Overview of convex function and optimization

Note Title 09-06-2010

Loss function $\sum_{i=1}^{N} \text{loss}(y_i, w^T x^i + w_0) \equiv \sum_{i=1}^{N} \text{loss}\left(y_i (w^T x^i + w_0)\right)$

$\hat{y} = \text{sign}(w^T x^i + w_0)$

$y_i \in \{+1, -1\}$

Training $\longrightarrow$

$F(W, w_0) = \sum_{i=1}^{N} \text{loss}\left(y_i (w^T x^i + w_0)\right) + R(w)$

$W \in \mathbb{R}^d \; ; \; w_0 = \mathbb{R}$

continuous & differ

Training goal $\quad w_1^{opt}, w_0^{opt} = \underset{w_1, w_0}{\operatorname{argmin}} F(w_1, w_0)$

$$\min_{w_1, \dots w_d} F(w_1, w_2, \dots w_d)$$

w is unconstraind

$w_j \in \mathbb{R}$

$$\min_{w} F(w)$$

$w \in \mathbb{R}^d$

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

$F(w)$ convex in w

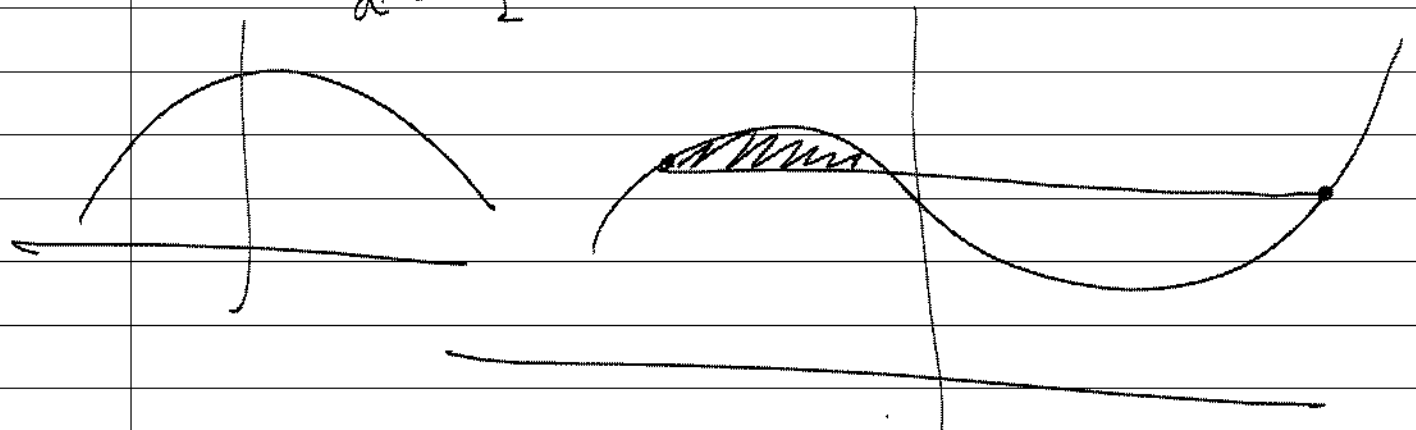
<u>Define</u>: A function $F(w)$ is a convex function of w iff

for any $w^1, w^2 \in \mathbb{R}^d$ and any

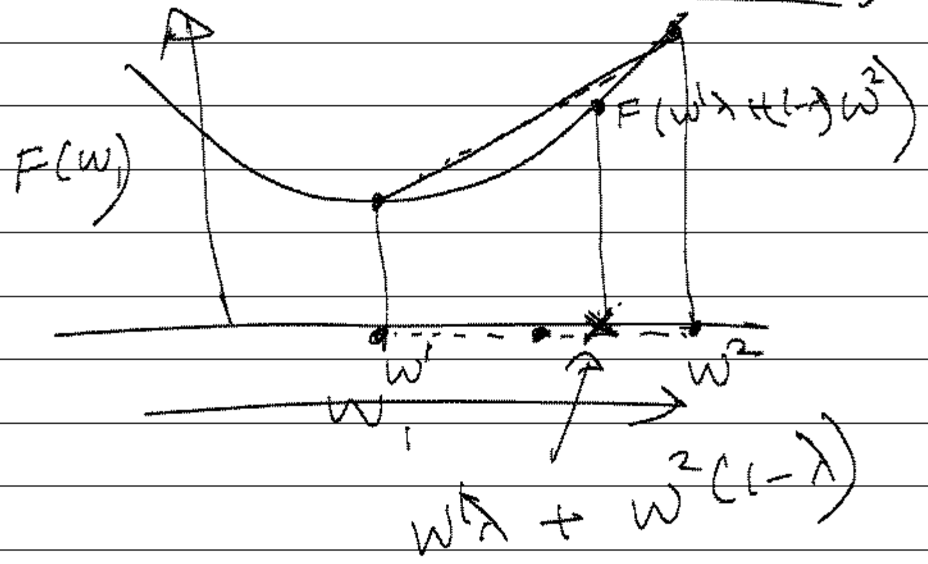
$0 \leq \lambda \leq 1 \quad \lambda \in \mathbb{R}$

