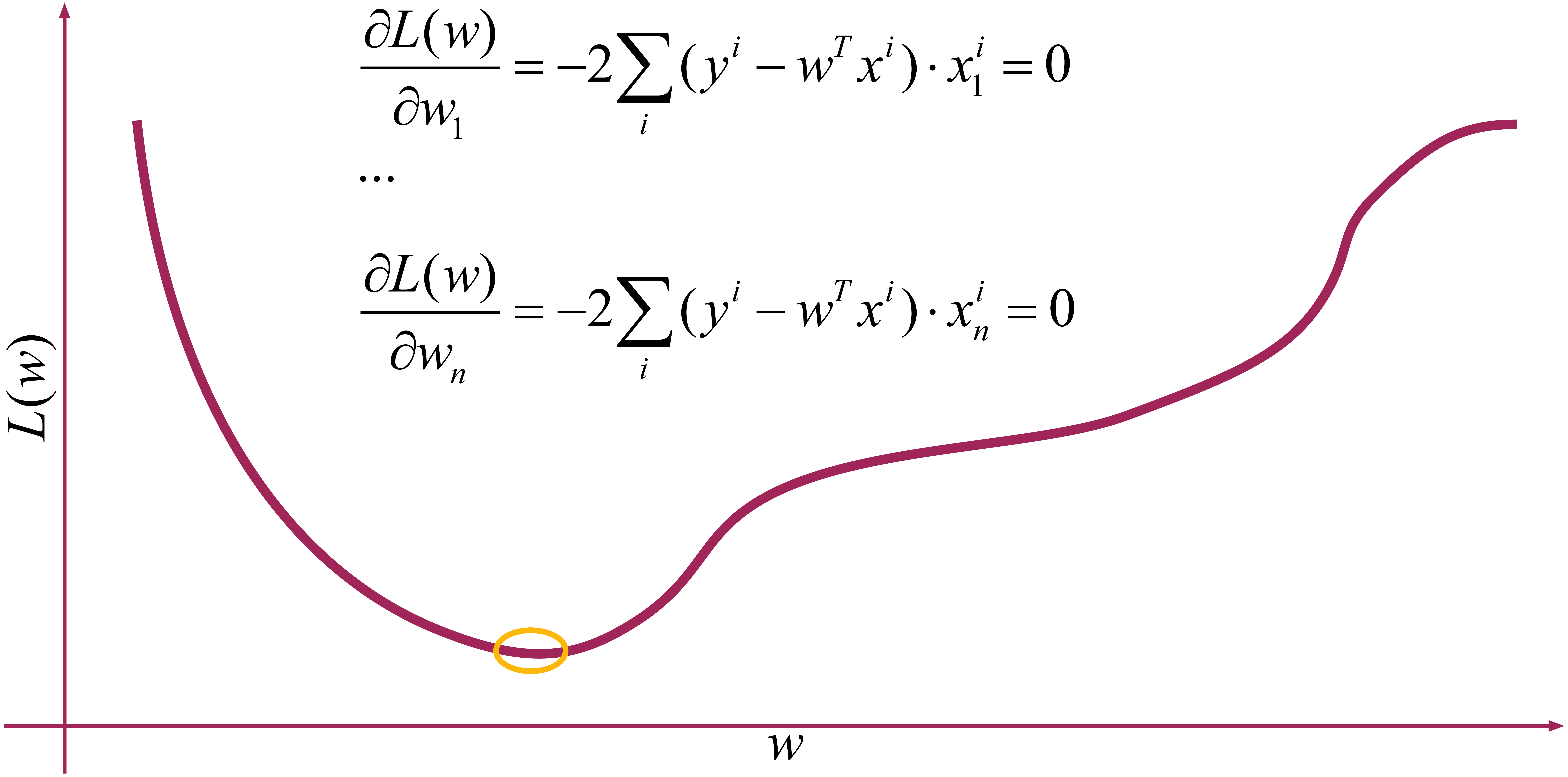


L(w)
w

$$\frac{\partial L(w)}{\partial w_1} = -2\sum_i (y^i - w^T x^i) \cdot x_1^i = 0$$

$$\cdots$$

$$\frac{\partial L(w)}{\partial w_n} = -2\sum_i (y^i - w^T x^i) \cdot x_n^i = 0$$



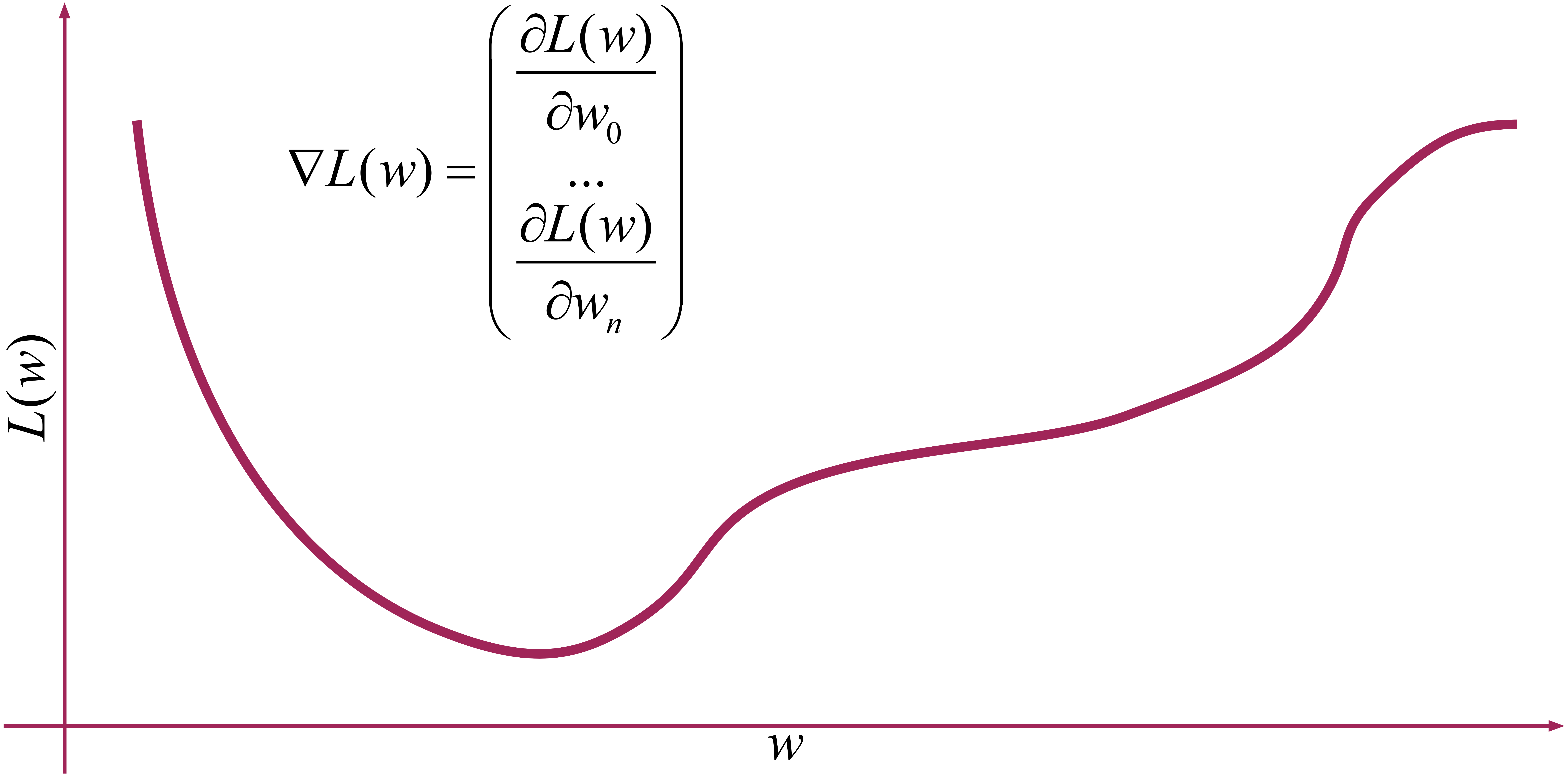
Problem: Complexity - $O(n^3)$

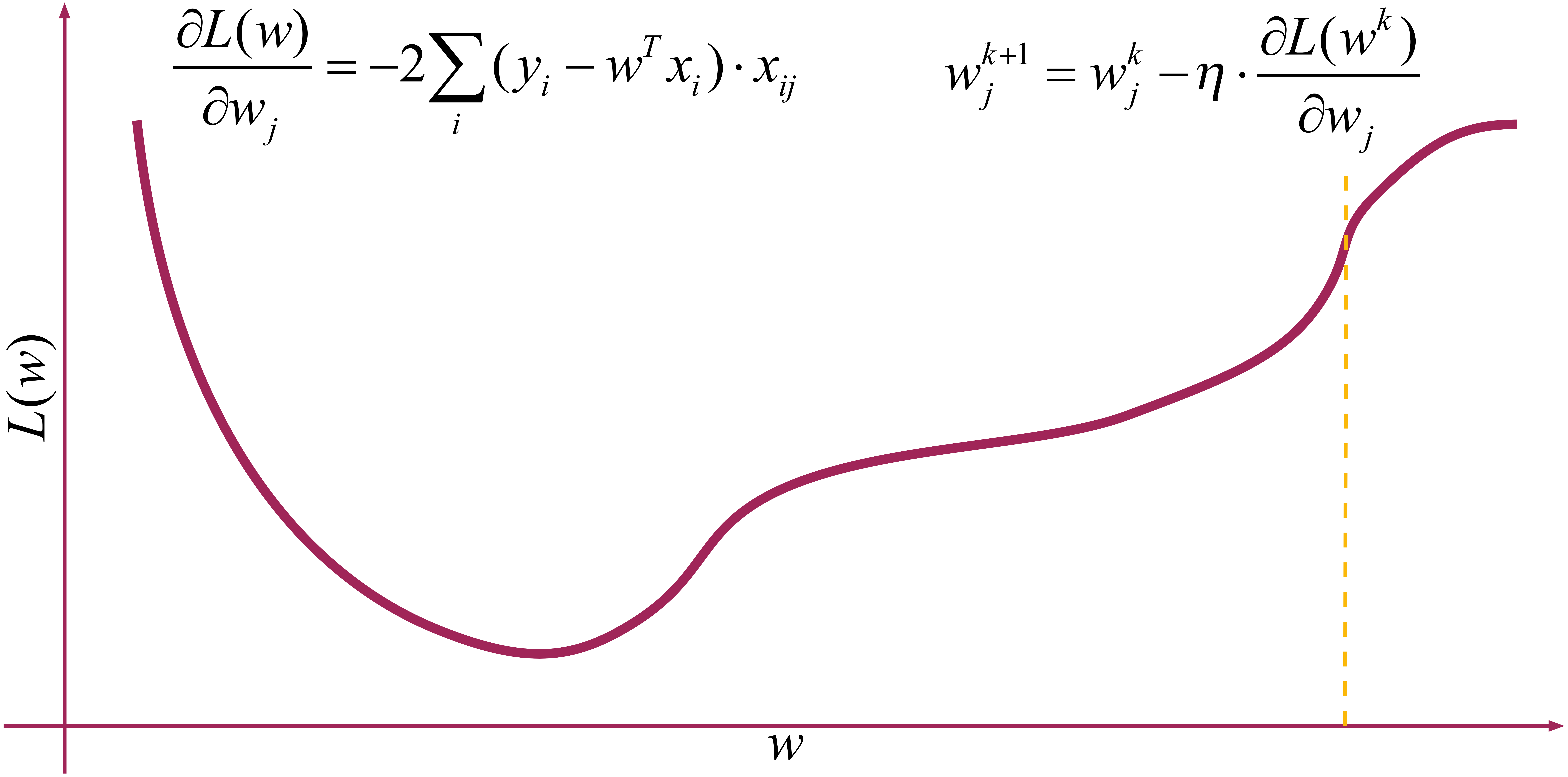
10	1000 operations
100	1000000 operations
1000	1000000000 operations

Problem 2 - method is not universal

Gradient Descend

$$\nabla L(w) = \begin{pmatrix} \dfrac{\partial L(w)}{\partial w_0} \\ \dots \\ \dfrac{\partial L(w)}{\partial w_n} \end{pmatrix}$$



