Lecture - 21

Training Neural Network - 2

From Last Lecture ...

Quiz questions in last class

Q1. Where will be the Minima of this:

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$w = -b/a$$

Quiz questions in last class (T/F)

2. The derivative of function is zero only at maxima or minima and no other points.

3. The second derivative of a function is negative at the minima.

4. Gradient ascent method can be used to find the maxima of a function.

5. Gradient descent never stuck at a local minima.

6. Backpropagation and Gradient Descent are the two different ways using anyone of them neural networks can be optimised.

7. Backpropogation is used to compute derivative of the error surface.

Quiz questions in last class (T/F)

8. In gradient descent, irrespective of initialisation, solution is always the same.

9. In gradient descent, initialisation does not matter if there is one and only one minima and no saddle point.

10. Neural network is a nested function of inputs and learnable weights.

The problem

$$W^* = \arg\min_W \sum_{i=1}^n \mathcal{DIV}(\mathcal{F}(x_i, W), \hat{y}_i)$$

Backprop

Let us learn it using paper and pen!

Backpropagation Algorithm

begin initialize network topology (# hidden units), $\mathbf{w}$, criterion $\theta, \eta, r \leftarrow 0$

    do $r \leftarrow r + 1$ (increment epoch)

        $m \leftarrow 0;\ \Delta w_{ij} \leftarrow 0;\ \Delta w_{jk} \leftarrow 0$

        do $m \leftarrow m + 1$

            $\mathbf{x}^m \leftarrow$ select pattern

            $\Delta w_{ij} \leftarrow \Delta w_{ij} + \eta \delta_j x_i;\quad \Delta w_{jk} \leftarrow \Delta w_{jk} + \eta \delta_k y_j^H$

        until $m = n$

        $w_{ij} \leftarrow w_{ij} + \Delta w_{ij};\quad w_{jk} \leftarrow w_{jk} + \Delta w_{jk}$

    until $\nabla \mathcal{L}(W) < \theta$

return w

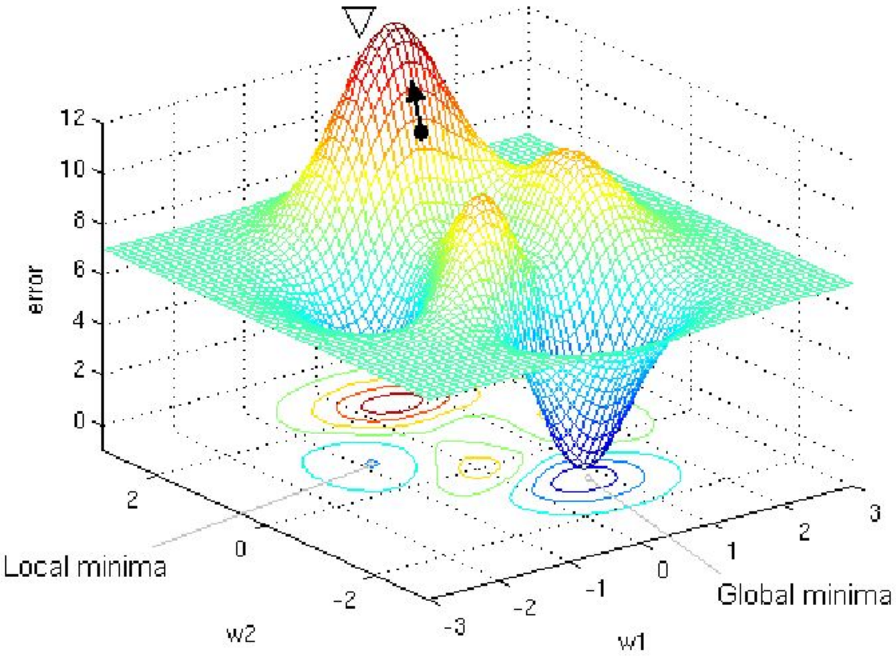
end

Module 1: The error surface, convergence, learning rate

So far ...

- Neural nets can be trained via gradient descent that minimizes a loss function
- Backpropagation can be used to derive the derivatives of the loss

The Error Surface



The Error Surface

Popular hypothesis: **In a large network**

- Saddle points are far more common than local minima
- Local minima are not too bad (many recent studies)

- Grzegorz Swirszcz, Wojciech Marian Czarnecki, Razvan Pascanu: **Local minima in training of deep networks.** CoRR abs/1611.06310 (2016)
- Anna Choromanska, Mikael Henaff, Michaël Mathieu, Gérard Ben Arous, Yann LeCun: **The Loss Surfaces of Multilayer Networks.** AISTATS 2015

The Error Surface

Popular hypothesis: **In a large network**

- Saddle points are far more common than local minima
- Local minima are not too bad

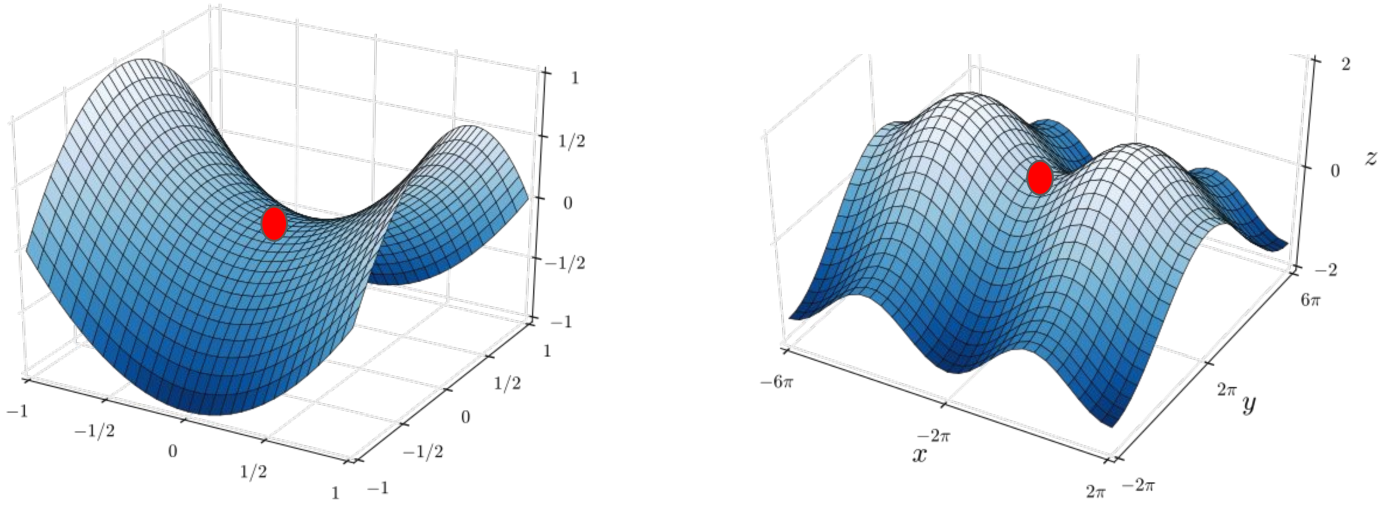
What is a Saddle Point

- A point where gradient is zero, and the value of the error surface increases in some directions but decreases in some other directions.

The Error Surface

What is a Saddle Point

- A point where gradient is zero, and the value of the error surface increases in some directions but decreases in some other directions.



The Error Surface

What is a Saddle Point

- A point where gradient is zero, and the value of the error surface increases in some directions but decreases in some other directions.
- Gradient descent often stuck at saddle point

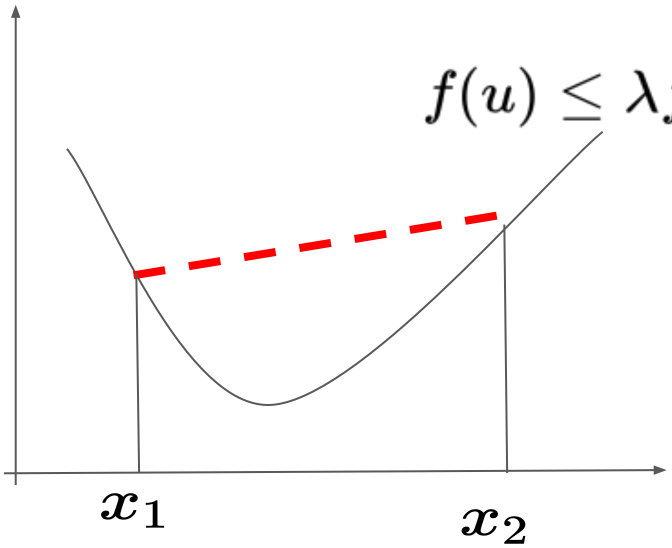
So far ...

