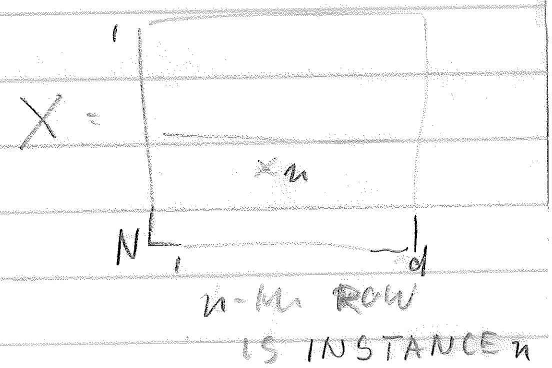
Lecture 7 242 F17

LINEAR LEAST SQUARES (LLS)

- FINDING SOL. FOR LINEAR REGRESSION (BATCH SETTING) BY SOLVING THE QUADRADIC EQUATIONS

$$\min_{w} \left(\underbrace{\sum_{n=1}^{N} (\underset{d\times 1}{\bar{x}_t} \cdot \underset{d\times 1}{\bar{w}} - \underset{1\times 1}{y_t})^2}_{\| \underset{T\times n}{X} \underset{n\times 1}{\bar{w}} - \underset{T\times 1}{\bar{y}} \|^2} \right)$$



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

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

$$= \bar{w}^T X^T X \bar{w} - \bar{w}^T X \bar{y} - \bar{y}^T X \bar{w} + \bar{y}^T \bar{y}$$

$$= \bar{w}^T X^T X \bar{w} - 2\bar{y}^T X \bar{w} + \bar{y}^T \bar{y} \qquad \text{QUADRADIC}$$

$$\nabla_w \| X\bar{w} - \bar{y} \|^2 = 2X^T X \bar{w} - 2X^T \bar{y} = 0 \qquad (*)$$

$$X^T X \vec{w} = X^T \bar{y} \qquad \text{NORMAL EQUATIONS}$$

$Xw = y$ MIGHT NOT HAVE SOLUTION

BUT NORMAL EQATIONS ALWAYS HAVE SOLUTION (Why?)

$$X^T X \bar{w}^* = X^T y$$

$$\underset{d\times 1}{\bar{w}^*} = \underset{d\times N\, N d}{(X^T X)^{-1}} \underset{d\times N\, N\times 1}{X^T y}$$

PROBLEM: $X^T X$ MIGHT NOT HAVE FULL RANK

$$r(X^T X) = r(X)$$

FIX 1: REGULARIZE

$$\inf_w \left(\lambda \|w\|^2 + \|X\bar{w} - \bar{y}\|^2 \right)$$ ← CALLED RIDGE REGRESSION

$$\nabla_w \left(\lambda \|w\|^2 + \|X\bar{w} - \bar{y}\|^2 \right)$$

$$= 2\lambda \bar{w} + 2X^T X \bar{w} - 2X^T \bar{y}$$

$$= 2\left((\lambda I + X^T X)\bar{w} - X^T \bar{y} \right)$$

$$= 0$$

$$\bar{w}^* = \underbrace{(\lambda I + X^T X)}_{\text{ALWAYS FULL RANK}}{}^{-1} X^T y$$

FIX 2: PSEUDO INVERSE

DIGRESSION

— MATRIX DECOMPOSITION:

A SYMMETRIC IF $A = A^T$

U ORTHOGONAL IF $UU^T = U^TU = I$

$\Big\}$ A, U SQUARE

$\forall$ SYM. $\underset{M \times M}{A}$: $\exists$ ORTH. MATR. U & REAL DIAG MATR. σ s.t.

$$A = U \sigma U^T$$

$$= \boxed{\quad U \quad} \underbrace{\begin{pmatrix} \ddots & 0 \\ 0 & \ddots \end{pmatrix}}_{\sigma} \boxed{\quad U \quad}^T$$

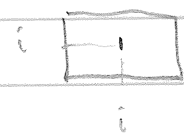
THE n COLUMNS u_i OF U
ARE THE EIGENVECTORS OF A
AND THE DIAGONAL ELEMENTS OF
σ THE EIGENVALS

u_i's ARE ORTHOGONAL:

$$u_i \cdot u_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \quad = \end{cases}$$

