

$$a^T x = x^T a \Rightarrow [a_1, a_2, \dots, a_n] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [x_1, x_2, \dots, x_n] \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \sum_{i=1}^n a_i x_i = f \quad (1)$$

$$\frac{\partial f}{\partial x} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right] = [a_1, a_2, \dots, a_n] = a^T$$

$$\Rightarrow \frac{\partial (a^T x)}{\partial x} = \frac{\partial (x^T a)}{\partial x} = a^T$$

$$m \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \ddots & \vdots & \vdots \\ A_{m1} & A_{m2} & \dots & A_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + \dots + A_{1n}x_n \\ A_{21}x_1 + A_{22}x_2 + \dots + A_{2n}x_n \\ \vdots \\ A_{m1}x_1 + A_{m2}x_2 + \dots + A_{mn}x_n \end{bmatrix}_{m \times 1} = \begin{bmatrix} \sum_{i=1}^n A_{1i}x_i \\ \sum_{i=1}^n A_{2i}x_i \\ \vdots \\ \sum_{i=1}^n A_{mi}x_i \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} = f \quad (2)$$

$$\frac{\partial f}{\partial x} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right] = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \ddots & \vdots & \vdots \\ A_{m1} & A_{m2} & \dots & A_{mn} \end{bmatrix}_{m \times n} = A$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \quad (3)$$

$$x^T A x = (\sum_{i=1}^n x_i a_{ii}) x_1 + (\sum_{i=1}^n x_i a_{i2}) x_2 + \dots + (\sum_{i=1}^n x_i a_{in}) x_n = f$$

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \sum_{i=1}^n x_i a_{ii} + a_{11} x_1 + a_{12} x_2 + \dots + a_{1n} x_n \\ &= (a_{11} + a_{11}) x_1 + (a_{21} + a_{12}) x_2 + (a_{31} + a_{13}) x_3 + \dots + (a_{n1} + a_{1n}) x_n \end{aligned}$$

$$\begin{aligned} \frac{\partial f}{\partial x_2} &= a_{21} x_1 + \sum_{i=1}^n x_i a_{i2} + a_{22} x_2 + a_{23} x_3 + \dots + a_{2n} x_n \\ &= (a_{21} + a_{12}) x_1 + (a_{22} + a_{22}) x_2 + (a_{23} + a_{32}) x_3 + \dots + (a_{2n} + a_{n2}) x_n \end{aligned}$$

دالة معرفة بمتغير واحد دالة متعددة المتغيرات

$$\frac{\partial f}{\partial x} = (A + A^T)x = \begin{bmatrix} a_{11} + a_{11} & a_{12} + a_{21} & \dots & a_{1n} + a_{n1} \\ a_{21} + a_{12} & a_{22} + a_{22} & \dots & a_{2n} + a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} + a_{1n} & a_{n2} + a_{2n} & \dots & a_{nn} + a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

~~$$\frac{\partial f}{\partial x} = \begin{bmatrix} (a_{11} + a_{11})x_1 + (a_{12} + a_{21})x_2 + \dots + (a_{1n} + a_{n1})x_n \\ (a_{21} + a_{12})x_1 + (a_{22} + a_{22})x_2 + \dots + (a_{2n} + a_{n2})x_n \\ \vdots \\ (a_{n1} + a_{1n})x_1 + (a_{n2} + a_{2n})x_2 + \dots + (a_{nn} + a_{nn})x_n \end{bmatrix} = (A + A^T)x$$~~

$$x^T A x = (\sum_{i=1}^n x_i a_{ii}) x_1 + (\sum_{i=1}^n x_i a_{i2}) x_2 + \dots + (\sum_{i=1}^n x_i a_{in}) x_n = f$$

$$\frac{\partial f}{\partial a_{ij}} = x_i x_j \Rightarrow \frac{\partial f}{\partial A} = \left\{ \begin{array}{l} \frac{\partial f}{\partial a_{11}} = x_1 x_1, \quad \frac{\partial f}{\partial a_{12}} = x_1 x_2, \dots, \frac{\partial f}{\partial a_{1n}} = x_1 x_n \\ \vdots \\ \frac{\partial f}{\partial a_{n1}} = x_n x_1, \quad \frac{\partial f}{\partial a_{n2}} = x_n x_2, \dots, \frac{\partial f}{\partial a_{nn}} = x_n x_n \end{array} \right\} = x x^T$$