- Neural nets can be trained via gradient descent that minimizes a loss function
- Backpropagation can be used to derive the derivatives of the loss
- For large networks, the loss function may have a large number of unpleasant saddle points
  – Which backpropagation may find

Convergence of gradient descent

- In the discussion so far we have assumed the training arrives at a local minimum
- Does it always converge?
- How long does it take?
- Hard to analyze for an MLP, but we can look at the problem through the lens of convex optimization

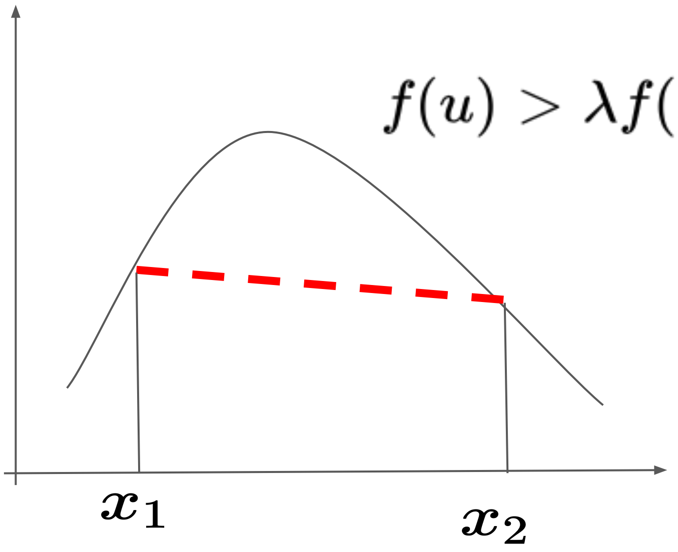
Convex Function



For any two points x_1 and x_2:

$$f(u) \leq \lambda f(x_1) + (1-\lambda) f(x_2), \forall u \in [x_1, x_2], \lambda \in \{0, 1\}$$

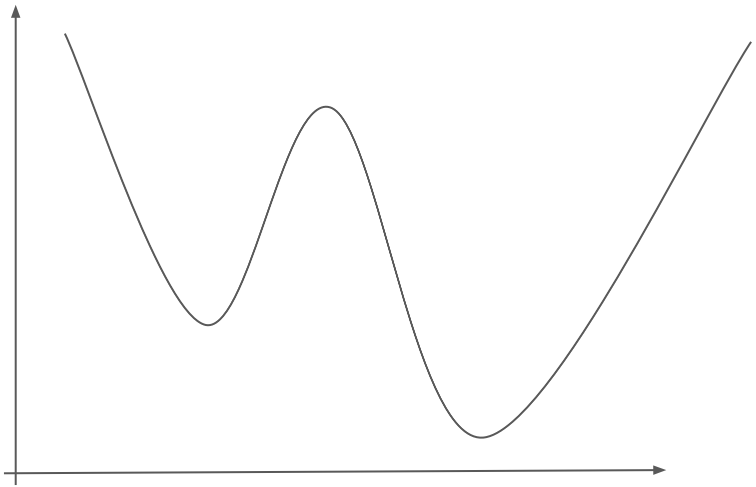
Concave Function



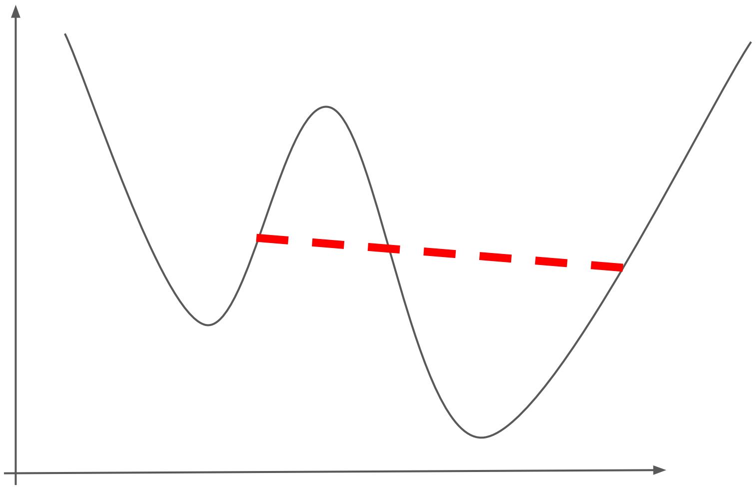
For any two points x_1 and x_2 :

$$f(u) > \lambda f(x_1) + (1-\lambda) f(x_2), \forall u \in [x_1, x_2], \lambda \in \{0, 1\}$$

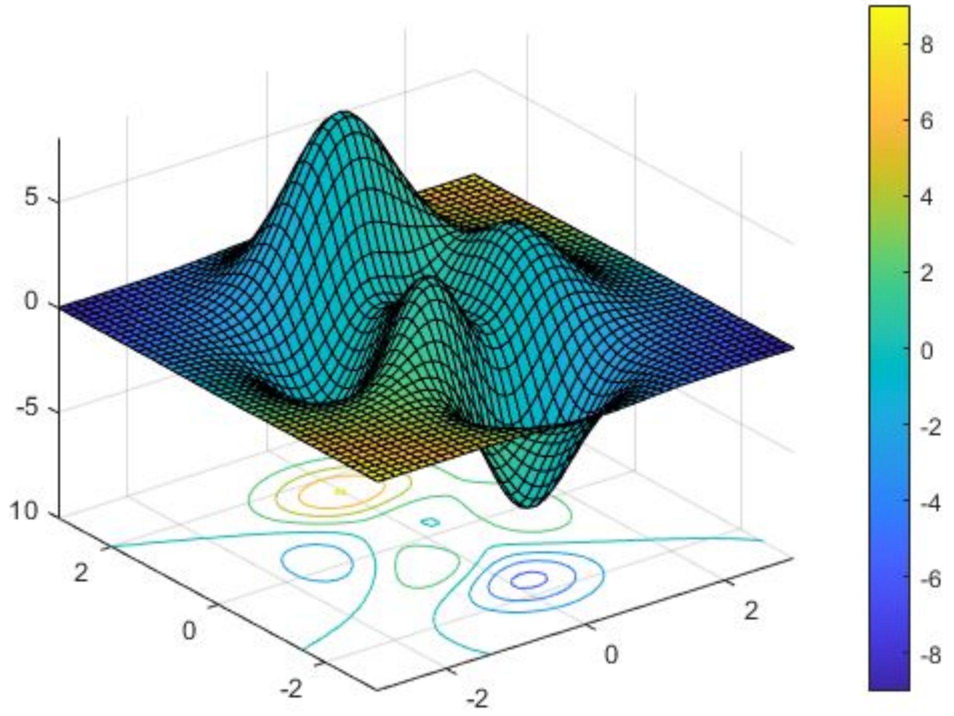
Non-convex Function



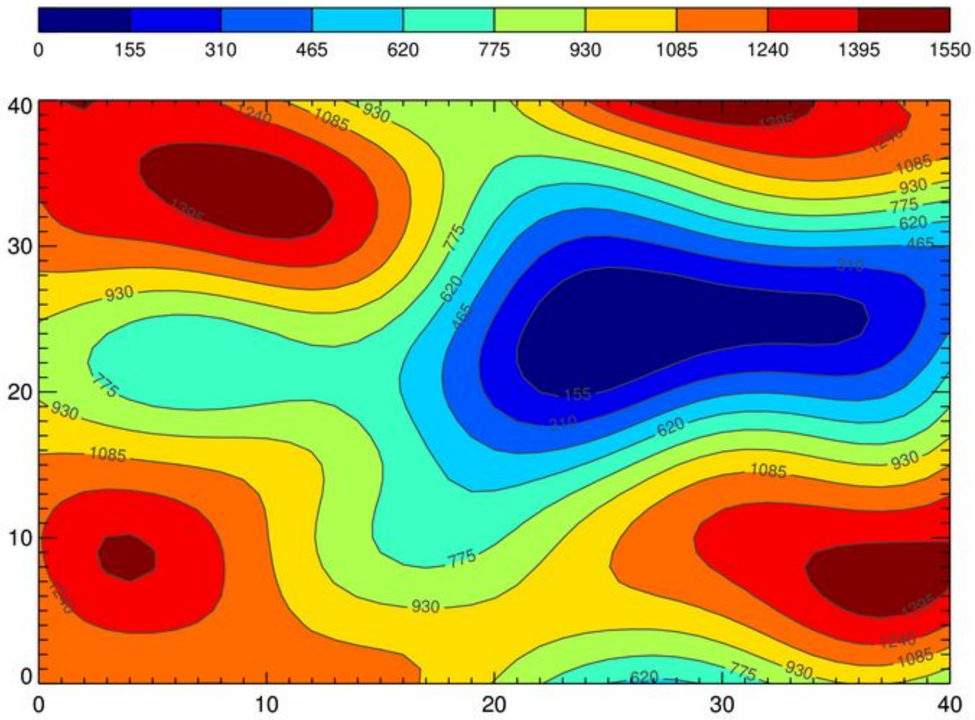
Non-convex Function



Contour representation



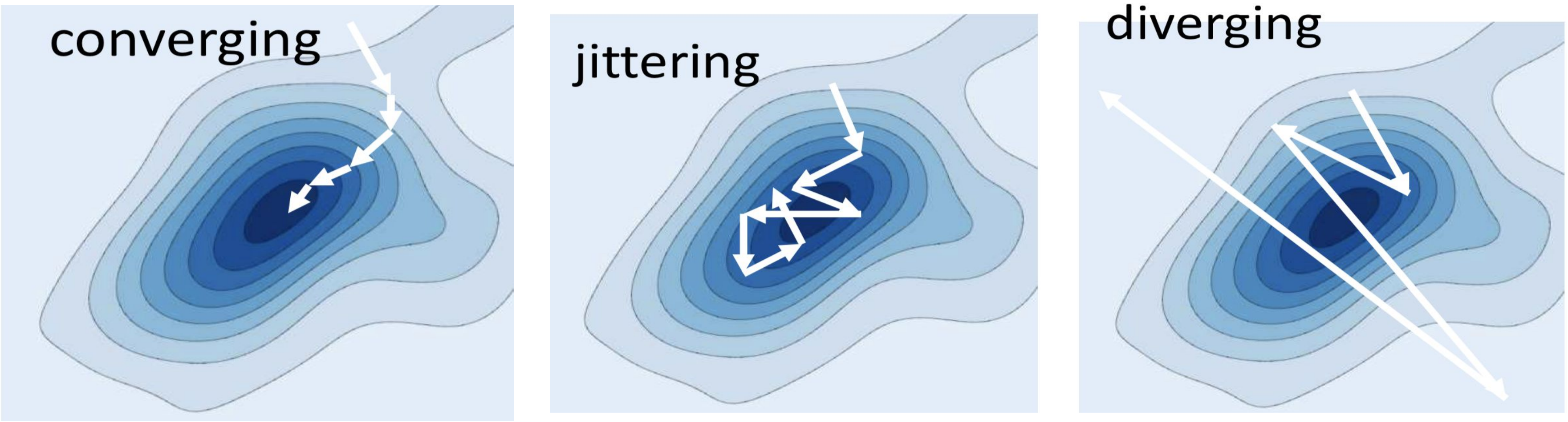
Contour representation



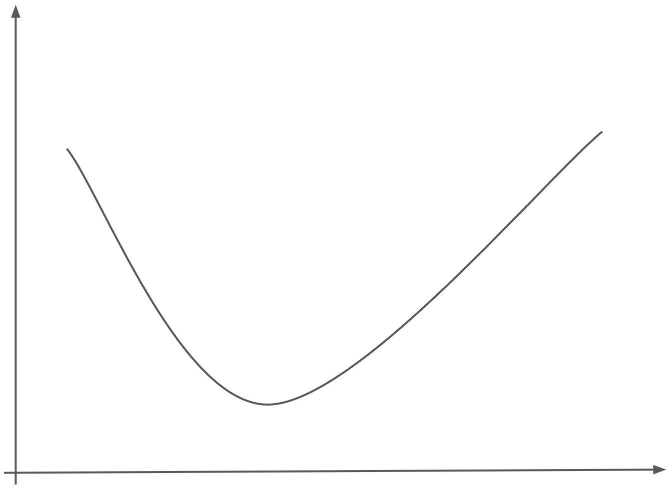
Convergence of Gradient Descent

- An iterative algorithm is said to converge to a solution if the value updates arrive at a fixed point
  – Where the gradient is 0 and further updates do not change the estimate