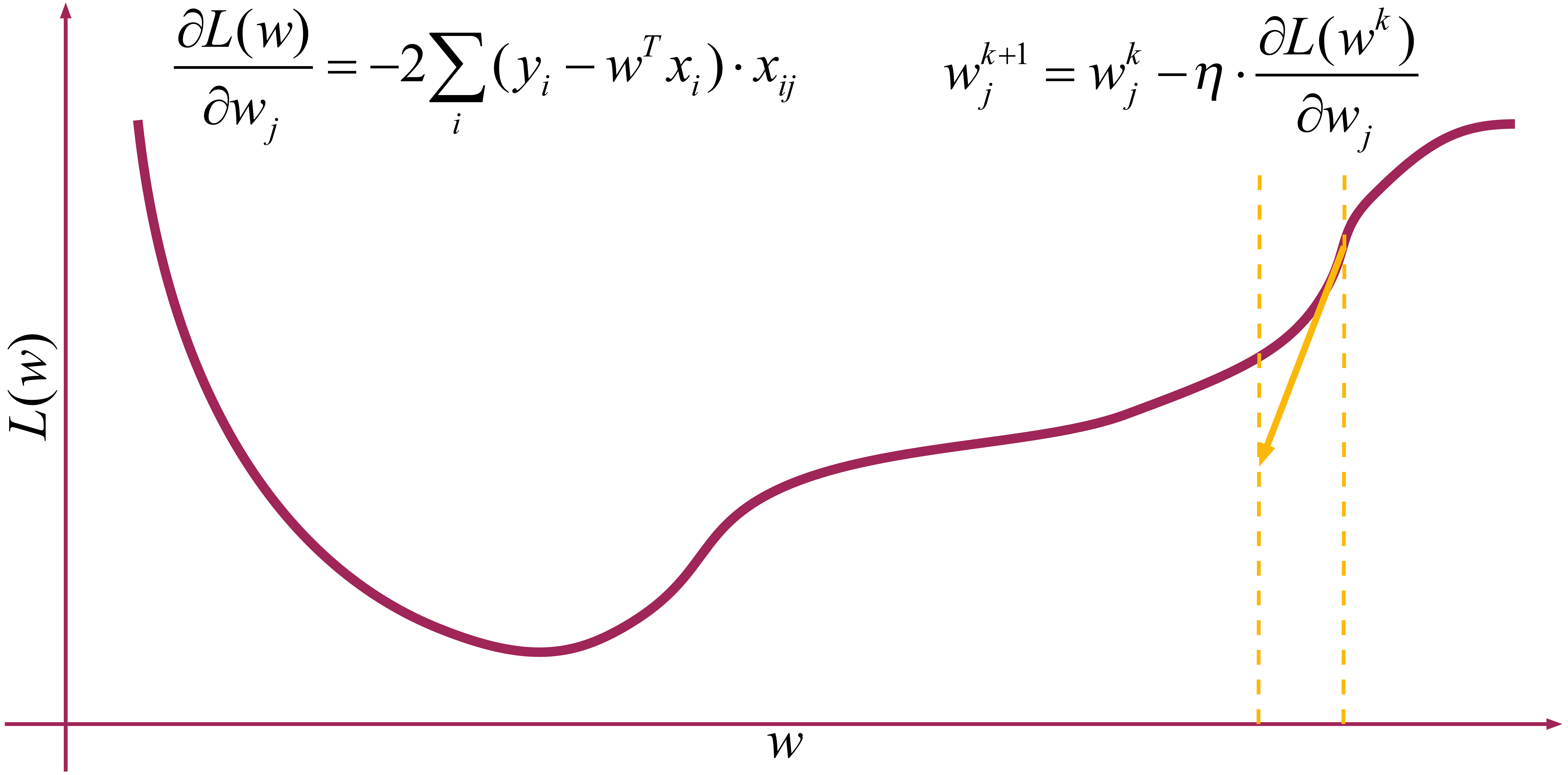


$$\frac{\partial L(w)}{\partial w_j} = -2\sum_i (y_i - w^T x_i)\cdot x_{ij}$$

$$w_j^{k+1} = w_j^k - \eta \cdot \frac{\partial L(w^k)}{\partial w_j}$$

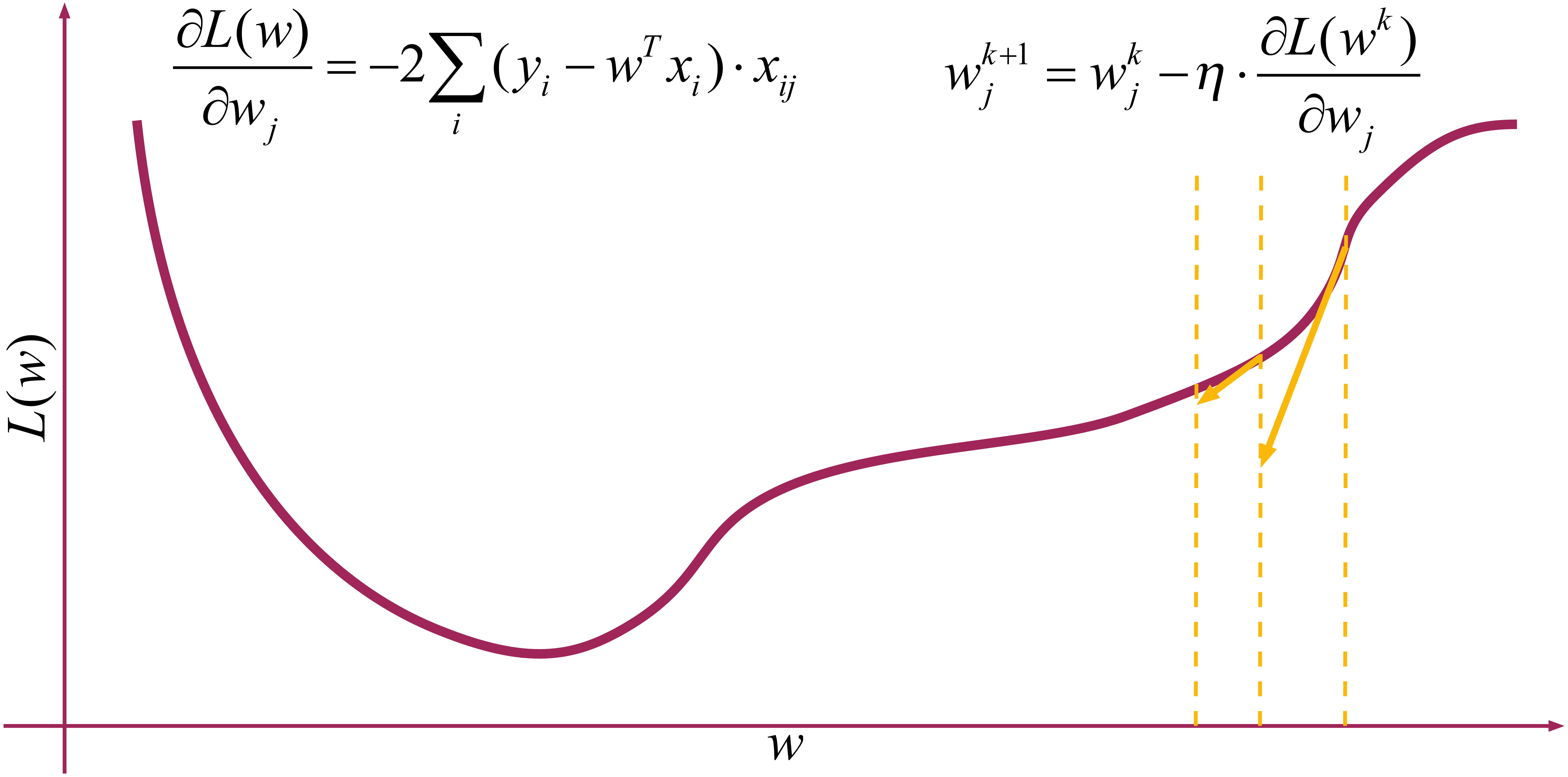
$$\frac{\partial L(w)}{\partial w_j} = -2\sum_i (y_i - w^T x_i) \cdot x_{ij}$$

$$w_j^{k+1} = w_j^k - \eta \cdot \frac{\partial L(w^k)}{\partial w_j}$$



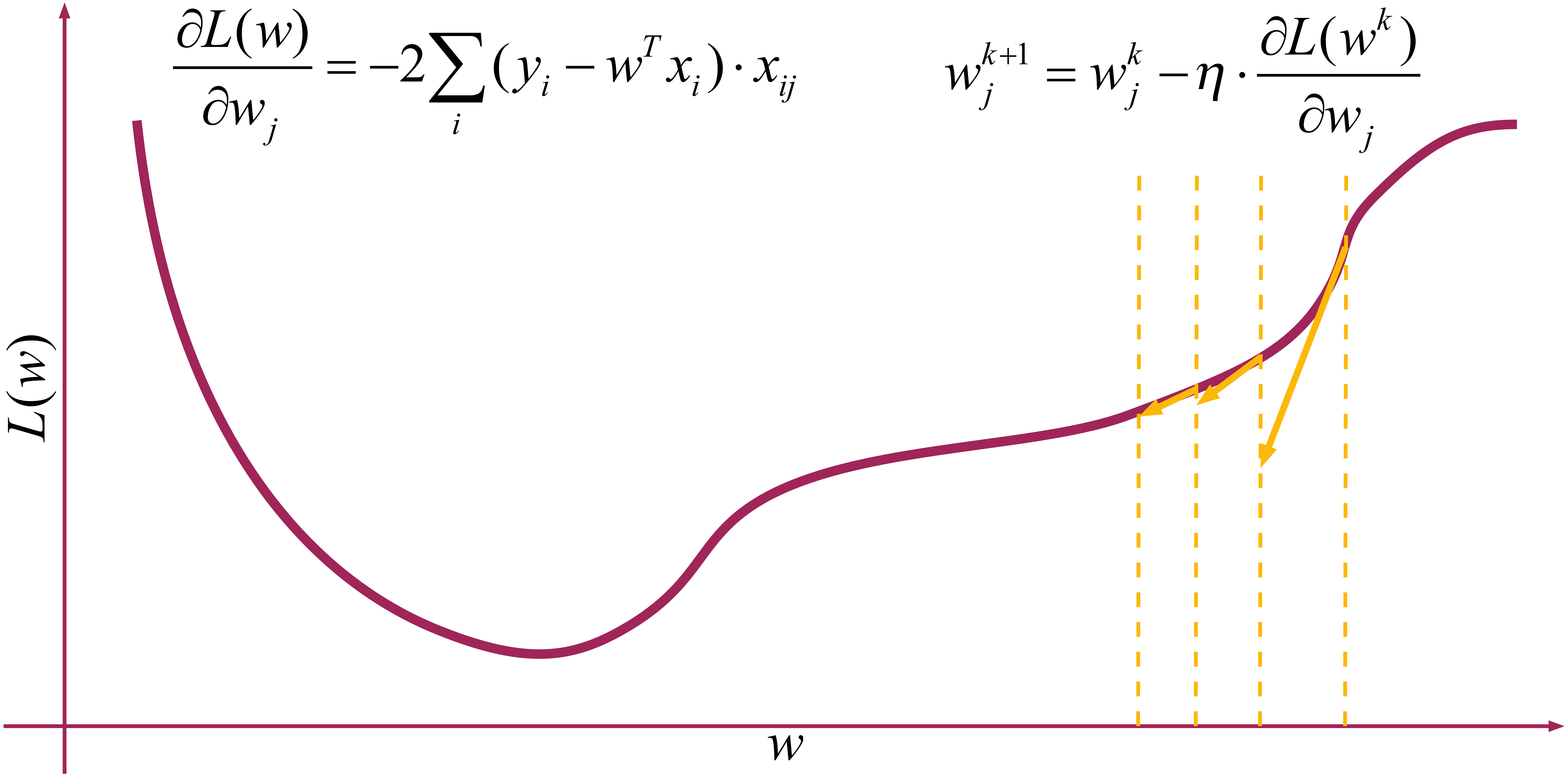
$$\frac{\partial L(w)}{\partial w_j} = -2\sum_i (y_i - w^T x_i)\cdot x_{ij}$$

