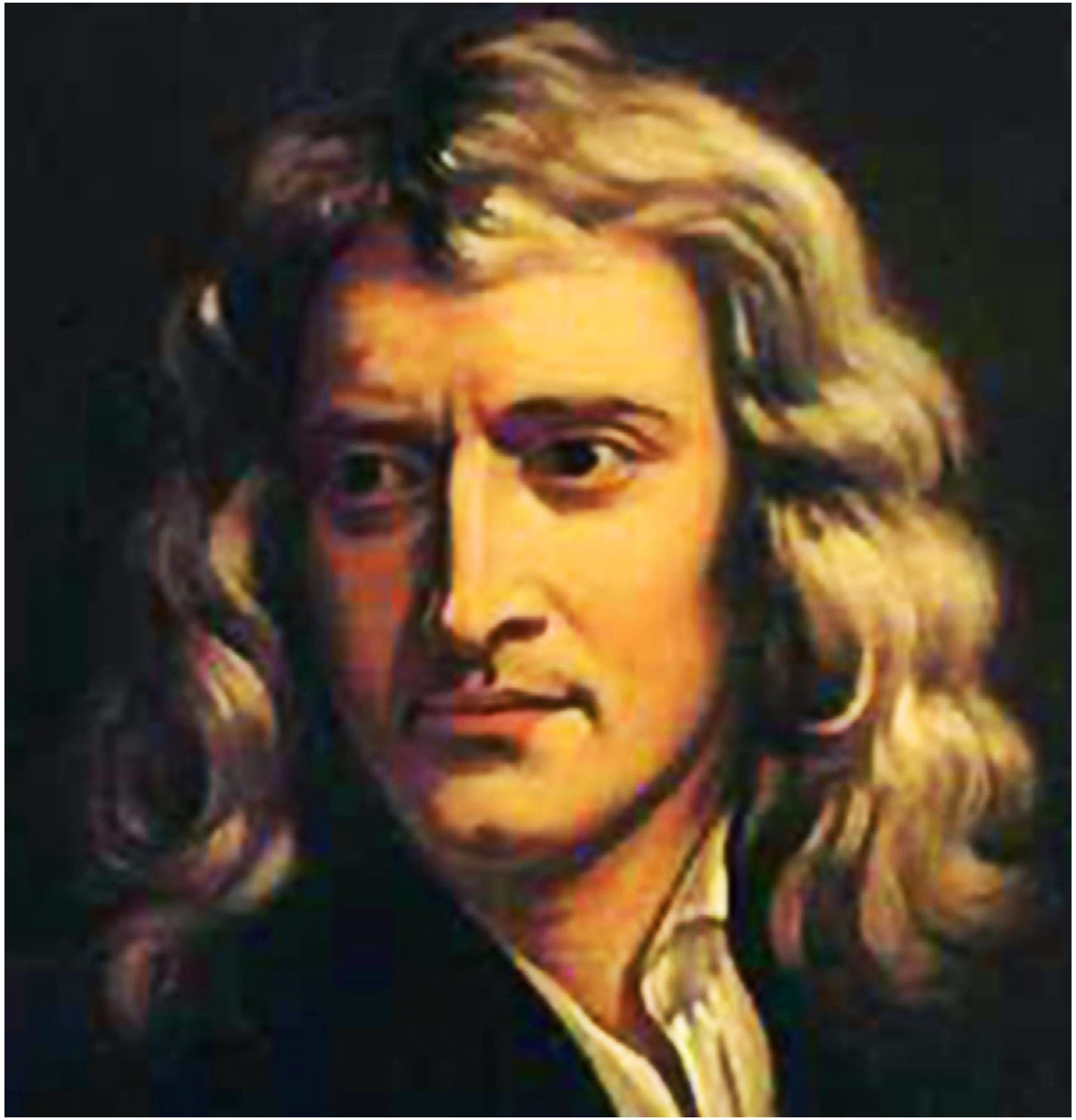
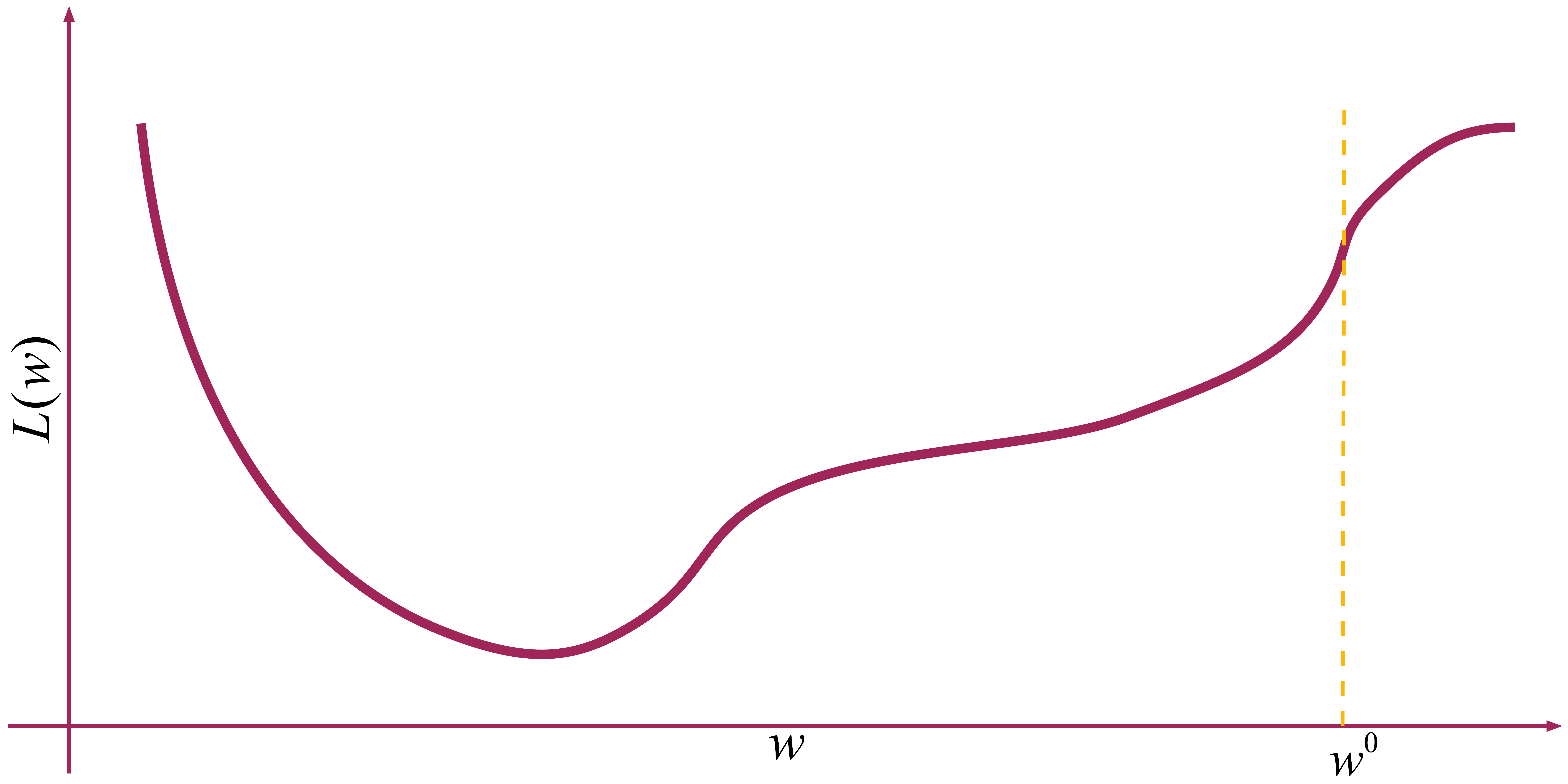
Second order optimization