$U \sigma U^T$

$$= U \sum_i \sigma_i \underbrace{e_i e_i^T} U^T$$



$$= \sum_i \sigma_i \underset{M\times M}{U} \underset{M\times 1}{e_i} \underset{1\times M}{e_i^T} \underset{M\times M}{U^T}$$

$$= \sum_i \sigma_i u_i u_i^T$$

u_j IS EIGENVEC j

$$\underbrace{\left(\sum_i \sigma_i u_i u_i^T\right)}_{A} u_j = \sum_i \sigma_i u_i (u_i^T u_j)$$

$$= \sigma_j u_j \underbrace{u_j^T u_j}_{1}$$

$$= \sigma_j u_j$$

$\uparrow$

EIGENVALUE

WHAT IF A non-square ?

$$\underset{M \times N}{\boxed{A}} = \underset{M \times M}{\boxed{U}} \; \underset{M \times N}{\overset{\sigma}{\boxed{\begin{matrix} \ddots & 0 & 0 \\ 0 & & \end{matrix}}}} \; \underset{N \times N}{\boxed{V^T}}$$

$UU^T = I_{M \times M}$

NON-NEG. REAL DIAGONAL MATRIX

$VV^T = I_{N \times N}$

- DIAGONAL ENTRIES OF σ ARE NON-NEGATIVE
- CALLED SINGULAR VALUES

$$= \sum_{i=1}^{\min(M,N)} \sigma_i \; u_i v_i^T$$

MATLAB: `[U, D] = eig(A)`

`[U, D, V] = svd(A)`

RETURN TO FIX 2:

$$\underset{d\times N}{X^T}\,\underset{N\times d}{X}\,\underset{d\times 1}{\bar{w}^*} = \underset{d\times N}{X^T}\,\underset{N\times 1}{\bar{y}}$$

SOLUTION: $w^* = X^+ \bar{y}$

↑ PSEUDO INVERSE OF X

If $X = U S V^T$

$X^+ := V S^+ U^T$

↑ INVERSE OF NON-ZERO ELMT'S OF DIAG

LEAVE 0'S AS 0'S

$$\begin{pmatrix} 6 & & & \\ & 2 & & \\ & & 0 & \boxed{} \end{pmatrix}^+ = \begin{pmatrix} \frac{1}{6} & & \\ & \frac{1}{2} & \\ & & 0 \\ & \boxed{} & \end{pmatrix}$$

CHECK NORMAL EQUATIONS!

$$\underbrace{V S^T U^T}_{X^T}\,\underbrace{U S V^T}_{X}\,\underbrace{V S^+ U^T \bar{y}}_{w^*}$$

$$= \underset{d\times d}{V}\ \underset{d\times N}{S^T}\ \underset{N\times d}{S}\ \underset{d\times N}{S^+}\ \underset{N\times N}{U^T}\ \underset{N\times 1}{\bar{y}}$$

$$= \underbrace{V\ \underset{d\times N}{S^T}\ U^T}_{X^T}\ \bar{y}$$

CONNECTION BETWEEN

FIX 1 & FIX 2

$$\lim_{\lambda \to 0} (X^T X + \lambda I)^{-1} X^T = X^+$$

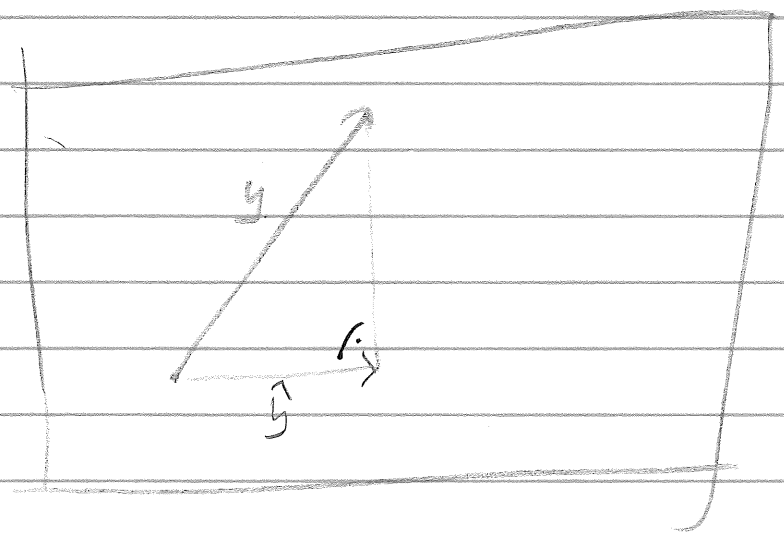
GEOMETRIC VIEW

$$\min_{w} \| \underbrace{Xw}_{\hat{y}} - y \|_2^2$$

$\hat{y} \in$ COLUMN SPACE OF X

$C(X)$

OPT $\hat{y}$ IS CLOSEST VECTOR TO Y THAT LIES IN $C(X)$



C(X)