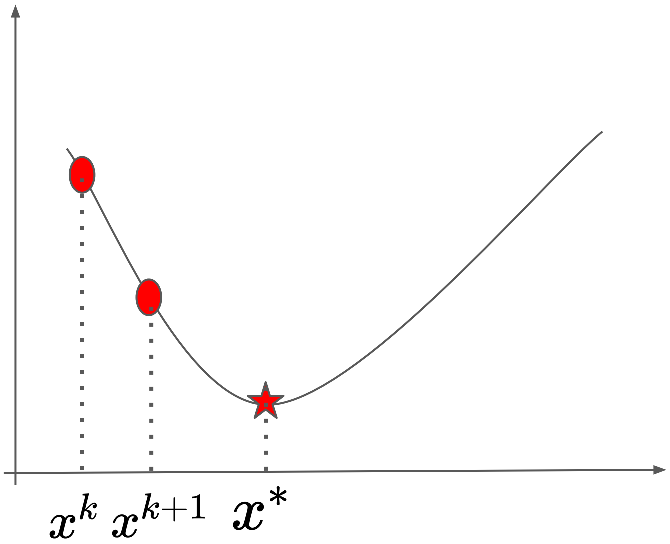
Convergence of Gradient Descent



Convergence Rate

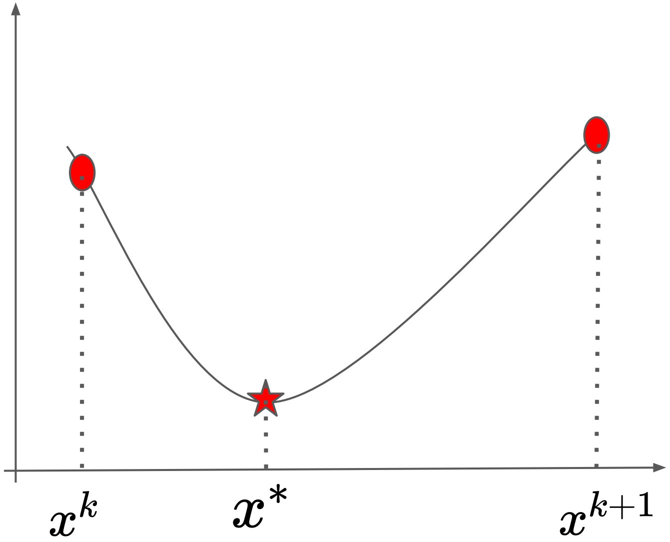


Convergence Rate



$$R = \frac{|f(x^*) - f(x^{k+1})|}{|f(x^*) - f(x^k)|}$$

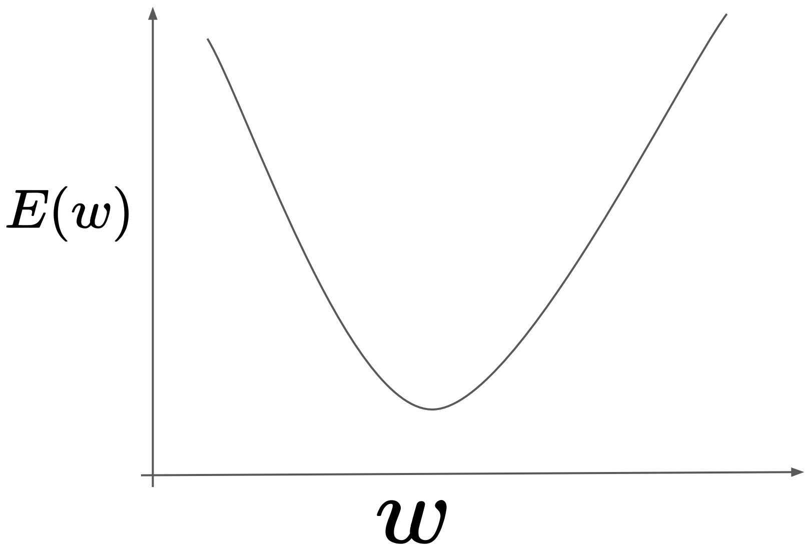
Convergence Rate



$$R = \frac{|f(x^*) - f(x^{k+1})|}{|f(x^*) - f(x^k)|}$$

Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

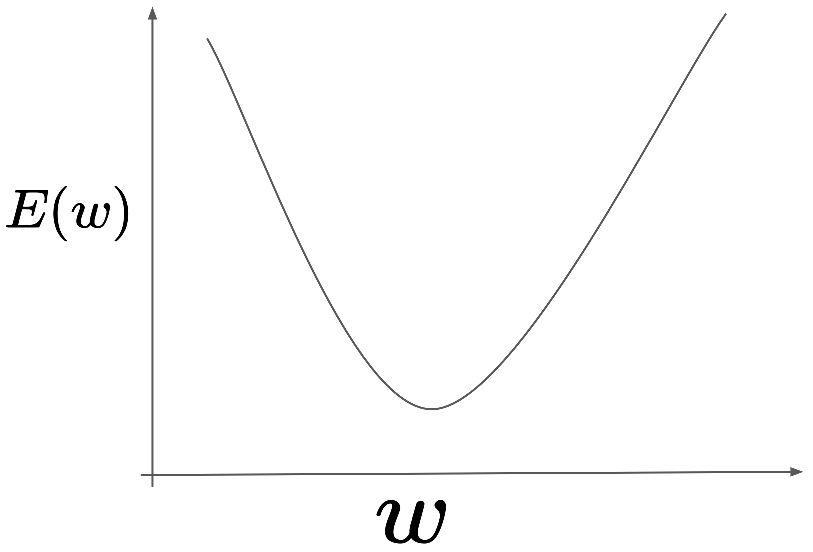


Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}$$

Gradient descent update rule

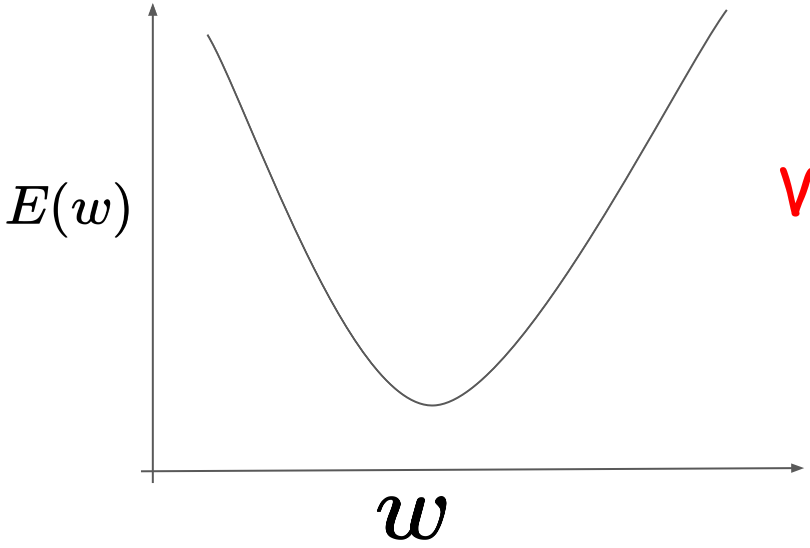


Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$w^{t+1} = w^t - \boxed{\eta}\frac{dE(w^t)}{dw}$$

Gradient descent update rule



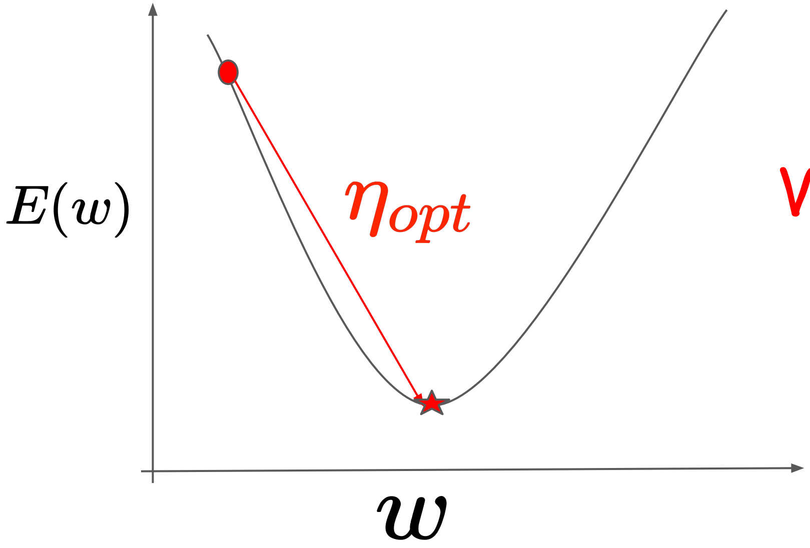
What is the best η?

Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$w^{t+1} = w^t - \boxed{\eta}\frac{dE(w^t)}{dw}$$

Gradient descent update rule

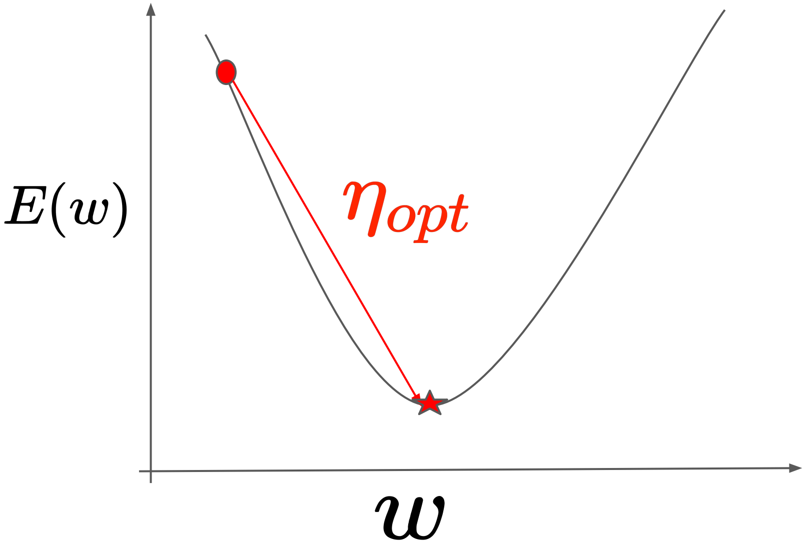


What is the best η?

Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$w^{t+1} = w^t - \eta\frac{dE(w^t)}{dw}$$



Let us find minima of E(w) using Newton's method.

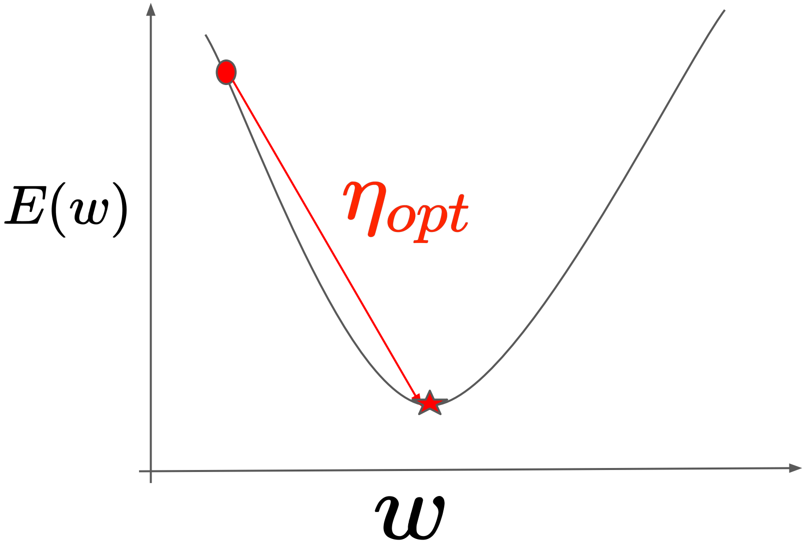
Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$w^{t+1} = w^t - \eta\frac{dE(w^t)}{dw}$$

Taylor series

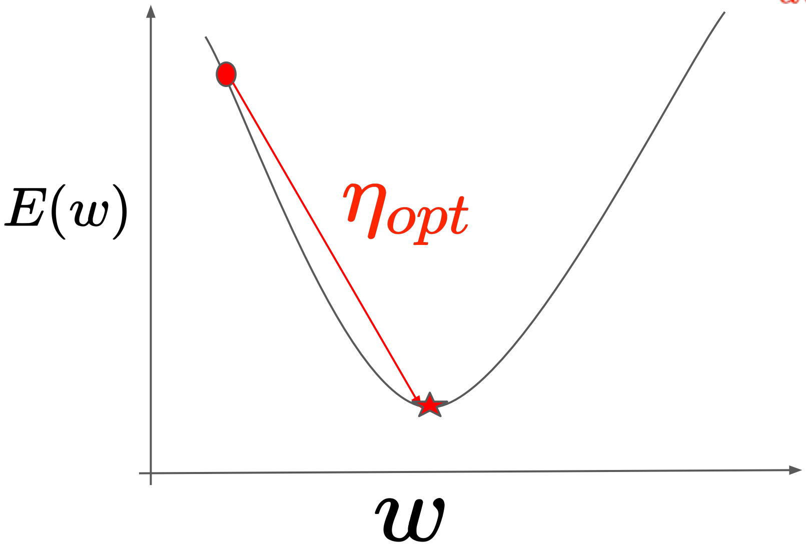
$$E(w) = E(w^t) + (w - w^t)E'(w^t) + \frac{(w-w^t)^2}{2}E''(w^t)$$



Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}$$



Taylor series

$$E(w) = E(w^t) + (w - w^t)E'(w^t) + \frac{(w-w^t)^2}{2}E''(w^t)$$

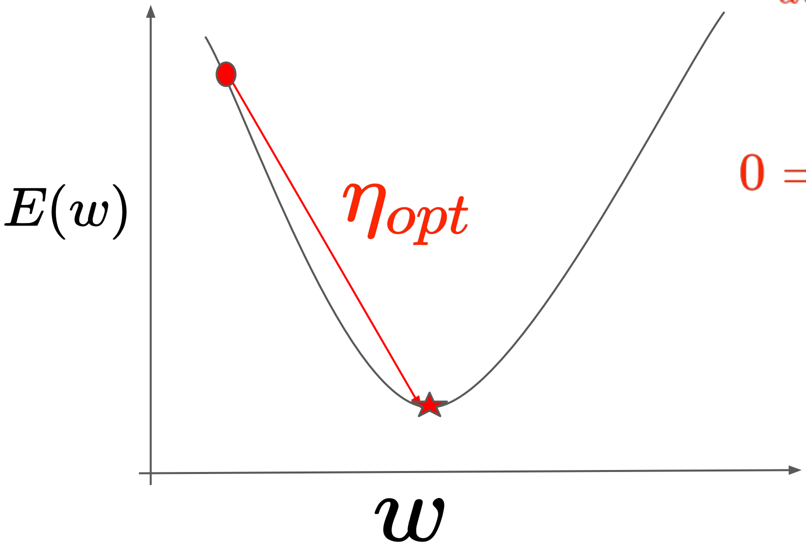
$$\frac{dE(w)}{dw} = E'(w^t) + (w - w^t)E''(w^t) + E'(w^t)$$

$$+ \frac{2(w-w^t)}{2}E''(w^t)$$

Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}$$



Taylor series

$$E(w) = E(w^t) + (w - w^t)E'(w^t) + \frac{(w-w^t)^2}{2}E''(w^t)$$

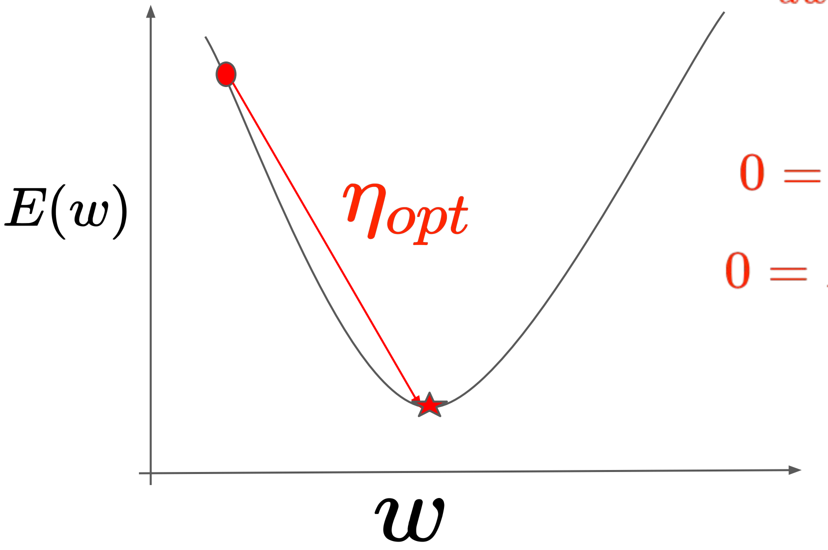
$$\frac{dE(w)}{dw} = E'(w^t) + (w - w^t)E''(w^t) + E'(w^t) + \frac{2(w-w^t)}{2}E''(w^t)$$

$$0 = 2E'(w^t) + 2(w - w^t)E''(w^t)$$

Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}$$



Taylor series

$$E(w) = E(w^t) + (w - w^t)E'(w^t) + \frac{(w-w^t)^2}{2}E''(w^t)$$

$$\frac{dE(w)}{dw} = E'(w^t) + (w - w^t)E''(w^t) + E'(w^t)$$

$$+ \frac{2(w-w^t)}{2}E''(w^t)$$

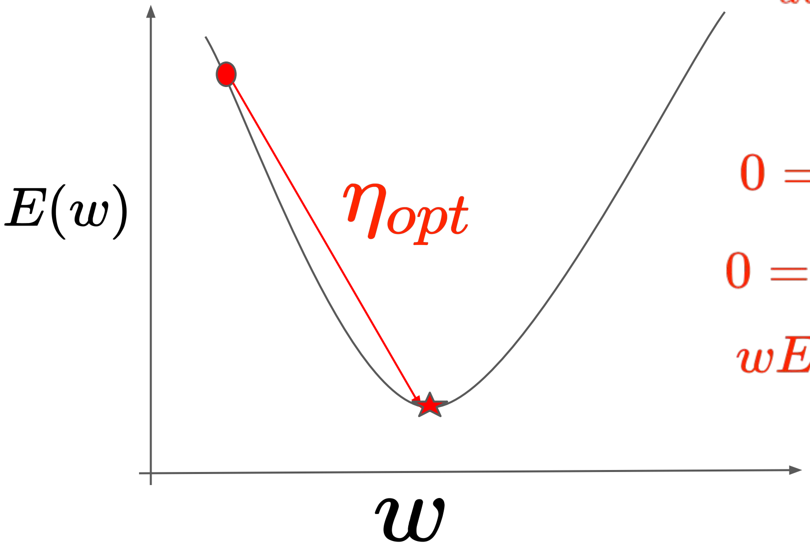
$$0 = 2E'(w^t) + 2(w - w^t)E''(w^t)$$

$$0 = E'(w^t) + wE''(w^t) - w^t E''(w^t)$$

Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$w^{t+1} = w^t - \eta\frac{dE(w^t)}{dw}$$



