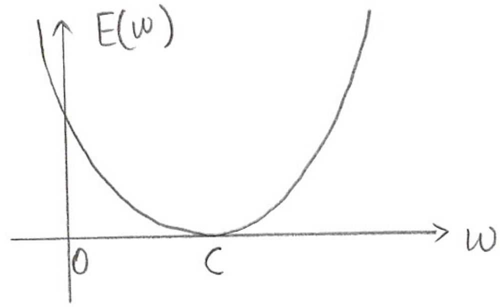
Gradient Descent

- Minimize an objective function.

$$E(w) = \frac{1}{2}(w - c)^2$$

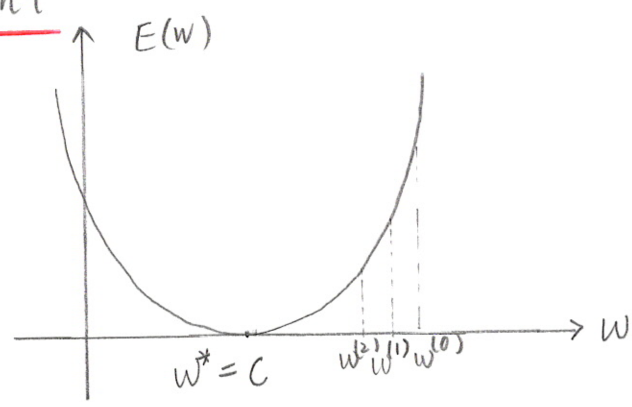


$$\frac{d\,E(w)}{d\,w} = \frac{1}{2} \cdot 2(w - c) = 0 \implies w^* = c.$$

A closed-form solution

For functions E(w) that are difficult to calculate the closed-form solution, or when the closed-form solution does not exist, we use another method to find w^*.

Gradient Descent



$w^{(0)}$: an initial value for w

Update the new $w^{(1)}$ by

$$w^{(1)} = w^{(0)} + \eta \cdot p$$

p: a direction ; $\eta > 0$: a step size , or learning rate

It's a constant.

Assume we know $E(w^{(0)})$, we want to find $E(w^{(1)})$

"Taylor Expansion" says:

$$E(w^{(1)}) \approx E(w^{(0)}) + \underbrace{\left.\frac{dE(w)}{dw}\right|_{w=w^{(0)}}}_{E'(w^{(0)})} \times (w^{(1)} - w^{(0)})$$

e.g. $E(w) = \frac{1}{2}(w-c)^2$

$$\frac{dE(w)}{dw} = \frac{1}{2} \cdot 2 \cdot (w-c) = w-c$$

$$\underbrace{\left.\frac{dE(w)}{dw}\right|_{w=w^{(0)}}}_{E'(w^{(0)})} = w^{(0)} - c$$

$$E(w^{(1)}) \approx E(w^{(0)}) + (w^{(0)}-c) \times (w^{(1)} - w^{(0)}) \leq E(w^{(0)})$$

$$(w^{(0)}-c) \times (w^{(1)} - w^{(0)}) \leq 0 \quad \cdots \quad (*)$$

since $w^{(1)} = w^{(0)} + \eta \cdot p \quad \cdots \quad (1)$

plug (1) into $(*)$

$$(w^{(0)}-c) \times (w^{(0)} + \eta \cdot p - w^{(0)}) = (w^{(0)}-c) \times \eta \cdot p \leq 0$$

$\because \eta > 0$

$\therefore \; p \cdot (w^{(0)} - c) \leq 0 \quad \cdots \quad (2)$

Let $p = -(w^{(0)}-c)$

then: LHS of $(2) = -(w^{(0)}-c)^2 \leq 0$

conclusion:

$$p = - \left. \frac{dE(w)}{dw} \right|_{w=w^{(0)}} = -E'(w^{(0)})$$

$$w^{(1)} = w^{(0)} - \eta \cdot \left. \frac{dE(w)}{dw} \right|_{w=w^{(0)}}$$

In the τ-th iteration, update $w^{(\tau)}$ by

$$w^{(\tau)} = w^{(\tau-1)} - \eta \cdot \left. \frac{d\,E(w)}{dw} \right|_{w = w^{(\tau-1)}}$$