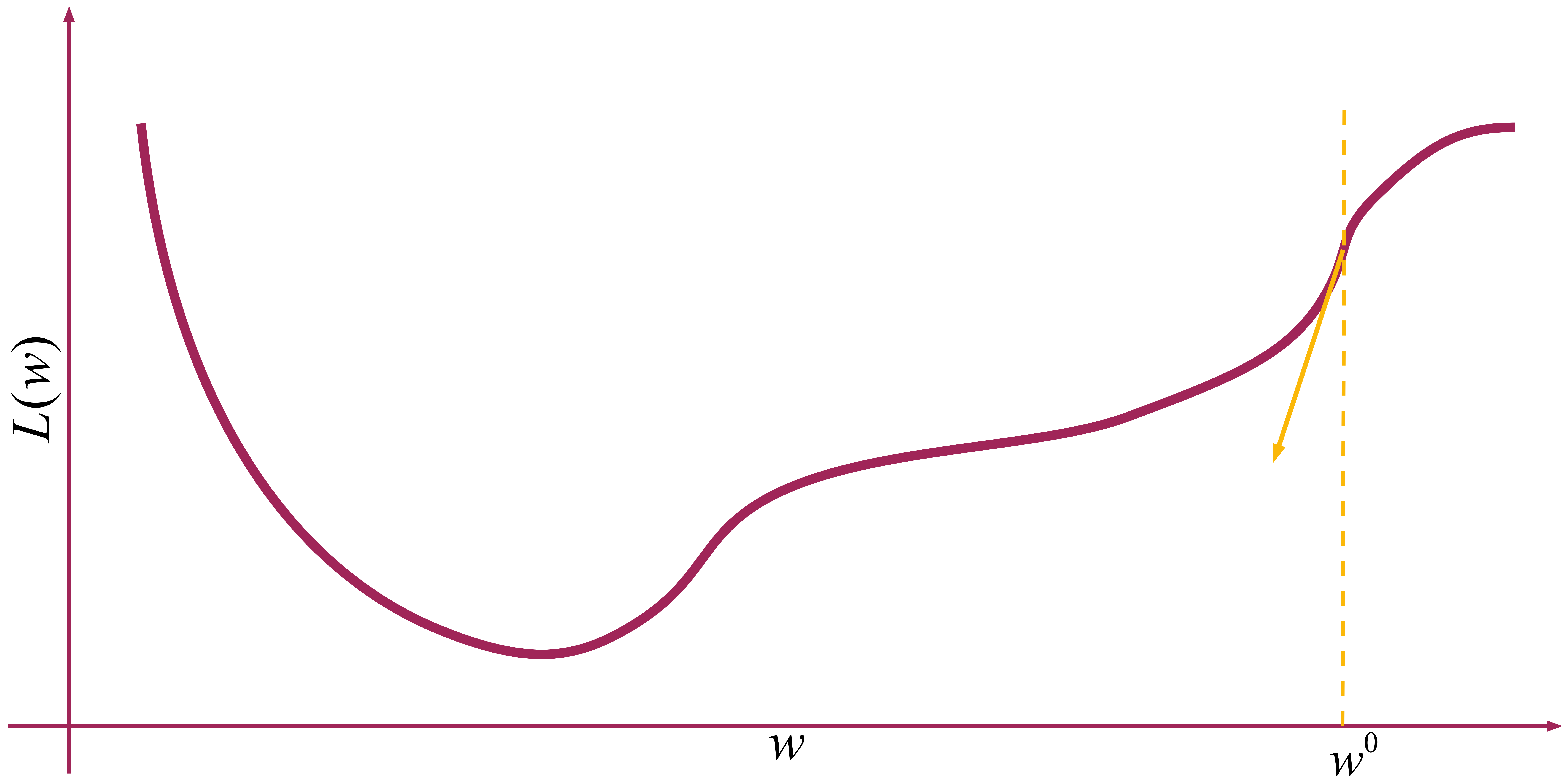
Today you will learn about

- The newton method
- The BFGS method
- The L-BFGS method
- The training of linear regression

