# Lecture 4

- weighted least squares:

$$\text{Loss}(\beta) = \frac{1}{2}\sum_{i=1}^{n} w_i (y_i - s_i)^2 = \frac{1}{2}\sum_{i=1}^{n} w_i \big(\overset{\text{Continuous}}{y_i} - \overbrace{x_i^T\beta}^{s_i}\big)^2$$

- Logistic regression:

$$p(y|x) = \frac{e^{ys}}{Z} = \frac{e^{ys}}{1+e^{s}} \qquad y \in \{0,1\}$$

Approximate iteratively with quadratic terms

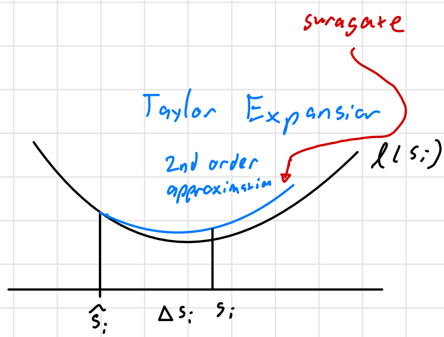
$$\text{Loglikelihood}(\beta) = \sum_{i=1}^{n} \Big[\underbrace{y_i s_i - \log(1+e^{s_i})}_{l(s_i)}\Big]$$

no analytical solution to maximize

- Iterative Reweighted Least Squares:

$$\begin{tikzcd}
\beta_t \arrow[r, "\Delta\beta"] \arrow[d, "x_i^T"'] & \beta_{t+1} \arrow[d] \\
\hat{s}_i \arrow[r, "\Delta s_i = x_i^T \Delta\beta"] & s_i = \hat{s}_i + \Delta s_i
\end{tikzcd}$$

$$l(s_i) \doteq l(\hat{s}_i) + l'(\hat{s}_i)\Delta s_i + \frac{1}{2} l''(\hat{s}_i)\Delta s_i^2$$




$$l'(s_i) = y_i - p_i = e_i$$

$$l''(s_i) = -p_i(1-p_i) = -w_i$$

$$l(s_i) \doteq l(\hat{s}_i) + \underbrace{l'(\hat{s}_i)}_{\hat{e}_i}\Delta s_i + \frac{1}{2}\underbrace{l''(\hat{s}_i)}_{-\hat{w}_i}\Delta s_i^2$$

$$\doteq \text{const} + \hat{e}_i \Delta s_i - \frac{1}{2}\hat{w}_i \Delta s_i^2$$

$$-l(s_i) = \frac{1}{2}\hat{w}_i\left(\Delta s_i^2 - 2\frac{\hat{e}_i}{\hat{w}_i}\Delta s_i\right) + \text{const}$$

$$-\ell(s_i) = \frac{1}{2}\hat{w}_i\left(\Delta s_i^2 - 2\frac{\hat{e}_i}{\hat{w}_i}\Delta s_i\right) + \text{const}$$

$$-\ell(s_i) \overset{\text{square form}}{=} \frac{1}{2}\hat{w}_i\left(\Delta s_i - \underbrace{\frac{\hat{e}_i}{\hat{w}_i}}_{\hat{y}_i}\right)^2 + \text{const}$$

$$= \frac{1}{2}\hat{w}_i\left(\hat{y}_i - x_i^T \Delta\beta\right)^2$$

So $\min\limits_{\Delta\beta} -\sum_{i=1}^{n} \ell(s_i)$

$$\Rightarrow \Delta\beta = \left(\sum \hat{w}_i x_i x_i^T\right)^{-1}\left(\sum_{i=1}^{n} \hat{w}_i x_i \hat{y}_i\right)$$

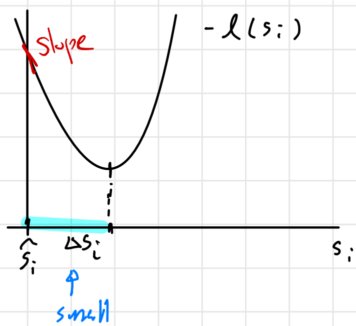
- This is a special case of the Newton-Raphson method.

