

1. Let $r(t) = \langle x_1(t), \dots, x_d(t) \rangle$

$$r(t_0) = x_0 \text{ for some } t_0 \in \mathbb{R}$$

$$r(t) \in L_f(x_0) \quad \forall t$$

$$f(x_1, \dots, x_d) = C \leftarrow \text{level curve}$$

$$\text{let } g(t) = f(x_1(t), \dots, x_d(t))$$

Since curve r lies in the surface level

$$r(t_0) = \langle x_{10}, \dots, x_{d0} \rangle \Rightarrow g(t) = f(x_1(t), \dots, x_d(t)) = C$$

Differentiate g

$$\frac{dg}{dt} = \frac{df}{dx_1} \Big|_{r(t_0)} \frac{dx_1}{dt} \Big|_{t_0} + \dots + \frac{df}{dx_d} \Big|_{r(t_0)} \frac{dx_d}{dt} \Big|_{t_0} = 0$$

In vector form

$$\underbrace{\left\langle \frac{df}{dx_1} \Big|_{r(t_0)}, \dots, \frac{df}{dx_d} \Big|_{r(t_0)} \right\rangle}_{\nabla f_0} \cdot \underbrace{\left\langle \frac{dx_1}{dt} \Big|_{t_0}, \dots, \frac{dx_d}{dt} \Big|_{t_0} \right\rangle}_{r'(t_0)} \leftarrow \begin{matrix} \text{tangent of } r \text{ at} \\ t_0 \end{matrix} = 0$$

$$\Rightarrow \nabla f_0 \cdot r'(t_0) = 0 \Rightarrow \nabla f_0 \text{ is orthogonal to tangent of } r(t) \text{ at } t_0$$

Because of the orthogonality, we know that the gradient points to the next surface level. In deep learning we want to find t_0 that minimizes the loss, represented by the surface curve. Gradient descent moves the opposite direction of the magnitude of the orthogonal vector, minimizing the function f .

2. $g : \mathbb{R}^n \rightarrow \mathbb{R}$

local min at some w^t if \exists some $\gamma > 0$ s.t. $\forall w \in \mathbb{R}^n$,
 $\|w^t - w\|_2 < \gamma \Rightarrow g(w^t) \leq g(w)$

g has local min at $w^t \Rightarrow \nabla g(w^t) = 0$:

$$w^t = \langle w_1, \dots, w_n \rangle$$

let $g_i(w^t)$ be the partial derivative of g w.r.t. w_i , $1 \leq i \leq n$

\Rightarrow for all i s.t. $1 \leq i \leq n$ $g_i(w^t) = 0$

$$\text{so } \nabla g(w^t) = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

$\nabla g(w^t) = 0 \not\Rightarrow g$ has local min at w^t :

If ∇g is 0, w^t could still be a saddle point

Consider $g = (x^2 - y^2)$ and $w^t = \langle 0, 0 \rangle$

$$\frac{\partial g}{\partial x}(w^t) = 2x = 2(0) = 0$$

$$\frac{\partial g}{\partial y}(w^t) = -2y = -2(0) = 0$$

$\langle 0, 0 \rangle$ is not a local min but $\nabla g(w^t) = 0$

It is not a local min. We can look at
the graph of $x^2 - y^2$ or check second
derivatives. $g'_{xx}(w^t) = 2$ $g'_{yy}(w^t) = -2$



3. Prove $g: \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and $\nabla g(w^*) = 0 \Rightarrow w^*$ is global min

Definition of convex:

$$g(\lambda \vec{x}_1 + (1-\lambda) \vec{x}_2) \leq \lambda g(\vec{x}_1) + (1-\lambda) g(\vec{x}_2) \quad \text{for arbitrary } \vec{x}_1, \vec{x}_2$$

$$g(\vec{x}_2 + \lambda(\vec{x}_1 - \vec{x}_2)) \leq g(\vec{x}_2) + \lambda(g(\vec{x}_1) - g(\vec{x}_2))$$

$$\lambda(g(\vec{x}_1) - g(\vec{x}_2)) \geq g(\vec{x}_2 + \lambda(\vec{x}_1 - \vec{x}_2)) - g(\vec{x}_2)$$

$$g(\vec{x}_1) - g(\vec{x}_2) \geq \frac{1}{\lambda} [g(\vec{x}_2 + \lambda(\vec{x}_1 - \vec{x}_2)) - g(\vec{x}_2)]$$

As $\lambda \rightarrow 0$, the point on the line segment parameterized by λ becomes \vec{x}_2
so right hand side becomes $\nabla g(\vec{x}_2)(\vec{x}_1 - \vec{x}_2)$

