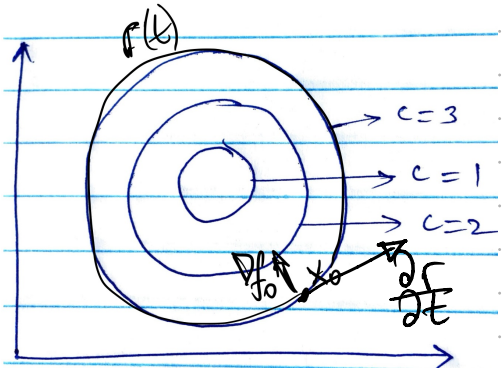
J COLLABORATORS:

Zhaokin Li

2.1

$r(t) = [x_1(t), x_2(t), \dots x_d(t)]$ lies within the level surface and passes through the gradient vector ∇f_0 i.e., $r(t) \in L_{f(x_0)}$ then:



intuitively, we know that the tangent will point out of the surface in the tangent direction of it, in the same plane as the level surface, while the gradient will point to the the direction of greatest increase of the function $\Rightarrow$ orthogonal

formally: $\frac{\partial r}{\partial t} = \left[\frac{d}{dt}x_1(t), \frac{d}{dt}x_2(t), \dots \frac{d}{dt}x_d(t)\right]$

$$\nabla f_0 = \left[\frac{\partial f_0}{\partial x_1}, \frac{\partial f_0}{\partial x_2}, \dots \frac{\partial f_0}{\partial x_d}\right]$$

at $t = t_0$: $\frac{\partial r(t_0)}{\partial t} = \left[\frac{d}{dt}x_1(t_0), \frac{d}{dt}x_2(t_0), \dots \frac{d}{dt}x_d(t_0)\right] =$

$$= \frac{\partial x_0}{\partial t} = \left[\frac{dx_{01}}{dt}, \frac{dx_{02}}{dt}, \frac{dx_{0d}}{dt}\right]$$

$$\left\langle \frac{\partial x_0}{\partial t}, \nabla f_0 \right\rangle = \left[\frac{dx_{01}}{dt}\frac{\partial f_0}{\partial x_1}, \frac{dx_{02}}{dt}\frac{\partial f_0}{\partial x_2}, \dots \frac{\partial x_{0d}}{\partial t}\frac{\partial f_0}{\partial x_d}\right]$$

$$= [0, 0, \dots, 0] \Rightarrow \frac{\partial x_0}{\partial t} \text{ \& } \nabla f_0 \text{ are ortho-gonal}$$

