COL865
Deep Learning
August 14, 2017

Last class:-

(a) Backpropagation (b) Regularization (means to avoid overfitting and/or improve generalization error).

L1 L2

Some linear Algebra:-

Let $A \in \mathbb{R}^{n \times n}$ symmetric.

(a) An eigenvector $v \in \mathbb{R}^n$ of A is defined as.

$$Av = v\lambda$$

Right eigenvector

eigenvalue

A square matrix A ($\mathbb{R}^{n \times n}$) has n such eigenvalues.

(b) A (real & symmetric) matrix can be decomposed as:-

$$A = Q \Lambda Q^T$$

where:-

Q:- orthonormal / orthogonal matrix composed of eigenvectors of A $\Rightarrow Q_{:i}$ → ith column of Q & columns of Q

Λ:- Diagonal matrix s.t. Λ_{ii} :- ith eigenvalue of A (Corresponding to corresponding column of Q).

$Q_i \cdot Q_j = 0 \quad i \neq j$

Similarly $Q_{i:} \cdot Q_{j:} = 0 \quad i \neq j$

} Rows & columns of Q are ⊥ to each other

Further, $\|Q_{:i}\|^2 = 1$

$\|Q_{i:}\|^2 = 1$

Thm:- columns of Q

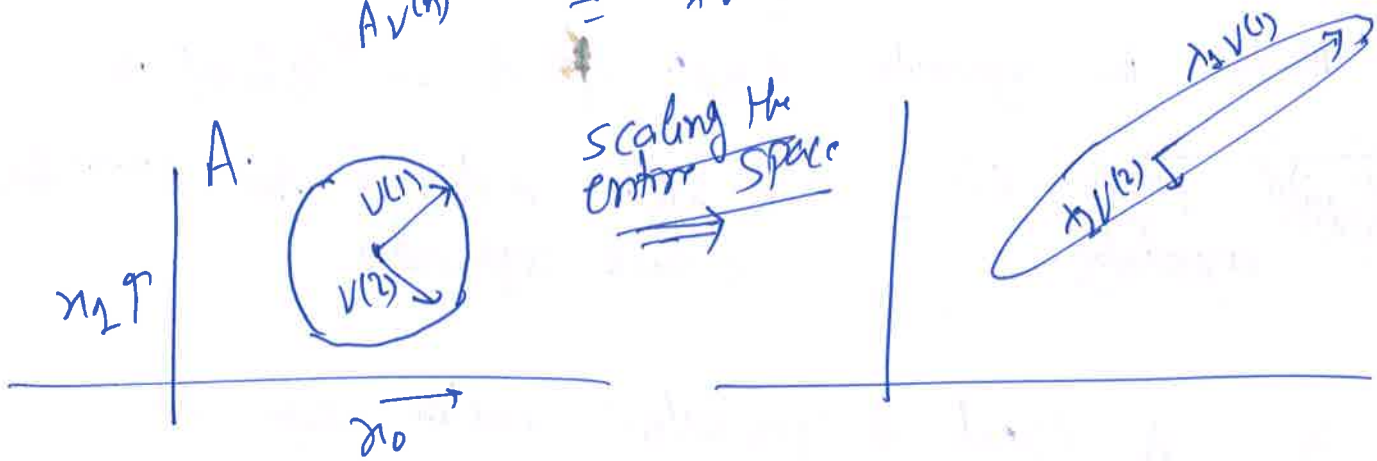
Let $v^{(1)}$ be an eigenvector of A.
$v^{(2)}$ be an eigenvector of A.

Property $Q^T = Q^{-1}$

$$\Rightarrow Av^{(1)} = \lambda v^{(1)}$$
$$\Rightarrow Av^{(2)} = \lambda v^{(2)}$$
$$\vdots$$
$$Av^{(n)} = \lambda v^{(n)}$$

But $v^{(1)}, v^{(2)} \ldots v^{(n)}$ span the entire space.

