①

H.W. 0

1.

Linear Regression Problem with n-training points and d-features when $n=d$, the feature matrix becomes $\mathbb{R}^{n \times n}$ where F has max singular value α and a very tiny min. singular value.

for $y = Fw^* + e$

$$\hat{\omega}_{inv} = F^{-1} y \Rightarrow \| \hat{\omega}_{inv} - w^* \|_2^2 = 10^{10}$$

Taking the Gradient Descent Approach,

$$l(w) = \frac{1}{2} \| Y - Fw \|^2$$

starting from $\omega_0 = 0$

Gradient Descent Update ⇒

$$\omega_t = \omega_{t-1} - \eta \left(F^T (F w_{t-1} - y) \right)$$

we are interested in minimizing $\| \omega_k - \omega^* \|_2^2$

Because

$F = V \Lambda V^{-1}$

$F^{-1} = V^{-1} \Lambda^{-1} V$

so,

$$\Lambda^{-1} = \begin{bmatrix} 1/r_1 & & & \\ & 1/r_2 & & \\ & & \ddots & \\ & & & 1/r_n \end{bmatrix}$$

⇓

Small singular value makes the reciprocal very large!

Show that: for $t > 0$, $\| \omega_t \|_2 \le \| \omega_{t-1} \|_2 + \eta \alpha \| y \|_2$

So, we know that ~~ω_{t-1}~~ $\omega_t = \omega_{t-1} - \eta F^T (F w_{t-1} - y))$

$$\omega_t = \omega_{t-1} - \eta F^T F w_{t-1} + \eta F^T y$$

$$\omega_t = \omega_{t-1} [I - \eta F^T F) + \eta F^T y$$

take norm (L-2) on both sides,

$$\| \omega_t \|_2 = \| \omega_{t-1} (I - \eta F^T F) \|_2 + \eta \| F^T y \|_2$$

$$\| \omega_t \|_2 \le \| \omega_{t-1} \|_2 \| I - \eta F^T F \|_2 + \eta \| F^T y \|_2$$

Recall, for gradient descent to converges we need to look more closely at $\| I - F^T F \eta \|$

If $F \in \mathbb{R}^{n \times n}$ then, $F = V \Lambda V^T$ where $\max(\Lambda) = \alpha$.

And, $F^T F = (V \Lambda V^T)^T (V \Lambda V^T)$

$= V \Lambda^T V^T V \Lambda V^T$ where $V =$ orthonormal basis

$F^T F = V \Lambda^2 V^T$

$\|F^T F\|_2 = \sqrt{\|\Lambda\|^4} = \|\Lambda\|_2^2$.

so, $\|I - F^T F \eta\| < 1$

$\|w_t\|_2 \leq \|w_{t-1}\|_2 + \eta \|F^T y\|_2$

$\|w_t\|_2 \leq \|w_{t-1}\|_2 + \eta \|V \Lambda V^T\|_2 \|y\|_2$

$\|w_t\|_2 \leq \|w_{t-1}\|_2 + \eta \alpha \|y\|_2$ since $\max \|V \Lambda V^T\|$ is α i.e., All diagonal entries have αs.

hence shown !

2. Show that

$$\boxed{\arg\min_w \|\hat{y} - \hat{X} w\|_2^2} \quad \text{— (I)}$$

has the same solution

as $$\boxed{\hat{w} = (X^T X + \Sigma_0^{-1})^{-1} X^T y} \quad \text{— (II)}.$$

and $$\hat{X} = \underset{(n+d) \times d}{\begin{bmatrix} X \\ \Gamma \end{bmatrix}}, \quad \hat{y} = \underset{(n+d) \times 1}{\begin{bmatrix} y \\ 0_d \end{bmatrix}}$$