$$X = \begin{bmatrix} | & | & | \\ x_1 & x_2 & \dots & x_n \\ | & | & \dots & | \end{bmatrix} = \begin{bmatrix} x_1^t \\ x_2^t \\ \vdots \\ x_n^t \end{bmatrix} \quad A = \begin{bmatrix} | & | & | \\ a_1 & a_2 & \dots & a_n \\ | & | & \dots & | \end{bmatrix}$$

$$X A X^T = \begin{bmatrix} x_1^t a_1 & x_1^t a_2 & \dots & x_1^t a_n \\ x_2^t a_1 & x_2^t a_2 & \dots & x_2^t a_n \\ \vdots & \vdots & \ddots & \vdots \\ x_n^t a_1 & x_n^t a_2 & \dots & x_n^t a_n \end{bmatrix} X^T = \begin{bmatrix} x_1^t A \\ x_2^t A \\ \vdots \\ x_n^t A \end{bmatrix} \rightarrow X^T$$

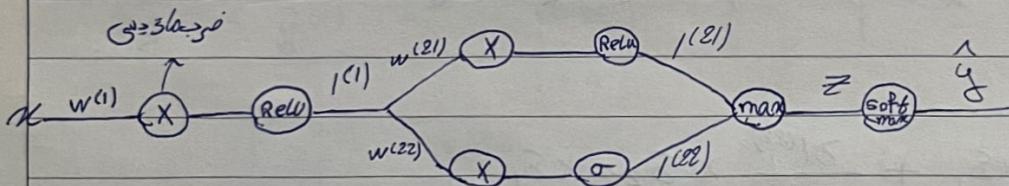
$$= \begin{bmatrix} x_1^t A \\ x_2^t A \\ \vdots \\ x_n^t A \end{bmatrix} \begin{bmatrix} | & | & | \\ x_1^{tT} & x_2^{tT} & \dots & x_n^{tT} \\ | & | & \dots & | \end{bmatrix}$$

$$= \begin{bmatrix} x_1^T A x_1^T \\ x_2^T A x_2^T \\ \vdots \\ x_n^T A x_n^T \end{bmatrix}$$

$$\operatorname{tr}(XAX^T) - \sum_{i=1}^n x_i^T A x_i^T = \begin{bmatrix} \frac{\partial}{\partial x_1^T} \sum_{i=1}^n x_i^T A x_i^T \\ \frac{\partial}{\partial x_2^T} \sum_{i=1}^n x_i^T A x_i^T \\ \vdots \\ \frac{\partial}{\partial x_n^T} \sum_{i=1}^n x_i^T A x_i^T \end{bmatrix}$$

$$= \begin{bmatrix} x_1^T (A + A^T) \\ x_2^T (A + A^T) \\ \vdots \\ x_n^T (A + A^T) \end{bmatrix} = X^T (A + A^T)$$

2 جسمی
①



$$\hat{y} \in \mathbb{R}^3 \quad z \in \mathbb{R}^3 \quad L = - \sum_{i=1}^3 y^{(i)} \log \hat{y}^{(i)}$$

$$\frac{\partial L}{\partial z_i} = \begin{cases} -\hat{y}_i y_j & \text{if } i \neq j \\ \hat{y}_j - y_j^2 & \text{if } i = j \end{cases} \Rightarrow \text{جاكوب جان} \rightarrow \text{مشكلة} \rightarrow \text{since} \text{ Ju} \rightarrow 3 \times 3$$

$$\frac{\partial L}{\partial z} = \frac{\partial \hat{y}}{\partial z} \times \frac{\partial L}{\partial \hat{y}} = \frac{\partial \hat{y}}{\partial z} \times \frac{y}{\hat{y}} = \delta_1 \in \mathbb{R}^{3 \times 1}$$

subject: _____

date: _____

$$\frac{\partial Z}{\partial I^{(21)}} = I(I^{(21)} > I^{(22)}) \cdot \delta_1 \in R^{3 \times 1}$$