$$w_j^{k+1} = w_j^k - \eta \cdot \frac{\partial L(w^k)}{\partial w_j}$$



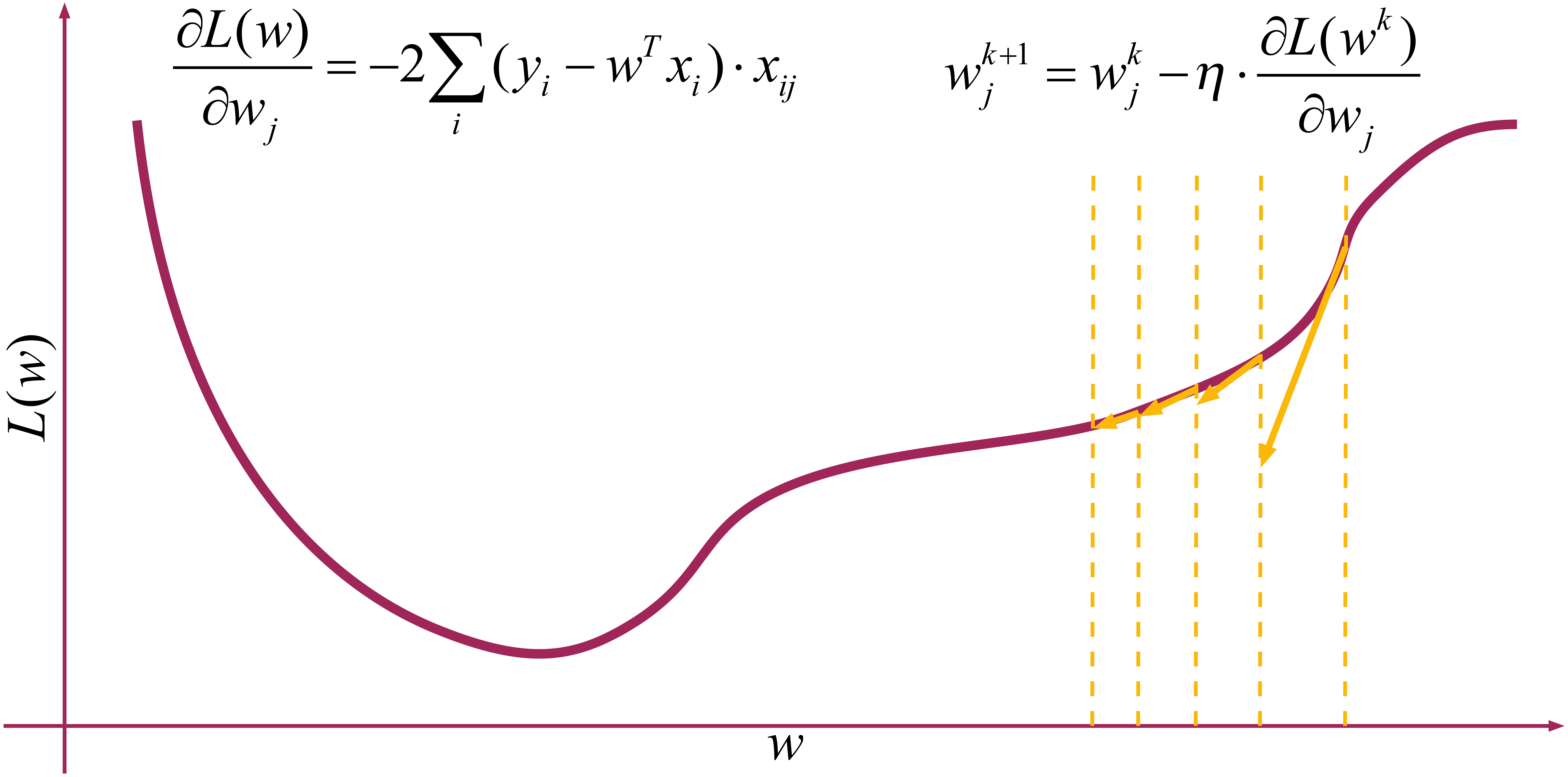
$$\frac{\partial L(w)}{\partial w_j} = -2\sum_i (y_i - w^T x_i) \cdot x_{ij}$$

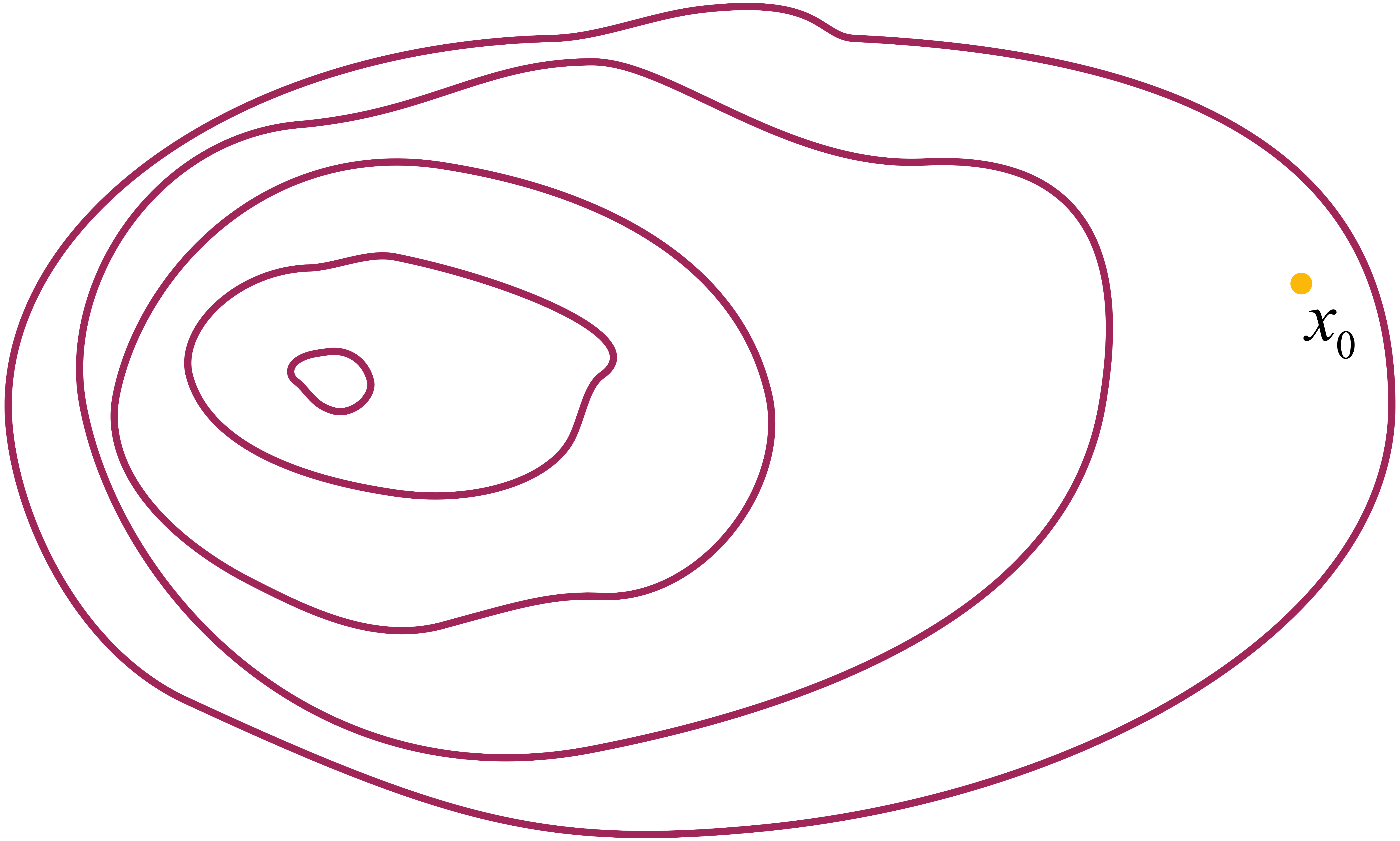
$$w_j^{k+1} = w_j^k - \eta \cdot \frac{\partial L(w^k)}{\partial w_j}$$

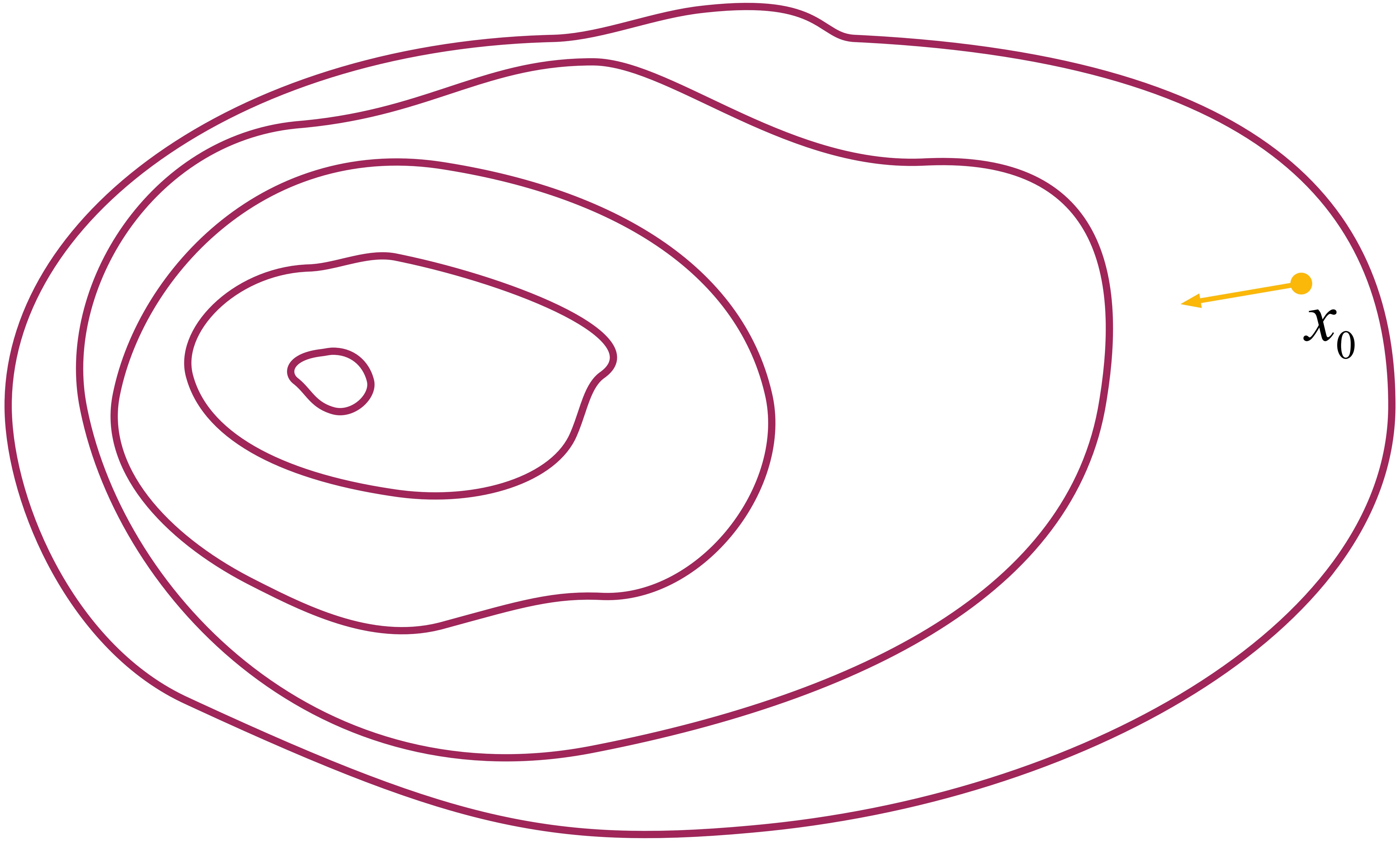


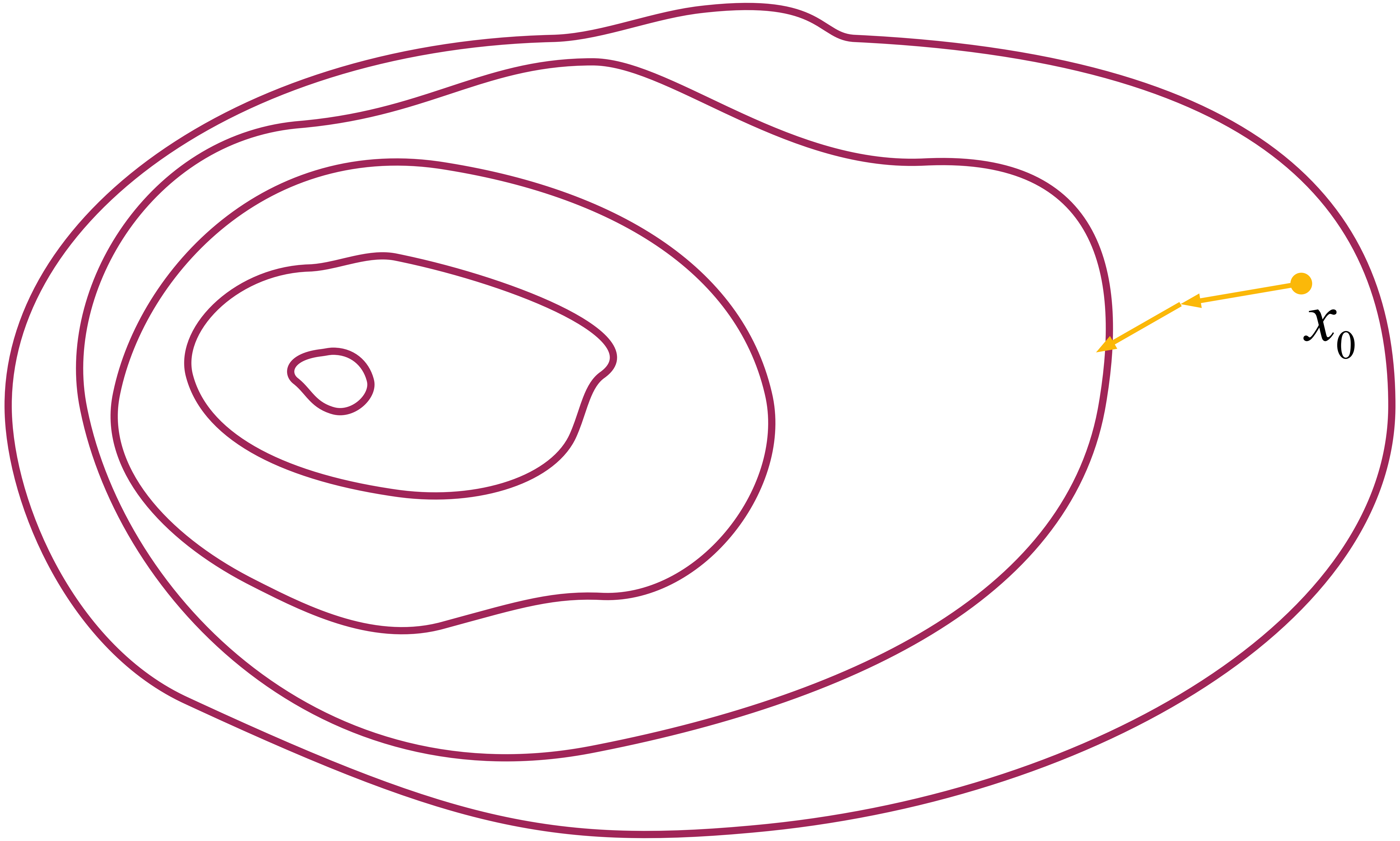
$$\frac{\partial L(w)}{\partial w_j} = -2\sum_i (y_i - w^T x_i)\cdot x_{ij}$$