from (I), if we differentiate w.r.t. w,

$$2(\hat{y} - \hat{X} w) \cdot \hat{X}^T = 0.$$

$$\hat{X}^T \hat{X} w = \hat{X}^T \hat{y}$$

$$w = (\hat{X}^T \hat{X})^{-1} \hat{X}^T \hat{y}$$

now, $$\left(\underset{d \times (n \times d)}{\begin{bmatrix} X^T & \Gamma^T \end{bmatrix}} \underset{(n \times d) \times d}{\begin{bmatrix} X \\ \Gamma \end{bmatrix}} \right)^{-1} \underset{d \times n \times d}{\begin{bmatrix} X^T & \Gamma^T \end{bmatrix}} \underset{n \times d \times 1}{\begin{bmatrix} y \\ 0_d \end{bmatrix}}$$

Performing blockwise matrix multiplication,

$$\hat{w} = (X^T X + \Gamma^T \Gamma)^{-1} [X^T y + \Gamma^T 0_d]$$

$$\hat{w} = (X^T X + \Sigma_i^{-1})^{-1} [X^T y + \bar{0}]$$

$$\hat{w} = (X^T X + \Sigma_i^{-1})^{-1} X^T y$$ - hence proven.

3. VECTOR CALCULUS

(a) show that $\frac{\partial}{\partial x} (\underset{1\times n}{X^T} \underset{n\times 1}{c}) = c^T$

$$= \frac{\partial}{\partial x} \left(\underset{1\times n}{\begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}} \underset{n\times 1}{\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{bmatrix}} \right)$$

$$= \frac{\partial}{\partial x} \underset{1\times 1}{\left[x_1 c_1 + x_2 c_2 + x_3 c_3 + \cdots x_n c_N \right]}$$

$\rightarrow$ Now, as $X \in \mathbb{R}^n$,

$$= \frac{\partial}{\partial x} \left[\frac{\partial}{\partial x_1} x_1 c_1 + x_2 c_2 + \cdots x_n c_N \quad \frac{\partial}{\partial x_2}(\quad) \cdots \frac{\partial}{\partial x_n}(\quad) \right]$$

$$= \begin{bmatrix} c_1 & c_2 & \cdots & c_N \end{bmatrix} = c^T$$ hence proven!

(b) $$\frac{\partial}{\partial x} \|x\|_2^2 = \frac{\partial}{\partial x} \langle \underset{1\times n}{x^T} \underset{n\times 1}{x} \rangle$$

$$= \frac{\partial}{\partial x} (x_1^2 + x_2^2 + \cdots x_n^2)$$

$$= [2x_1, 2x_2, \cdots 2x_n]$$

$$= 2X^T$$

hence proven.

(c) $\frac{\partial}{\partial x}(Ax) = A$

$$\frac{\partial}{\partial x}\underbrace{\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & A_{2N} \\ \vdots & & & \\ A_{N1} & A_{N2} & \cdots & A_{NN} \end{bmatrix}}_{n\times n}\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}}_{n\times 1} = \frac{\partial}{\partial x}\underbrace{\begin{bmatrix} A_{11}x_1 + A_{12}x_2 + \cdots A_{1N}x_N \\ A_{21}x_1 + A_{22}x_2 + \cdots A_{2N}x_N \\ \vdots \\ A_{N1}x_1 + A_{N2}x_2 + \cdots A_{NN}x_N \end{bmatrix}}_{n\times 1}$$

$$= \frac{\partial}{\partial x} Ax = \begin{bmatrix} \partial/\partial x_1 (\text{first row}) & \partial/\partial x_2 (\text{first row}) & \cdots \\ \partial/\partial x_2 (\text{first row}) & \cdots & \\ & & \cdots \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{N} \\ A_{21} & A_{22} & \cdots & A_{2N} \\ & \vdots & & \\ A_{N1} & A_{N2} & \cdots & A_{NN} \end{bmatrix} = A.$$

(d) $$\frac{\partial}{\partial x}(\underset{1\times n}{x^T}\underset{n\times n}{A}\underset{n\times 1}{x}) = x^T(A + A^T)$$