$$\frac{\partial Z}{\partial I^{(22)}} = I(I^{(22)} > I^{(21)}) \cdot \delta_{22} \in R^{3 \times 1}$$

$$\frac{\partial L}{\partial I^{(21)}} = \delta_{21} \odot \delta_1 \in R^{3 \times 1}$$

$$\frac{\partial L}{\partial I^{(22)}} = \delta_{22} \odot \delta_1 \in R^{3 \times 1}$$

$$\frac{\partial \tilde{Z}^{(21)}}{\partial w^{(21)}} = I(w^{(21)} I^{(1)} > 0) \times (I^{(1)})^T$$

$$\frac{\partial L}{\partial w^{(21)}} = I(w^{(21)} I^{(1)} > 0) \odot \delta_{21} \odot \delta_1 \times (I^{(1)})^T \in R^{3 \times 2}$$

$$\frac{\partial I^{(22)}}{\partial w^{(22)}} = \sigma(w^{(22)} I^{(1)}) \odot (1 - \sigma(w^{(22)} I^{(1)})) \times (I^{(1)})^T$$

$$\frac{\partial L}{\partial w^{(22)}} = \underbrace{\sigma(w^{(22)} I^{(1)}) \odot (1 - \sigma(w^{(22)} I^{(1)}))}_{R^{3 \times 1}} \cdot \underbrace{\delta_{22} \odot \delta_1 \times (I^{(1)})^T}_{R^{3 \times 1} \times R^{1 \times 2}} \in R^{3 \times 2}$$

$$\frac{\partial L}{\partial I^{(1)}} = \frac{\partial I^{(21)}}{\partial I^{(1)}} \times \delta_{21} \times \delta_1 + \frac{\partial I^{(22)}}{\partial I^{(1)}} \times \delta_{22} \times \delta_1$$

$$= (w^{(21)})^T \times I(w^{(21)} I^{(1)} > 0) \odot I(I^{(21)} > I^{(22)}) \odot \delta_1 + (w^{(22)})^T \times \underbrace{\sigma(w^{(22)} I^{(1)}) \odot (1 - \sigma(w^{(22)} I^{(1)}))}_{R^{3 \times 4}} \times \delta_{22}$$

$$\odot I(I^{(22)} > I^{(21)}) \odot \delta_1 = \delta_3 \in R^{2 \times 1}$$

$$\frac{\partial L}{\partial w^{(1)}} = \frac{\partial I^{(1)}}{\partial w^{(1)}} \times \delta_3 \xrightarrow{\text{approx}} \text{where } \frac{\partial L}{\partial I^{(1)}} = \frac{\partial L}{\partial I^{(1)}}$$

$$I(w^{(1)} > 0) \odot \delta_3 \times x^T \in R^{2 \times 4}$$

(3)

feed forward

$$l^{(1)} = \text{ReLU}(w^{(1)}x) = \text{ReLU}\left(\begin{bmatrix} -1 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$l^{(2)} = \text{ReLU}\left(\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) = \text{ReLU}\left(\begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}$$

$$l^{(2)} = \text{sigmoid}\left(\begin{bmatrix} 2 & -1 \\ 1 & -2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) = \text{sigmoid}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$$

$$Z = \max\left(\begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}\right) = \begin{bmatrix} 3 \\ 0.5 \\ 1 \end{bmatrix}$$

$$g = \text{softmax}(Z) = \frac{e^{z_i}}{\sum_{i=1}^3 e^{z_i}} \Rightarrow \hat{g} = \text{softmax}\left(\begin{bmatrix} 3 \\ 0.5 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 0.82 \\ 0.07 \\ 0.11 \end{bmatrix}$$