$$F(w^1\lambda + w^2(1-\lambda)) \le \underline{\lambda F(w^1) + (1-\lambda) F(w^2)}$$

$\alpha = 1$



neither convex nor concave



Convex

Define: If $F(w)$ is convex, $-F(w)$ is concave

In general, pick any k points $w^1, w^2, \ldots, w^k$, pick $\underline{p_1, p_2, \ldots, p_k}$

s.t.

$\sum p_i = 1 \quad 0 \leq p_i \leq 1 \; ; \quad F\left(\sum_k w^k p_k\right) \leq \sum_k p_k F(w^k)$

Differentiable functions $\quad w^0 \in \mathbb{R}^d, \quad w \in \mathbb{R}^d$

$$F(w) = F(w^0) + \nabla F(w^0)^T (w - w^0) + \|w - w^0\| \alpha(w^0, w - w^0)$$

where $\alpha(w^0, w - w^0) \to 0$ as $w \to w^0$

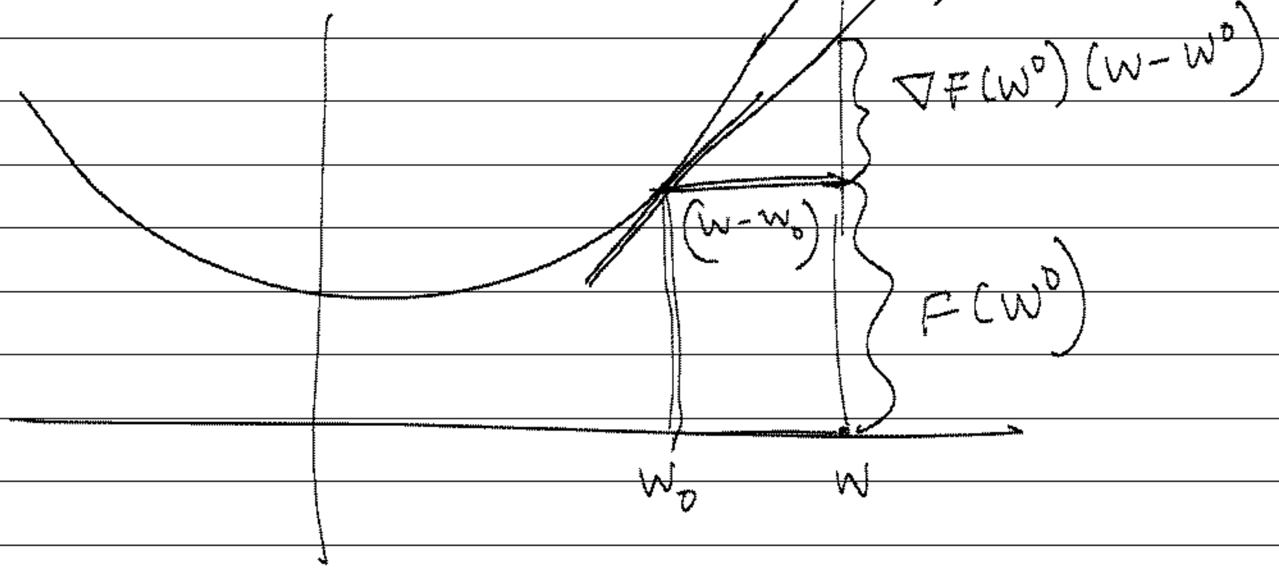
First order Taylor expansion of F

$$\nabla F(w^0) = \left.\begin{bmatrix} \frac{\partial F(w)}{\partial w_1} \\ \vdots \\ \frac{\partial F(w)}{\partial w_d} \end{bmatrix}\right|_{w = w_0}$$

gradient vector

A differentiable function is convex iff

$$F(W) \geq F(W^0) + \nabla F(W^0)(W - W^0) \quad \forall\, W^0, W$$



Twice differentiable function

Define Hessian matrix of a twice differentiable function $F(w)$

$$H_F = \begin{array}{cc} & \\ i & \begin{bmatrix} \frac{\partial^2 F}{\partial w_i \partial w_j} \end{bmatrix}_{d \times d} \\ & j \end{array}$$