Using the definition of fundamental theorem:

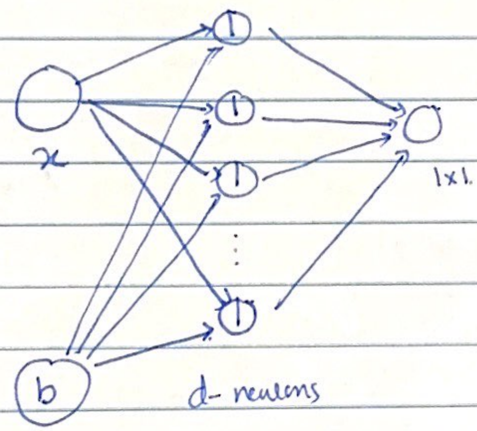
$$f(x+\Delta) = (x+\Delta)^T A (x+\Delta)$$
$$= x^T A x + \Delta^T A x + x^T A \Delta + \Delta^T A \Delta$$
$$= f(x) + (x^T A^T + x^T A)\Delta$$

which yields the derivative as $x^T A^T + x^T A$
$\Rightarrow (A^T + A)x^T$

(e) If $A^T A$ is symmetric.

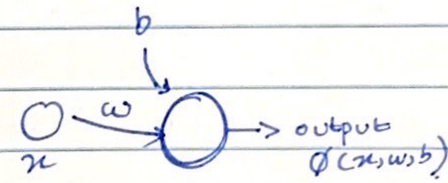
4. RELU Elbow Update Under SGD.

$$\underset{1\times1}{\hat{f}(x)} = \underset{1\times d}{W^{(2)}}\, \Phi\big(\underset{d\times1}{W^{(1)}}\,\underset{1\times1}{x} + \underset{d\times1}{b}\big)$$



$\}$ loss $\frac{1}{2}\|y-\hat{f}(x)\|_2^2$

(a) Starting with ONE RELU: $\rightarrow$



$\Downarrow$

$$\phi(x) = \begin{cases} wx+b & wx+b>0 \\ 0 & \text{o.w.} \end{cases}$$

loss $= \frac{1}{2}\|\phi(x)-y\|_2^2$

(i) Location of the **elbow** e of the function where it transitions from 0 to something else.

$$\phi(x) = 0 = wx+b$$

$$x = -\frac{b}{w}$$

(ii) $\frac{\partial l}{\partial \phi} \begin{cases} (\phi(x) - y) & \phi(x) > 0 \\ 0 & \text{o.w.} \end{cases}$

(iii) $l(x, w, b) = \frac{1}{2} \| \phi(x) - y \|^2.$

$$\frac{\partial l}{\partial w} = \frac{\partial l}{\partial \phi} \cdot \frac{\partial \sigma}{\partial w} = \begin{cases} (\phi(x) - y) \circ x & wx + b > 0 \\ 0 & \text{o.w.} \end{cases}$$

(iv) $\frac{\partial l}{\partial b} = \begin{cases} (\phi(x) - y) & wx + b > 0 \\ 0 & \text{o.w.} \end{cases}$

(b) $\phi(x) - y = 1$

(i) $\phi(x) = 0$

Gradient Descent:

$$w_t = w_{t-1} - \eta (\nabla_w l)$$

If $\phi(x) = 0 \Rightarrow \frac{\partial l}{\partial w} = 0 \Rightarrow$ Gradient remains unchanged.

So, slope remains unchanged and since $\frac{\partial l}{\partial \phi} = 0$ thus no change in elbow either.

7

(ii) $\omega > 0, \; x > 0, \; \phi(x) > 0.$

now, $\phi(x) \geqslant 0$ which means $\omega x + b > 0$

$$\frac{\partial l}{\partial \phi} = \phi(x) - y = 1$$

Gradient Descent $\Rightarrow$ $\omega_{new} = \omega_{old} - \eta \frac{\partial l}{\partial \omega}.$

$\omega_{new} = \omega_{old} - \eta \left((\phi(x) - y) \cdot x \right)$

since $x > 0, \; \omega > 0,$

$\omega_{new} = \omega_{old} - \eta (1) \cdot x$

when $\omega_{new} < \omega_{old},$

$\omega_{new} x + b$ would decrease

so slope will decrease!