Backward pass

$$\frac{\partial L}{\partial g} = -\frac{y}{\hat{g}^{(i)}} = -\frac{\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}}{\begin{bmatrix} 0.82 \\ 0.07 \\ 0.11 \end{bmatrix}} = \begin{bmatrix} 0 \\ -14.28 \\ 0 \end{bmatrix}$$

$$\frac{\partial \hat{g}}{\partial z} = \begin{bmatrix} \frac{\partial \hat{g}^{(1)}}{\partial z^{(1)}} & \frac{\partial \hat{g}^{(1)}}{\partial z^{(2)}} & \frac{\partial \hat{g}^{(1)}}{\partial z^{(3)}} \\ \frac{\partial \hat{g}^{(2)}}{\partial z^{(1)}} & \frac{\partial \hat{g}^{(2)}}{\partial z^{(2)}} & \frac{\partial \hat{g}^{(2)}}{\partial z^{(3)}} \\ \frac{\partial \hat{g}^{(3)}}{\partial z^{(1)}} & \frac{\partial \hat{g}^{(3)}}{\partial z^{(2)}} & \frac{\partial \hat{g}^{(3)}}{\partial z^{(3)}} \end{bmatrix} = \begin{bmatrix} 0.1476 & -0.0574 & -0.0902 \\ -0.0574 & 0.0651 & -0.0077 \\ -0.0902 & -0.077 & 0.0979 \end{bmatrix}$$

$$\frac{\partial L}{\partial z} = \frac{\partial \hat{g}}{\partial z} \times \frac{\partial L}{\partial \hat{g}} = \begin{bmatrix} 0.82 \\ -0.93 \\ 0.11 \end{bmatrix} = s_1$$

$$\frac{\partial L}{\partial I^{(21)}} = \frac{\partial z}{\partial I^{(21)}} \times S_1 = I(I^{(21)} > I^{(22)}) \times S_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \odot \begin{bmatrix} 0.82 \\ -0.93 \\ 0.11 \end{bmatrix} = \begin{bmatrix} 0.82 \\ 0 \\ 0.11 \end{bmatrix}$$

$$\frac{\partial L}{\partial I^{(22)}} = \frac{\partial z}{\partial I^{(22)}} \times S_1 = I(I^{(22)} > I^{(21)}) \times S_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \odot \begin{bmatrix} 0.82 \\ -0.93 \\ 0.11 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.93 \\ 0 \end{bmatrix}$$

$$\begin{aligned} \frac{\partial L}{\partial w^{(21)}} &= \frac{\partial I^{(21)}}{\partial w^{(21)}} \times \frac{\partial z}{\partial I^{(21)}} \times S_1 = I(w^{(21)} > 0) \odot \frac{\partial z}{\partial I^{(21)}} \times S_1 \times (I^{(1)})^T \\ &= \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \odot \begin{bmatrix} 0.82 \\ 0 \\ 0.11 \end{bmatrix} \times \begin{bmatrix} 1 & 2 \end{bmatrix} = \begin{bmatrix} 0.82 & 1.64 \\ 0 & 0 \\ 0.11 & 0.22 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial w^{(22)}} &= \frac{\partial I^{(22)}}{\partial w^{(22)}} \times \frac{\partial z}{\partial I^{(22)}} \times S_1 = \cancel{\sigma(w^{(22)} > 0)} \odot (1 - \sigma(w^{(22)} > 0)) \odot \frac{\partial z}{\partial I^{(22)}} \times S_1 \times (I^{(1)})^T \\ &= \begin{bmatrix} 0 \\ 0.5 \\ 0 \end{bmatrix} \odot \begin{bmatrix} 0 \\ 0.5 \\ 0 \end{bmatrix} \odot \begin{bmatrix} 0 \\ -0.93 \\ 0 \end{bmatrix} \times \begin{bmatrix} 1 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -0.23 & -0.46 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial I^{(1)}} &= \frac{\partial I^{(21)}}{\partial I^{(1)}} \times \frac{\partial L}{\partial I^{(21)}} + \frac{\partial I^{(22)}}{\partial I^{(1)}} \times \frac{\partial L}{\partial I^{(22)}} \\ &= (w^{(21)})^T \times I(w^{(21)} > 0) \odot \frac{\partial L}{\partial I^{(21)}} + (w^{(22)})^T \sigma(w^{(22)} > 0) \odot (1 - \sigma(w^{(22)} > 0)) \times S_{22} \\ &= \begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \odot \begin{bmatrix} 0.82 \\ 0 \\ 0.11 \end{bmatrix} + \begin{bmatrix} 2 & 4 & -2 \\ -1 & -2 & 1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0.25 \\ 0 \end{bmatrix} \odot \begin{bmatrix} 0 \\ -0.93 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.22 \\ 1.39 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial w^{(1)}} &= \frac{\partial I^{(1)}}{\partial w^{(1)}} \times I = I(w^{(1)} > 0) \odot \cancel{G} \times x^T \\ &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \odot \begin{bmatrix} -0.22 \\ 1.39 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 3 & 1 \end{bmatrix} = \begin{bmatrix} -0.22 & -0.44 & -0.66 & -0.22 \\ 1.39 & 2.78 & 4.17 & 1.39 \end{bmatrix} \end{aligned}$$