Sub in w^* for \vec{x}_2

$$\therefore g(\vec{x}_1) - g(w^*) \geq 0$$

$$\Rightarrow g(\vec{x}_1) \geq g(w^*)$$

\vec{x}_1 is all possible points on g so it is w

$$\Rightarrow g(w^*) \leq g(w) \forall w \text{ on } g$$

$\Rightarrow w^*$ is the global minima \blacksquare

$$4. \quad S_i = \frac{e^{z_i}}{\sum_k e^{z_k}}$$

$$\text{Gradient of } S_i \text{ w.r.t. } z_j \leftarrow \stackrel{j^{\text{th}} \text{ input}}{=} \frac{ds_i}{dz_j} = \frac{d}{dz_j} \frac{e^{z_i}}{\sum_k e^{z_k}}$$

$$\text{Quotient Rule: } f(x) = \frac{g(x)}{h(x)}$$

$$g(x) = e^{z_i}$$

$$h(x) = \sum_k e^{z_k}$$

$$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h(x)^2}$$

Case $i=j$:

$$\frac{ds_i}{dz_i} = \frac{(e^{z_i} \sum_k e^{z_k}) - (e^{z_i} e^{z_i})}{(\sum_k e^{z_k})^2}$$

$$= \underbrace{\frac{e^{z_i}}{\sum_k e^{z_k}}}_{S_i} - \frac{e^{z_i} e^{z_i}}{(\sum_k e^{z_k})^2}$$

$$= S_i - S_i \cdot S_i = S_i(1 - S_i)$$

Case $i \neq j$:

$$\frac{ds_i}{ds_j} = \frac{0 - e^{z_i} e^{z_j}}{(\sum_k e^{z_k})^2} = -S_i S_j$$

$$\therefore \frac{ds_i}{ds_j} = \begin{cases} S_i(1 - S_j) & \text{if } i=j \\ -S_i S_j & \text{if } i \neq j \end{cases}$$

$$\text{Jacobian} = \begin{bmatrix} S_1(1 - S_1) & -S_1 S_2 & \dots & -S_1 S_n \\ -S_2 S_1 & \ddots & & \\ \vdots & & \ddots & \\ -S_n S_1 & & & S_n(1 - S_n) \end{bmatrix}$$

$$5. S(x) = \underset{y \in \mathbb{R}^d}{\operatorname{argmin}} -x^T y - H(y) \text{ s.t. } 1^T y = 1, 0 \leq y_i \leq 1 \forall i$$

$$H(y) = -\sum_i y_i \log(y_i)$$

Lagrangian: $L(y, v) = -x^T y - H(y) + v(1^T y - 1)$

KKT conditions of L gives us:

$$\nabla_y L(y, v) = 0, \quad v(1^T y - 1) = 0$$

$$\nabla_y L(y, v) = \log(y_i) + 1 - x_i + v = 0$$

$$\Rightarrow y_i = e^{-1+x_i-v} = e^{x_i-1-v}$$

$\log(y_i)$ guarantees $y_i > 0 \Rightarrow 0 < y_i \leq 1$

constraint $1^T y = 1$:

$$1 = \sum y_i = \sum (e^{x_i-1-v}) = \sum (e^{x_i} e^{-1} e^{-v})$$

$$\Rightarrow v = \log \sum e^{x_i-1}$$

sub in v :

$$y_i = e^{x_i-1-v} = e^{x_i} e^{-1} e^{-\log \sum e^{x_i-1}}$$

$$= \frac{e^{x_i}}{e^{\sum e^{x_i-1}}} = \frac{e^{x_i}}{e e^{-1} \sum e^{x_i}} = \frac{e^{x_i}}{\sum e^{x_i}} = 1 = 1^T y$$



6. If G DAG $\Rightarrow G$ has topological ordering

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

Topological ordering of $G = \{v_1, \dots, v_i, \dots, v_n\}$ s.t. for every edge $e_{ij} = (v_i, v_j)$ $i < j$

Proof by induction:

Base case: $n=1$ 1 node then that single node is the topological order

Inductive Step: Given G_n of size n is a DAG, G_n has topological order
Shows if G_{n+1} is a DAG $\Rightarrow G_{n+1}$ has topological order

G_{n+1} is DAG \Rightarrow exists a node in G with no incoming edges

subproof: assume G_{n+1} is DAG and all nodes have incoming edges
then there has to be a cycle some where contradicting the DAG property

Find the node, m , in G_{n+1} with no incoming edges

$(G_{n+1} - m)$ is a DAG of size n because removing a node with no incoming preserves the DAG property. $(G_{n+1} - m)$ is a DAG of size n and has a topological ordering t_n . Then the topological ordering of G_{n+1} is

$\{m, t_n\}$ Since no incoming edges to m , there will be no node with a higher number pointing to m \blacksquare

7. If G has topological order $\Rightarrow G$ is DAG

Proof by contradiction:

suppose not: G has topological order but is not DAG, aka a directed cyclic graph

let topological ordering of G be $t = \{v_1, \dots, v_n\}$

since G is cyclic, there exists at least 1 pair of nodes v_i, v_j
s.t. there is a path $v_i \rightarrow v_j$ and a path $v_j \rightarrow v_i$

In t , either $i < j$ or $j < i$

case $i < j$:

let the path $v_j \rightarrow v_i$ be $v_j \rightarrow v_{j1} \rightarrow \dots \rightarrow v_{jm} \rightarrow v_i$

$j < jx \quad \forall x \in (1, m)$ in t

we have $i < j$ so there will be a contradiction

at v_{jm} that points to v_i where

$i < j, j < jm, \text{ and } jm < i$

e.g. $i=1, j=2, jm=3 \rightarrow jm < i \rightarrow 3 < 1$

contradiction

case $j < i$:

let path $v_i \rightarrow v_j$ be $v_i \rightarrow v_{i1} \rightarrow \dots \rightarrow v_{im} \rightarrow v_j$

$i < ix \quad \forall x \in (1, m)$ in t

we have $j < i$ and like the other case $im < j$

\therefore we have $j < i, i < im, \text{ and } im < j$

contradiction

All possible cases have been shown to contradict,

∴ if G has topological order $\Rightarrow G$ is a DAG

