Homework 0

1. $y = F w^* + \epsilon$

define $L(w) = \frac{1}{2} \| y - Fw \|^2$

using gradient descent $\frac{\partial L}{\partial w} = 2 \cdot \frac{1}{2} (y - Fw)^T \frac{\partial (y - Fw)}{\partial w}$

$= (y - Fw)^T (-F)$

therefore, $W_t = W_{t-1} - \eta \frac{\partial L}{\partial w_{t-1}}^T = W_{t-1} + \eta F^T (y - F W_{t-1})$

$I - \eta F^T F$ $\qquad = W_{t-1} - \eta F^T (F W_{t-1} - y)$

$F^T F$ is the inner correlation matrix of F

$F^T F = V \hat{\Sigma} \hat{U}^T \hat{U} \hat{\Sigma} V^T = V \hat{\Sigma}^2 V^T$

therefore $\phi = I - \eta F^T F = V I V^T - \eta V \hat{\Sigma}^2 V^T = V (I - \eta \hat{\Sigma}^2) V^T$

$\phi V = V (I - \eta \hat{\Sigma}^2) = VD$, which means the eigenvalue is $I - \eta \hat{\Sigma}^2$

given that η is a learning rate small enough, each item of $D > 0$

$F^T y = V \Sigma^T U^T y = [v_1 \cdots v_n] \begin{bmatrix} \alpha_1 u_1^T \\ \alpha_2 u_2^T \\ \vdots \\ \alpha_n u_n^T \end{bmatrix} y = [v_1 v_2 \cdots v_n] \begin{bmatrix} \alpha_1 u_1^T y \\ \vdots \\ \alpha_n u_n^T y \end{bmatrix} = \sum_{i=1}^{n} v_i \cdot \alpha_i u_i^T y$

therefore, $W_t = (I - \eta F^T F) W_{t-1} + \eta F^T y = V D V^T W_{t-1} + \eta \sum_{i=1}^{n} \alpha_i v_i u_i^T y$?

therefore $\|W_t\|_2 \leq \|W_{t-1}\|_2 + \eta \alpha \|y\|_2$

$\|W_t\|_2 \leq \|W_{t-1}\|_2 + \eta \alpha \|y\|_2$

$\|W_t\|_2 - \|W_{t-1}\|_2 \leq \eta \alpha \|y\|_2$

$\|W_k\|_2 - \|W_0\|_2 \leq k \eta \alpha \|y\|_2$

$\|W_k\|_2 \leq k \eta \alpha \|y\|_2$

therefore, $\| W_k - w^* \|_2 \leq \|W_k\|_2 + \|w^*\|_2 = k \eta \alpha \|y\|_2 + \|w^*\|_2$

Singular vector decomposition

$\sigma_1 \vec{u}_1 \vec{v}_1^T$ rank 1

$\sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_2 \vec{v}_2^T$ rank 2

that's right

$$\underset{m\times n}{X} = \underset{m\times m}{U} \underset{m\times n}{\Sigma} \underset{n\times n}{V^T} = \sum_{i=1}^{n} \sigma_i \vec{u}_i \vec{v}_i^T$$

economy form

$$\approx \tilde{U} \tilde{\Sigma} \tilde{V}^T = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T$$

truncated.

$$X^T X = V \hat{\Sigma} \hat{U}^T \hat{U} \hat{\Sigma} V^T = V \hat{\Sigma}^2 V^T$$

$$X^T X V = V \hat{\Sigma}^2$$

$$X X^T \hat{U} = \hat{U} \hat{\Sigma}^2$$

$Ax = \lambda x$

$AV = VD$

if a matrix is unitary and orthogonal

$$A = \hat{U} \hat{\Sigma} V^T$$

$$A^+ = V \hat{\Sigma}^{-1} \hat{U}^T$$

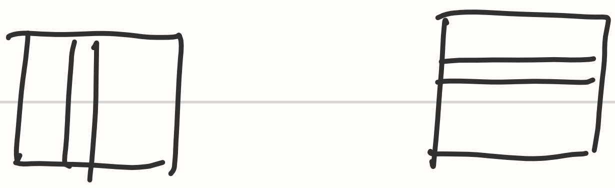
Notes about Matrix Differential

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

$$\frac{\partial(y^T A x)}{\partial y} = x^T A^T \qquad \text{proof is:}$$

$$\frac{\partial(x^T A x)}{\partial x} = \frac{\partial \sum \sum a_{ij} x_i x_j)}{\partial x}$$

$$\frac{\partial(x^T A x)}{\partial x_k} = \sum_{i=1}^{m} a_{ik} x_i + \sum_{j=1}^{n} a_{kj} x_j$$



$$\frac{\partial \alpha}{\partial x} = \frac{\partial(x^T A x)}{\partial x} = x^T A + (Ax)^T$$

$$= x^T A + x^T A^T = x^T (A + A^T)$$

$$\frac{\partial(y^T A x)}{\partial z} = y^T A \frac{\partial x}{\partial z} + x^T A^T \frac{\partial y}{\partial z}$$

$$\frac{\partial(A^{-1} A)}{\partial \alpha} = \frac{\partial(A^{-1})}{\partial \alpha} A + A^{-1} \frac{\partial(A)}{\partial \alpha} = 0$$

α ↓ Scalar

$$\frac{\partial(A^{-1})}{\partial \alpha} = -A^{-1} \frac{\partial A}{\partial \alpha} A^{-1}$$