(3) بیان

لمس نظریه 2، گوییم ①

$$f(x+s) = f(x) + \nabla f(x)^T s + \frac{s^2}{2} \nabla^2 f(x) s^T$$

$$f(x+t\Delta t) = f(x) + t \nabla f(x)^T \Delta x + \frac{t^2}{2} \Delta x \nabla^2 f(x) \Delta x^T$$

اگر $\nabla^2 f(x) \leq mI$ $\rightarrow \Delta x \nabla^2 f(x) \Delta x^T \leq m \|\Delta x\|_2^2$

upper bound of $f(x+t\Delta t)$:

$$f(x+t\Delta t) \leq f(x) + t \nabla f(x)^T \Delta x + \frac{t^2}{2} m \|\Delta x\|_2^2 \quad ①$$

برای کمینه کردن این معادله، t را بروز کنید

$$f(x+t\Delta t) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$$

~~①~~ $f(x) + t \nabla f(x)^T \Delta x + \frac{t^2}{2} m \|\Delta x\|_2^2 \leq f(x) + \alpha t \nabla f(x)^T \Delta x$

$$(1-\alpha) t \nabla f(x)^T \Delta x + \frac{t^2}{2} m \|\Delta x\|_2^2 \leq 0$$

$$\div t \rightarrow (1-\alpha) \nabla f(x)^T \Delta x + \frac{t}{2} m \|\Delta x\|_2^2 \leq 0$$

$$t \leq -\frac{2(1-\alpha) \nabla f(x)^T \Delta x}{m \|\Delta x\|_2^2}$$

$$\Rightarrow t \leq -\frac{\nabla f(x)^T \Delta x}{m \|\Delta x\|_2^2}$$

 $\alpha \in (0, 0.5]$ if $\alpha = 0.5$

لـ iteration الـ t ، نـ $\nabla f(\alpha)$ بـ β ، α لـ iteration t ، $\Delta \alpha$ بـ β (2)
از رـ $\nabla f(\alpha)$ بـ β بـ $\Delta \alpha$

$$t = (\beta)^x t$$

$$\frac{\text{لـ } \nabla f(\alpha) \text{ بـ } \beta}{\text{لـ } \alpha} \rightarrow (\beta)^x t \leq - \frac{\nabla f(\alpha)^T \Delta \alpha}{M \|\Delta \alpha\|_2^2}$$

$$\frac{\div t}{\div t} \rightarrow (\beta)^x \leq - \frac{\nabla f(\alpha)^T \Delta \alpha}{M \|\Delta \alpha\|_2^2 t} \xrightarrow{\log \beta} \boxed{x \leq \log - \frac{\nabla f(\alpha)^T \Delta \alpha}{M \|\Delta \alpha\|_2^2 t}}$$

$$v_0 = 0$$

جـ v

①

$$v_1 = \beta_2 v_0 + (1-\beta_2) g_1^2$$

$$v_2 = \beta_2 v_1 + (1-\beta_2) g_2^2 = \beta_2 (\beta_2 v_0 + (1-\beta_2) g_1^2) + (1-\beta_2) g_2^2$$

$$v_3 = \beta_2 v_2 + (1-\beta_2) g_3^2 = \beta_2 [\beta_2 (\beta_2 v_0 + (1-\beta_2) g_1^2) + (1-\beta_2) g_2^2] + (1-\beta_2) g_3^2 \\ = \beta_2^2 (1-\beta_2) g_1^2 + \beta_2 (1-\beta_2) g_2^2 + (1-\beta_2) g_3^2$$