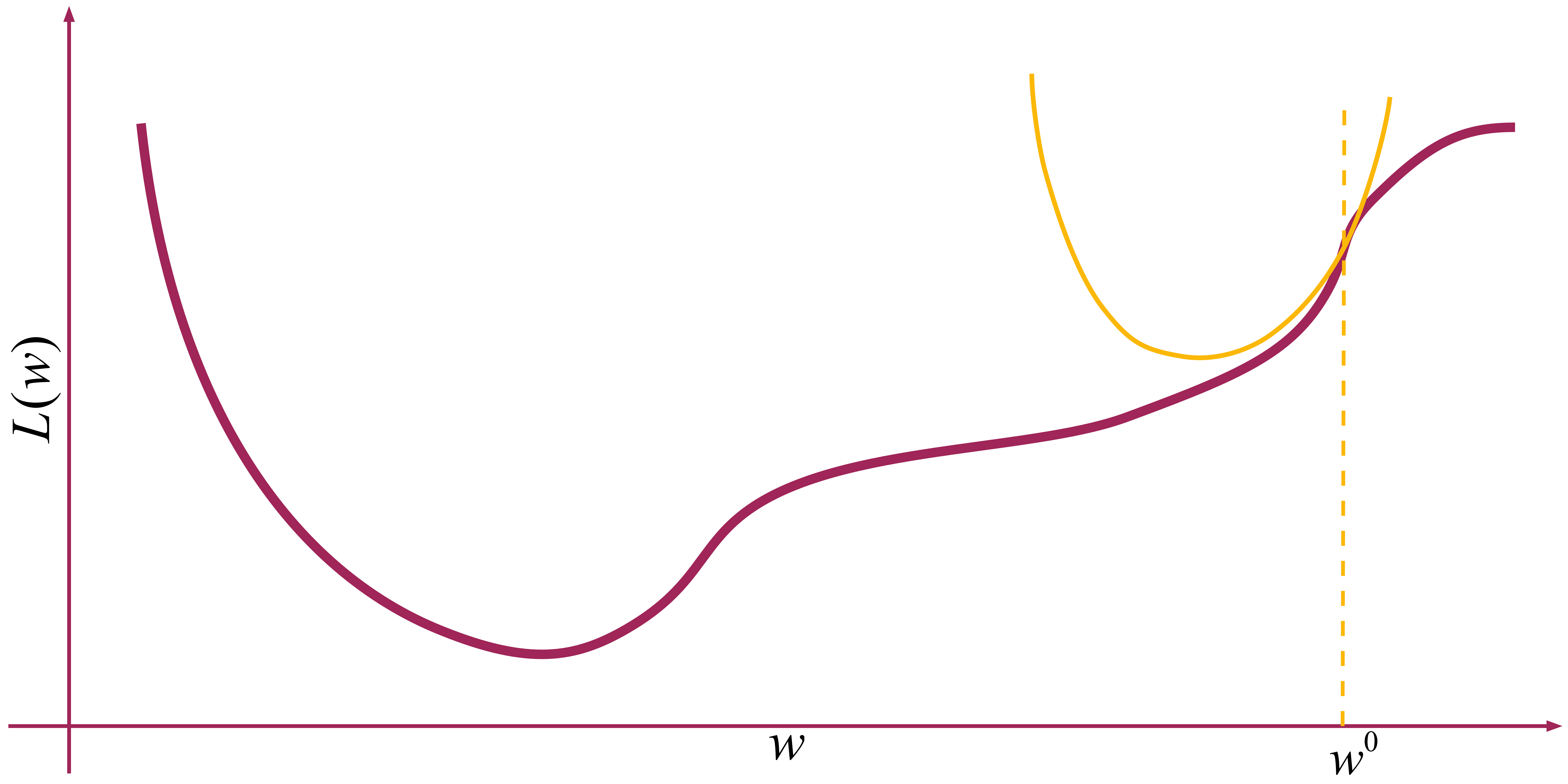
Newton method



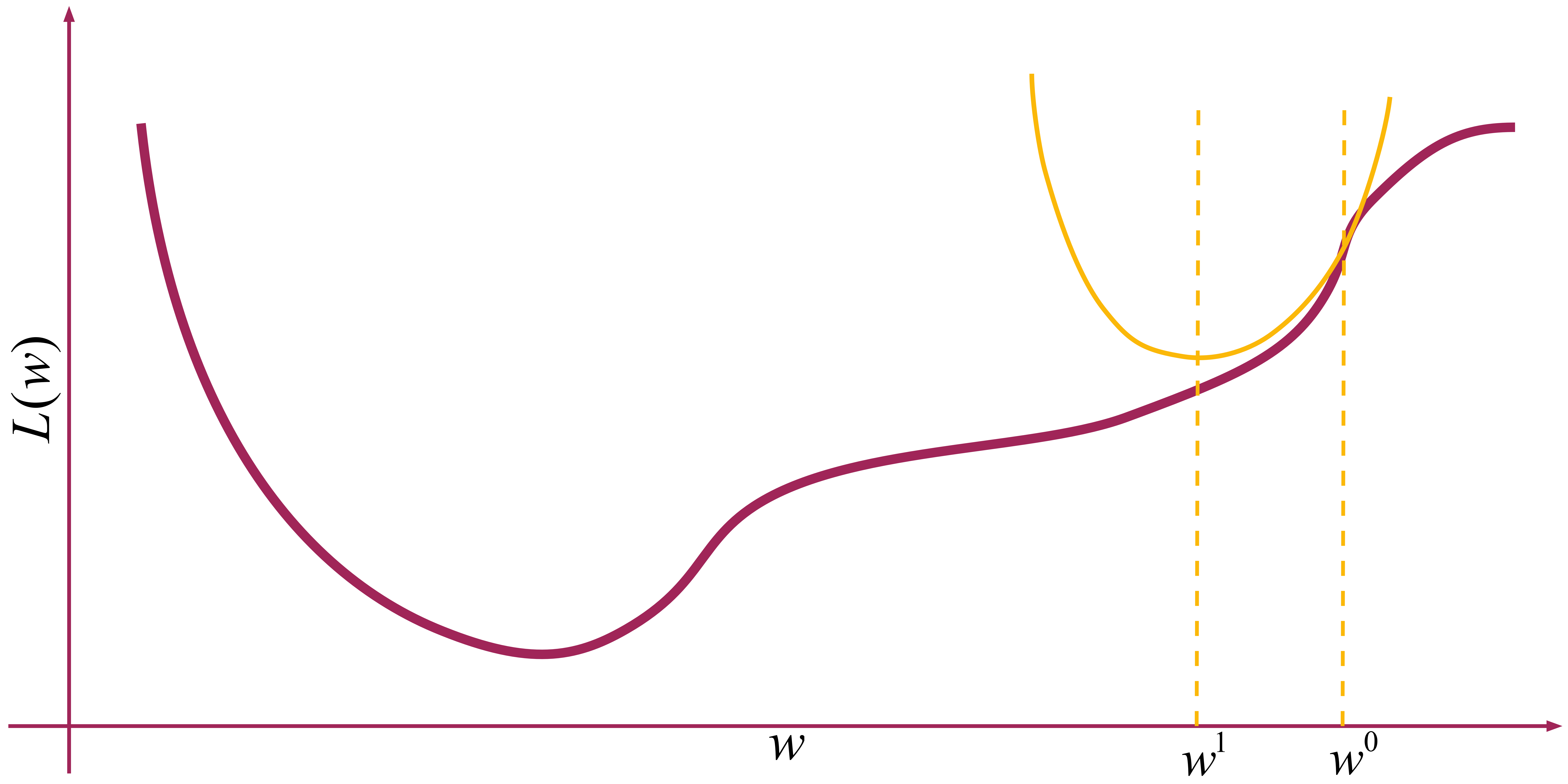
$L(w)$
w
w^0



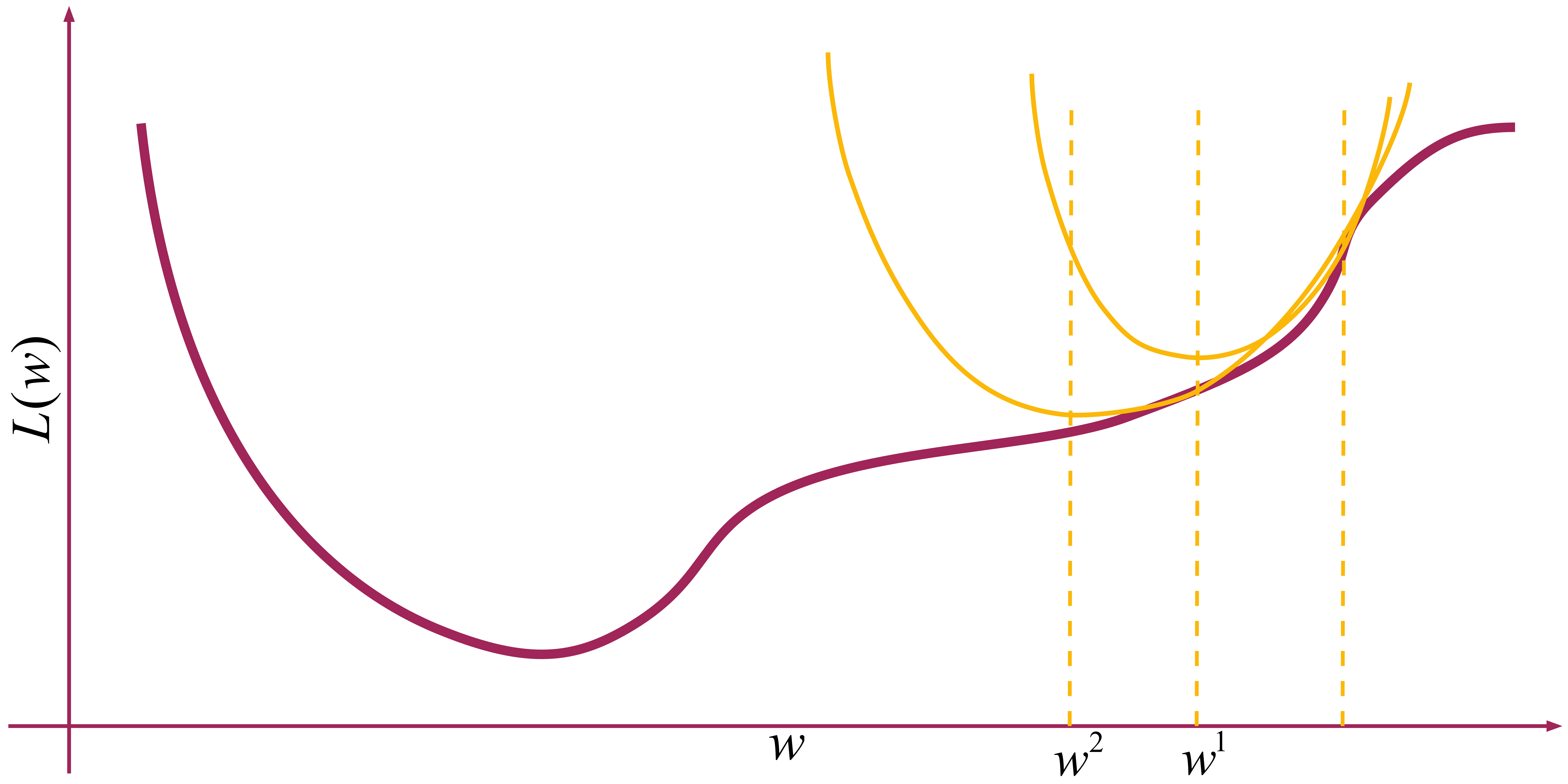
$L(w)$
w
w^0

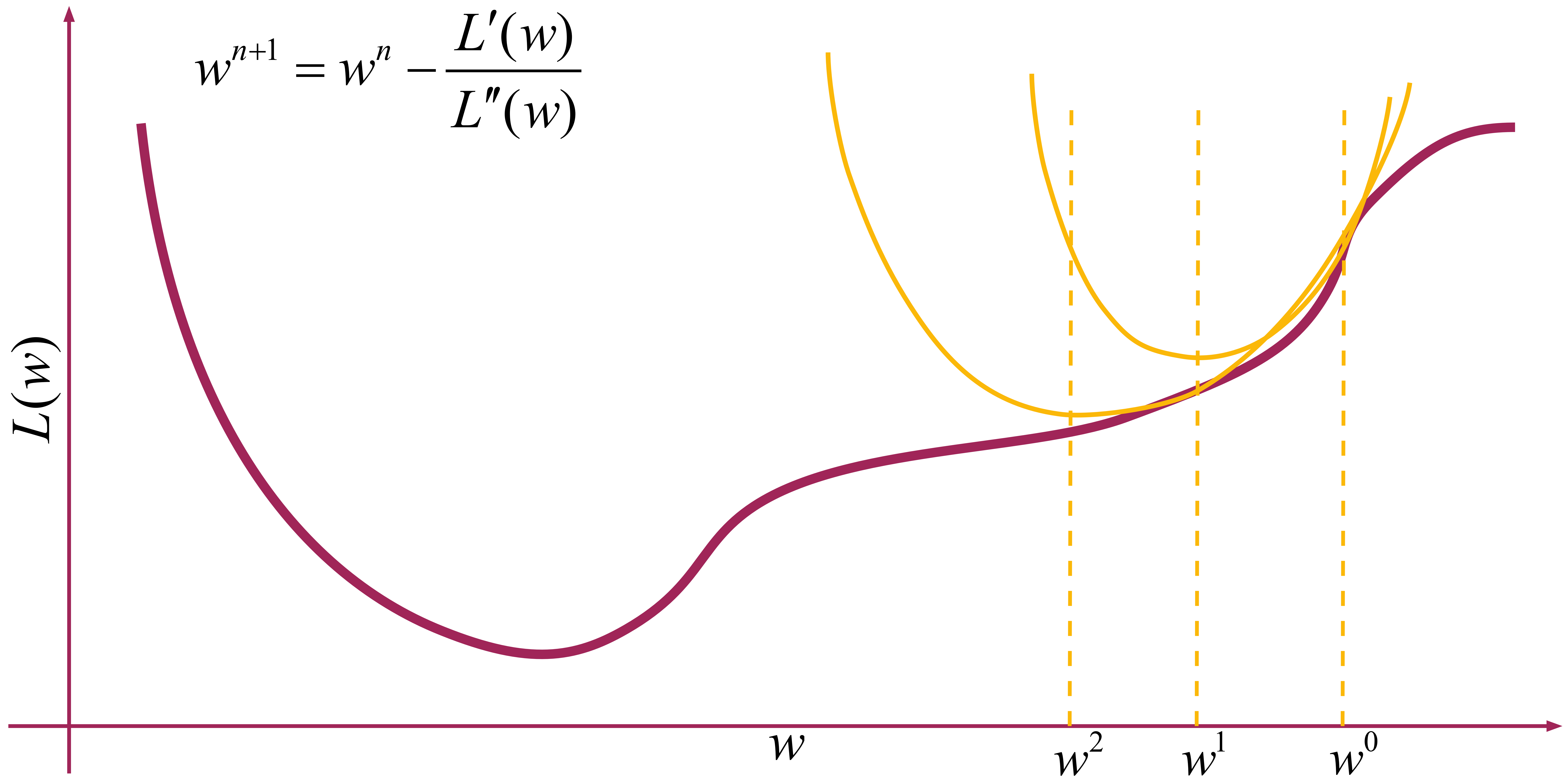


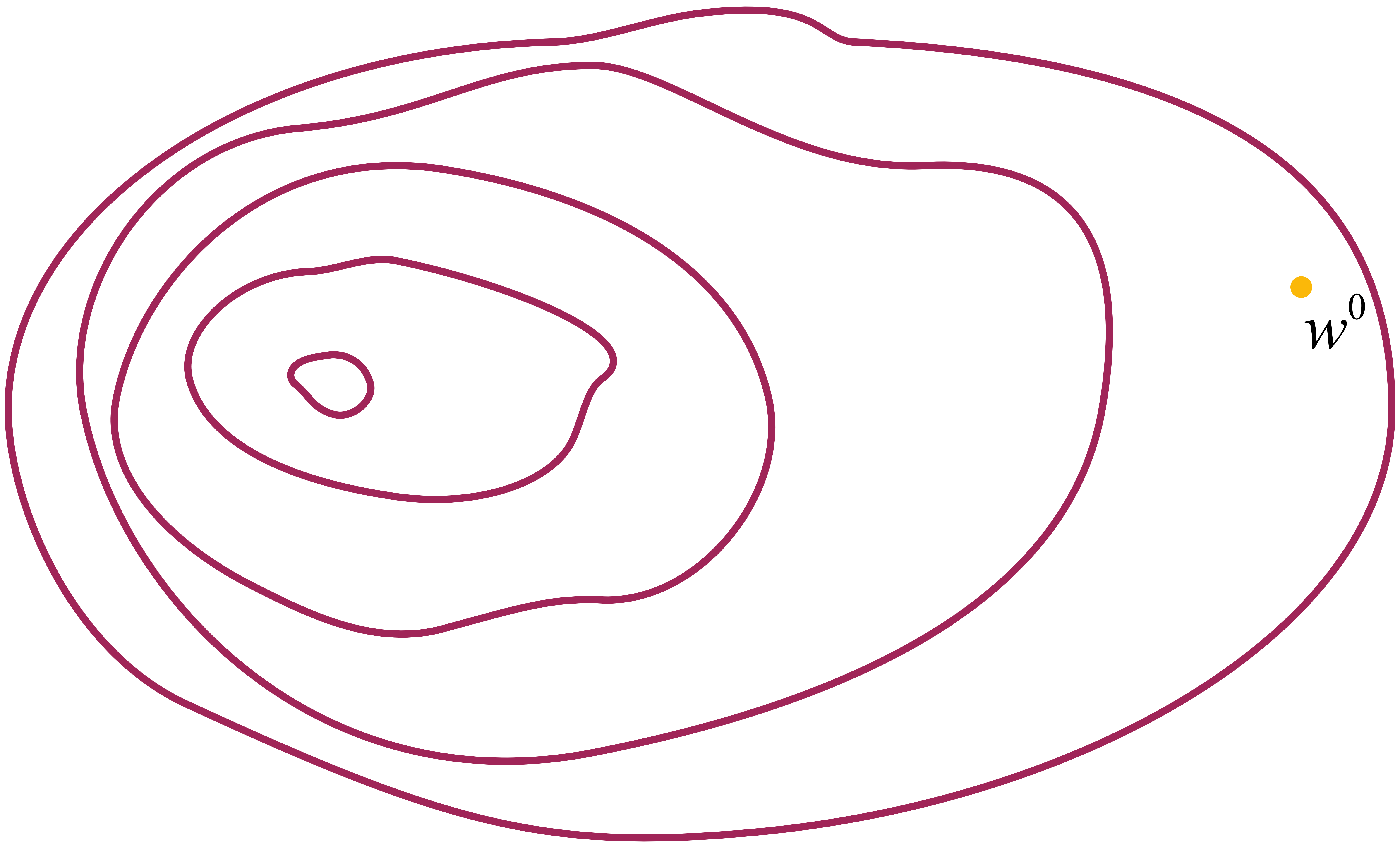
$L(w)$
w
w^0

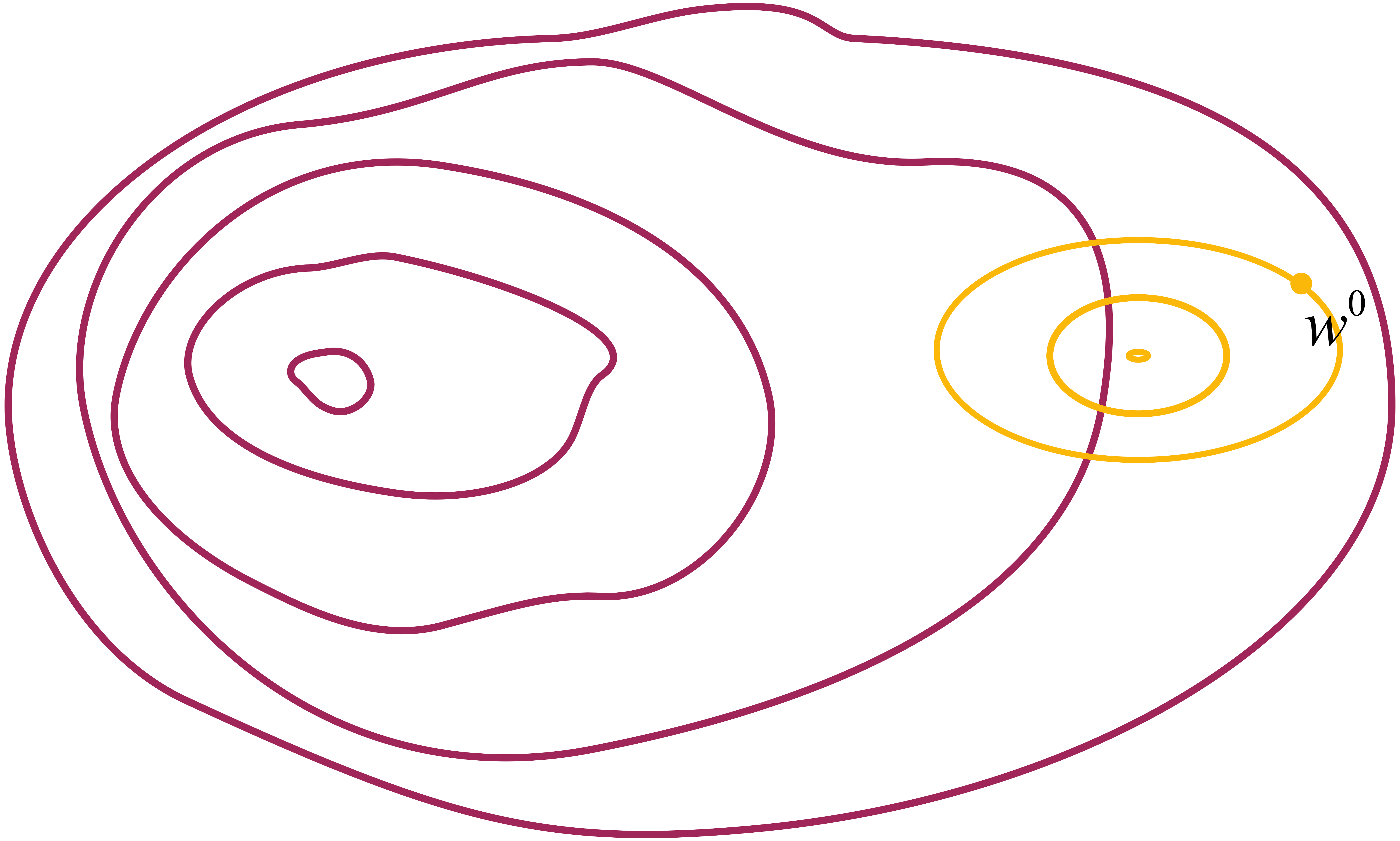


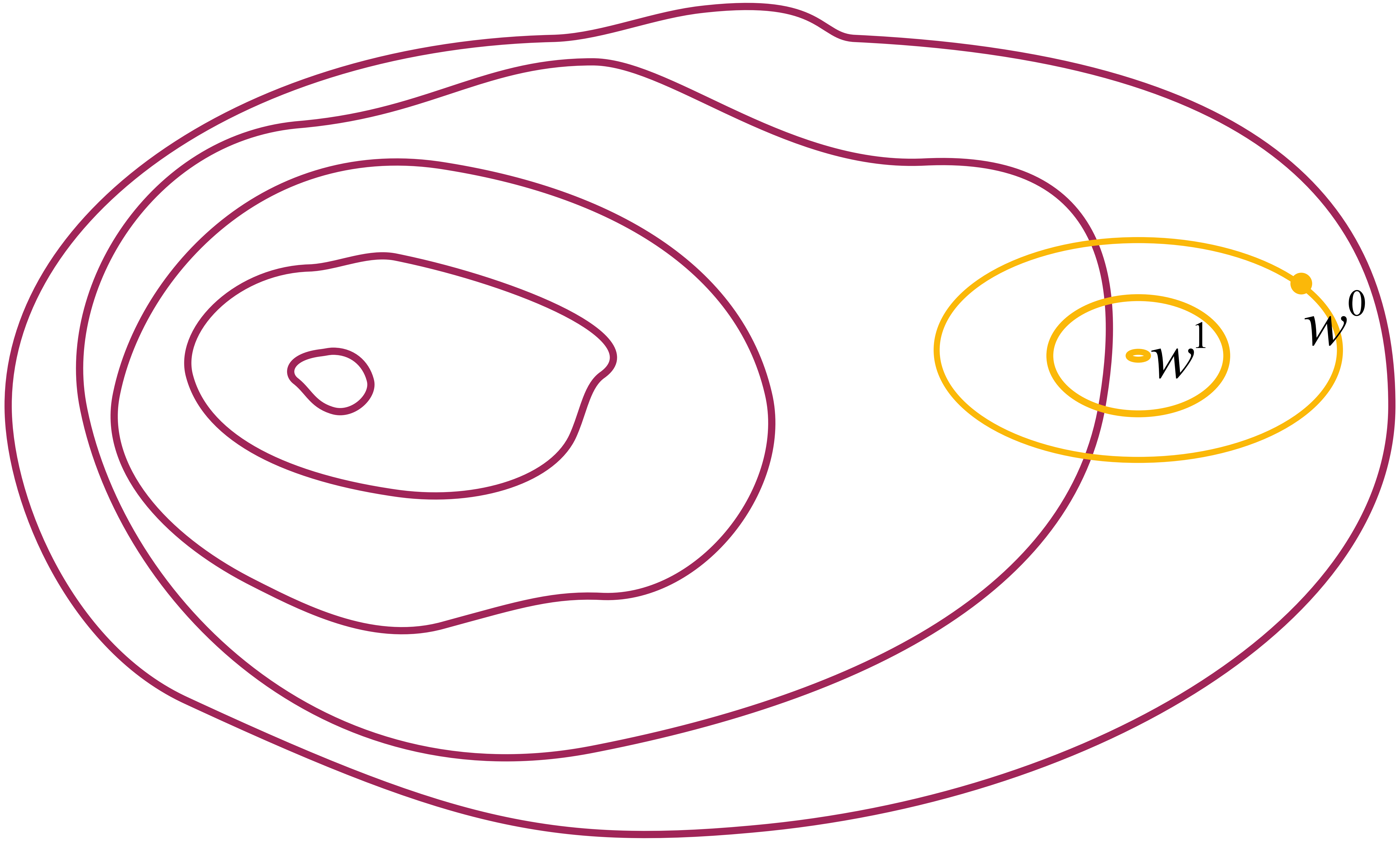
$L(w)$
w
w^1
w^0

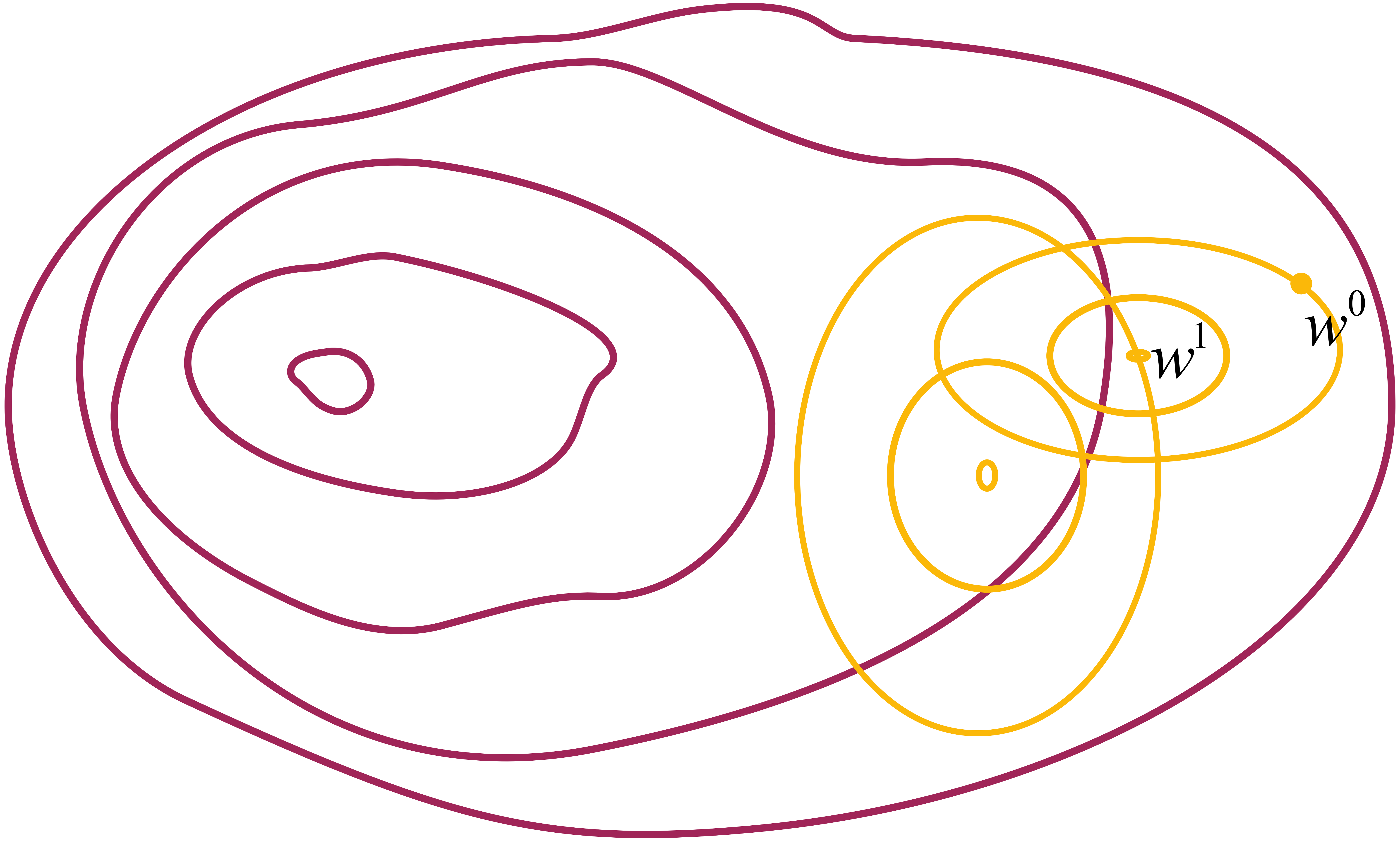