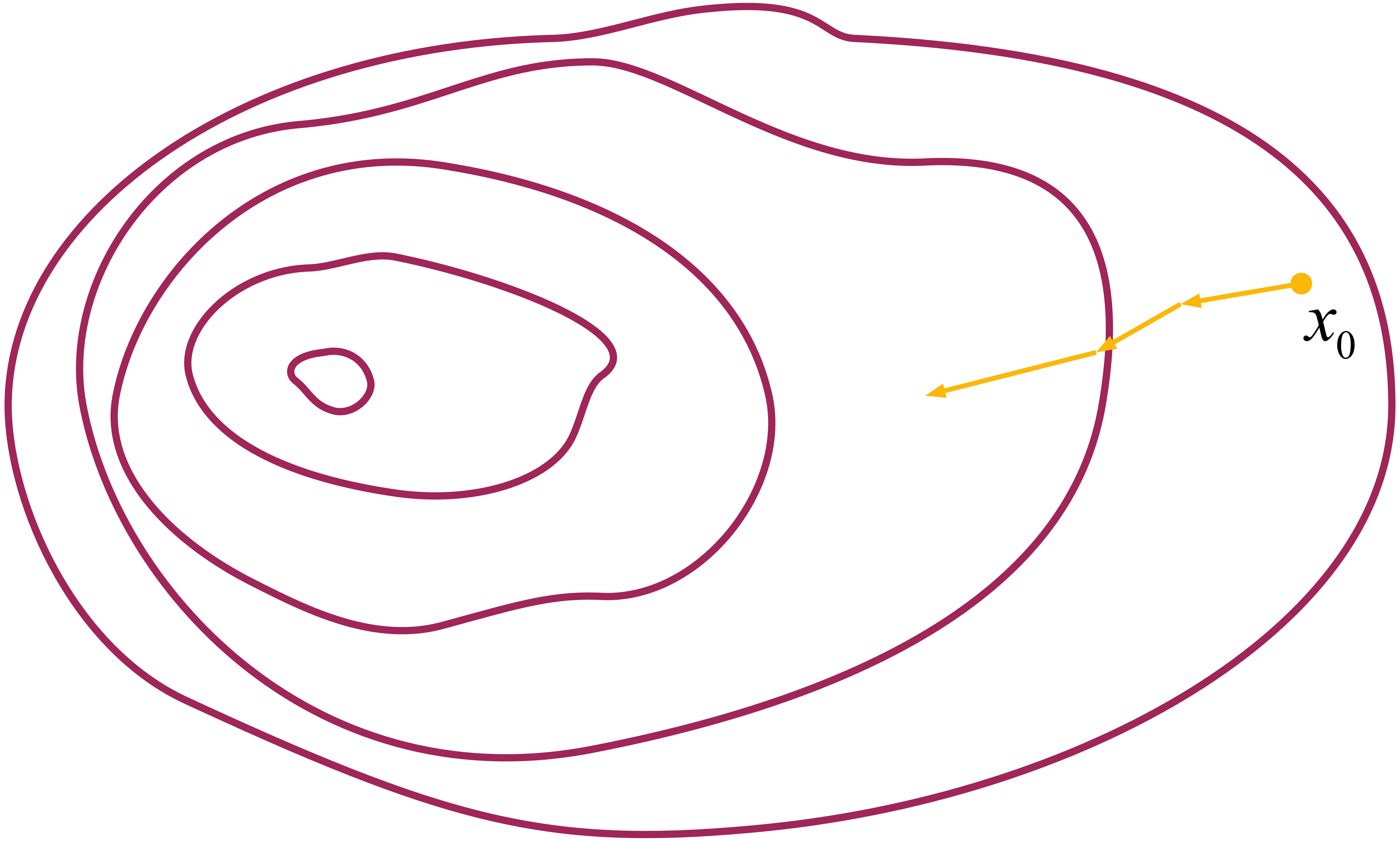
$$w_j^{k+1} = w_j^k - \eta\cdot\frac{\partial L(w^k)}{\partial w_j}$$