لـ v_t اـ $v_t = (1-\beta_2) [\beta_2^{t-1} g_1^2 + \beta_2^{t-2} g_2^2 + \dots + \beta_2^0 g_t^2]$

$$\boxed{v_t = (1-\beta_2) \sum_{i=1}^t \beta_2^{t-i} g_i^2}$$

لـ $E[g_i^2]$ $E[g_i^2] = E[g]$ لـ i i.i.d g_i لـ $E[g_i^2] = E[g^2]$ (2)

$$E[v_t] = 1-\beta_2 \sum_{i=1}^t \beta_2^{t-i} E[g_i^2] = (1-\beta_2) E[g^2] \sum_{i=1}^t \beta_2^{t-i}$$

لـ $\sum_{i=0}^{t-1} \beta_2^i = \frac{1-\beta_2^t}{1-\beta_2}$ converges to β_2

$$E[v_t] = (1-\beta_2) E[g^2] \frac{1-\beta_2^t}{1-\beta_2} = \boxed{E[g^2] (1-\beta_2^t)}$$

در اینجا از فریم های متنی که در گذشته ایدیوم

$$U_{t+1,m} = \lim_{P \rightarrow \infty} \left(1 - \beta_2 \sum_{i=1}^t \beta_2^{P(t-i)} \lg_i(P) \right)^{\frac{1}{P}}$$

ما وحده ایک کم $P \rightarrow \infty$ می توان از قسم B_2 -ا صرف نظر کردن زیرا مقدار دسیمارین جزو دارد

حالی دفعه بزرگ عبارت dominate و تعدد و مجموعات آن را بزرگترین قدر می‌دانیم.

$$\sum_{i=1}^t \beta_2^{P(t-i)} |g_i|^P = \max_{1 \leq i \leq t} (\beta_2^{t-i} |g_i|)^P$$

$$a_t = \lim_{P \rightarrow \infty} \left(\max_{1 \leq i \leq t} (B_2 |g_i|)^P \right)^{\frac{1}{P}} = \max_{1 \leq i \leq t} B_2^{t-i} |g_i|$$

$$U_t = \max_{1 \leq i \leq t} (\beta_2 \cdot \max_{j \geq 1} \beta_2^{(t-1)-i} |g_i|, |g_j|)$$

اگر یعنی در راه اکسیم می‌گذرد 2 نمایم ریواج و گردان
باشد one scaled بز بزرگتر max -1

١٩٦١ مارس - ٢

$$u_t = \max(B_2 u_{t-1}, |g_t|)$$

5

acetars glycols fatty

در آنکه آنها را با این نسبتی که آنها را در آن می‌دانند، می‌دانند و آنها را با این نسبتی که آنها را در آن می‌دانند، می‌دانند.

Sparse Gradients (2)

در کنایه از ما نیز NLP گوایان را بسیار سازگار می‌کنند این کارها را می‌توان در مکانی اسناد گوایان نیز داشت این مکانی همانند دستگاهی است که با توجه به مقدار داده شده از گوایان می‌تواند آن را برآورد کند. این دستگاه از آنکه آن را با آنکه آن را برآورد کند و برآورده کنند برآورده کنند

(٣) مادی و مکانیزم نیاز به تغییر پارامترها

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِيْمِ

$$f(x) \approx f(x_0) + (x - x_0)^T \nabla f(x_0) + \frac{1}{2} (x - x_0)^T H (x - x_0)$$

11

$$f(x_0 - \epsilon g) \approx f(x_0) + (x_0 - \epsilon g + x_0)^T \nabla f(x_0) + \frac{1}{2} (-\epsilon g)^T H(\epsilon g) + O(\epsilon^3)$$

$$f(x_0 - \epsilon g) \approx f(x_0) + \epsilon g^T g + \frac{1}{2} \epsilon^2 g^T H g + O(\epsilon^3)$$

2

ج) هنف با شد متابعی دارد بالعوایتی و مقدار نابغه فوراً ماده کاهشی می‌باشد. هنف بوسیله
حتمی یعنی // log_e و هنف با نامهای است که این یعنی مقدار زیستی (minima) یعنی
جهنم ندارد.