~~eff~~ elbow $= \frac{-(b - \Delta b)}{(\omega - \Delta \omega)}$.

$= \frac{-\left(b - \frac{\partial l}{\partial b}\right)}{\omega - \frac{\partial l}{\partial \omega}}$

(iii) $\omega > 0, \; x < 0, \; \phi(x) > 0$

$$\frac{\partial l}{\partial \phi} = \phi(x) - y = 1$$

$\frac{\partial l}{\partial \omega} = (\phi(x) - y)(x) = 1(x) < 0$

so, $\omega_{new} = \omega_{old} - \eta(1)(x)$

$\omega_{old} < \omega_{new}.$

~~(wx+b~~)

(8)

$\omega_{new} x \cancel{+b} + b > \omega_{old} + b$

so, slope increases.

$$e' = \frac{-b + \eta \frac{\partial l}{\partial b}}{\omega - \eta \frac{\partial l}{\partial \omega}}$$ } elbow should move to left as error decreases.

(v) $\omega < 0, x > 0, \phi(x) > 0.$

$\omega_{new} = \omega_{old} - \eta(\phi(x) - y) \cdot x$

$\omega_{new} = \omega_{old} - \eta(1)(x)$

$\omega_{new} < \omega_{old}$

becomes more -ve.

so, slope gets smaller

elbow moves to left.

(c) $\hat{f}(x) = W^{(2)} \Phi(W^{(1)} x + b)$

$\text{Loss} = \frac{1}{2} \| \hat{f}(x) - y \|_2^2.$

~~$\text{Loss}_i = \frac{1}{2}\left[W_i^2 \Phi(W_i^1 x + b)\right]^2.$~~

~~$\frac{\partial \text{Loss}_i}{\partial W_i^2} = \frac{2 \Phi}{2}(W_i^1 x + b)^2$~~

~~$\frac{\partial \text{Loss}}{\partial \phi_i} = (W_i^1 x + b), \quad \frac{\partial \text{Loss}}{\partial W_i^1} = (W_i^1 x + b) x.$~~

elbow for i^{th} neuron would be

$$e_i = \frac{-b}{\omega_i^1} \qquad \omega_i^1 x + b = 0$$

$$x = \frac{-b}{\omega_i^1}$$

$$x = \frac{-b}{W_i^1}$$

(d) $e_i' \Rightarrow$ new elbow

$$= \frac{-(b - \Delta b)}{(W_i^1 - \Delta W_i^1)}$$

we need to know what is Δb and ΔW_i^1 after one SGD iteration.

$$\Delta b = \frac{\partial l}{\partial b} \qquad Loss = \frac{1}{2}\|W^2 \Phi(W^1 x + b) - y\|^2$$

$$\frac{\partial L}{\partial W^2} = (W^2 \Phi(W^1 x + b) - y) \cdot \Phi(Wx+b)$$

$$\frac{\partial L}{\partial W^1} = (W^2 \Phi(W^1 x + b) - y) \cdot x$$

$$\frac{\partial L}{\partial W_i^1} = W_i^2 [\Phi(W_i^1 x + b) - y] \cdot x.$$

$$\frac{\partial L}{\partial b} = W_i^2 [\Phi(W_i^1 x + b) - y]$$

$$e_i = \frac{-(b_i - \eta W_i^2 (\Phi(W_i^1 x + b) - y))}{W_i^1 - \eta W_i^2 x [\Phi(W_i^1 x + b) - y]}.$$

6 Homework Process & study Group.

(a) sources => eecs127_reader.pdf , EECS16B SVD Notes

(b) worked Individually.

(c) Number of hours => 5