- Note $w_i$ measures the curvature since

$$\ell''(s_i) = -p_i(1-p_i) = -w_i$$

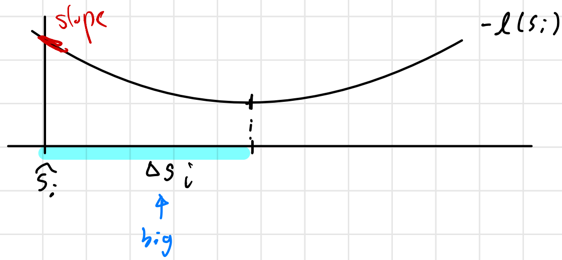
and is maximized when $p_i = \frac{1}{2}$ (uncertain examples)

- $w_i$ big $\Rightarrow$




golf last shot
care about accuracy

- $w_i$ small $\Rightarrow$




golf first shot

- Review of Multivariate Calculus:

$$\underbrace{y}_{m\times 1} = F(\underbrace{x}_{n\times 1})$$

$$F'(x) = \left(\frac{\partial y}{\partial \underbrace{x^T}_{1\times n}}\right)_{m\times n} = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ m \end{array} \left(\begin{array}{c} \\ \\ \partial y_i \\ \\ \\ \end{array}\right)_{m\times 1} \begin{array}{c} 1 \cdots j \cdots n \\ \left(\quad \frac{1}{\partial x_j} \quad\right)_{1\times n} \end{array}$$

$$= \begin{array}{c} \\ i \\ m \end{array} \begin{array}{c} j \quad\quad n \\ \left(\quad \frac{\partial y_i}{\partial x_j} \quad\right) \end{array}$$

- example:

$$y = A\,x$$

$$\underbrace{y}_{m\times 1} \ni y_i, \quad \underbrace{A}_{m\times n} \ni a_{ij}, \quad \underbrace{x}_{n\times 1} \ni x_j$$

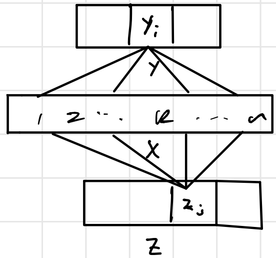
$$y_i = \sum_{j=1}^{n} a_{ij}\, x_j$$

$$\frac{\partial y}{\partial x^T} = \begin{array}{c} \\ i \\ \\ \end{array} \begin{array}{c} j \\ \left(\quad a_{ij} \quad\right) \end{array} = A$$

- Chain Rule:

$$\underset{m\times 1}{Y} = F(\underset{n\times 1}{X}) \qquad \underset{n\times 1}{X} = G(\underset{\ell\times 1}{Z})$$

$$\underset{m\times \ell}{\frac{\partial Y}{\partial Z^T}} = \underset{m\times n}{\frac{\partial Y}{\partial X^T}} \; \underset{n\times \ell}{\frac{\partial X}{\partial Z^T}}$$




$$\frac{\partial Y}{\partial z_j} = \sum_{k=1}^{n} \frac{\partial Y_i}{\partial x_k} \frac{\partial x_k}{\partial z_j}$$

$$\begin{array}{c} \\ i \\ m \end{array}\!\!\left( \begin{array}{c} j \quad \ell \\ \dfrac{\partial Y_i}{\partial z_j} \end{array} \right) = \begin{array}{c} \\ i \\ m \end{array}\!\!\left( \begin{array}{c} k \quad n \\ \dfrac{\partial Y_i}{\partial x_k} \end{array} \right) \left( \begin{array}{c} j \quad \ell \\ \dfrac{\partial x_k}{\partial z_j} \end{array} \right)_{n}$$