Q:- orthonormal



Now, let us revisit L2-Norm :-

~~Assumption :- No~~

Question:- $\Omega(\theta) = \frac{1}{2} w^T w$ $\equiv$ (L2-Norm)^2

Why to ignore bias parameter. Not very useful to regularize

(w requires large amount of data)

$y = w^T x + b$ $\quad z = wx$

$z = w^T x + b$ → Not difficult to learn (less data required) → Does not capture correlation among how to ~~how~~ variables interact

of only H-1 z (output)

②

Model No bias :- Only for exposition

① Quadratic approximation to a function :-

approximation $\leftarrow f(w) = f(w^*) + f'(w^*)(w - w^*) + f''(w^*)\frac{(w-w^*)^2}{2}$ $\quad x \simeq x^*$, $f: R \to R$

Now, suppose :- $f: R^n \to R$

$f'(w) = f(w^*) + \nabla_w f(w^*) + \frac{1}{2}(w-w^*)^T H (w-w^*)$

H - Hessian evaluated at w^*

$\Rightarrow J'(w) =$

∴ Consider $J(w)$:- error function

Let w^* be a minima of $J(w)$

(□) Let w be a point close to w^*.

Note :- ① $\nabla_w J(w^*) = 0$

② H is +ve semi-definite (symmetric)

(at w^*) :- since function should concave at w^*

$\Rightarrow J'(w^*) = J(w^*) + \nabla_w J(w^*) + \frac{(w-w^*)^T H (w-w^*)}{2}$

$\nabla_w J(w^*) = 0$

$-2(w) = \frac{1}{2} w^T w$

~~We want to minimi~~

Further :- $\tilde{J}(w^*) = J(w) + \alpha \cdot \frac{1}{2} w^T w$

$\Rightarrow \tilde{J}'(w) = J'(w) + \frac{\alpha}{2} w^T w$

quadratic Approx.

Now. we want to minimize $\tilde{J}(w)$

$\Rightarrow$ minimize $\tilde{J}'(w)$

$\Rightarrow \nabla_w \tilde{J}'(w) = 0$

$\Rightarrow \nabla_w \tilde{J}(w) + \frac{2\alpha}{2} w = 0$

$\Rightarrow$ =

$(\tilde{w} - w^*) H$

$$H(\tilde{w} - w^*) + \alpha \tilde{w} = 0$$

$\nabla_w w^T H w = 2Hw$

$\nabla_w w^T H w = 2Hw$

$\Rightarrow H\tilde{w} + \alpha I \tilde{w} = Hw^*$

$\Rightarrow (H + \alpha I)\tilde{w} = Hw^*$

$\Rightarrow \tilde{w} = (H + \alpha I)^{-1} Hw^*$

Now:- H is symmetric-

$H = Q \Lambda Q^T$ → Orthonormal

$\Rightarrow Q = (Q \Lambda Q^T + \alpha Q I Q^T)^{-1} Q \Lambda Q^T w^*$

If H is diagonal-

$= [Q (\Lambda + \alpha I) Q^T]^{-1} Q \Lambda Q^T w^*$

$= [Q (\Lambda + \alpha I)^{-1} \Lambda Q^T] w^*$

H'

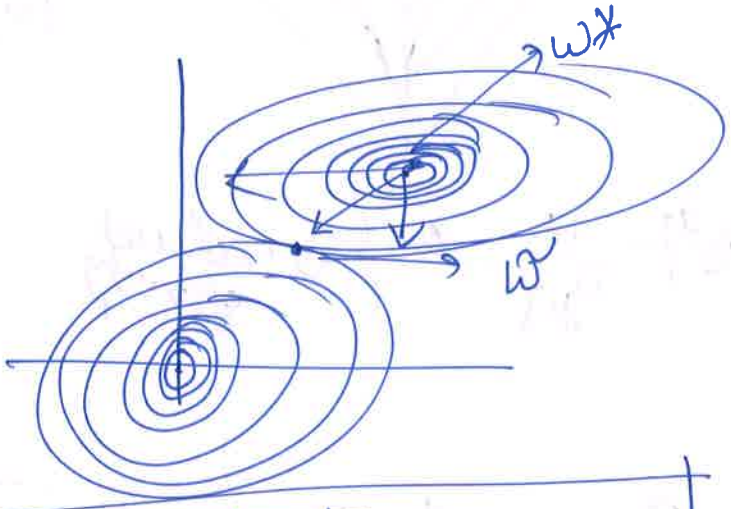
with eigenvector same as H & eigenvalues $\frac{\lambda_i}{\lambda_i + \alpha}$ if λ_i is an eigenvalue of H.

⑧

Now Re-scales the w space by $\frac{\lambda_i}{\lambda_i + \alpha}$ along each dimension corresponding to eigenvector of H

$\Rightarrow \tilde{w}$ is w^* re-scaled on the re-scaled space.