$\hat{y}$ is PROJECTION OF y ONTO $C(X)$

$$\hat{y} = Py = \underbrace{X^+X^T}_{\text{PROJ. MATRIX}} y$$

$$X = U\sigma V'$$

$$X^+X^T = U\sigma^+ \cancel{V'V} \sigma U' = \tilde{U}\tilde{U}^T = \sum_{i:\lambda_i \neq 0} u_i u_i^T$$

ALL COL. WITH NON-ZERO EIGEN VAL

BREGMAN PROJ WRT $\| \|_2^2$

Recall GD update for minimizing batch loss

$$L(w) = \sum_n L_{y_n}(w \cdot x_n)$$

$$w_{q+1} = w_q - \eta \sum_n L'_{y_n}(w_q \cdot x_n)\, x_n$$

Motivation:

$$w_{q+1} = \operatorname{argmin}_w \left(\frac{1}{2} \| w - w_q \|^2 + \eta \sum_n L_{y_n}(w \cdot x_n) \right)$$

$$\nabla \text{''} \Big|_{w = w_{q+1}} = 0$$

$$w_{q+1} - w_q + \eta \sum_n L'_{y_n}(w_{q+1} \cdot x_n)\, x_n$$

$$w_{q+1} = w_q - \eta \sum_n L'_{y_n}(w_{q+1} \cdot x_n)\, x_n$$

IMPLICIT UPDATE

$$\approx w_q - \eta \sum_n L'_{y_n}(w_q \cdot x_n)\, x_n$$

APPROXIMATION

EXPLICIT UPDATE

IMPLICIT USES GRADIENT OF LOSS AT CURRENT ITERATION

EXPLICIT " LAST "

PRECISE WAY TO MOTIVATE EXPLICIT UPDATE:

$$w_{t+1} = \arg\inf_{w} \underbrace{\|w - w_t\|^2}_{\substack{\text{INERTIA TERM} \\ \text{NEEDED BECAUSE} \\ \text{LOSS LINEAR}}} + \eta \underbrace{\left(L(w_t) + (w - w_t)^T \Delta L(w_t) \right)}_{\substack{\text{FIRST ORDER APPROX.} \\ \text{OF BATCH LOSS} \\ \text{AT } w_t}}$$

BATCH GD VS STOCHASTIC GD BASED ON SINGLE EXAMPLE

TIME 1 ITERATION ≈ 1 PASS

converges faster because it uses "more recent" gradients

IDEA: WHY NOT USE 2ND ORDER APPROXIM. OF LOSS

$$L(w) \approx L(w_q) + \underset{1\times d}{(w-w_q)^T} \underset{d\times 1}{\nabla L(w_q)}$$

$$+ \frac{1}{2}(w-w_q)^T \underset{d\times d}{\nabla^2 L(w_q)} (w-w_q)$$

quadratic approx. of $L(w)$ at w_q

$\hat{L}_q(w)$

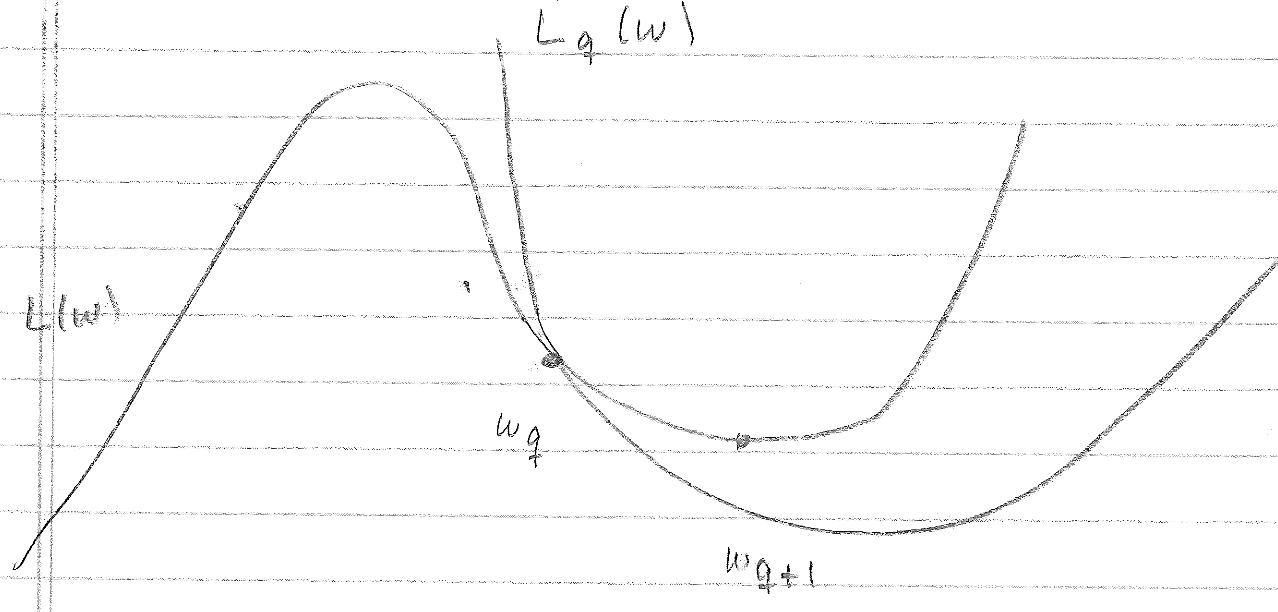
$$w_{q+1} = \underset{w}{\operatorname{argmin}} \; \hat{L}_q(w) \qquad (1)$$