١) دفع خرض : Coronal convexity

روش خوبی و روش های مرتبه دوم غیر رایجی که مارکس دانست همچنان این است که در میان صنایع که بعدها مطالعه کردند اینها را در تصور خود می‌دانند اما با وجود این فقرات در تصور خود می‌دانند

: saddle point موجو دار کردن diverge $\int f(z) dz$ (2)

لما كانت المجموعة غير محددة فإنها تدعى كنقطة سaddle point أي هي نقطة لا يحيى لها اتجاه واحد

۳) از $O(\ell^3)$ که در مجموع ℓ تا کوچکتر از ℓ است که با ℓ برابر نباشد

$$f(x - \epsilon g) = f(x_0) - \epsilon g^T g + \frac{1}{2} \epsilon^2 g^T H g + O(\epsilon^3) \xrightarrow{\text{ignore}} \text{ignore}$$

$$\frac{\partial f(x - \epsilon g)}{\partial g} = -g^T g + \epsilon g^T H g = 0$$

$$\Rightarrow \mathcal{E}^* = -\frac{\vec{g}^\top \vec{g}}{\vec{g}^\top H \vec{g}}$$

فرضیه کنیم H مثبت می باشد است

v_1, v_2, \dots, v_n بطریق ورثه ای درستند

$\lambda_1, \lambda_2, \dots, \lambda_n$ بطریق ورثه ای درستند

گذاشتن رساله برای داشتار دهنم $g = \sum_{i=1}^n \alpha_i v_i$

$$g^T g = \left(\sum_{i=1}^n \alpha_i v_i \right)^T \left(\sum_{i=1}^n \alpha_i v_i \right) = \sum_{i=1}^n \alpha_i^2$$

$$Hg = \sum_{i=1}^n \alpha_i H v_i = \sum_{i=1}^n \alpha_i \lambda v_i$$

$$g^T H g = \left(\sum_{i=1}^n \alpha_i v_i \right)^T \left(\sum_{i=1}^n \alpha_i \lambda v_i \right) = \sum_{i=1}^n \alpha_i^2 \lambda$$

$$\varepsilon^* = \frac{g^T g}{g^T H g} = \frac{\sum \alpha_i^2}{\sum \alpha_i^2 \lambda}$$

از تعریف ε^* نتایج رخواهد کرد فقط در اینجا کوچکتر مقدار و خوب تر

$$\varepsilon^* = \frac{\lambda_{\min}}{\alpha_{\min}^2 \lambda_{\min}} = \frac{1}{\lambda_{\min}}$$

آنچه می بینیم کوچکتر ε^* نباشد رخواهد کرد قبلاً در اینجا رخواهد کرد

$$\varepsilon^* = \frac{\lambda_{\max}^2}{\alpha_{\max}^2 \lambda_{\max}} = \frac{1}{\lambda_{\max}}$$

$$\frac{1}{\lambda_{\max}} \leq \varepsilon^* \frac{g^T g}{g^T H g} \leq \frac{1}{\lambda_{\min}}$$

جوسن 6

①

$$f(x) = w \cdot x$$

$$\xrightarrow[\text{Lipschitz}]{\text{لابسچيت}} \|w\alpha_1 - w\alpha_2\| \leq L \|\alpha_1 - \alpha_2\| \quad \forall \alpha_1, \alpha_2$$

$$\xrightarrow{\alpha_1 - \alpha_2 = \delta} L = \max_{\delta \neq 0} \frac{\|w\delta\|}{\|\delta\|} \quad \text{I}$$

مقدار مطلق $\|w\delta\|$ مطلق $\|w\|_2$ معرفه شده است

$$\|w\|_2 = \max_{\delta \neq 0} \frac{\|w\delta\|}{\|\delta\|} \quad \text{II} \quad \xrightarrow{\text{I}, \text{II}} L = \|w\|_2$$

②

ویرایشی کردن SVD $w \in \mathbb{R}^{m \times n}$ می‌کند

$$w = U \Sigma V^T$$

$U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ \rightarrow ماتریسی که مجموعه ای از مداری را درست می‌کند