Taylor series

$$E(w) = E(w^t) + (w - w^t)E'(w^t) + \frac{(w-w^t)^2}{2}E''(w^t)$$

$$\frac{dE(w)}{dw} = E'(w^t) + (w - w^t)E''(w^t) + E'(w^t)$$

$$+ \frac{2(w-w^t)}{2}E''(w^t)$$

$$0 = 2E'(w^t) + 2(w - w^t)E''(w^t)$$

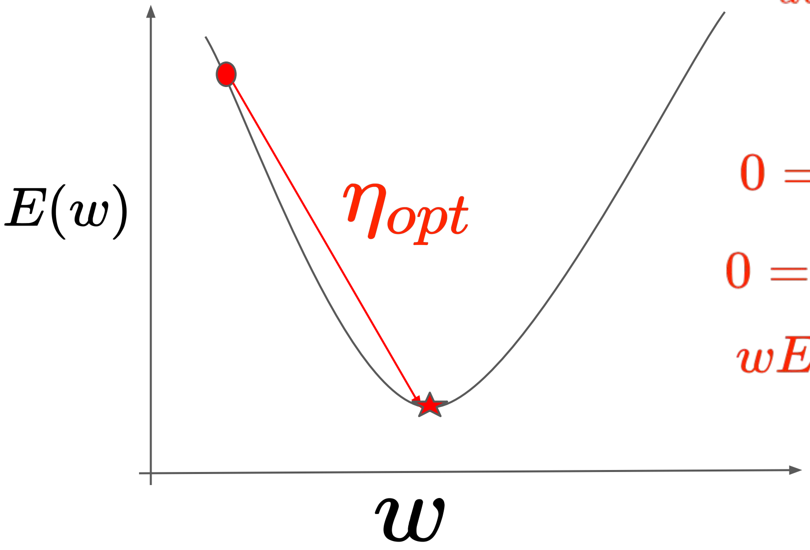
$$0 = E'(w^t) + wE''(w^t) - w^t E''(w^t)$$

$$wE''(w^t) = w^t E''(w^t) - E'(w^t)$$

Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$w^{t+1} = w^t - \eta\frac{dE(w^t)}{dw}$$



Taylor series

$$E(w) = E(w^t) + (w - w^t)E'(w^t) + \frac{(w-w^t)^2}{2}E''(w^t)$$

$$\frac{dE(w)}{dw} = E'(w^t) + (w - w^t)E''(w^t) + E'(w^t)$$

$$+\frac{2(w-w^t)}{2}E''(w^t)$$

$$0 = 2E'(w^t) + 2(w - w^t)E''(w^t)$$

$$0 = E'(w^t) + wE''(w^t) - w^t E''(w^t)$$

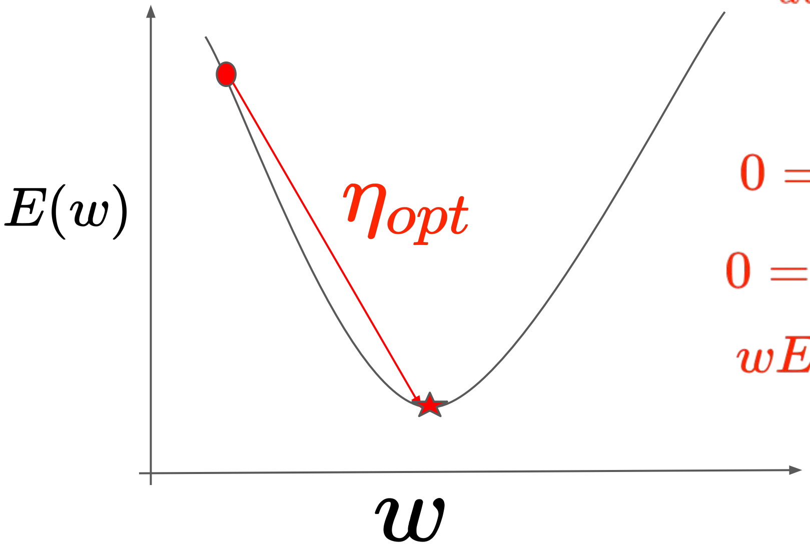
$$wE''(w^t) = w^t E''(w^t) - E'(w^t)$$

$$w = w^t - \frac{E'(w^t)}{E''(w^t)}$$

Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$\boxed{w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}}$$



Taylor series

$$E(w) = E(w^t) + (w - w^t)E'(w^t) + \frac{(w-w^t)^2}{2}E''(w^t)$$

$$\frac{dE(w)}{dw} = E'(w^t) + (w - w^t)E''(w^t) + E'(w^t)$$

$$+ \frac{2(w-w^t)}{2}E''(w^t)$$

$$0 = 2E'(w^t) + 2(w - w^t)E''(w^t)$$

$$0 = E'(w^t) + wE''(w^t) - w^t E''(w^t)$$

$$wE''(w^t) = w^t E''(w^t) - E'(w^t)$$

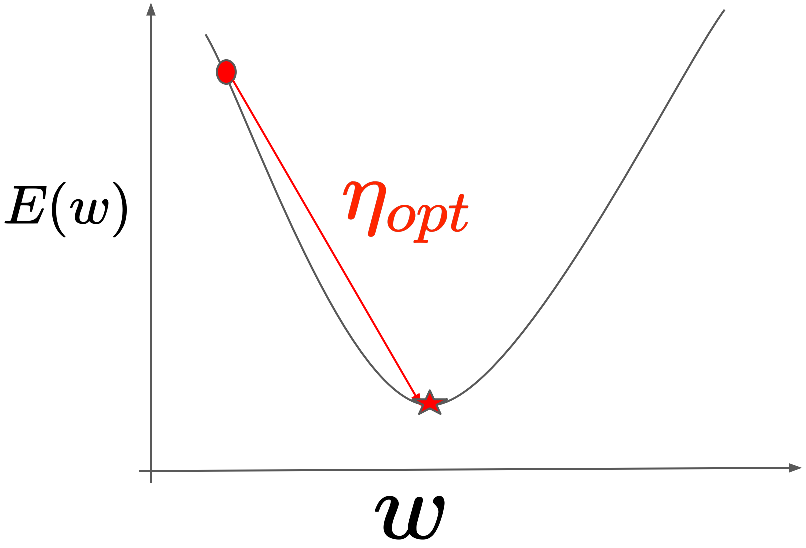
$$\boxed{w = w^t - \frac{E'(w^t)}{E''(w^t)}}$$

Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$\boxed{w^{t+1} = w^t - \eta\frac{dE(w^t)}{dw}}$$

$$\boxed{w = w^t - \frac{E'(w^t)}{E''(w^t)}}$$



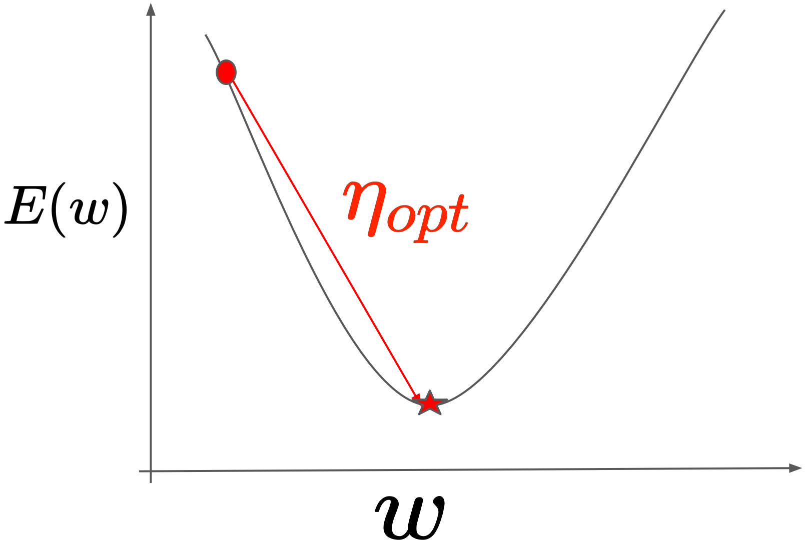
Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$\boxed{w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}}$$