Example: $d = 2$

$$F(w_1, w_2) = \underline{2w_1 + 6w_2 - 2w_1^2 - 3w_2^2 + 4w_1 w_2}$$

$$w^0 = [0, 0]$$

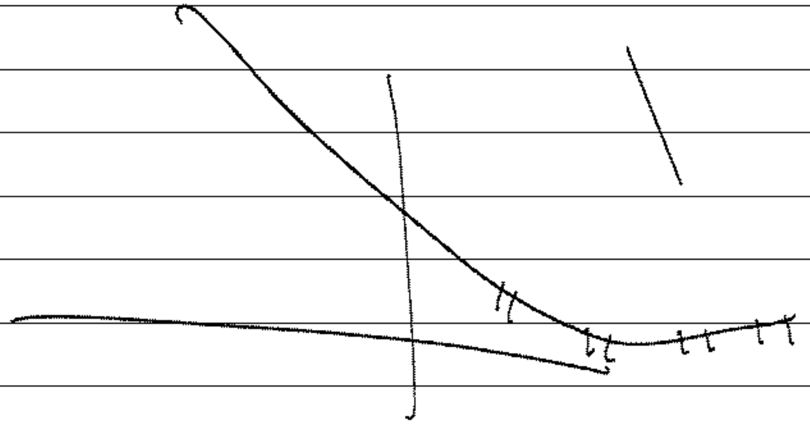
$$\nabla F = \begin{bmatrix} \frac{\partial F}{\partial w_1} \\ \frac{\partial F}{\partial w_2} \end{bmatrix} = \begin{bmatrix} 2 - 4w_1 + 4w_2 \\ 6 - 6w_2 + 4w_1 \end{bmatrix} \qquad \nabla F(w^0) = \begin{bmatrix} 2 \\ 6 \end{bmatrix}$$

$$H_F(w) = \begin{bmatrix} \frac{\partial^2 F}{\partial w_1^2} & \frac{\partial^2 F}{\partial w_1 \partial w_2} \\ \frac{\partial^2 F}{\partial w_1 \partial w_2} & \frac{\partial^2 F}{\partial w_2^2} \end{bmatrix} = \begin{bmatrix} -4 & 4 \\ 4 & -6 \end{bmatrix}$$

Second order Taylor's expansion of a twice differentiable function $F(w)$

$$F(w) = F(w^0) + \nabla F(w^0)^T (w - w^0) + \frac{1}{2}(w-w^0)^T H_p(w_0)(w-w^0) + \|w-w^0\|^2 \alpha(w^0, w-w^0)$$

$$\lim_{w \to w_0} \alpha(w^0, w-w^0) \to 0$$



$$F(w_1, w_2) = 0 + \begin{bmatrix} 2 \\ 6 \end{bmatrix}^T [w_1, w_2]^T + \frac{1}{2}[w_1 \;\; w_2] \begin{bmatrix} -4 & 4 \\ 4 & -6 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= 2w_1 + 6w_2 + \frac{1}{2}\begin{bmatrix} -4w_1+4w_2 & 4w_1-6w_2 \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= 2w_1 + 6w_2 + \frac{1}{2}\left(-4w_1^2 + 4w_1w_2 + 4w_1w_2 - 6w_2^2\right)$$

$$= 2w_1 + 6w_2 - 2w_1^2 + 4w_1w_2 - 3w_2^2$$

Example 2

$$F(w_1, w_2) = e^{2w_1+3w_2} \qquad w^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad F(w^0) = 1$$

$$\nabla F = \begin{bmatrix} e^{2w_1+3w_2}(2) \\ e^{2w_1+3w_2}(+3) \end{bmatrix} \qquad \nabla F(w^0) = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$H_p = \begin{bmatrix} 4e^{2w_1+3w_2} & 6e^{2w_1+3w_2} \\ 6e^{2w_1+3w_2} & 9e^{2w_1+3w_2} \end{bmatrix} \qquad H_p(w^0) = \begin{bmatrix} 4 & 6 \\ 6 & 9 \end{bmatrix}$$