$L(w)$
w
w^2
w^1


$w^{n+1} = w^n - \frac{L'(w)}{L''(w)}$
$L(w)$
w
w^2
w^1
w^0

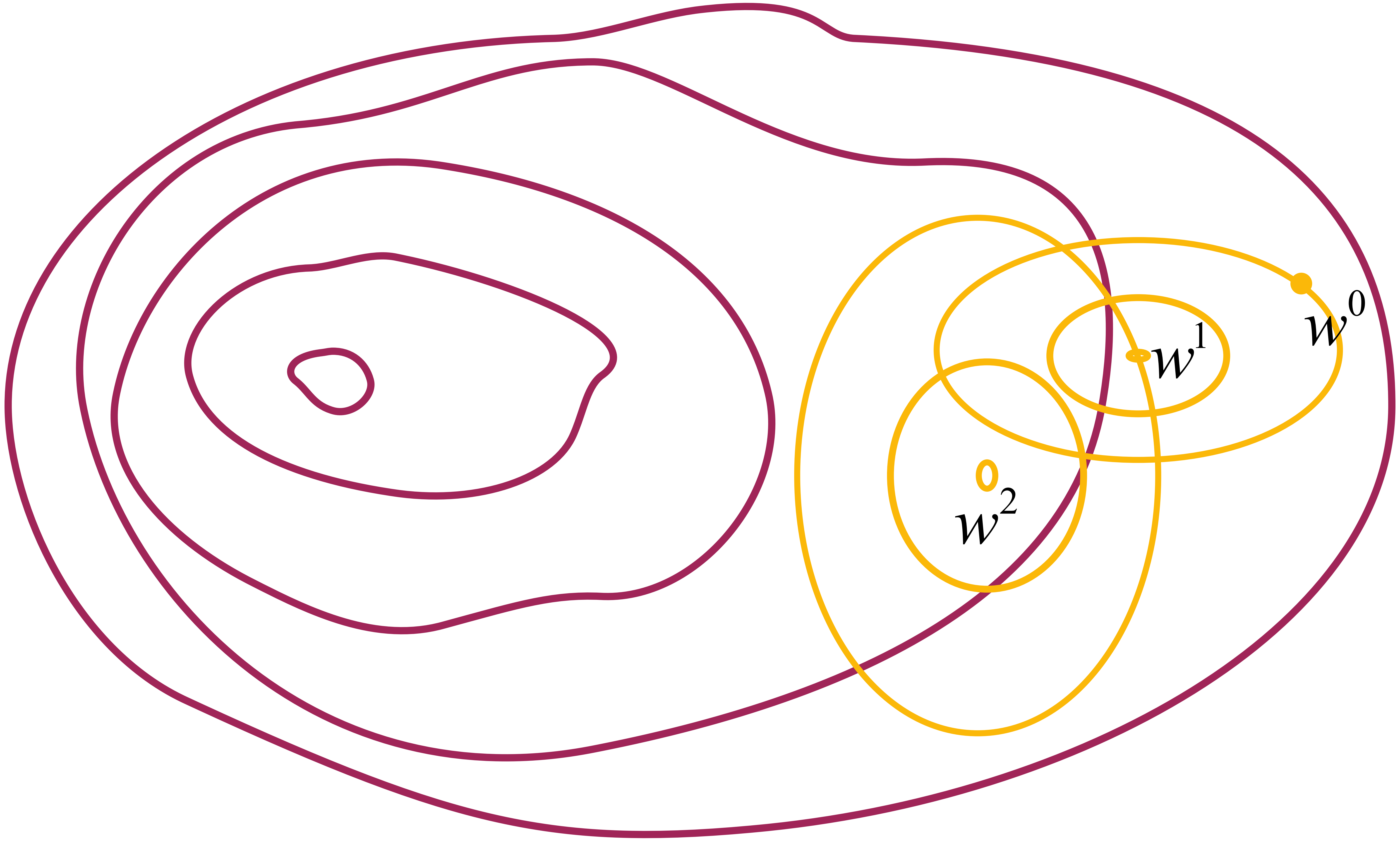

w^0


w^0


w^1
w^0



w^1
w^0


w^0
w^1
w^2

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

$$L'(w) \Rightarrow \nabla L(w) = \begin{pmatrix} \frac{\partial L(w)}{\partial w_0} \\ \cdots \\ \frac{\partial L(w)}{\partial w_n} \end{pmatrix}$$

$$L''(w) \Rightarrow H\big(L(w)\big) = \begin{pmatrix} \frac{\partial^2 L(w)}{\partial w_0^2} & \cdots & \frac{\partial^2 L(w)}{\partial w_n \partial w_0} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 L(w)}{\partial w_0 \partial w_n} & \cdots & \frac{\partial^2 L(w)}{\partial w_n^2} \end{pmatrix}$$

$$L'(w) \Rightarrow \nabla L(w) = \begin{pmatrix} \frac{\partial L(w)}{\partial w_0} \\ \cdots \\ \frac{\partial L(w)}{\partial w_n} \end{pmatrix}$$

$$L''(w) \Rightarrow H\big(L(w)\big) = \begin{pmatrix} \frac{\partial^2 L(w)}{\partial w_0^2} & \cdots & \frac{\partial^2 L(w)}{\partial w_n \partial w_0} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 L(w)}{\partial w_0 \partial w_n} & \cdots & \frac{\partial^2 L(w)}{\partial w_n^2} \end{pmatrix}$$

$$w^{n+1} = w^n - \frac{L'(w)}{L''(w)} \Rightarrow w^{n+1} = w^n - H^{-1}\big(L(w)\big) \cdot \nabla L(w)$$