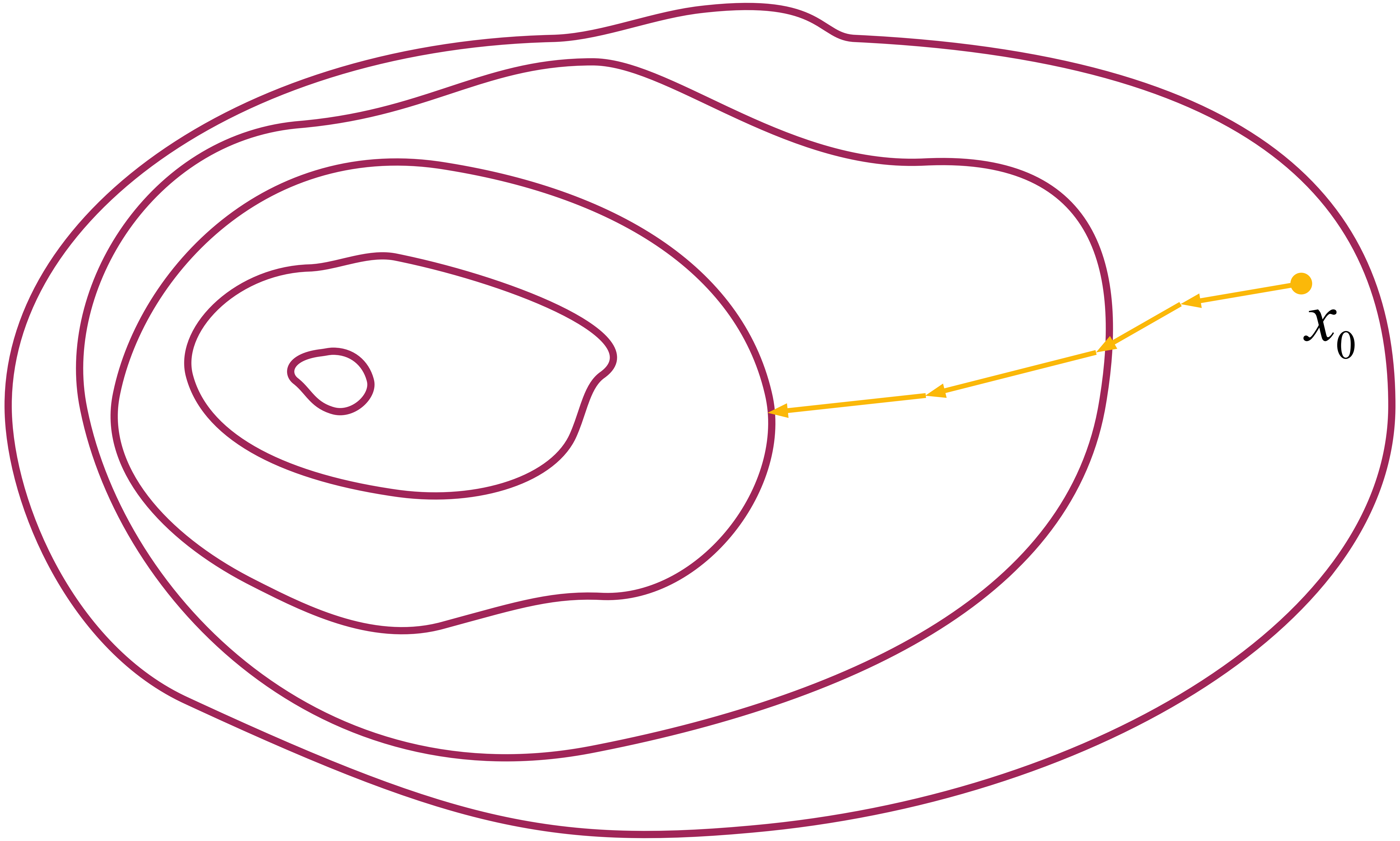


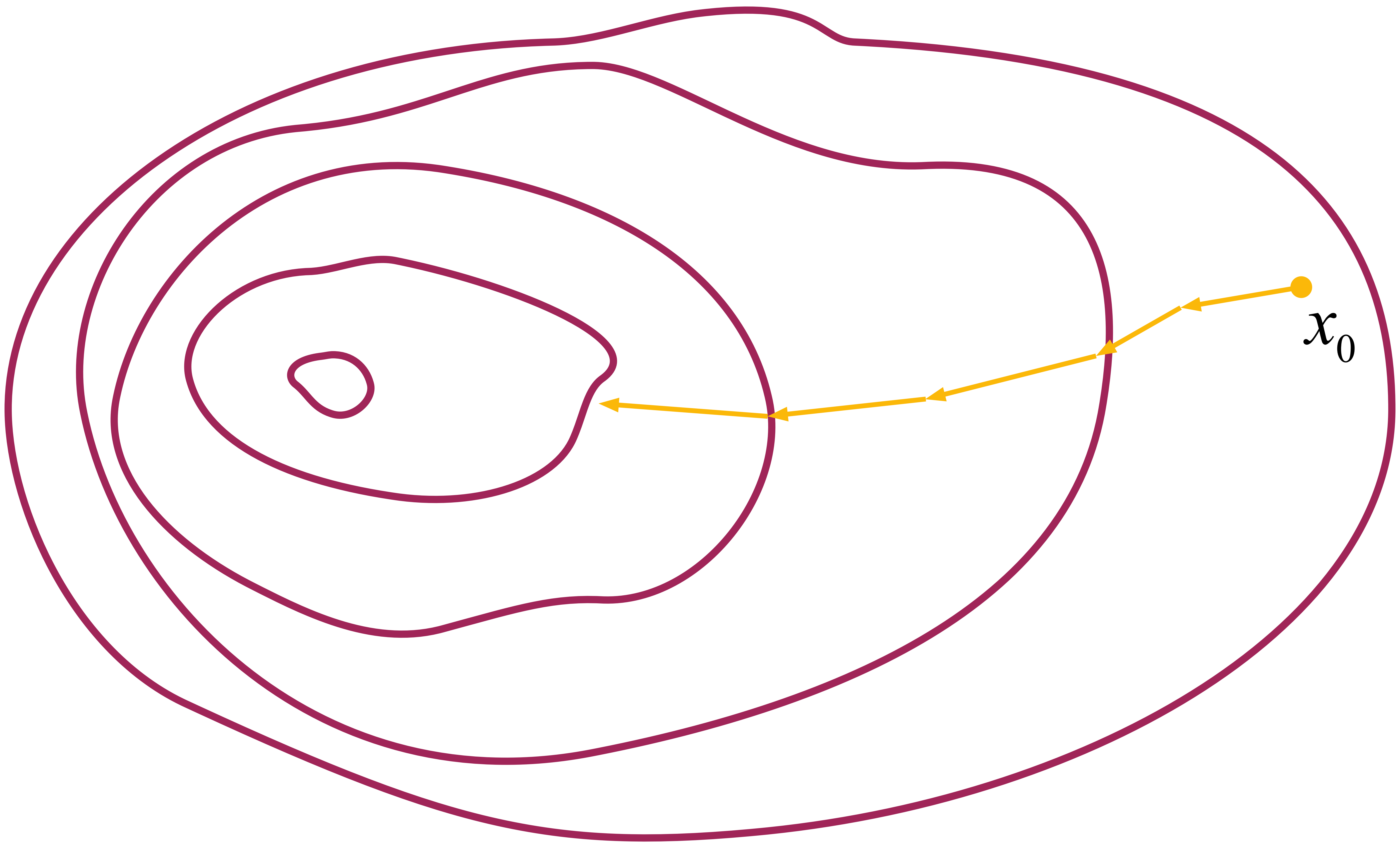

x_0

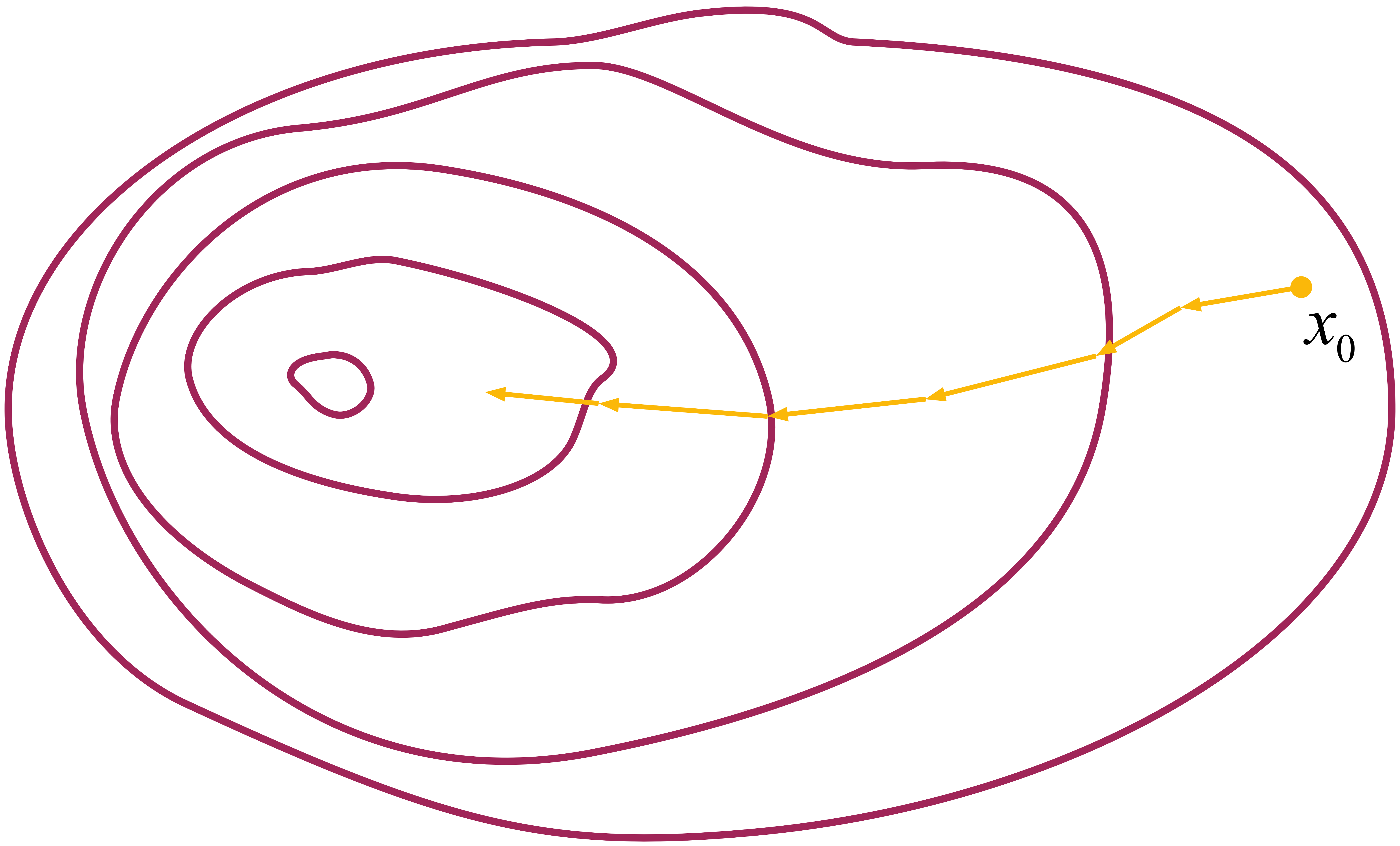

x_0


x_0


x_0


x_0


x_0


x_0

	features	label
0	[1.0, 0.0, 1.0, 0.0, 6.0, 2.0, 0.344167, 0.363...	985
1	[1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.150833, 0.150...	1321
2	[2.0, 0.0, 4.0, 0.0, 0.0, 2.0, 0.426667, 0.426...	2895
3	[2.0, 0.0, 4.0, 0.0, 1.0, 2.0, 0.595652, 0.565...	3348
4	[2.0, 0.0, 4.0, 0.0, 3.0, 2.0, 0.4125, 0.41728...	2162
...	...	...
505	[2.0, 0.0, 4.0, 0.0, 1.0, 1.0, 0.573333, 0.542...	3115
506	[2.0, 0.0, 4.0, 0.0, 2.0, 2.0, 0.414167, 0.398...	1795
507	[2.0, 0.0, 4.0, 0.0, 3.0, 1.0, 0.390833, 0.387...	2808
508	[2.0, 0.0, 4.0, 0.0, 5.0, 2.0, 0.335833, 0.324...	1471
509	[2.0, 0.0, 4.0, 0.0, 6.0, 2.0, 0.3425, 0.34152...	2455

	features	label
0	[1.0, 0.0, 1.0, 0.0, 6.0, 2.0, 0.344167, 0.363...	985
1	[1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.150833, 0.150...	1321
2	[2.0, 0.0, 4.0, 0.0, 0.0, 2.0, 0.426667, 0.426...	2895
3	[2.0, 0.0, 4.0, 0.0, 1.0, 2.0, 0.595652, 0.565...	3348
4	[2.0, 0.0, 4.0, 0.0, 3.0, 2.0, 0.4125, 0.41728...	2162
...	...	...
505	[2.0, 0.0, 4.0, 0.0, 1.0, 1.0, 0.573333, 0.542...	3115
506	[2.0, 0.0, 4.0, 0.0, 2.0, 2.0, 0.414167, 0.398...	1795
507	[2.0, 0.0, 4.0, 0.0, 3.0, 1.0, 0.390833, 0.387...	2808
508	[2.0, 0.0, 4.0, 0.0, 5.0, 2.0, 0.335833, 0.324...	1471
509	[2.0, 0.0, 4.0, 0.0, 6.0, 2.0, 0.3425, 0.34152...	2455

x

	features	label
0	[1.0, 0.0, 1.0, 0.0, 6.0, 2.0, 0.344167, 0.363...	985
1	[1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.150833, 0.150...	1321
2	[2.0, 0.0, 4.0, 0.0, 0.0, 2.0, 0.426667, 0.426...	2895
3	[2.0, 0.0, 4.0, 0.0, 1.0, 2.0, 0.595652, 0.565...	3348
4	[2.0, 0.0, 4.0, 0.0, 3.0, 2.0, 0.4125, 0.41728...	2162
...	...	...
505	[2.0, 0.0, 4.0, 0.0, 1.0, 1.0, 0.573333, 0.542...	3115
506	[2.0, 0.0, 4.0, 0.0, 2.0, 2.0, 0.414167, 0.398...	1795
507	[2.0, 0.0, 4.0, 0.0, 3.0, 1.0, 0.390833, 0.387...	2808
508	[2.0, 0.0, 4.0, 0.0, 5.0, 2.0, 0.335833, 0.324...	1471
509	[2.0, 0.0, 4.0, 0.0, 6.0, 2.0, 0.3425, 0.34152...	2455

x y

$$(y - w^T x) \cdot x_j$$

	features	label
0	[1.0, 0.0, 1.0, 0.0, 6.0, 2.0, 0.344167, 0.363...	985
1	[1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.150833, 0.150...	1321
2	[2.0, 0.0, 4.0, 0.0, 0.0, 2.0, 0.426667, 0.426...	2895
3	[2.0, 0.0, 4.0, 0.0, 1.0, 2.0, 0.595652, 0.565...	3348
4	[2.0, 0.0, 4.0, 0.0, 3.0, 2.0, 0.4125, 0.41728...	2162
...	...	...
505	[2.0, 0.0, 4.0, 0.0, 1.0, 1.0, 0.573333, 0.542...	3115
506	[2.0, 0.0, 4.0, 0.0, 2.0, 2.0, 0.414167, 0.398...	1795
507	[2.0, 0.0, 4.0, 0.0, 3.0, 1.0, 0.390833, 0.387...	2808
508	[2.0, 0.0, 4.0, 0.0, 5.0, 2.0, 0.335833, 0.324...	1471
509	[2.0, 0.0, 4.0, 0.0, 6.0, 2.0, 0.3425, 0.34152...	2455

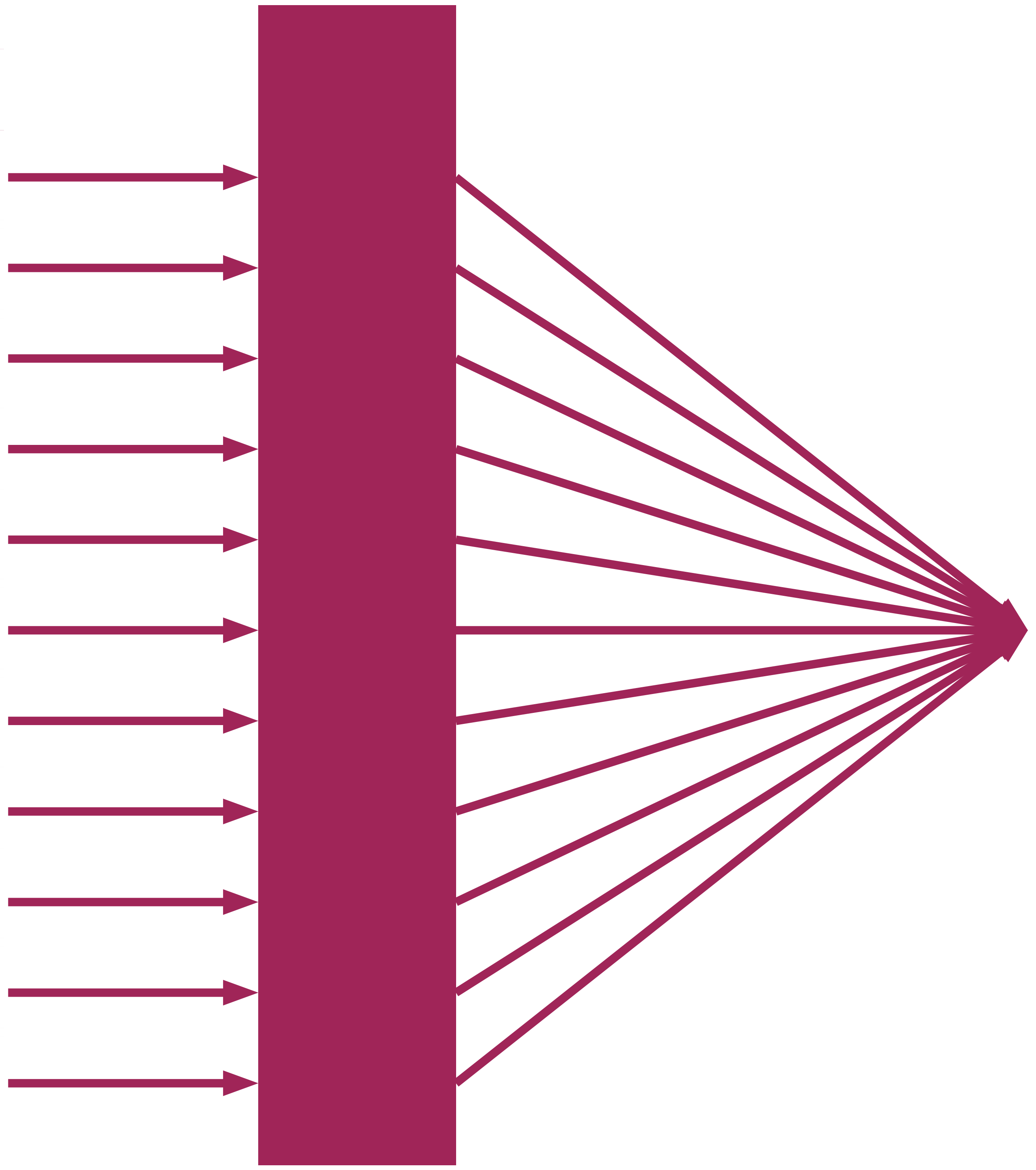
x y

$$(y - w^T x) \cdot x_j$$

sum

	features	label
0	[1.0, 0.0, 1.0, 0.0, 6.0, 2.0, 0.344167, 0.363...	985
1	[1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.150833, 0.150...	1321
2	[2.0, 0.0, 4.0, 0.0, 0.0, 2.0, 0.426667, 0.426...	2895
3	[2.0, 0.0, 4.0, 0.0, 1.0, 2.0, 0.595652, 0.565...	3348
4	[2.0, 0.0, 4.0, 0.0, 3.0, 2.0, 0.4125, 0.41728...	2162
...	...	...
505	[2.0, 0.0, 4.0, 0.0, 1.0, 1.0, 0.573333, 0.542...	3115
506	[2.0, 0.0, 4.0, 0.0, 2.0, 2.0, 0.414167, 0.398...	1795
507	[2.0, 0.0, 4.0, 0.0, 3.0, 1.0, 0.390833, 0.387...	2808
508	[2.0, 0.0, 4.0, 0.0, 5.0, 2.0, 0.335833, 0.324...	1471
509	[2.0, 0.0, 4.0, 0.0, 6.0, 2.0, 0.3425, 0.34152...	2455