WITH INERTIA TERM:

$$w_{q+1} = \underset{w}{\operatorname{argmin}} \; \|w - w_q\|^2 + \eta \hat{L}_q(w) \qquad (2)$$

(1) SIMPLER

FOCUS ON (1) FIRST



$$\underset{d\times 1}{\nabla \hat{L}_q(w)} = \underset{d\times 1}{\nabla L(w_q)} + \underset{d\times d}{\nabla^2 L(w_q)} \underset{d\times 1}{(w - w_q)}$$

$$\nabla \hat{L}(w)\Big|_{w = w_{q+1}} = 0$$

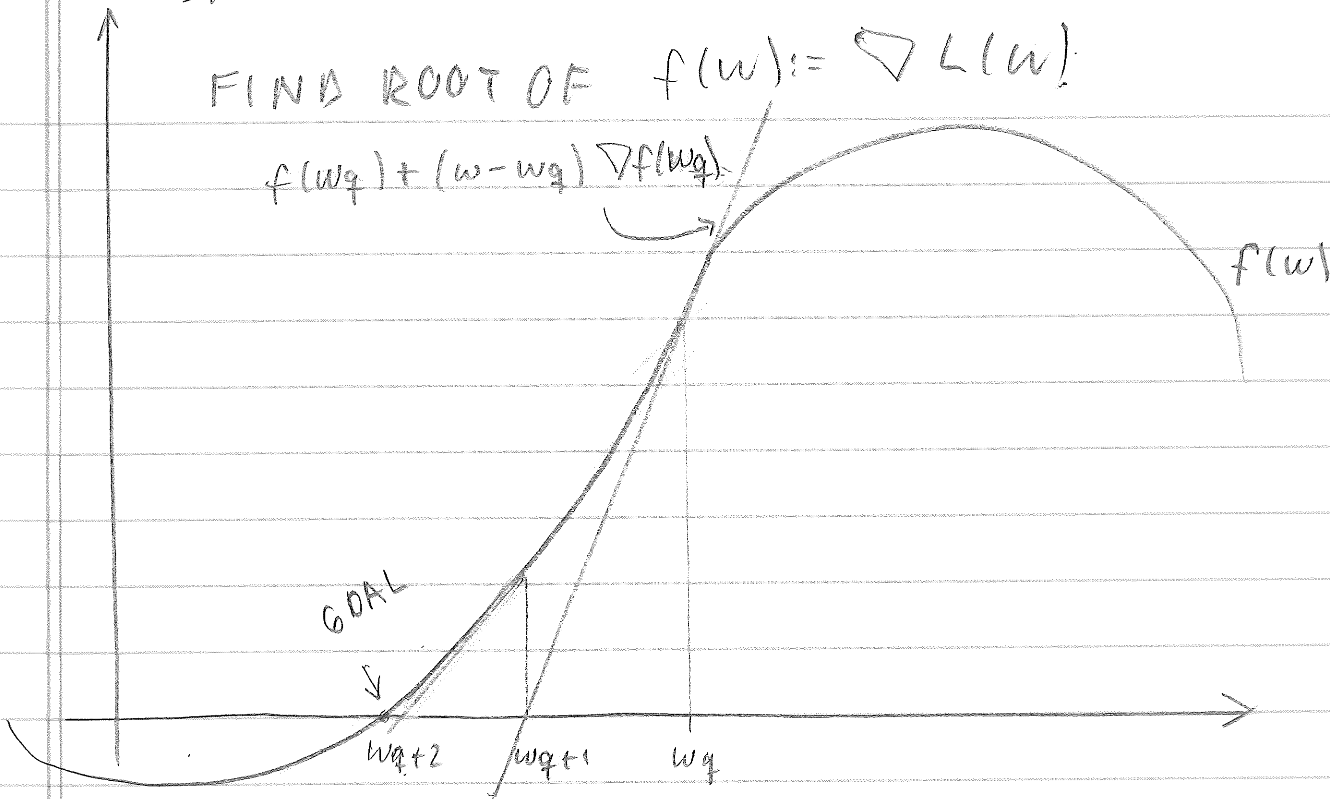
$$\nabla L(w_q) + \nabla^2 L(w_q)(w_{q+1} - w_q) = 0$$