Value of w^* is shrunk by a factor of $\frac{\lambda_i}{\lambda_i + \alpha}$ in each direction. → constant



if λ_i is large
less shrinking.

λ_i is large :- High second order derivative.

High curvature
⇒ less shrinkage.

$J(w)$ changes very fast in direction corresponding to λ_i. can't afford to have ~~large regularizer~~ much regularization.

L2- Regularization.-

Consider quadratic approximation as before.

$$J^q(w) = J(w^*) + (w-w^*)^T \nabla_w J(w^*) + \frac{(w-w^*)^T H (w-w^*)}{2}$$

$(w-w^*)\nabla_w J(w^*) \downarrow 0$

$$-2(w) = \alpha \|w\|_2$$

$\downarrow$

$\updownarrow w_j$

$\downarrow$

Regularized

$$\tilde{J}^q(w) = J^q(w) + \|w\|_2$$

$$\nabla_w \tilde{J}^q(w) = \nabla_w J^q(w) + \nabla_w \|w\|_2$$

$$\nabla_w \tilde{J}^q(w) = (w_j - w_j^*) H_{jj} + \alpha \operatorname{sign}(w_j)$$

Equating gradient to zero, we get:-

$$(\tilde{w} + w^*) H = \alpha \nabla_w \|w\|_1$$

Assume H is diagonal
Taking componentwise derivative.

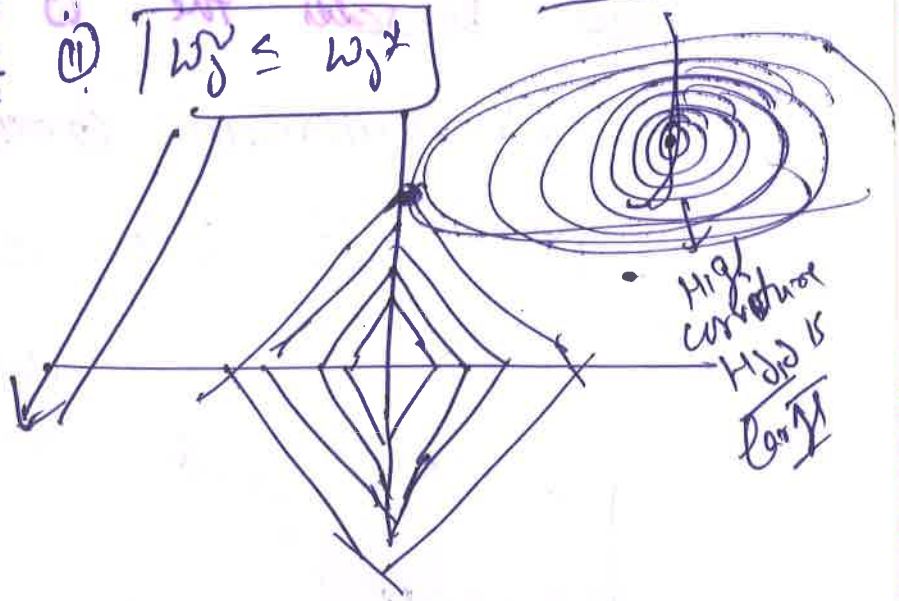
$$[-\tilde{w}_j + w_j^*] H_{jj} = \alpha \cdot \operatorname{sign}(\tilde{w}_j) \quad H_{jj} \geq 0$$

$$-\tilde{w}_j + w_j^* = \frac{\alpha}{H_{jj}} \operatorname{sign}(\tilde{w}_j)$$

then first

$|\tilde{w}_j| \le |w_j^*|$:- not ① $\text{sign}(\tilde{w}_j) = \text{sign}(w_j^*)$

② $\tilde{w}_j \le w_j^*$



if $w_j^* \ge 0$

$\Rightarrow 0 \le w_j \le w_j^*$

if $w_j^* \le 0$

$\Rightarrow w_j^* \le w_j \le 0$

~~Further~~ ~~$w_j^* \ge 0$~~ :-

$$w_j = \text{sign}(w_j^*) \max\left\{ |w_j^*| - \frac{\alpha}{H_{jj}}, 0 \right\} \Rightarrow \text{induces sparsity}$$

$w_j^* \ge 0$:-

if $w_j^* > \frac{\alpha}{H_{jj}} \Rightarrow \tilde{w}_j = w_j^* - \frac{\alpha}{H_{jj}}$ Non-zero

else $\tilde{w}_j = 0$

$\Rightarrow$ L1-Norm induces sparsity.