$$\boxed{w = w^t - \frac{E'(w^t)}{E''(w^t)}}$$

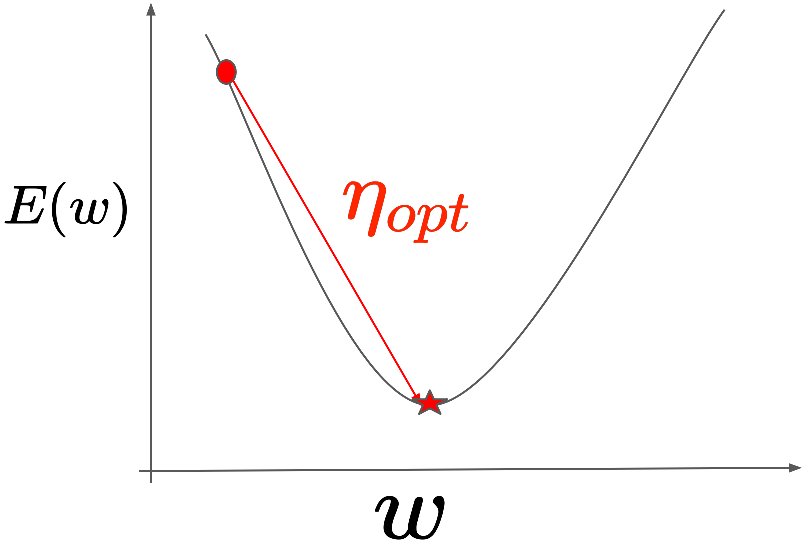
$$\eta_{opt} = \frac{1}{E''(w^t)}$$



Convergence for quadratic surface

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

$$\boxed{w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}}$$



$$\boxed{w = w^t - \frac{E'(w^t)}{E''(w^t)}}$$

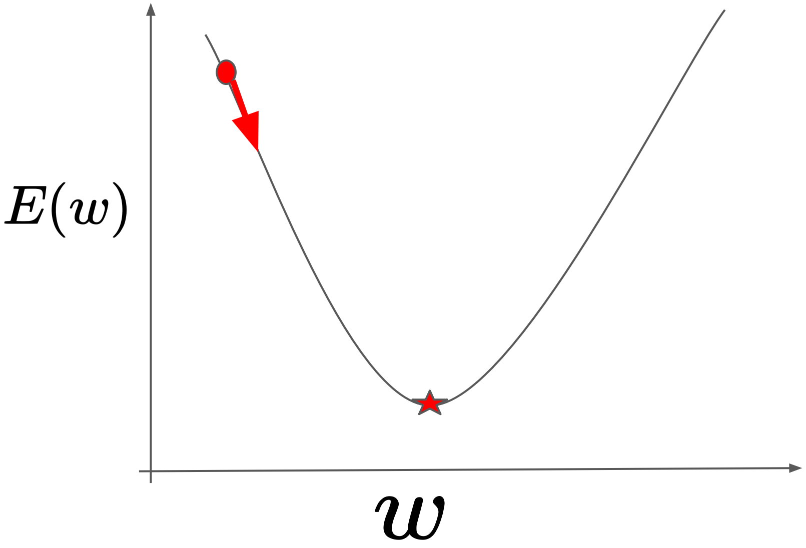
$$\eta_{opt} = \frac{1}{E''(w^t)}$$

$$\eta_{opt} = \frac{1}{a}$$

Case 1: $\eta < \eta_{opt}$

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

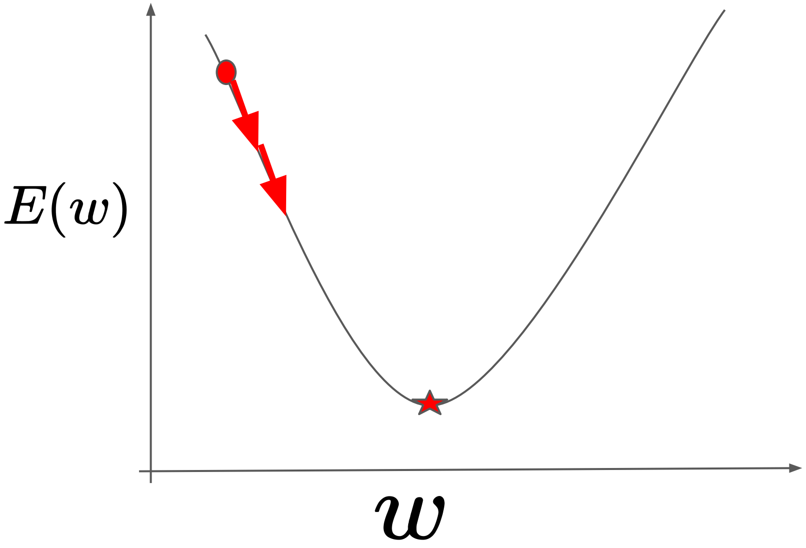
$$\boxed{w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}}$$



Case 1: $\eta < \eta_{opt}$

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

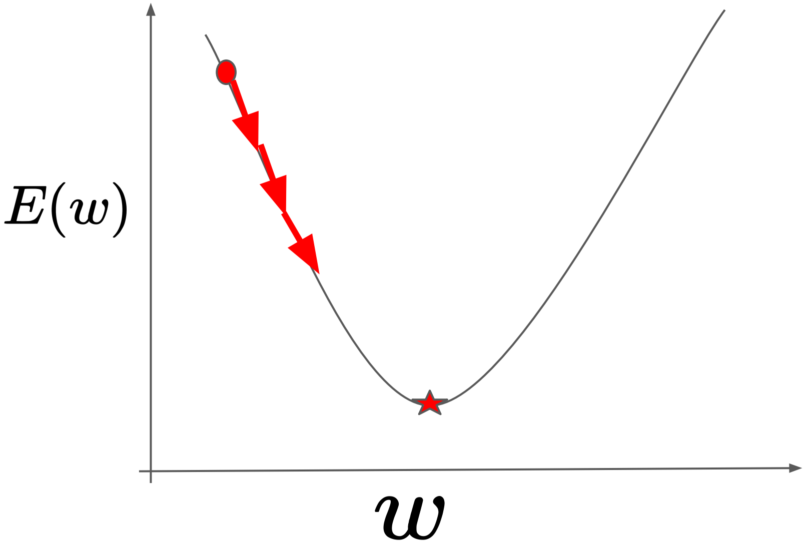
$$\boxed{w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}}$$



Case 1: $\eta < \eta_{opt}$

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

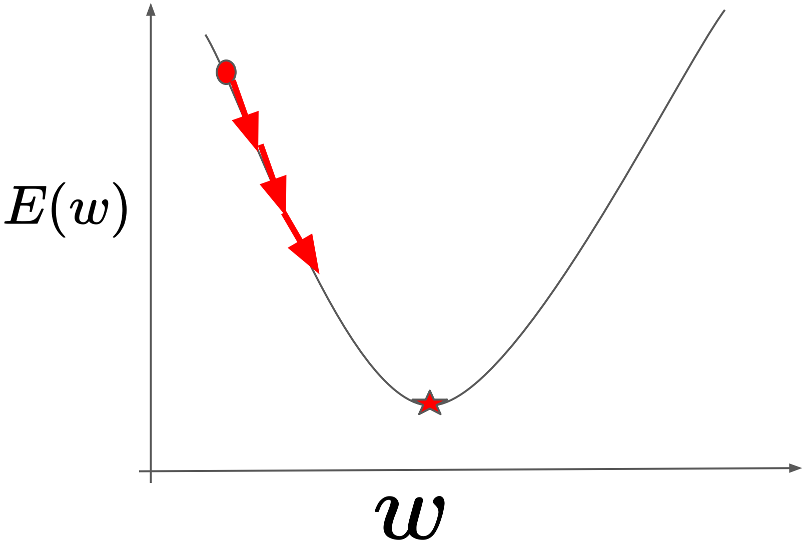
$$\boxed{w^{t+1} = w^t - \eta\frac{dE(w^t)}{dw}}$$



Case 1: $\eta < \eta_{opt}$

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

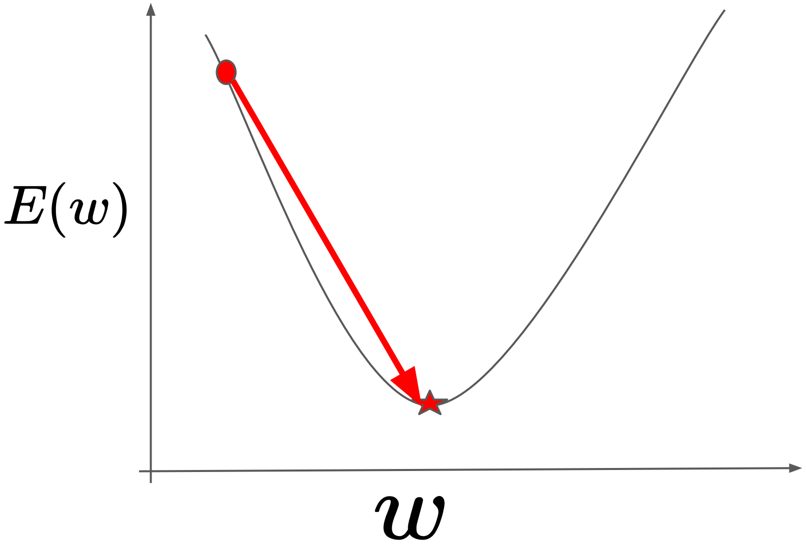
$$\boxed{w^{t+1} = w^t - \eta\frac{dE(w^t)}{dw}}$$



Case 2: $\eta = \eta_{opt}$

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

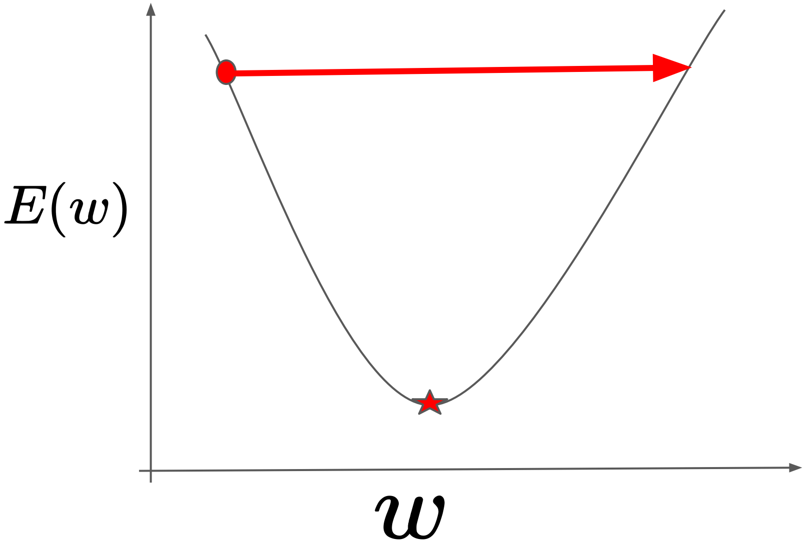
$$\boxed{w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}}$$