y



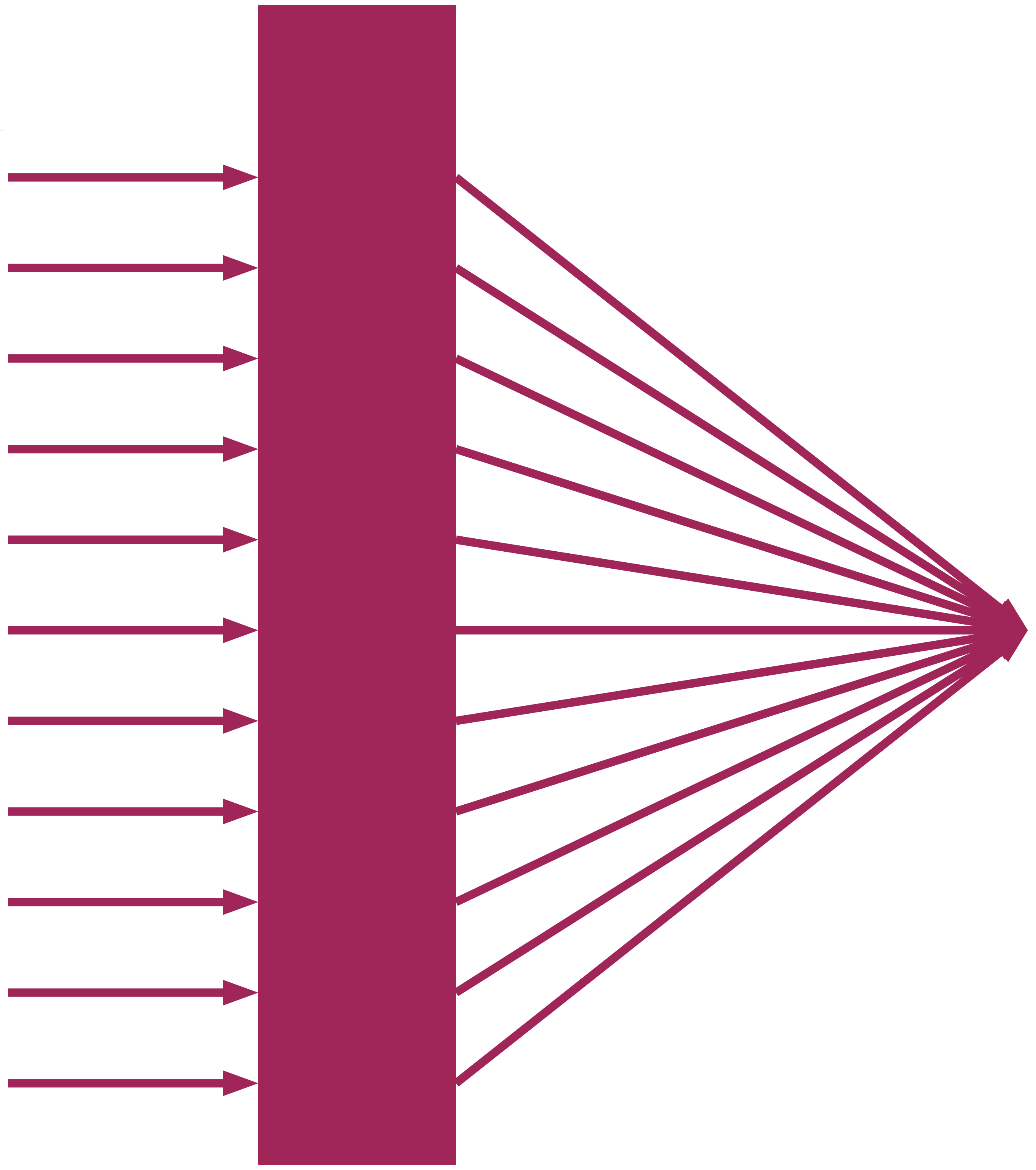
$$\frac{\partial L(w^k)}{\partial w_j}$$

$$(y - w^T x) \cdot x_j$$

sum

	features	label
0	[1.0, 0.0, 1.0, 0.0, 6.0, 2.0, 0.344167, 0.363...	985
1	[1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.150833, 0.150...	1321
2	[2.0, 0.0, 4.0, 0.0, 0.0, 2.0, 0.426667, 0.426...	2895
3	[2.0, 0.0, 4.0, 0.0, 1.0, 2.0, 0.595652, 0.565...	3348
4	[2.0, 0.0, 4.0, 0.0, 3.0, 2.0, 0.4125, 0.41728...	2162
...	...	...
505	[2.0, 0.0, 4.0, 0.0, 1.0, 1.0, 0.573333, 0.542...	3115
506	[2.0, 0.0, 4.0, 0.0, 2.0, 2.0, 0.414167, 0.398...	1795
507	[2.0, 0.0, 4.0, 0.0, 3.0, 1.0, 0.390833, 0.387...	2808
508	[2.0, 0.0, 4.0, 0.0, 5.0, 2.0, 0.335833, 0.324...	1471
509	[2.0, 0.0, 4.0, 0.0, 6.0, 2.0, 0.3425, 0.34152...	2455

y



$$\frac{\partial L(w^k)}{\partial w_j}$$

$$w_j^{k+1} = w_j^k - \eta \cdot \frac{\partial L(w^k)}{\partial w_j}$$

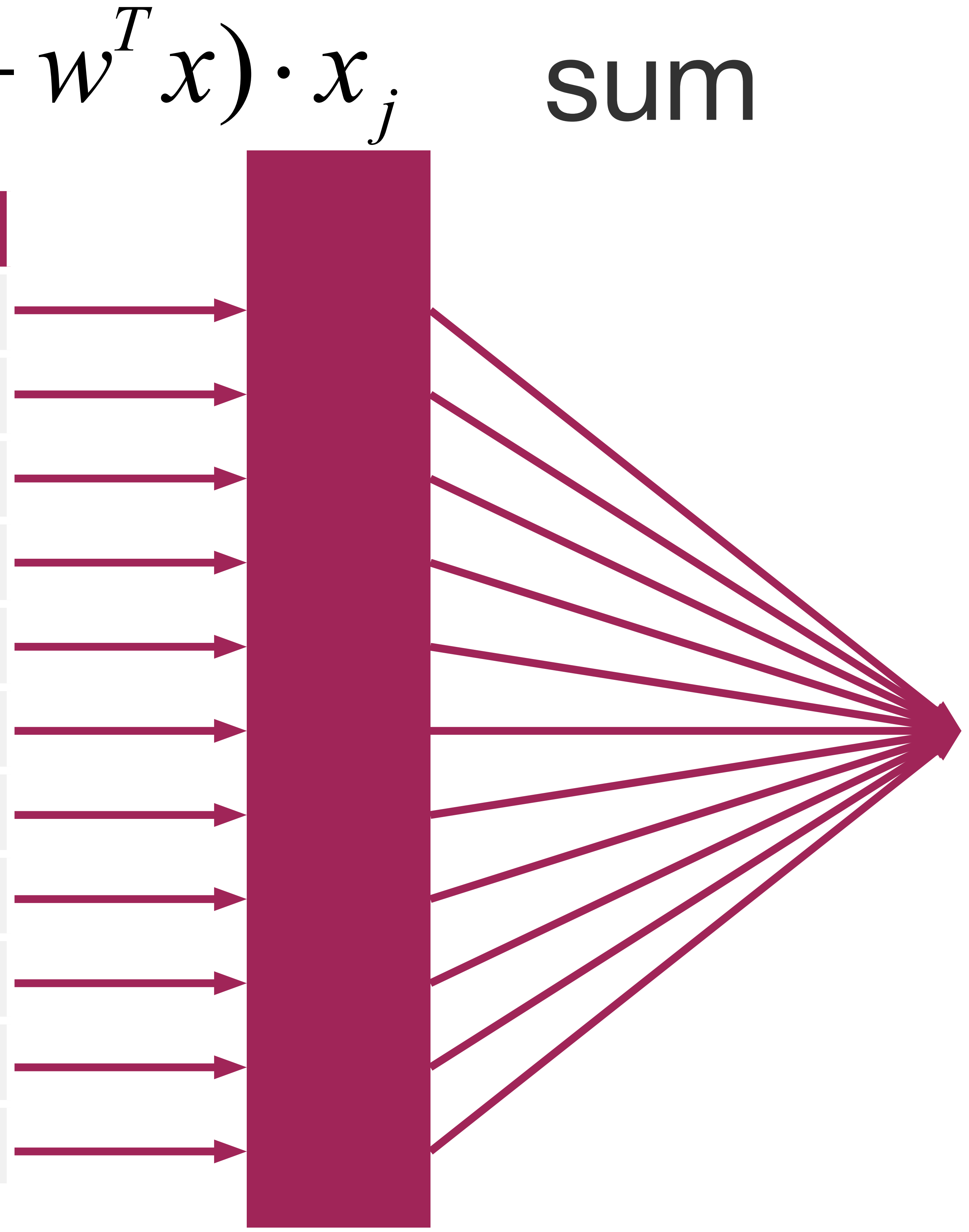
Pros and cons

- Method is universal
- Complexity is proportional to number of features
- Problem: can get stuck in local minimum

Stochastic Gradient Descend

	features	label
0	[1.0, 0.0, 1.0, 0.0, 6.0, 2.0, 0.344167, 0.363...	985
1	[1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.150833, 0.150...	1321
2	[2.0, 0.0, 4.0, 0.0, 0.0, 2.0, 0.426667, 0.426...	2895
3	[2.0, 0.0, 4.0, 0.0, 1.0, 2.0, 0.595652, 0.565...	3348
4	[2.0, 0.0, 4.0, 0.0, 3.0, 2.0, 0.4125, 0.41728...	2162
...	...	...
505	[2.0, 0.0, 4.0, 0.0, 1.0, 1.0, 0.573333, 0.542...	3115
506	[2.0, 0.0, 4.0, 0.0, 2.0, 2.0, 0.414167, 0.398...	1795
507	[2.0, 0.0, 4.0, 0.0, 3.0, 1.0, 0.390833, 0.387...	2808
508	[2.0, 0.0, 4.0, 0.0, 5.0, 2.0, 0.335833, 0.324...	1471
509	[2.0, 0.0, 4.0, 0.0, 6.0, 2.0, 0.3425, 0.34152...	2455