$$w^{n+1} = w^n - H^{-1}\left(L(w)\right) \cdot \nabla L(w)$$

Problem

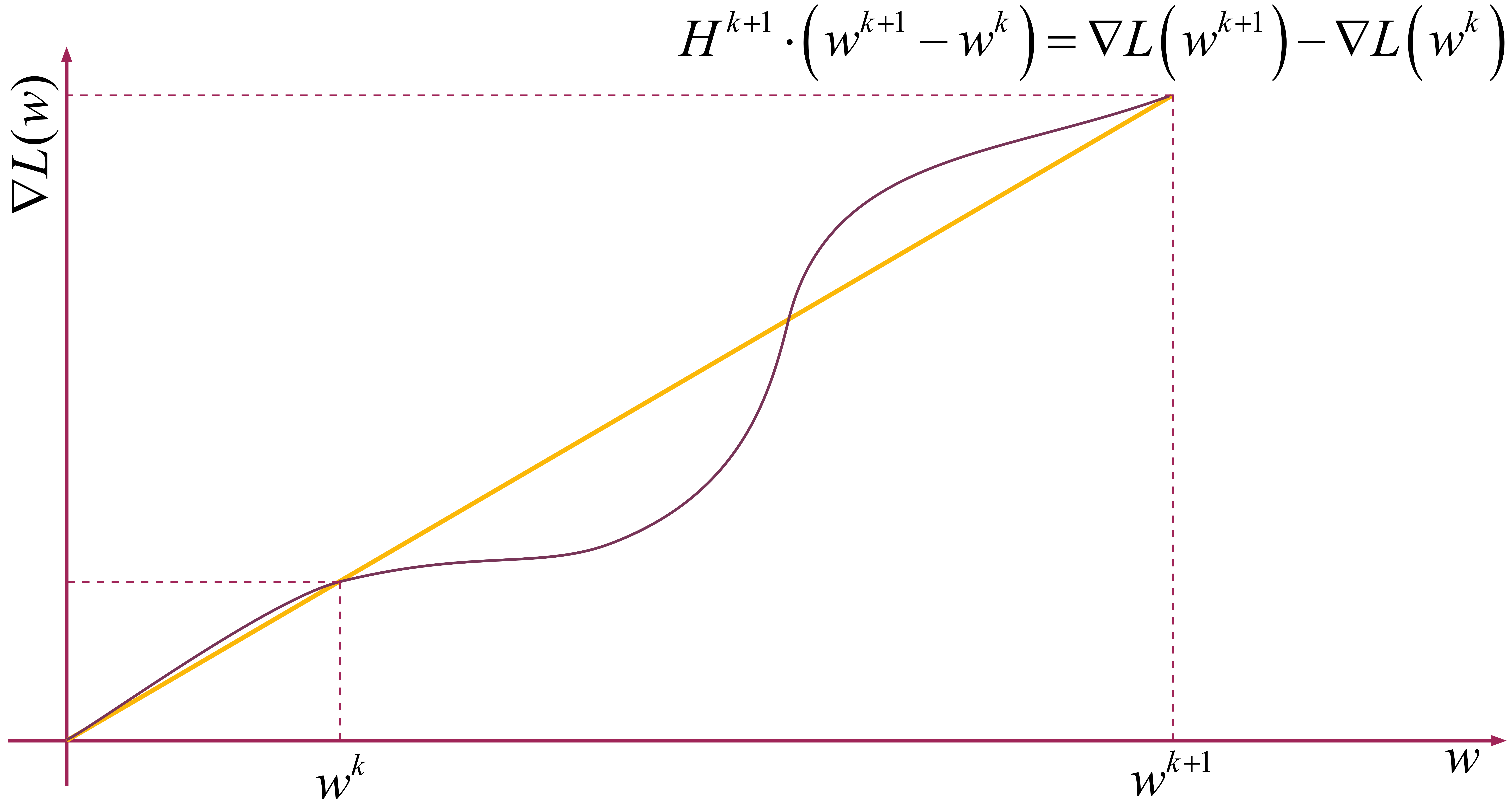
$H^{-1}\left(L(w)\right)$	n^3 operations
10 features	1 000 operations
100 features	1 000 000 operations
1 000 features	1 000 000 000 operations

BFGS

Broyden–Fletcher–Goldfarb–Shanno algorithm

Approximate value

$$H^{-1}\left(L\left(w^{n+1}\right)\right) = H^{-1}\left(L\left(w^{n}\right)\right) + \delta \quad = U \; V^{T}$$

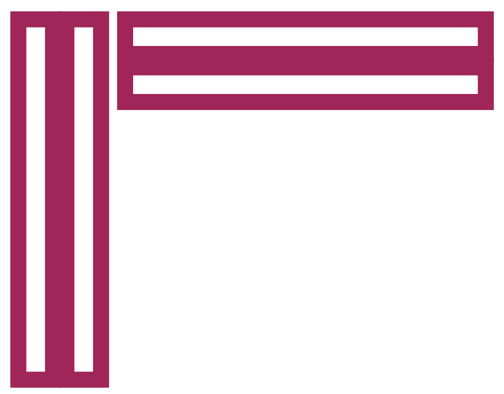

$H^{k+1} \cdot \left(w^{k+1} - w^{k}\right) = \nabla L\left(w^{k+1}\right) - \nabla L\left(w^{k}\right)$
$\nabla L(w)$
w^{k}
w^{k+1}
w

$H^{-1}\left(L(w)\right)$	n^2 operations
10 features	100 operations
100 features	10 000 operations
1 000 features	1 000 000 operations

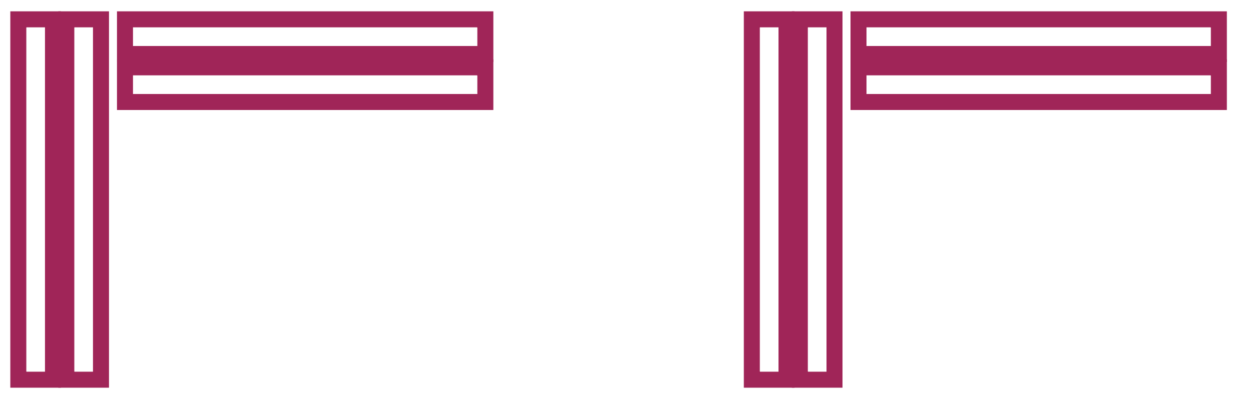
L-BFGS

Limited memory
Broyden–Fletcher–Goldfarb–Shanno
algorithm

$$H^{-1}\left(L\left(w^{k+1}\right)\right) = U^{k} \cdot V^{k^{T}} + H^{-1}\left(L\left(w^{k}\right)\right)$$



$$H^{-1}\left(L\left(w^{k+1}\right)\right) = U^{k} \cdot V^{k^T} + U^{k-1} \cdot V^{(k-1)^T} + H^{-1}\left(L\left(w^{k-1}\right)\right)$$



$$H^{-1}\left(L\left(w^{k+1}\right)\right)=U^{k}\cdot V^{k^{T}}+U^{k-1}\cdot V^{(k-1)^{T}}+\ldots+U^{1}\cdot V^{1^{T}}+H^{-1}\left(L\left(w^{0}\right)\right)$$

$$= \alpha \cdot I$$

$$\begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

$$H^{-1}\left(L\left(w^{k+1}\right)\right)=U^{k}\cdot V^{k^{T}}+U^{k-1}\cdot V^{(k-1)^{T}}+\ldots+U^{1}\cdot V^{1^{T}}+\alpha\cdot I$$

$$H^{-1}\left(L\left(w^{k+1}\right)\right) = U^{k} \cdot V^{k^{T}} + U^{k-1} \cdot V^{(k-1)^{T}} + \ldots + \cancel{U^{1} \cdot V^{1^{T}}} + \alpha \cdot I$$

last N = 10

$$H^{-1}\left(L\left(w^{k+1}\right)\right) = U^{k} \cdot V^{k^{T}} + U^{k-1} \cdot V^{(k-1)^{T}} + \ldots + U^{(k-9)} \cdot V^{(k-9)^{T}} + \alpha \cdot I$$

$$w^{n+1} = w^n - H^{-1}\big(L(w)\big) \cdot \nabla L(w)$$

$$w^{n+1} = w^n - \left(U^k \cdot V^{k^T} + U^{k-1} \cdot V^{(k-1)^T} + \ldots + U^{(k-9)} \cdot V^{(k-9)^T} + \alpha \cdot I\right) \cdot \nabla L(w)$$