Case 3: $\eta = 2\eta_{opt}$

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

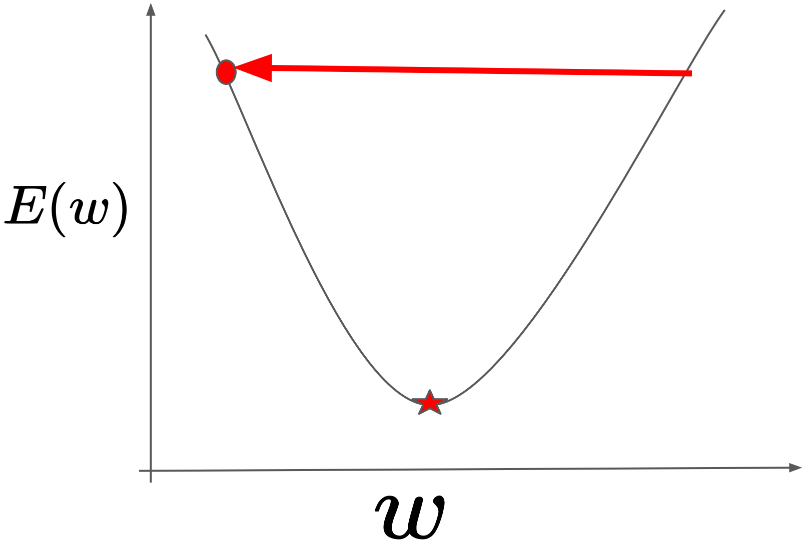
$$\boxed{w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}}$$



Case 3: $\eta = 2\eta_{opt}$

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

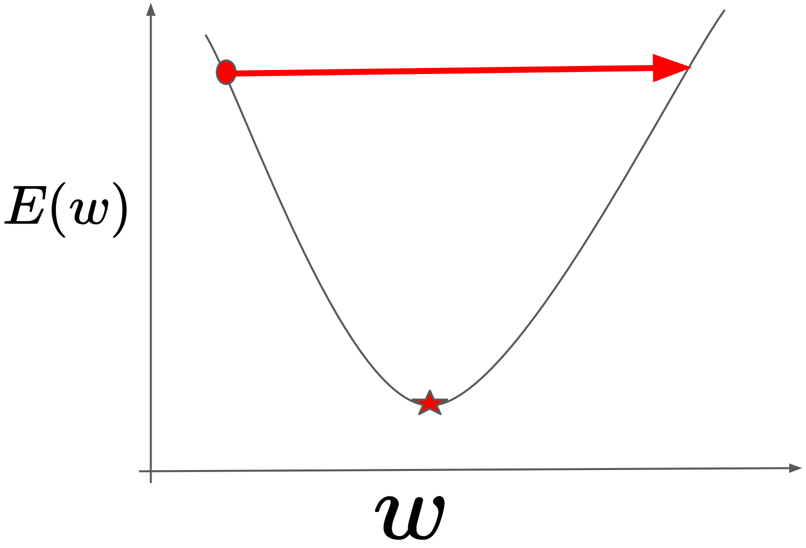
$$\boxed{w^{t+1} = w^t - \eta\frac{dE(w^t)}{dw}}$$



Case 3: $\eta = 2\eta_{opt}$

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

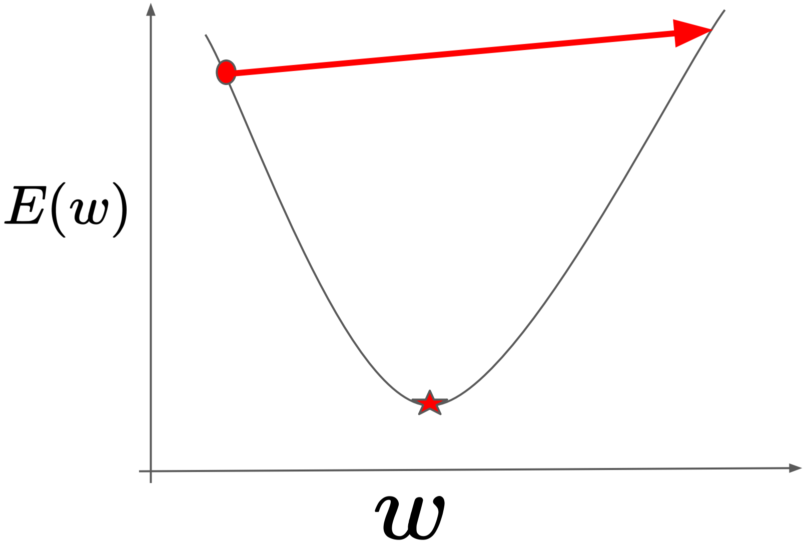
$$\boxed{w^{t+1} = w^t - \eta\frac{dE(w^t)}{dw}}$$



Case 4: $\eta > 2\eta_{opt}$

$$E(w) = \frac{1}{2}aw^2 + bw + c$$

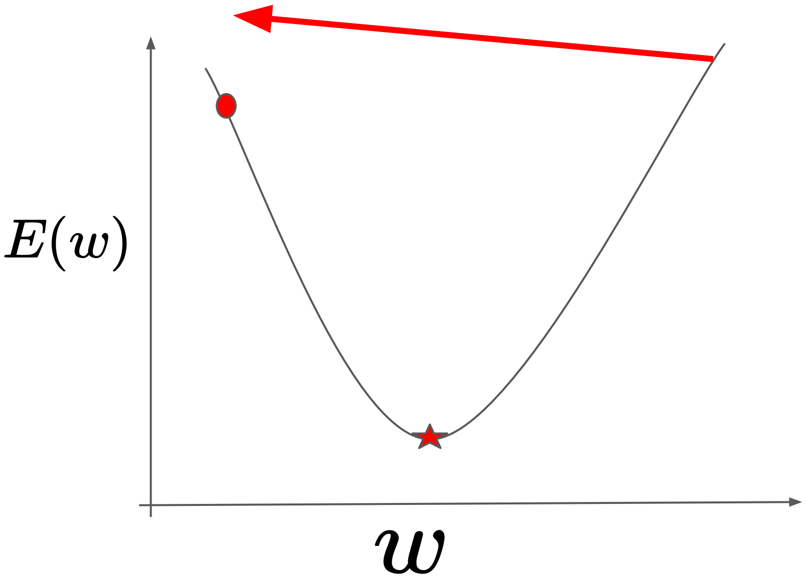
$$\boxed{w^{t+1} = w^t - \eta\frac{dE(w^t)}{dw}}$$



Case 4: $\eta > 2\eta_{opt}$

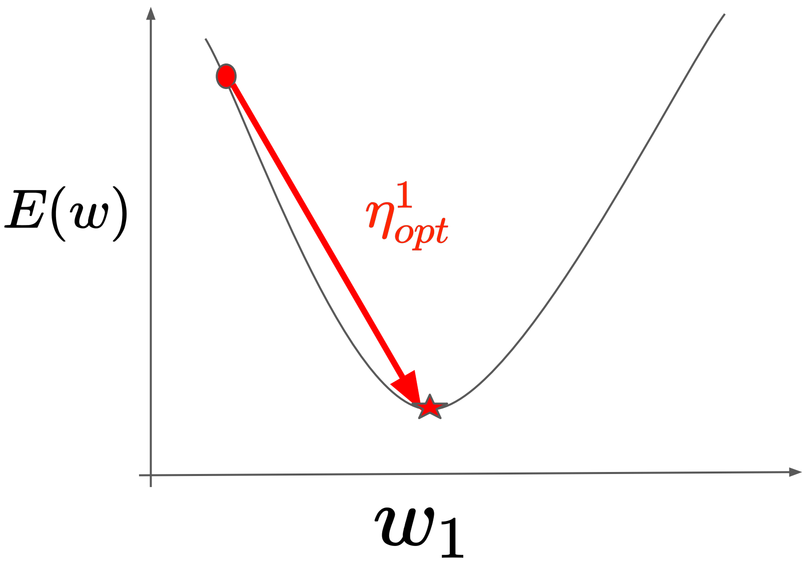
$$E(w) = \frac{1}{2}aw^2 + bw + c$$

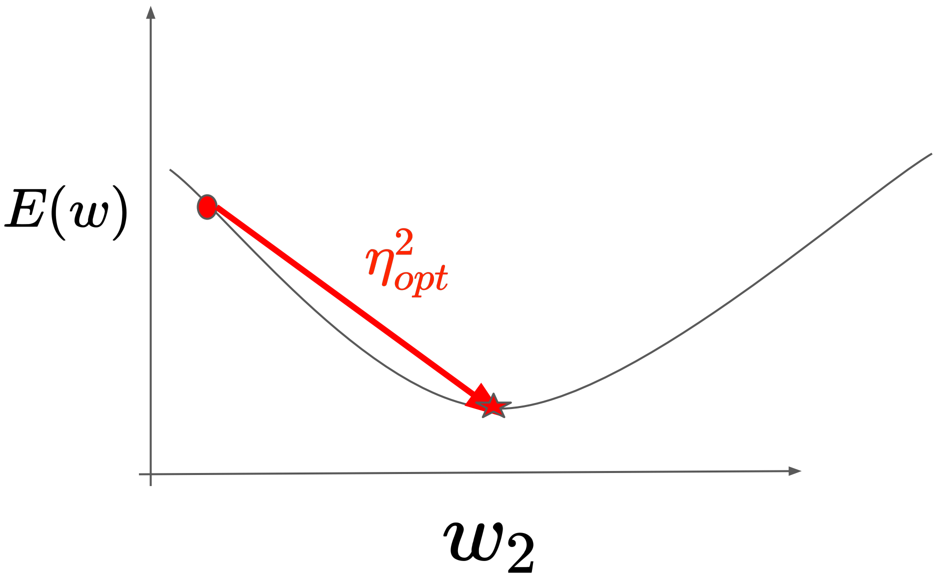
$$\boxed{w^{t+1} = w^t - \eta \frac{dE(w^t)}{dw}}$$