$$F(w) = 1 + [2\ \ 3]\cdot[w_1\ \ w_2] + \frac{1}{2}[w_1\ \ w_2]\begin{bmatrix} 4 & 6 \\ 6 & 9 \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= 1 + 2w_1 + 3w_2 + \frac{1}{2}\begin{bmatrix} 4w_1 + 6w_2 & 6w_1 + 9w_2 \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$= 1 + 2w_1 + 3w_2 + \frac{1}{2}\left(4w_1^2 + 6w_1w_2 + 6w_1w_2 + 9w_2^2\right)$$

$$= 1 + 2w_1 + 3w_2 + 2w_1^2 + 6w_1w_2 + \frac{9}{2}w_2^2$$

A twice differentiable function $F(w)$ is convex if and only if the Hessian of $F(w)$ is positive semi-definite.

<u>Defn</u>

A matrix M is positive semi-definite if $w^T M w \geq 0 \quad \forall w$.

A " M is negative semi-definite if $w^T M w \leq 0 \quad \forall w$.

$$\text{loss}(y w^T x) = \text{squareloss}(y w^T x) = (1 - y w^T x)^2$$

$$\min_w F(w) = \min_w \sum_{i=1}^{N} (1 - y_i w^T x^i)^2$$

$$w^T x^i = w_1 x_1^i + w_2 x_2^i + \cdots w_d x_d^i$$

Is $F(w)$ convex in w?

$$\underline{\nabla F(w)} = \begin{bmatrix} \sum_{i=1}^{N} 2(1 - y_i w^T x^i)(-y_i x_1^i) \\ \vdots \\ \sum_{i=1}^{N} 2(1 - y_i w^T x^i)(-y_i x_d^i) \end{bmatrix}$$

$$\frac{\partial F}{\partial w_j}$$

$$= \sum_{i=1}^{N} 2(1 - y_i w^T x^i)(-y_i) X^i$$

$$H_F = \begin{array}{r c}
 & \begin{bmatrix} \sum_i 2(-y_i x_1^i)(-y_i x_1^i) & & \\ & \sum_i 2(-y_i x_k^i)(-y_i)(x_j^i) & \\ & & \end{bmatrix} \\
j^{th} & \\
 & k^{th}
\end{array}$$

$$= 2 X^T X$$

$$\underset{N \times d}{X} = \begin{bmatrix} x^{1^T} y_1 \\ \vdots \\ x^{N^T} y_N \end{bmatrix}$$

If H_F psd? $\quad w^T H w \geq 0 \quad \forall w$

$$w^T X^T X w \overset{?}{\geq} 0 \qquad H = X^T X$$

$$= \underbrace{(Xw)}_{=}^T (Xw) \overset{?}{\geq} 0$$

$$z^T z = z_1^2 + z_2^2 + \cdots z_d^2 \geq 0$$

$\Rightarrow$ H_F is psd.

$D = \{(x^i, y^i) : y^i = +1 \text{ or } -1 \}_{i=1 \cdots N}$; $C(x) = \text{sign}(W^T x)$

min logistic loss over training loss

$$\min_W \sum_{i=1}^{N} \text{logloss}(\underline{y_i W^T x^i})$$

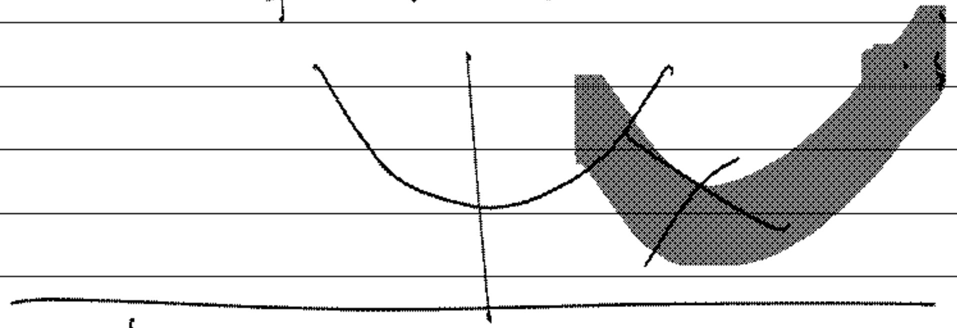
$$= \min_W \underbrace{\sum_{i=1}^{N} \log(1 + e^{-y_i W^T x^i})}_{F(W)}$$

Is this convex in W?