2, Regularization from the Augumentation Perspective

given that for $y = X\beta$

$$\hat{\beta} = \underset{\beta}{\arg\min} \sum_{i=1}^{n} |y_i - \sum X_{ij}\beta_j|^2 = \underset{\beta}{\arg\min} \|y - X\beta\|_2^2 = (X^TX)^{-1}X^Ty$$

set $X := \hat{X} = \begin{bmatrix} X \\ \Gamma \end{bmatrix}$ $\quad y := \hat{y} = \begin{bmatrix} y \\ 0_d \end{bmatrix}$ $\quad w := \beta$

therefore. $\underset{w}{\arg\min} \|\hat{y} - \hat{X}w\|_2^2 = (\hat{X}^T\hat{X})^{-1}\hat{X}^T\hat{y}$

$$= (X^TX + \Gamma^T\Gamma)^{-1} Xy$$

$$= (X^TX + \Sigma^{-1})^{-1} Xy$$

3. Vector Caculus Review

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

(b) $\frac{\partial}{\partial x}\|x\|^2 = \frac{\partial}{\partial x}(x^T x) = x^T\frac{\partial x}{\partial x} + x^T\frac{\partial x}{\partial x} = 2x^T$

(c) $y_i = \sum_{k=1}^{n} a_{ik} x_k$

$\partial y_i / \partial x_j = a_{ij}$

$\partial Ax / \partial x = A$

(d) $\alpha = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij} x_i x_j$ $\quad \frac{\partial \alpha}{\partial x_k} = \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ik} x_i$

therefore $\frac{\partial \alpha}{\partial x} = x^T A^T + x^T A = x^T(A^T + A)$

(e) $A = A^T$

4. ReLU Elbow Update under SGD

(a)(i) $e = -b/w$

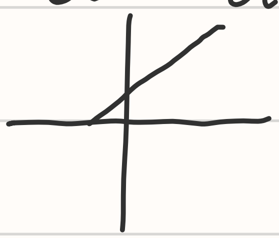
(ii) $\frac{dL}{d\phi} = (\phi(x) - y)$

(iii) $\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \phi}\frac{\partial \phi}{\partial w} = (\phi(x) - y)\cdot x$

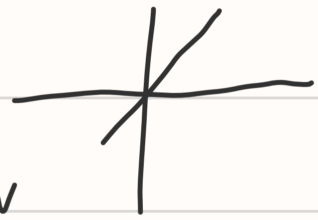
(iv) $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \phi}\frac{\partial \phi}{\partial b} = (\phi(x) - y)$

(b) $w_t = w_{t-1} - \eta\frac{\partial L}{\partial w_{t-1}} = w_{t-1} - \eta(\phi(x) - y)\frac{\partial \phi(x)}{\partial w_{t-1}} = w_{t-1} - \eta x_{t-1}$

$b_t = b_{t-1} - \eta\frac{\partial L}{\partial b_{t-1}} = b_{t-1} - \eta(\phi(x) - y)\frac{\partial \phi(x)}{\partial b_{t-1}} = b_{t-1} - \eta$

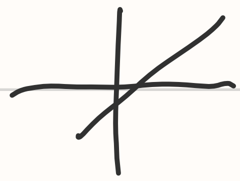


(i) $x_{t-1} = -w^{-1}b$ $\quad$ $w_t = w_{t-1} + \eta w^{-1}b > 0$

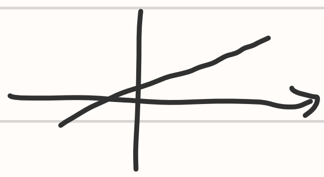


$b_{t-1} < b_t$ $\quad$ $elbow' = -(w_{t-1} + \eta w^{-1}b)^{-1}(b_{t-1} - \eta) > elbow$

(ii) $w_t > w_{t-1}$ $\quad$ $b_t < b_{t-1}$ $\quad$ $elbow' > elbow$



(iii) $w_t < w_{t-1}$ $\quad$ $b_t < b_{t-1}$ $\quad$ $elbow' > elbow$



$-(w + \eta w^{-1}b)w^{-1}b + b - \eta$

$= -b - \eta w^{-2}b^2 + b - \eta = -\eta(1+b^2) < 0$

(iv) $w_t < w_{t-1}$ $\quad$ $b_t < b_{t-1}$

$\phi(elbow) = -(w - \eta w^{-1}b)w^{-1}b + b - \eta = -\eta(1 + w^{-2}b^2)$ $\quad$ hard to decide.

(c) (d)

5. Using Pytorch to Learn the Color Organ

(a) find that R value is 200

(b) R value should be near 200

cutoff frequency should be 792 Hz

(c) lr	initial	final	time
200	300	200	3:51
200	200	200	0
200	100	198	3:01
200	1000	200	14:37

Lr reduce the learning time

(d) the leaned resistor value is 309 (wrong) and it varied to the lr or inital R

(e) the binary data is not proper for the previous Loss function

we can change the loss function by converting negtive samples from 0 to 0.3 which raise the mean expectation indirectly right drag the curve thus converging R to 196

(f) Lowpass: $H = \frac{1}{j\omega C} / (\frac{1}{j\omega C} + R) = 1/1+j\omega R$

$|H| = 1/\sqrt{1+w^2R^2}$

highpass: $H = R/(\frac{1}{j\omega C}+R) = j\omega CR / 1+j\omega RC$

$|H| = wCR / \sqrt{1+w^2C^2R^2}$

use the train data generated by new high Pass model and find that R is 34 Ω (I think correct)

(g) the result is that $R\_low = 44$

$R\_high = 159$

(h). for lowpass. $f_l = \frac{1}{2\pi RC} = 1000$

for highpass: $f_h = \frac{1}{2\pi RC} = 4000$

here, after iteration $f_l = 1039$. $f_h = 3816$

it basicly matches

(i) it takes no longer than calculating seperately

it takes longer time if the inital value is set far away

6. Homework Process and Study Group

(a) sources: Stackoverflow. wikipedia

(b) myself but exchange opinions with classmates on discussion after finishing them

(c) about 16 hours