$$w_{q+1} - w_q = -\left(\nabla^2 L(w_q)\right)^{-1} \nabla L(w_q)$$

$$w_{q+1} = w_q - \left(\nabla^2 L(w_q)\right)^{-1} \nabla L(w_q)$$

SECOND VIEW OF NEWTON

FIND ROOT OF $f(w) := \nabla L(w)$



$$\left. f(w_q) + (w - w_q)\nabla f(w_q) \right|_{w = w_{q+1}} = 0$$

$$w_{q+1} - w_q = -\left(\nabla f(w_q)\right)^{-1} f(w_q)$$

$$\underset{n \times 1}{w_{q+1}} = \underset{n \times 1}{w_q} - \underset{n \times n}{\left(\nabla f(w_q)\right)^{-1}} \underset{n \times 1}{f(w_q)}$$

NEWTON W. REGULARIZATION

$$w_{q+1} = \min_{w} \left(\frac{\lambda}{2} \|w - w_q\|^2 + L(w_q) + (w - w_q)^T \nabla L(w_q) + \frac{1}{2}(w - w_q)^T \nabla^2 L(w_q)(w - w_q) \right)$$

$$\nabla '' = \lambda (w - w_q) + \nabla L(w_q) + \nabla^2 L(w_q)(w - w_q)$$

$$\nabla '' \Big|_{w = w_{q+1}} = 0$$

$$\nabla L(w_q) + (\nabla^2 L(w_q) + \lambda I)(w_{q+1} - w_q) = 0$$

$$w_{q+1} = w_q - (\nabla^2 L(w_q) + \lambda I)^{-1} \nabla L(w_q)$$

LLS = 1 STEP OF NEWTON

RIDGE REGRESSION
= 1 STEP OF REG. NEWTON

NEWTON-RAPSON FOR LOGISTIC REGRESSION

BATCH LOSS $\quad L(w) = \sum_n L_{y_n}(w \cdot x_n)$

WHERE $\quad L_y(a) = \ln(1+e^a) - ya$

$$\nabla L(w) = X^T(\hat{y} - y) = \sum_n x_n(\hat{y}_n - y_n)$$

$$\begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_N \end{pmatrix} \quad \begin{pmatrix} y_1 \\ \\ y_N \end{pmatrix}$$

X HAS EXAMPLE AS ROWS
X^T " " COLS

$$\nabla_w \; x(\sigma(\underbrace{w \cdot x}_{a}) - y)$$

$$= x x' \sigma'(a)$$
$$= x x' \sigma(a)(1-\sigma(a))$$

$$\sigma(a) = \frac{e^a}{1+e^a} \qquad \sigma'(a) = \frac{e^a(1+e^a) - e^{2a}}{(1+e^a)^2}$$

$$= \sigma(a)(1-\sigma(a))$$

$$\nabla \nabla L(w)$$

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

$$= \nabla \sum_n x_n (\hat{y}_n - y_n)$$

$$= \sum_n x_n x_n^T \, \sigma(w x_n)(1 - \sigma(w x_n))$$

$$= X^T R X$$

$$\uparrow$$

$$\begin{pmatrix} \cancel{\hat{y}_n (1 - \hat{y}_n)} \end{pmatrix}$$

LINEAR REGR.

$$\nabla^2 L(w) = X^T X$$

R depends on weight vector

NEWTON RAPSON

$$w_{q+1} = w_q - (X^T R_q X)^{-1} X^T (\hat{y}_q - y)$$

$$= (X^T R_q X)^{-1} (X^T R_q X w_q - X^T (\hat{y}_q - y))$$

$$= (X^T R_q X)^{-1} X^T R_q z_q$$

WHERE $z_q = X w_q - R_q^{-1} (\hat{y}_q - y)$

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

→ CALLED ITERATED REWEIGHTED LEAST SQUARES

MANY NUMERICAL ISSUES

EXPLICIT → GD USES OLD GRADIENTS

NEWTON " " + HESSIANS

IMPLICIT GD USES CURRENT GRADIENTS

Conjecture: IMPLICIT GD BEATS NEWTON