$$w^{n+1} = w^n - U^k \cdot V^{k^T} \cdot \nabla L(w) - U^{k-1} \cdot V^{(k-1)^T} \cdot \nabla L(w) - \ldots - U^{(k-9)} \cdot V^{(k-9)^T} \cdot \nabla L(w) - \alpha \cdot \nabla L(w)$$

n

Linear regression

Linear regression

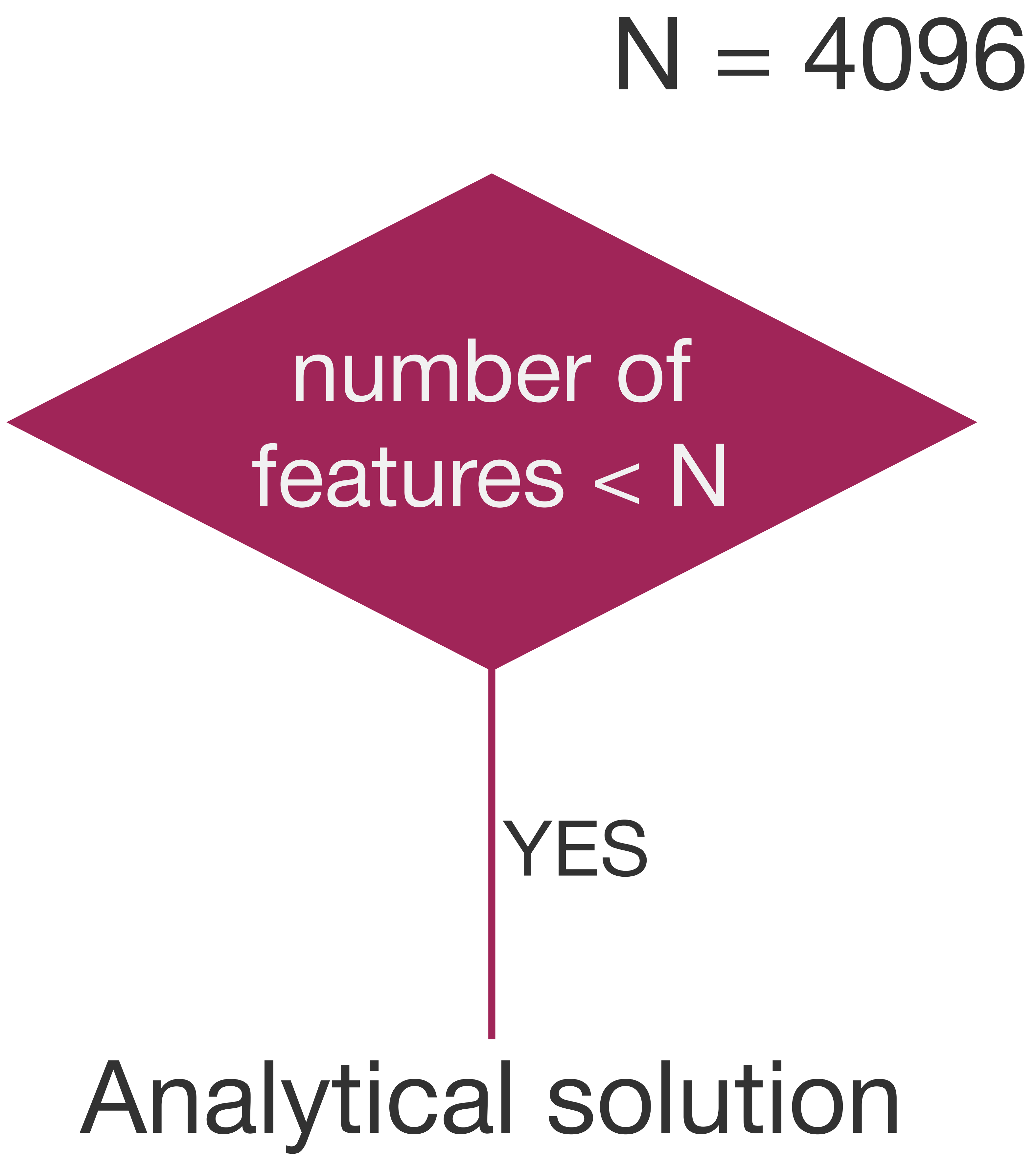
number of features < N

Linear regression

$N = 4096$

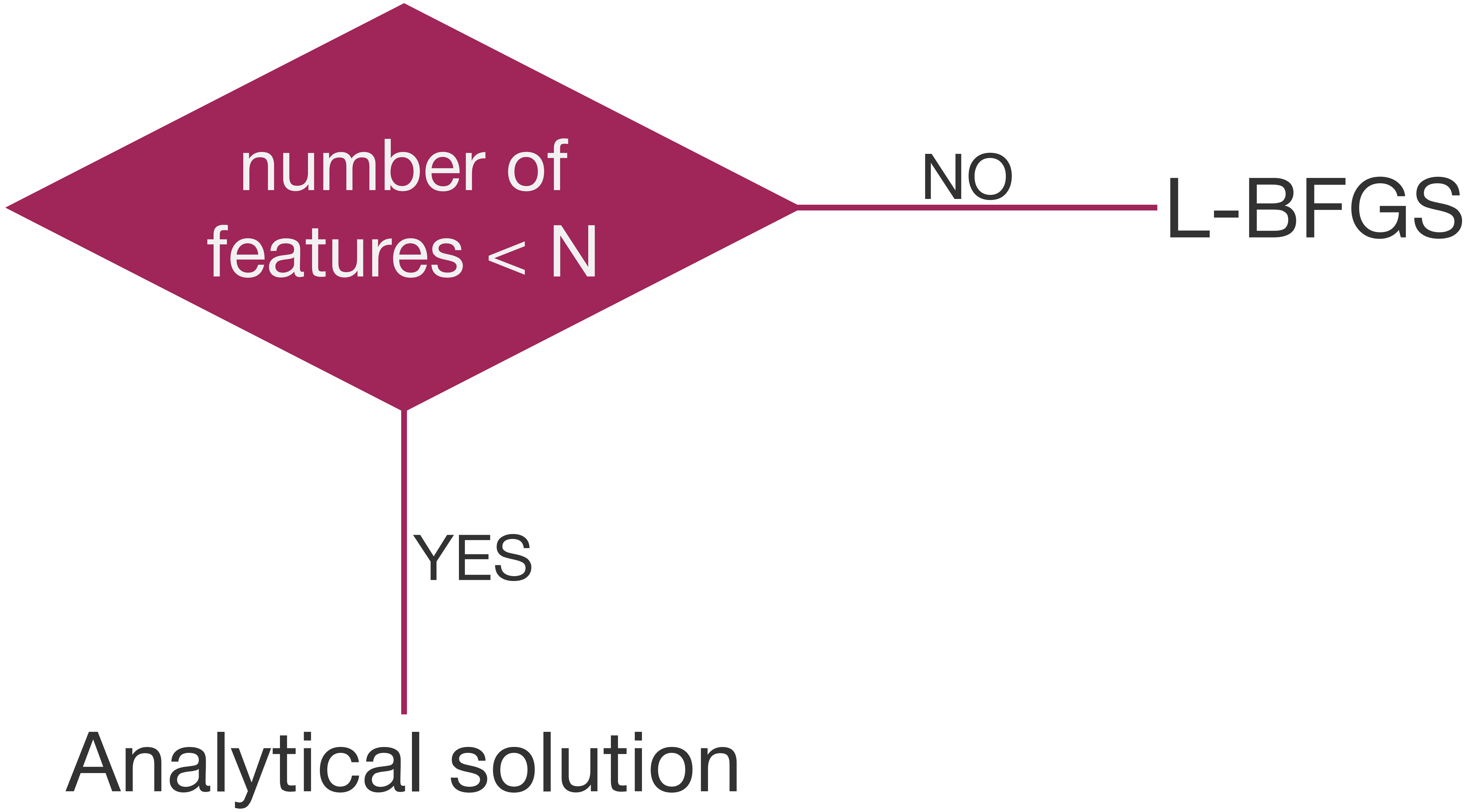
number of
features $< N$

Linear regression



Linear regression

N = 4096



Today you have learned about:

- The newton method
- The BFGS method
- The L-BFGS method
- The training of linear regression