x y



$$w_j^{k+1} = w_j^k - \lambda \cdot \frac{\partial L(w^k)}{\partial w_j}$$

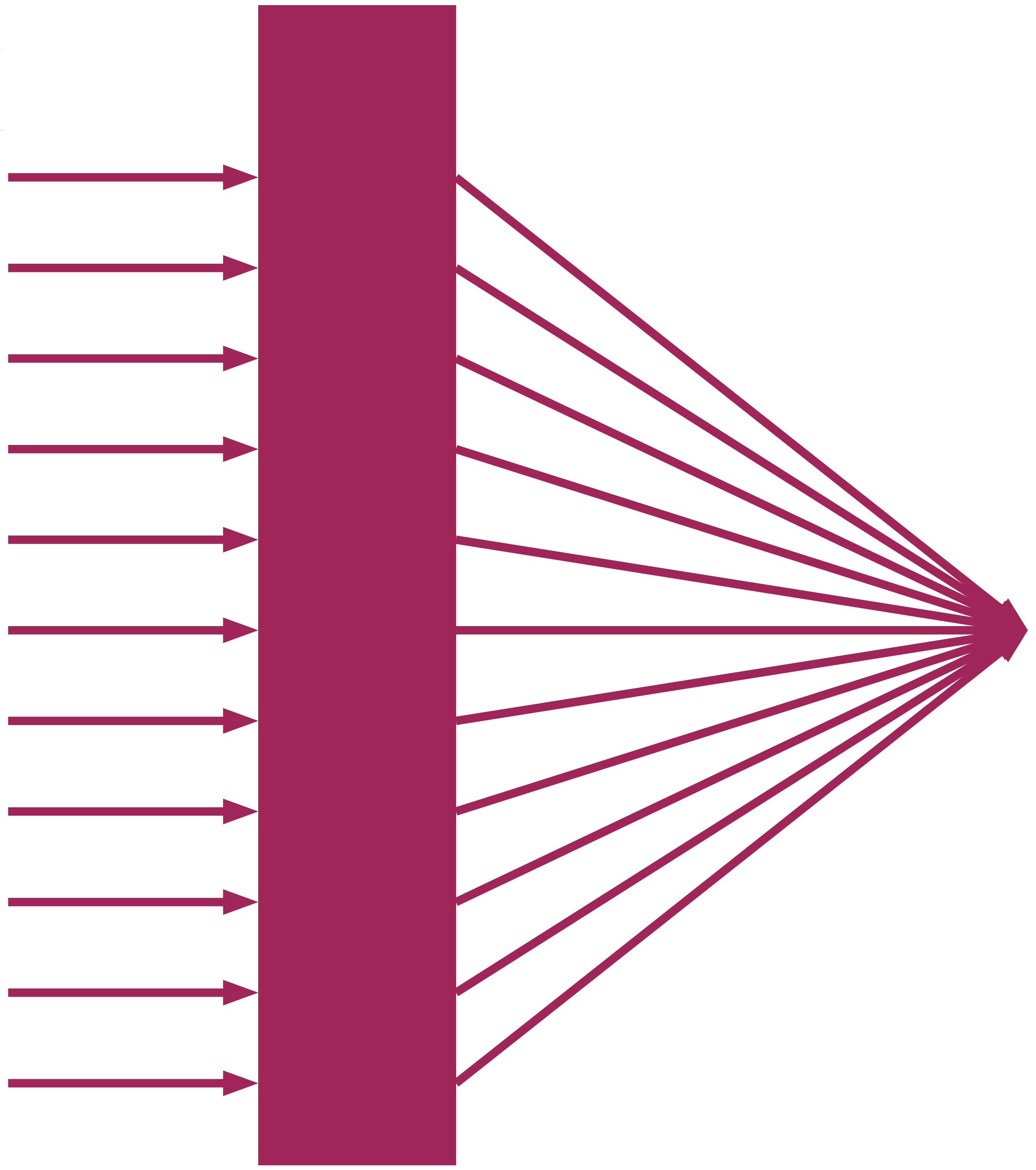
random shuffle

	features	label
0	[1.0, 0.0, 1.0, 0.0, 6.0, 2.0, 0.344167, 0.363...	985
1	[1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.150833, 0.150...	1321
2	[2.0, 0.0, 4.0, 0.0, 0.0, 2.0, 0.426667, 0.426...	2895
3	[2.0, 0.0, 4.0, 0.0, 1.0, 2.0, 0.595652, 0.565...	3348
4	[2.0, 0.0, 4.0, 0.0, 3.0, 2.0, 0.4125, 0.41728...	2162
...	...	...
505	[2.0, 0.0, 4.0, 0.0, 1.0, 1.0, 0.573333, 0.542...	3115
506	[2.0, 0.0, 4.0, 0.0, 2.0, 2.0, 0.414167, 0.398...	1795
507	[2.0, 0.0, 4.0, 0.0, 3.0, 1.0, 0.390833, 0.387...	2808
508	[2.0, 0.0, 4.0, 0.0, 5.0, 2.0, 0.335833, 0.324...	1471
509	[2.0, 0.0, 4.0, 0.0, 6.0, 2.0, 0.3425, 0.34152...	2455

x y

$$(y - w^T x) \cdot x_j$$

sum



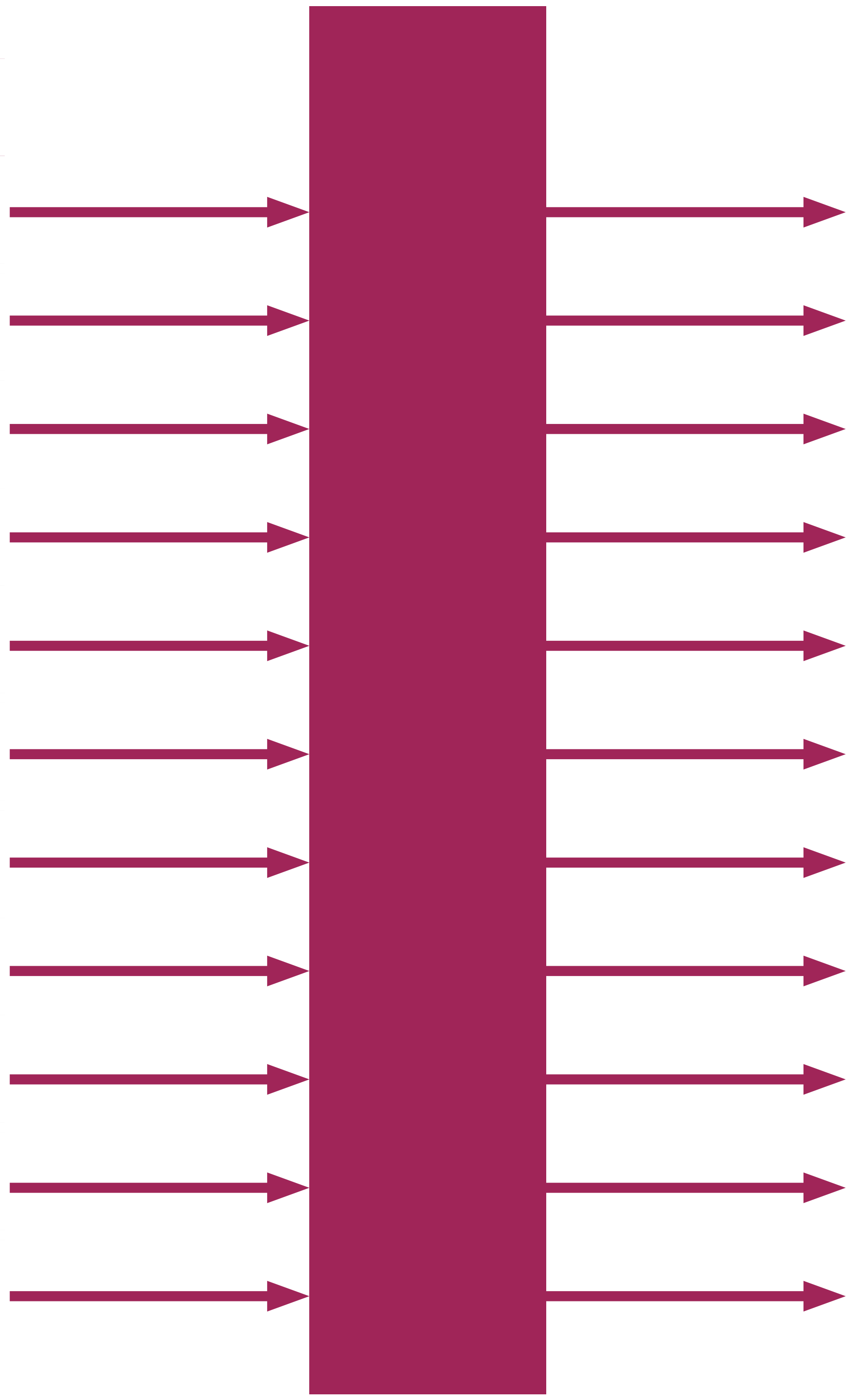
$$w_j^{k+1} = w_j^k - \lambda \cdot \frac{\partial L(w^k)}{\partial w_j}$$

$$(y - w^T x) \cdot x_j$$

sum

	features	label
0	[1.0, 0.0, 1.0, 0.0, 6.0, 2.0, 0.344167, 0.363...	985
1	[1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.150833, 0.150...	1321
2	[2.0, 0.0, 4.0, 0.0, 0.0, 2.0, 0.426667, 0.426...	2895
3	[2.0, 0.0, 4.0, 0.0, 1.0, 2.0, 0.595652, 0.565...	3348
4	[2.0, 0.0, 4.0, 0.0, 3.0, 2.0, 0.4125, 0.41728...	2162
...	...	...
505	[2.0, 0.0, 4.0, 0.0, 1.0, 1.0, 0.573333, 0.542...	3115
506	[2.0, 0.0, 4.0, 0.0, 2.0, 2.0, 0.414167, 0.398...	1795
507	[2.0, 0.0, 4.0, 0.0, 3.0, 1.0, 0.390833, 0.387...	2808
508	[2.0, 0.0, 4.0, 0.0, 5.0, 2.0, 0.335833, 0.324...	1471
509	[2.0, 0.0, 4.0, 0.0, 6.0, 2.0, 0.3425, 0.34152...	2455