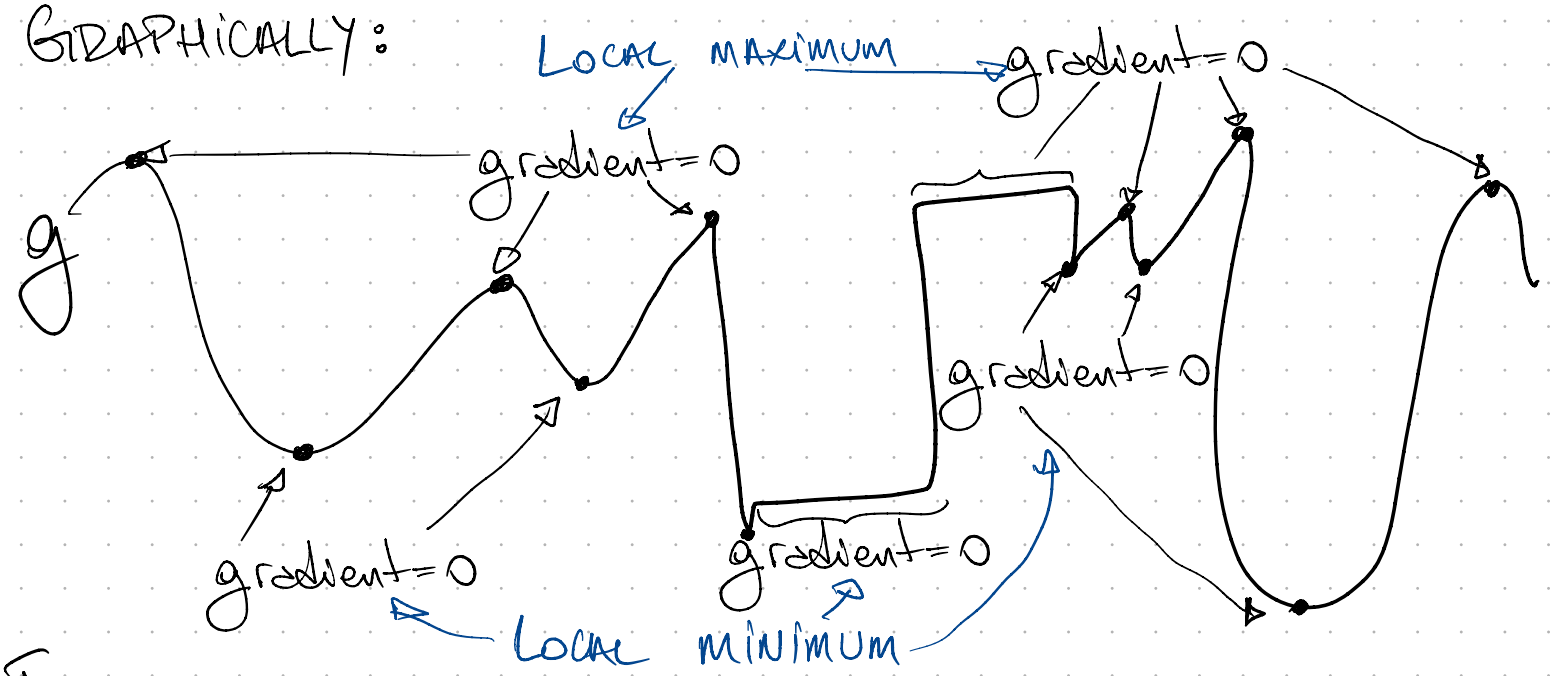
2.1 — continued

I've just proven that the gradient of an n-d function points towards where the func. increases / decreases the most, i.e. where there is most change per step relative to level surfaces.

This is key in DL because it all boils down to optimization, which is basically finding minima. With that info in hand we know that the gradient of a function is what we want to compute to be able to find said minima/um.

2.2 if $\exists\ \gamma > 0$ s.t. $\|w^T - w\|_2 < \gamma\ \forall\ w^T \in \mathbb{R}^d$ and $\|w^T - w\|_2 < \gamma \Rightarrow g(w^T) \leq g(w) \Rightarrow \exists$ local minimum

GRAPHICALLY:



FORMALLY:

let $w^T = [w_1, w_2, \dots w_n]^T$ take the gradient:

$$\nabla g(w^T) = \begin{bmatrix} g_{w_1}(w_1, w_2, \dots w_n) \\ g_{w_2}(w_1, w_2, \dots w_n) \\ \vdots \\ g_{w_n}(w_1, w_2, \dots w_n) \end{bmatrix}$$

if g has local minimum at w^T, by the definition of derivatives & minima:

$$\nabla g(w^T) = \begin{bmatrix} g_{w_1}(w_1, w_2, \dots w_n) \\ g_{w_2}(w_1, w_2, \dots w_n) \\ \vdots \\ g_{w_n}(w_1, w_2, \dots w_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

2.2 continued

which proves that if $\exists$ local minimum at some w^T then the gradient at $w^T = 0$.

the contrary is not true because $\nabla g(w^T) = 0$ at local maximums and saddle points, e.g.:

$g = -w^2 \longrightarrow w^T = 0 \Rightarrow \nabla g = -2w \Rightarrow \nabla g = 0$

$\therefore \nabla g(w^T) = 0$ is sufficient, but does not guarantee that w^T is a local minimum ■

2.3 We know the definition of local minimum irrespective of convexity from the previous question:

if $\exists\ \gamma > 0$ s.t. $\|w^* - w\|_2 < \gamma\ \forall\ w^T \in \mathbb{R}^d$ and $\|w^* - w\|_2 < \gamma \Rightarrow g(w^*) \leq g(w) \Rightarrow \exists$ local minimum

$\nabla g(w^*) = 0$ (conv. properties)

$\forall\ \delta \in (0, 1]$:

$$g((1-\delta)w^* + \delta w) \leq (1-\delta)g(w^*) + \delta g(w)$$

$$< (1-\delta)g(w^*) + \delta g(w^*) = g(w^*)$$

for small $\delta \Rightarrow \|(1-\delta)w^* + \delta w - w^*\| < \gamma \Rightarrow$ global minimum ■