$\tau = 1, 2, 3, \cdots$

$\boxed{\text{For example:}}$ $E(w) = \frac{1}{2}(w-4)^2$

Initialize $w^{(0)} = 3$, let $\eta = 1e-2$, $\varepsilon = 1e-8$

For $\tau = 1, 2, 3, \cdots$

$$w^{(\tau)} = w^{(\tau-1)} - \eta \cdot \left[w^{(\tau-1)} - 4 \right]$$

update $E(w^{(\tau)}) = \frac{1}{2}(w^{(\tau)} - 4)^2$

check stopping criterion:

if $\left| E(w^{(\tau)}) - E(w^{(\tau-1)}) \right| < \varepsilon$

break

Return w^*

Code: gradient_descent_parabola.py

Summary : Scalar-version Gradient-Descent Algorithm

Objective : $w^* = \arg\min_{w} E(w)$

Initialize: $w^{(0)}$, $E(w^{(0)})$, $\eta > 0$, $\varepsilon > 0$, max Iter.

For $\tau = 1, 2, 3, \cdots,$ max Iter.

① Update gradient (or derivative)

$$\nabla_w E(w)\Big|_{w=w^{(\tau-1)}} = \cdots$$

② update weight

$$w^{(\tau)} = w^{(\tau-1)} - \eta \cdot \nabla_w E(w)\Big|_{w=w^{(\tau-1)}}$$

③ update objective function

$$E(w^{(\tau)}) = \cdots$$

④ Check stopping criterion.

$$\text{If} \quad \left| E(w^{(\tau)}) - E(w^{(\tau-1)}) \right| < \varepsilon.$$

break.

Return $w^* = w^{(\tau_{stop})}$

When $E(\vec{w})$ is a function of a vector $\vec{w} = [w_0 \ w_1 \ \cdots \ w_m]^T$

update $\vec{w}$ by

$$\vec{w}^{(\tau)} = \vec{w}^{(\tau-1)} + \eta \cdot \vec{p}$$

$$\vec{p} = -\nabla_{\vec{w}} E(\vec{w}) \Big|_{\vec{w} = \vec{w}^{(\tau-1)}}$$

The gradient of the function $E(\cdot)$ with respect to $\vec{w}$ is:

$$\nabla_{\vec{w}} E(\vec{w}) \triangleq \begin{bmatrix} \dfrac{\partial E(\vec{w})}{\partial w_0} \\ \dfrac{\partial E(\vec{w})}{\partial w_1} \\ \vdots \\ \dfrac{\partial E(\vec{w})}{\partial w_M} \end{bmatrix}$$

, where $\dfrac{\partial E(\vec{w})}{\partial w_m}$ is called the partial derivative

$m = 0, 1, 2, \cdots, M$

$$\nabla_{\vec{w}} E(\vec{w}) \Big|_{\vec{w} = \vec{w}^{(\tau-1)}} \triangleq \begin{bmatrix} \dfrac{\partial E(\vec{w})}{\partial w_0} \Big|_{\vec{w} = \vec{w}^{(\tau-1)}} \\ \dfrac{\partial E(\vec{w})}{\partial w_1} \Big|_{\vec{w} = \vec{w}^{(\tau-1)}} \\ \vdots \\ \dfrac{\partial E(\vec{w})}{\partial w_M} \Big|_{\vec{w} = \vec{w}^{(\tau-1)}} \end{bmatrix}$$

Example: $\vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

$$E(\vec{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(w_0 + w_1 x_n - t_n\right)^2$$

$$\frac{\partial E(\vec{w})}{\partial w_0} = \frac{1}{2}\sum_{n=1}^{N} 2\left(w_0 + w_1 x_n - t_n\right) = \sum_{n=1}^{N}\left(w_0 + w_1 x_n - t_n\right)$$

$$\frac{\partial E(\vec{w})}{\partial w_1} = \frac{1}{2}\sum_{n=1}^{N} 2\left(w_0 + w_1 x_n - t_n\right)\cdot x_n = \sum_{n=1}^{N}\left(w_0 + w_1 x_n - t_n\right)\cdot x_n$$

$$\nabla_{\vec{w}} E(\vec{w}) = \begin{bmatrix} \dfrac{\partial E(\vec{w})}{\partial w_0} \\[2ex] \dfrac{\partial E(\vec{w})}{\partial w_1} \end{bmatrix}$$