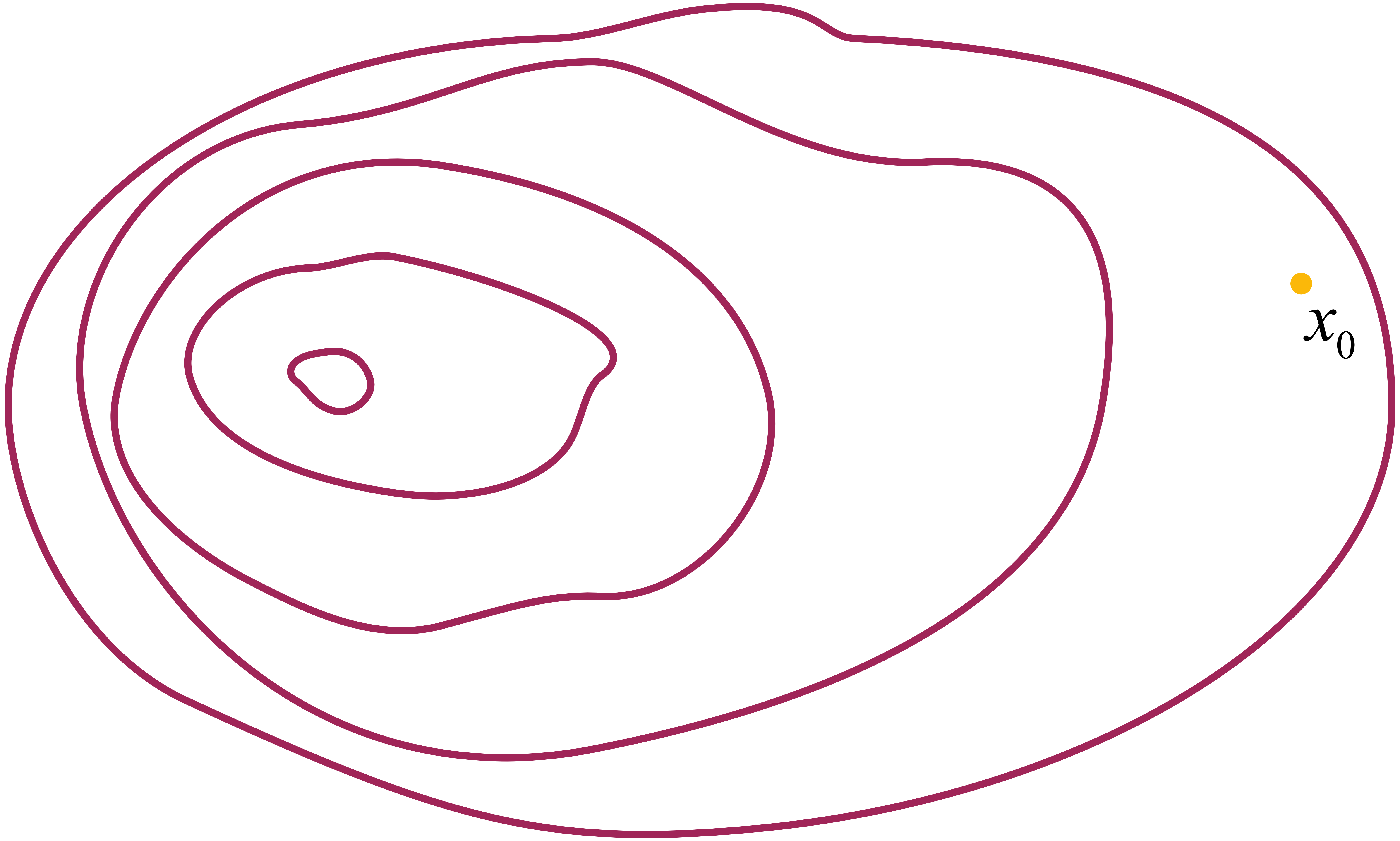
x y

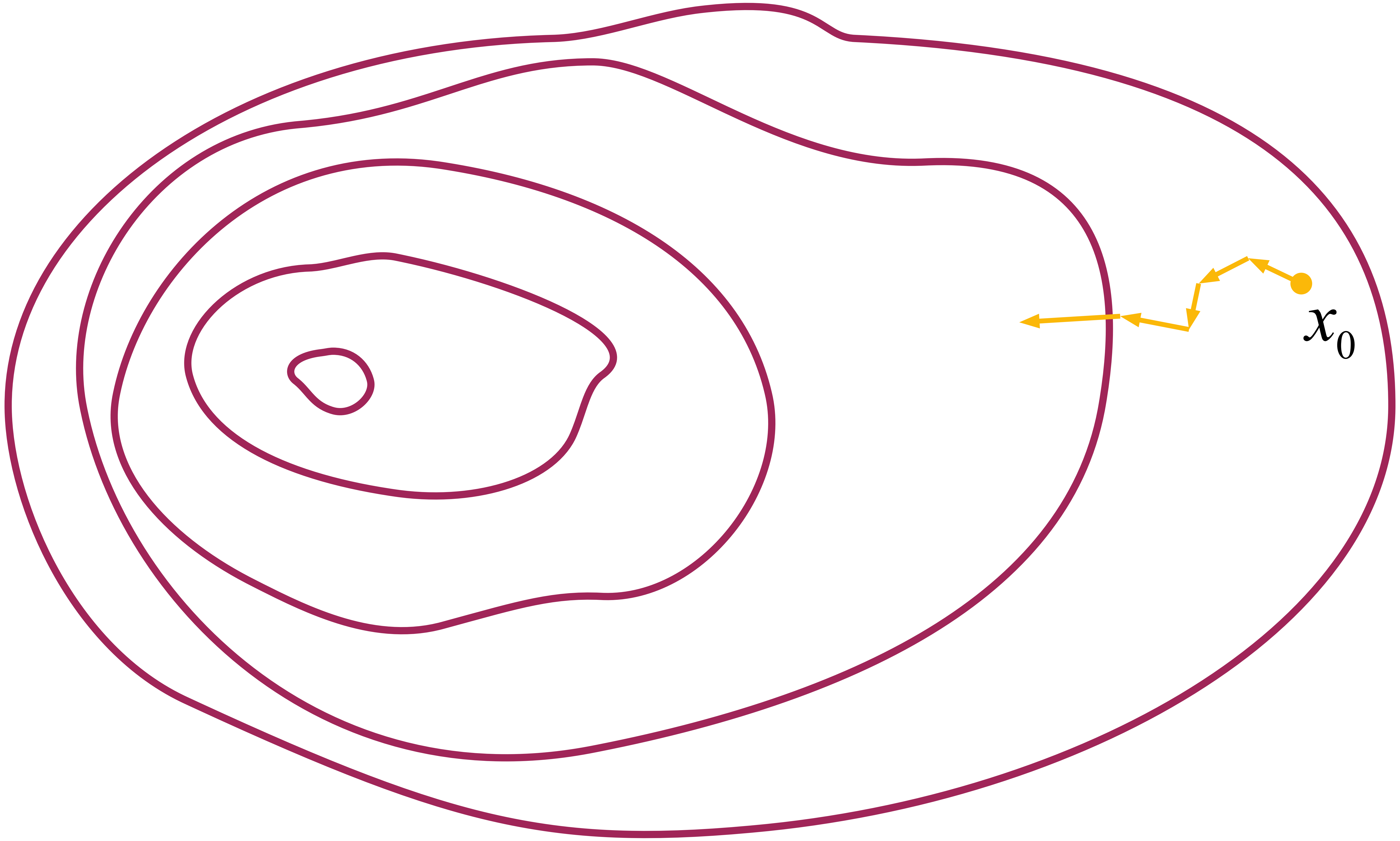


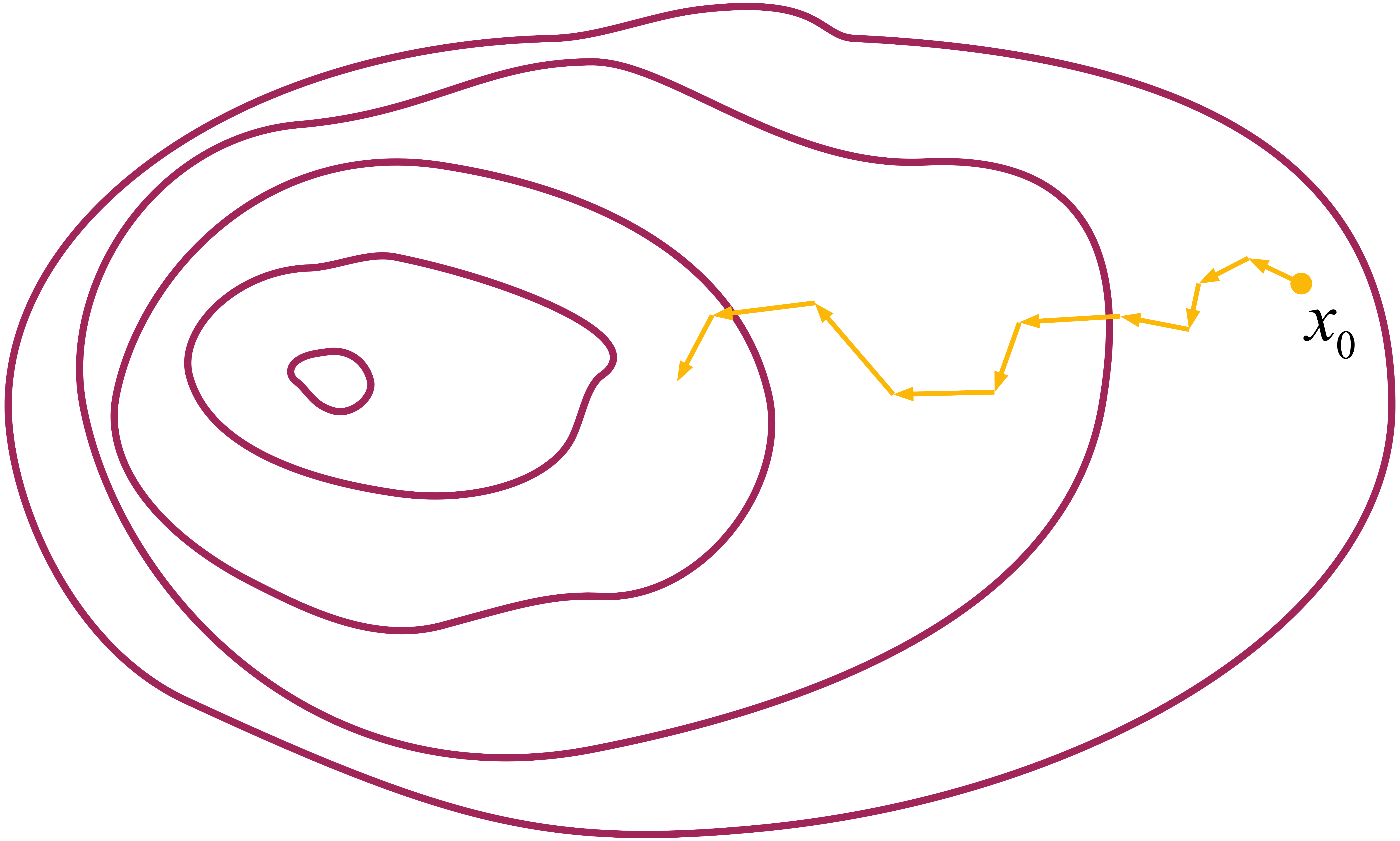
$$w_j^{k+1} = w_j^k - \eta \cdot \frac{\partial L(w^k)}{\partial w_j}$$

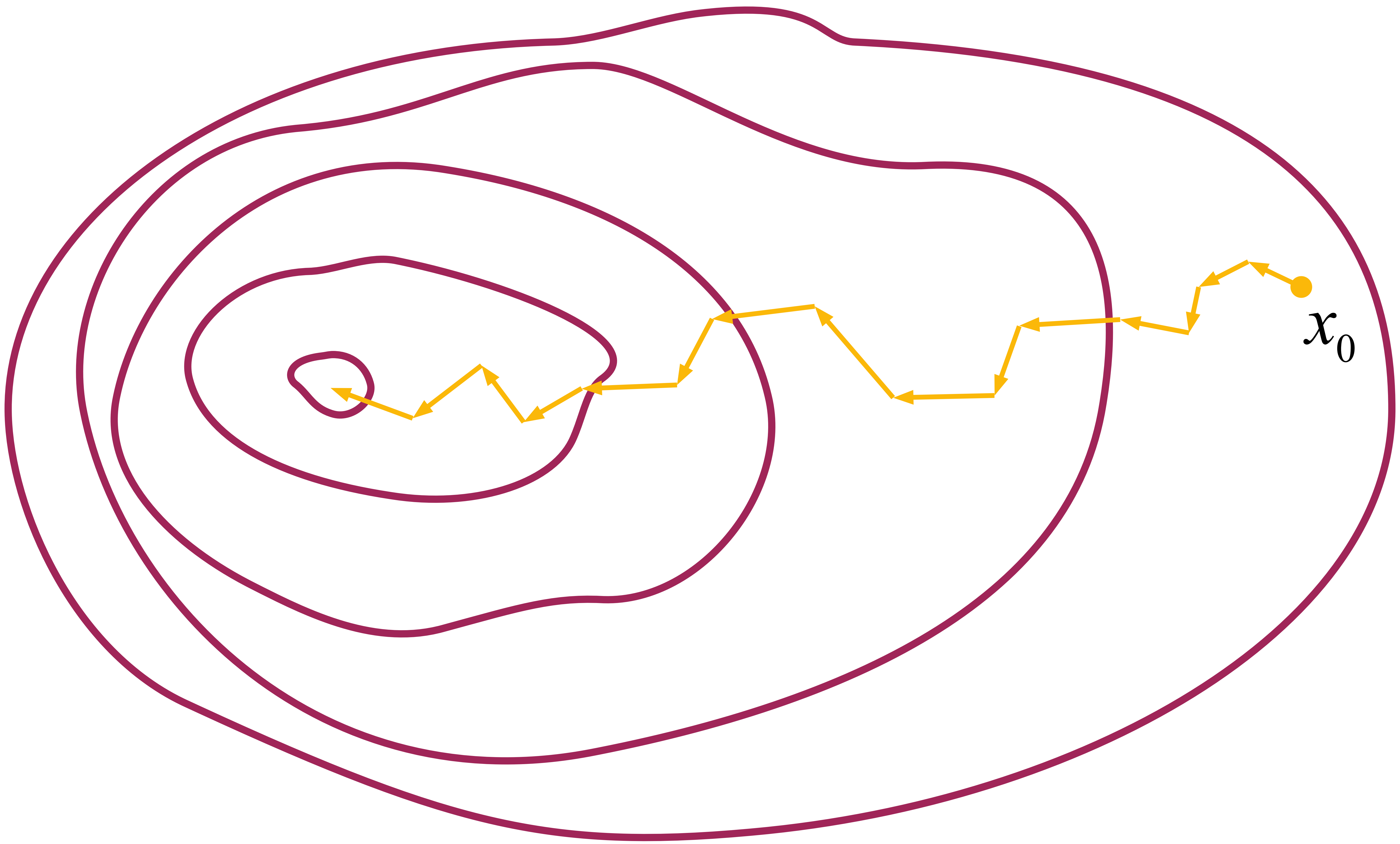
$$w_j^{k+2} = w_j^{k+1} - \eta \cdot \frac{\partial L(w^{k+1})}{\partial w_j}$$

...


x_0


x_0


x_0


x_0

- The acceleration of convergence
- Global minimum
- Online learning

$$(y - w^T x) \cdot x_j$$

sum

	features	label
0	[1.0, 0.0, 1.0, 0.0, 6.0, 2.0, 0.344167, 0.363...	985
1	[1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.150833, 0.150...	1321
2	[2.0, 0.0, 4.0, 0.0, 0.0, 2.0, 0.426667, 0.426...	2895
3	[2.0, 0.0, 4.0, 0.0, 1.0, 2.0, 0.595652, 0.565...	3348
4	[2.0, 0.0, 4.0, 0.0, 3.0, 2.0, 0.4125, 0.41728...	2162
...	...	...
505	[2.0, 0.0, 4.0, 0.0, 1.0, 1.0, 0.573333, 0.542...	3115
506	[2.0, 0.0, 4.0, 0.0, 2.0, 2.0, 0.414167, 0.398...	1795
507	[2.0, 0.0, 4.0, 0.0, 3.0, 1.0, 0.390833, 0.387...	2808
508	[2.0, 0.0, 4.0, 0.0, 5.0, 2.0, 0.335833, 0.324...	1471
509	[2.0, 0.0, 4.0, 0.0, 6.0, 2.0, 0.3425, 0.34152...	2455
	x	y

$$w_j^{k+1} = w_j^k - \eta \cdot \frac{\partial L(w^k)}{\partial w_j}$$

$$w_j^{k+2} = w_j^{k+1} - \eta \cdot \frac{\partial L(w^{k+1})}{\partial w_j}$$

...

$$(y - w^T x) \cdot x_j$$

sum

	features	label
0	[1.0, 0.0, 1.0, 0.0, 6.0, 2.0, 0.344167, 0.363...	985
1	[1.0, 0.0, 1.0, 0.0, 1.0, 1.0, 0.150833, 0.150...	1321
2	[2.0, 0.0, 4.0, 0.0, 0.0, 2.0, 0.426667, 0.426...	2895
3	[2.0, 0.0, 4.0, 0.0, 1.0, 2.0, 0.595652, 0.565...	3348
4	[2.0, 0.0, 4.0, 0.0, 3.0, 2.0, 0.4125, 0.41728...	2162
...	...	...
505	[2.0, 0.0, 4.0, 0.0, 1.0, 1.0, 0.573333, 0.542...	3115
506	[2.0, 0.0, 4.0, 0.0, 2.0, 2.0, 0.414167, 0.398...	1795
507	[2.0, 0.0, 4.0, 0.0, 3.0, 1.0, 0.390833, 0.387...	2808
508	[2.0, 0.0, 4.0, 0.0, 5.0, 2.0, 0.335833, 0.324...	1471
509	[2.0, 0.0, 4.0, 0.0, 6.0, 2.0, 0.3425, 0.34152...	2455

x y

$$w_j^{k+1} = w_j^k - \eta \cdot \frac{\partial L(w^k)}{\partial w_j}$$

$$w_j^{k+2} = w_j^{k+1} - \eta \cdot \frac{\partial L(w^{k+1})}{\partial w_j}$$

$$w_j^{k+3} = w_j^{k+2} - \eta \cdot \frac{\partial L(w^{k+2})}{\partial w_j}$$

What have you learned:

- How to formulate machine learning problem
- How to solve it analytically
- How to solve it by gradient descend
- How to solve it by stochastic gradient descend