2.4 $S(z)$: $s_i = \frac{e^{z_i}}{\sum_k e^{z_k}}$ Derive $\frac{\partial S}{\partial z}$

$$\frac{\partial s_i}{\partial z_j} = \frac{\partial \frac{e^{z_i} g}{\sum_k e^{z_k} h}}{\partial z_k} \longrightarrow$$

$$g' = \begin{cases} e^{z_k} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \qquad f' = \frac{g'h - h'g}{h^2}$$

$$h' = e^{z_k}$$

(Jacobian) MATRIX. Diag. $\Rightarrow$ $\Rightarrow i = j$

$$= \frac{g' \sum_k e^{z_k} - e^{z_k} e^{z_i}}{\left(\sum_k e^{z_k}\right)^2} \quad (I)$$

First, let's consider $i = j$

then from (I) and g': $\frac{\partial s_i}{\partial z_j} = \frac{e^{z_j} \sum_k e^{z_k} - e^{z_j} e^{z_i}}{\left(\sum_k e^{z_k}\right)^2}$

$$= \frac{e^{z_j}\left(\sum_k e^{z_k} - e^{z_i}\right)}{\left(\sum_k e^{z_k}\right)^2} = \underbrace{\frac{e^{z_j}}{\sum_k e^{z_k}}}_{\text{softmax}(j)} \cdot \underbrace{\frac{\sum_k e^{z_k} - e^{z_i}}{\sum_k e^{z_k}}}_{1 - \text{softmax}(i)}$$

$\frac{\partial s_i}{\partial z_j}$ (for $i = j$) $= (1 - S(j))\, S(i)$ $\rightsquigarrow$ Diag. of Jacobian

For $i \neq j$:

$$\frac{\partial s_j}{\partial z_j} = \frac{e^{z_i} \sum_k e^{z_k} - e^{z_j} e^{z_i}}{\left(\sum_k e^{z_k}\right)^2} \quad (e^{z_i} \to 0)$$

$$= -\underbrace{\frac{e^{z_j}}{\sum_k e^{z_k}}}_{\text{softmax}(j)} \cdot \underbrace{\frac{e^{z_i}}{\sum_k e^{z_k}}}_{\text{softmax}(i)} = -S(j)\, S(i)$$

2.4 — continued

$$\therefore \frac{\partial s}{\partial z} = \begin{cases} (1 - s(j))\, s(i) & \text{if } i = j \\ -s(j)\, s(i) & \text{else} \end{cases}$$

3.6

$G = (V, E)$; $V = \{v_1, v_2, \dots v_i, \dots v_n\}$; $E = \{e_{ij} = (v_i, v_j) \mid v_i, v_j \in V\}$

$G_+ = G + v_{n+1}$ is also DAG. Simmilarly, $G_- = G - v_n$ is also DAG

$G_+ = G - v_{n+1} \Rightarrow$ Edge $e_{n,n+1} = (v_n, v_{n+1})$ has to be added to E, now that V was added with v_{n+1}

$G_- = G - v_{n-1} \Rightarrow e_{n-1,n} = (v_{n-1}, v_n)$ has to be deleted from E because V was updated by deleting v_n

$G_- = G - v_{i-1} \Rightarrow e_{i-2,i-1} = (v_{i-2}, v_{i-1})$ and $e_{i-1,i} = (v_{i-1}, v_i)$ are deleted ; $e_{i-2,i} = (v_{i-2}, v_i)$ is added

Let $i = 1 \Rightarrow e_{0,1}$ and $e_{1,2}$ are deleted

$v_0 \notin V \Rightarrow e_{0,2} = (v_0, v_2)$ can not be added $\Rightarrow$

$\Rightarrow G_-$ doesn't have incomming edge $\Rightarrow$ there is topological order ■

3.7 Given the previous proof:

- each time a node is added, E_+ increases
- " " " " deleted, E_- decreases

the same goes for V_+ & V_- (in # of elements)

∴ When a node is added, it has a topological place. If it did not then it would be possible that the added edges are cyclical, e.g. $e_{new} = (v_i, v_i) \Rightarrow G$ is a DAG ■