$$\nabla_{\vec{w}} E(\vec{w}) \Big|_{\vec{w} = \vec{w}^{(\tau-1)}} = \begin{bmatrix} \sum_{n=1}^{N}\left(w_0^{(\tau-1)} + w_1^{(\tau-1)} \cdot x_n - t_n\right) \\[2ex] \sum_{n=1}^{N}\left(w_0^{(\tau-1)} + w_1^{(\tau-1)} \cdot x_n - t_n\right)\cdot x_n \end{bmatrix}$$

$$= \begin{bmatrix} N \cdot w_0^{(\tau-1)} + w_1^{(\tau-1)} \cdot \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} t_n \\[2ex] w_0^{(\tau-1)} \cdot \sum_{n=1}^{N} x_n + w_1^{(\tau-1)} \cdot \sum_{n=1}^{N} x_n^2 - \sum_{n=1}^{N} t_n \cdot x_n \end{bmatrix}$$

Summary : vector-version Gradient Descent Algorithm

Objective : $\vec{w}^* = \arg\min_{\vec{w}} E(\vec{w})$

Initialize : $\vec{w}^{(0)}$, $E(\vec{w}^{(0)})$, $\eta > 0$, $\varepsilon > 0$, maxIter

For $\tau = 1, 2, 3, \cdots,$ maxIter

① Update gradient

$$\nabla_{\vec{w}} E(\vec{w}) \Big|_{\vec{w} = \vec{w}^{(\tau-1)}} = \cdots$$

② update weights

$$\vec{w}^{(\tau)} = \vec{w}^{(\tau-1)} - \eta \cdot \nabla_{\vec{w}} E(\vec{w}) \Big|_{\vec{w} = \vec{w}^{(\tau-1)}}$$

③ update objective function

$$E(\vec{w}^{(\tau)}) = \cdots$$

④ Check stopping criterion:

If $\left| E(\vec{w}^{(\tau)}) - E(\vec{w}^{(\tau-1)}) \right| < \varepsilon$

break

Return $\vec{w}^* = \vec{w}^{(\tau_{stop})}$

Vector-version gradient descent for Linear Regression.

$$E(\vec{w}) = \frac{1}{2} \sum_{n=1}^{N} (\vec{x}_n^T \cdot \vec{w} - t_n)^2$$

$$\vec{w}^{(\tau)} = \vec{w}^{(\tau-1)} - \eta \cdot \nabla_{\vec{w}} E(\vec{w}) \Big|_{\vec{w} = \vec{w}^{(\tau-1)}}$$

Recall: $\dfrac{d\, \vec{a}^T \cdot \vec{w}}{d \vec{w}} = \vec{a}$

Hence, $\nabla_{\vec{w}} E(\vec{w}) = \frac{1}{2} \cdot \sum_{n=1}^{N} 2 \cdot (\vec{x}_n^T \cdot \vec{w} - t_n) \cdot \dfrac{d\, \vec{x}_n^T \cdot \vec{w}}{d \vec{w}}$

$$= \sum_{n=1}^{N} (\vec{x}_n^T \cdot \vec{w} - t_n) \cdot \vec{x}_n = \sum_{n=1}^{N} \vec{x}_n \cdot (\vec{x}_n^T \cdot \vec{w} - t_n)$$

$$\nabla_{\vec{w}} E(\vec{w}) = \left(\sum_{n=1}^{N} \vec{x}_n \cdot \vec{x}_n^T \right) \vec{w} - \sum_{n=1}^{N} \vec{x}_n \cdot t_n$$

Recall: $X^T = \begin{bmatrix} \vec{x}_1 & \vec{x}_2 & \cdots & \vec{x}_N \end{bmatrix}$, $X = \begin{bmatrix} \vec{x}_1^T \\ \vec{x}_2^T \\ \vdots \\ \vec{x}_N^T \end{bmatrix}$, $\vec{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$

Hence, $\nabla_{\vec{w}} E(\vec{w}) = X^T \cdot X \cdot \vec{w} - X^T \cdot \vec{t}$

$$\nabla_{\vec{w}} E(\vec{w}) \Big|_{\vec{w} = \vec{w}^{(\tau-1)}} = X^T \cdot X \cdot \vec{w}^{(\tau-1)} - X^T \cdot \vec{t}$$

$$\Longrightarrow \boxed{\vec{w}^{(\tau)} = \vec{w}^{(\tau-1)} - \eta \cdot \left[X^T \cdot X \cdot \vec{w}^{(\tau-1)} - X^T \cdot \vec{t} \right] \quad \tau = 1, 2, 3, \cdots}$$