$\Rightarrow$ Feature selection mechanism

as opposed to:-

$\tilde{w}_j = w_j^*$

LASSO:- Linear Model + Least Squares objective + L1-Penalty

Least Absolute Shrinkage & selection Operator

(4)

Other Variations:- 7.2 - 7.8 Read

Constrained optimization:-

① $\Omega(\theta) \le k$ Enforce Explicitly

$w^T w \le k$ rather than having a reg/ penalty

Solve using method of Lagrangians. (Dual)

or solving using gradient descent along with projection of ② $w^T w = k$ (L2 Ball) with radius $\sqrt{k}$.

$\underset{\theta}{\operatorname{argmin}}\ J(\theta)$

$\Omega(\theta) \le k$ How to solve it?

Under Constrained Problems

X:- Design matrix

$X^T X$:- ~~Dec~~

often need to invert $X^T X$.

Moore-Penrose pseudoinverse

Invert:- $(X^T X + \alpha I)$ $(X^T X + \alpha I)^{-1} X$

Linear Regression-

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

$(X^T X + \alpha I)^{-1} X y$ ↓ equivalent to L2 regularization

$J(w) = \frac{1}{2}(X^T w - y)^T (X^T w - y)$

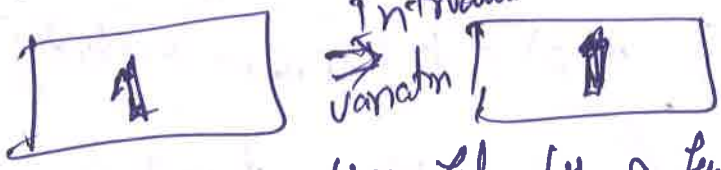
Linear Regression with quadratic loss

$\tilde{J}(\theta) = J(\theta) + \frac{\alpha w^T w}{2}$

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

$w = (X^T X + \alpha I)^{-1} X y$

Dataset Augmentation:-

(1) 

(translate by a few pixels)

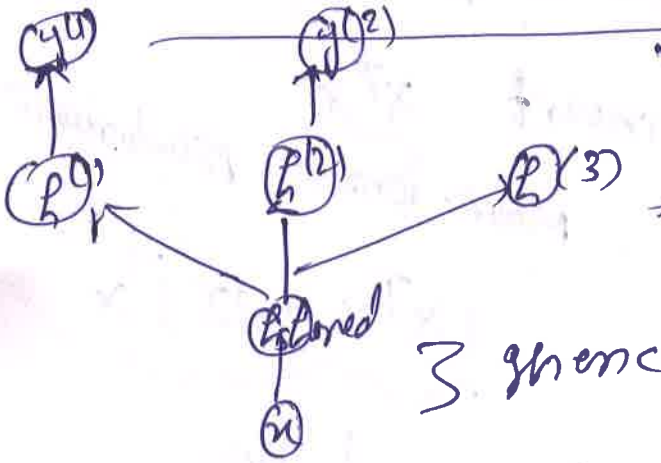
(2) Explicitly introduce noise. (random noise)
↳ can apply to hidden units as well

Noise Introduction in weights :-

$w' \rightarrow w + \epsilon_w$ for each example we randomly perturbed weights

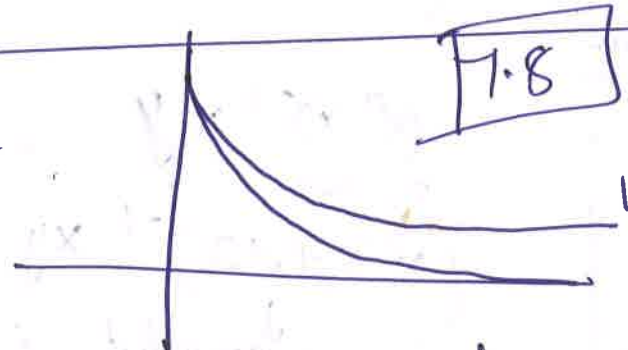
$\epsilon_w \sim N(\epsilon; 0, \eta I)$

↓ stochasticity



Multi-task learning | Ensembles
(1) Ensembles
(2) Parameter Tying

Early Stopping:-



$\alpha = \frac{1}{\epsilon_h}$ ⇒ learning rate

$\alpha = \frac{1}{\epsilon_h}$

iteration