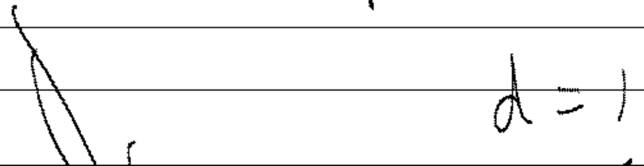
Rule 1: positive weighted sum of convex functions is convex

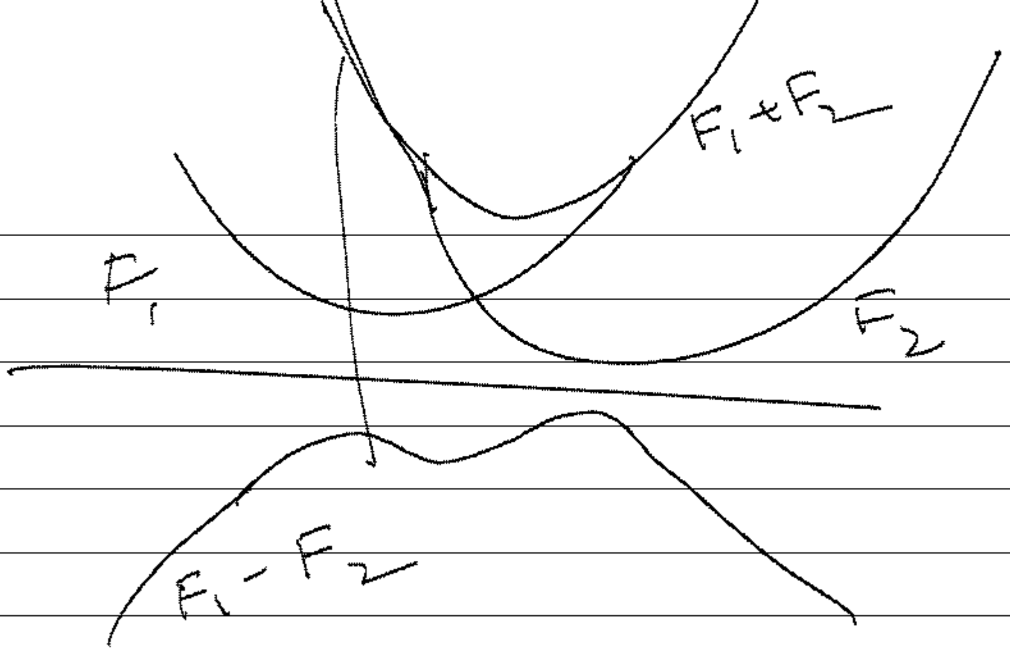
$$F_1(W) + F_2(W) + \cdots F_k(W) = \underline{\underline{F(W)}}$$

$F_1(W) - F_2(W)$ is not necessarily convex if $F_1(W)$ & $F_2(W)$ is convex.



$d = 1$




F1+F2
F1
F2
F1-F2

Rule 2: If $g(z)$ is convex where $z \in \mathbb{R}$

$\underline{F(W)} = g(W^T x)$ where $W, x \in \mathbb{R}^d$

eg: $d = 2$

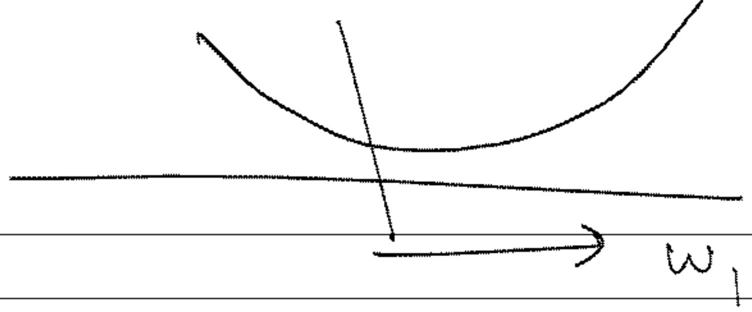
$F(w_1, w_2) = g(w_1 x_1 + w_2 x_2)$

then $F(W)$ is convex in W.

<u>Proof:</u> $g(z)$ is convex $\Longleftrightarrow$ $g''(z) \geq 0$

$$z g''(z) z \geq 0$$

$$z^2 g''(z) \geq 0$$



then $H(F(w))$

$$\nabla F(w) = \nabla_w g(w^T x)$$

$$= \underbrace{g'(w^T x)}\, x \qquad \| \; d \times 1$$

$$H(F(w)) = \underbrace{g''(w^T x)}_{\geq 0}\, x^T x$$

$$\alpha x^T x$$

$$W^T H_F W = \alpha W^T X^T X W \geq 0$$

~~Now~~ Apply proof $\sum_i \log(1 + e^{-W^T x^i y_i})$ is convex.