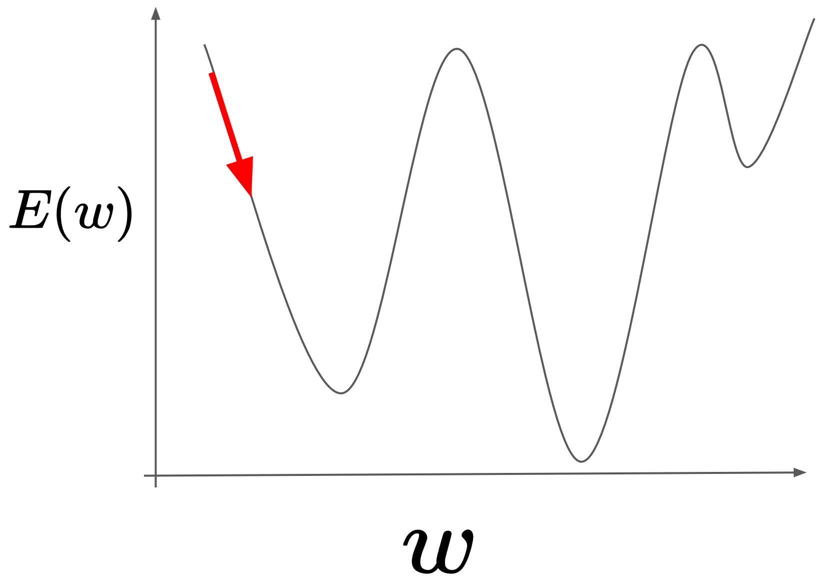
So far we have analyzed only single variable and convex functions

Problem 1: Multi-variable cost function

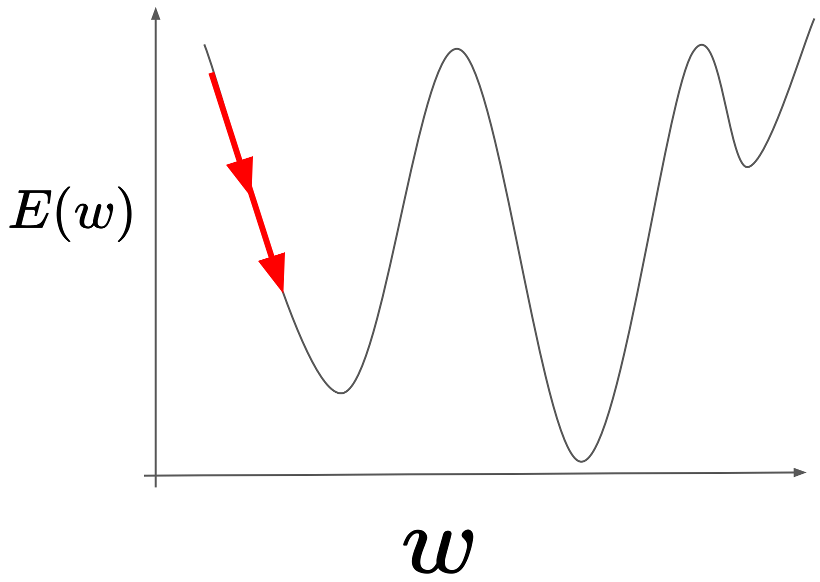




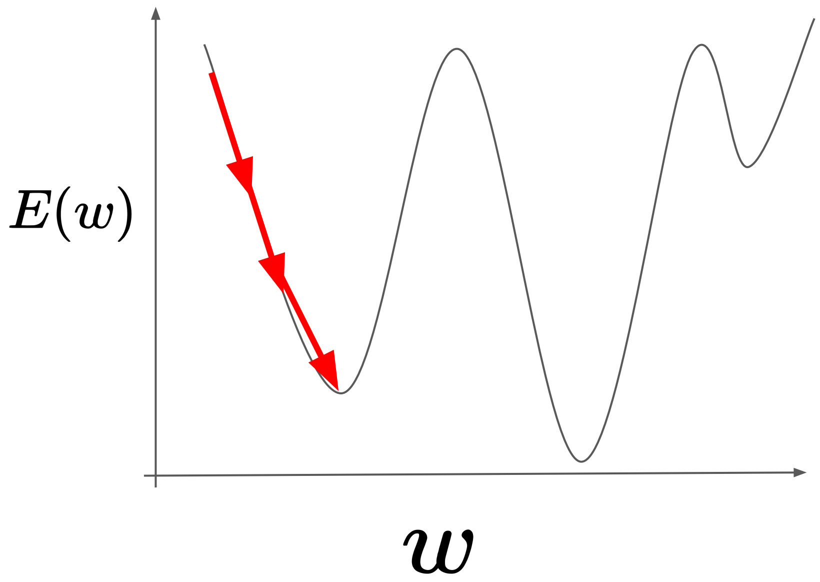
Problem 2: Non-convex cost function



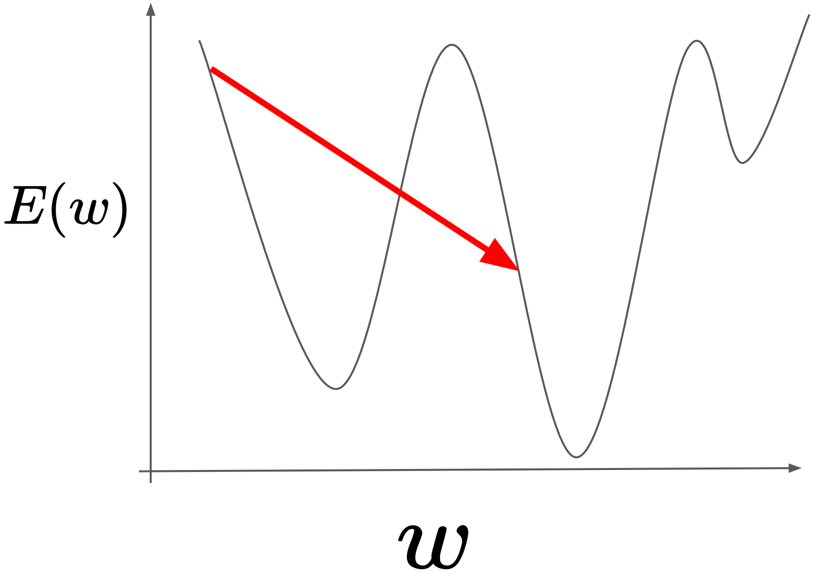
Problem 2: Non-convex cost function



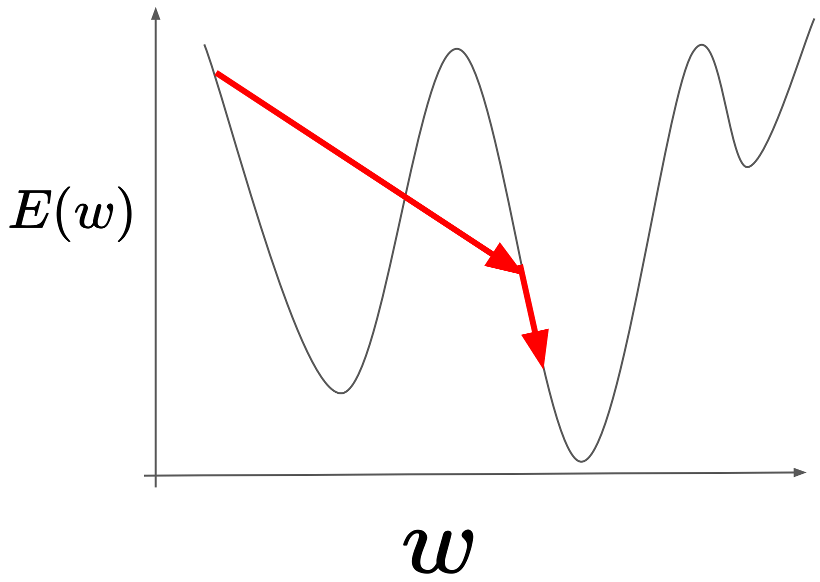
Problem 2: Non-convex cost function



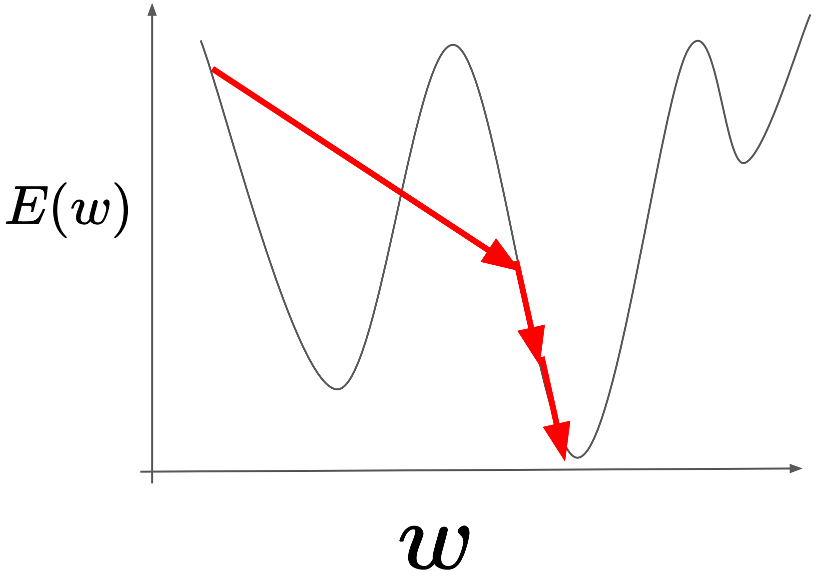
Problem 2: Non-convex cost function



Problem 2: Non-convex cost function



Problem 2: Non-convex cost function



Decaying Learning rate

- Linear decay

$$\eta_t = \frac{\eta_0}{t+1}$$

- Quadratic decay

$$\eta_t = \frac{\eta_0}{(t+1)^2}$$

- Exponential decay

$$\eta_t = \eta_0 e^{-\beta t}, \beta > 0$$