$\Sigma \in \mathbb{R}^{m \times n} \rightarrow \sigma_1 > \sigma_2 > \dots > \sigma_{\min(m, n)}$ ماتریسی که مجموعه ای از مداری را درست می‌کند

$$\|w\delta\|_2 = \|U \Sigma V^T \delta\|_2$$

برای داشتن این مقدار $\|w\delta\|_2 = \|\Sigma \delta\|_2 = \sqrt{\sum_{i=1}^n \sigma_i^2 \delta_i^2} \leq \sigma_{\max} \|\delta\|_2$

$$\rightarrow \|w\delta\|_2 \leq \sigma_{\max} \|\delta\|_2 \rightarrow \sigma_{\max} = \max_{\delta \neq 0} \frac{\|w\delta\|_2}{\|\delta\|_2}$$

$L = \sigma_{\max}$

$\delta \in \mathbb{R}^n$ معرفه شده است

لipschitz ثابت ③

$$L = \sup_{\alpha} |f'(\alpha)|$$

$$\text{ReLU}'(z) = \frac{\partial \text{ReLU}(z)}{\partial z} = \begin{cases} 1 & z > 0 \\ 0 & z < 0 \end{cases}$$

$$\text{ReLU}'(z) \leq 1 \rightarrow L = \sup_z |\text{ReLU}'(z)| = 1$$

$$\begin{aligned} \|\text{ReLU}(z_2) - \text{ReLU}(z_1)\| &\leq 1 \times \|z_2 - z_1\| \\ &\leq \frac{1 \times \|w\|_2 \times \|x_2 - x_1\|}{L} \end{aligned}$$

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad \frac{\partial \tanh(z)}{\partial z} = 1 - \tanh^2(z)$$

$$\tanh(z) \in (-1, 1) \rightarrow \|1 - \tanh^2(z)\| \leq 1$$

$$L = \sup_z \|1 - \tanh^2(z)\| = 1$$

$$\begin{aligned} \|\tanh(z_2) - \tanh(z_1)\| &\leq \|z_2 - z_1\| \\ &\leq \frac{1 \cdot \|w\|_2 \cdot \|x_2 - x_1\|}{L} \end{aligned}$$

$$\sigma(z) = \frac{1}{1+e^{-z}} \rightarrow \frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1-\sigma(z))$$

$$L = \sup_z \|\sigma(z)(1-\sigma(z))\| = \frac{1}{4}$$

$$\begin{aligned} \|\sigma(z_2) - \sigma(z_1)\| &\leq \frac{1}{4} \|z_2 - z_1\| \\ &\leq \frac{1}{4} \frac{\|w\|_2 \|x_2 - x_1\|}{L} \end{aligned}$$

$\text{ReLU}^{(1)} = \text{ReLU}(w^{(1)} x^{(0)})$ $\|w^{(1)}\|_2$ لـ $w^{(1)}$ $x^{(0)}$ ReLU $\|w^{(1)}\|_2$ $\|w^{(1)}\|_2$ $\|w^{(1)}\|_2$ $\|w^{(1)}\|_2$

$$\text{layer}(1) = \|w^{(0)}\|_2 \cdot 1 = \|w^{(0)}\|$$

$$L_{\text{total}} = \prod_{i=1}^n L_i = \prod_{i=1}^n \|w^{(i)}\|_2$$

۵) دریچه هر داد با نوزیری یعنی $y + x = \alpha$ باعثی شود که مانند زیرشود.

$$w\tilde{x} = w(x+\epsilon) = wx + w\epsilon$$

که در راسته نظرخواست و مواردیم در
آنچه مذکور شود فرموده باشد اینجا از آنها

$$\|\omega e\| \leq \|\omega\| \cdot \|e\|$$

بایلر کوچک کردن // س || فعالیتی از نظر اطمینانی رعد.

تابع $h(wx)$ با h که activation function نام دارد و Lipschitz است -
با w معرفی شده است

$\|w\|_2 \cdot b$, (spectral norm of $w \times a^T$'s Lipschitz constant)

با خصوصیت Lipschitz و مطابق با فرضیه ثابت 2 بهمینجا باید $\|w\|_2$ بینتر (نماینده کوچکتر) باشد.