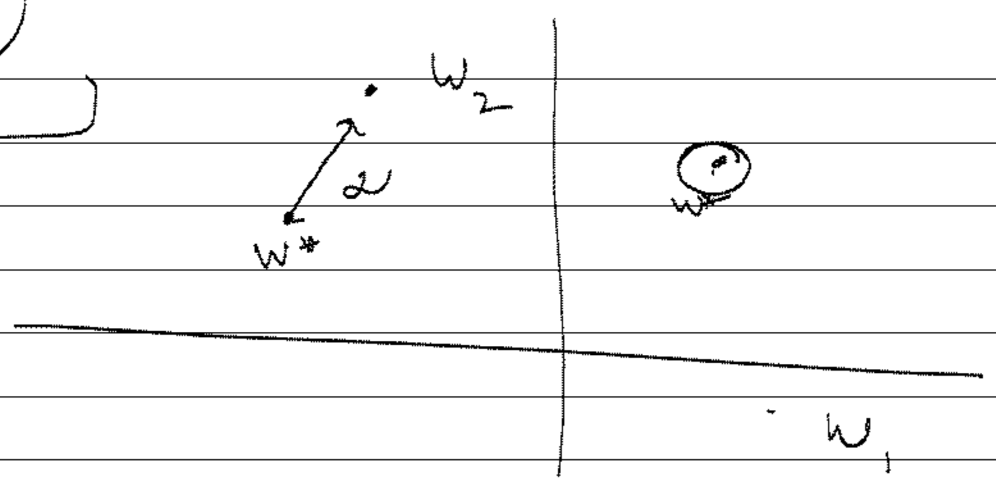
$$\equiv \sum_i F_i(W) \quad \text{where } F_i(W) = g(W^T x^i y_i) \quad \text{where}$$

$$g(z) = \log(1 + e^{-z})$$

Now just show that $g(z)$ is convex.

$$g'(z) = \frac{e^{-z}(-1)}{1+e^{-z}} = \frac{-1}{1+e^{z}}$$

$$g''(z) = \frac{-1 \; (-1) e^{z}}{(1+e^{z})^2} = \underbrace{\frac{e^{z}}{(1+e^{z})^2}} \geq 0$$



Optimization

$$\min_{w} F(w)$$

A function $F(w)$ is locally minimum at a w^* if
$F(w^*) \leq F(w)$ for all w s.t
$\|w - w^*\| \leq \varepsilon$ for some
$\varepsilon \geq 0$

Thm If W^* is a local optima of $F(W)$ and $F(W)$ is differentiable than $\nabla F(W^*) = 0$

Proof: follows from 1^{st} order Taylor expansion of $F(W)$

$$F(\bar{W}) \approx F(W^*) + \nabla F(W^*)(W - W^*) \quad \text{when } \|W - W^*\| \leq \varepsilon$$

$$\geq F(W^*) \quad \forall W$$

Pick a $W' = W^* + \alpha d$ $\quad \alpha \in \mathbb{R}$

d = a unit vector in

$$d = \frac{W' - W^*}{\|W' - W^*\|}$$

$$w^2 = w^* - \alpha d \qquad \text{If } \|w^1 - w^*\| \leq \varepsilon$$
$$\|w^2 - w^*\| \leq \varepsilon$$

$F(w^1) \geq F(w^*)$ since w^* is a local minimum

$F(w^2) \geq F(w^*)$

$$F(w^*) + \nabla F(w^*)\alpha \cdot d \geq F(w^*)$$
$$\& \quad F(w^*) + \nabla F(w^*)(-\alpha d) \geq F(w^*)$$

Only possible if $d = 0$ or $\nabla F(w^*) = 0$

$d \neq 0$ since $w^1 \neq w^*$

$$\Rightarrow \quad \nabla F(w^*) = 0$$

<u>Thm</u> For convex functions a w^* is a global minimum if and if $\nabla F(w^*) = 0$

Proof: $F(w) \geq F(w^*) + \underline{\nabla F(w^*)(w - w^*)} \quad \forall w$

$= F(w^*)$ if $\nabla F(w^*) = 0$

(only if)

(If) part proved earlier.