- So consider: $L(\beta) = \frac{1}{2}|e|^2$

$$\frac{\partial L}{\partial \beta^T} = \frac{\partial L}{\partial e^T} \frac{\partial e}{\partial \beta^T}$$

Note: $\dfrac{\partial L}{\partial e} = \left( \dfrac{\partial L}{\partial e_i} \right) = \left( e_i \right) = e$

Note: $e = Y - X\beta$ so $\dfrac{\partial L}{\partial \beta^T} = -X$

Thus $\dfrac{\partial L}{\partial \beta^T} = -e^T X$

$$\frac{\partial L}{\partial \beta} = -X^T e = 0$$

$$-X^T(Y - X\beta) = 0$$

$$X^T Y - X^T X \beta = 0$$

$$(X^T X)^{-1} X^T Y = \hat{\beta}$$

- $\underset{1\times 1}{y} = \underset{n\times 1}{F(x)}$

multidimensional

$\underset{n\times 1}{F'} = F'(x) = \frac{\partial y}{\partial x}_{n\times 1}$ instead of $1\times n$ } Gradient

$F''(x) = \frac{\partial F'}{\partial x^T}_{n\times n} = \frac{\partial \frac{\partial y}{\partial x}}{\partial x^T} = \frac{\partial^2 y}{\partial x \, \partial x^T}$ } Hessian

$$\left(\frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial x^T}\right) y$$

$$\begin{pmatrix} \vdots \\ \frac{\partial}{\partial x_i} \\ \vdots \end{pmatrix} \left( \cdots \frac{\partial}{\partial x_j} \cdots \right) y = \begin{array}{cc} & j \\ i & \begin{pmatrix} & \vdots & \\ \cdots & \frac{\partial^2 y}{\partial x_i \partial x_j} & \cdots \\ & \vdots & \end{pmatrix} \end{array}$$

- Consider quadratic form :

$$y = x^T \underset{n\times n}{A} \underset{n\times 1}{x}$$

$$= \sum_{i,j} a_{ij} x_i x_j = \text{const} + a_{ii} x_i^2 + \sum_{j\neq i} a_{ij} x_i x_j + \sum_{j\neq i} a_{ji} x_i x_j$$

$$\frac{\partial y}{\partial x}$$

$$\begin{array}{c} 1 \\ \vdots \\ n \end{array} \left( \frac{\partial y}{\partial x_i} \right) \qquad \frac{\partial y}{\partial x_i} = 2 a_{ii} x_i + \sum_{j\neq i} (a_{ij} + a_{ji}) x_j =$$

- Consider quadratic form :

$$y = x^T \underset{n\times n}{A} \underset{n\times 1}{x}$$

$$= \sum_{i,j} a_{ij} x_i x_j = const + a_{ii} x_i^2 + \sum_{j\neq i} a_{ij} x_i x_j + \sum_{j\neq i} a_{ji} x_i x_j$$

$$\frac{\partial y}{\partial x}$$

$$\begin{array}{c} 1 \\ \vdots \\ n \end{array} \begin{pmatrix} \\ \frac{\partial y}{\partial x_i} \\ \\ \end{pmatrix} \qquad \frac{\partial y}{\partial x_i} = 2a_{ii}x_i + \sum_{j\neq i} (a_{ij} + a_{ji}) x_j =$$

$$= \begin{array}{c} \\ i \\ \\ \end{array} \begin{array}{c} \hphantom{a_{ij} + a_{ji}} j \\ \begin{pmatrix} & & \\ & a_{ij} + a_{ji} & \\ & & \end{pmatrix} \end{array} \begin{pmatrix} \\ x_j \\ \\ \end{pmatrix}$$

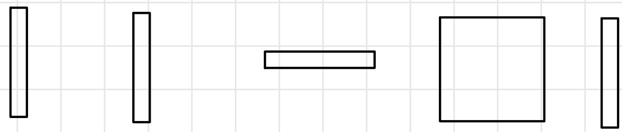
So $$\frac{\partial y}{\partial x} = (A + A^T)\, x$$

&

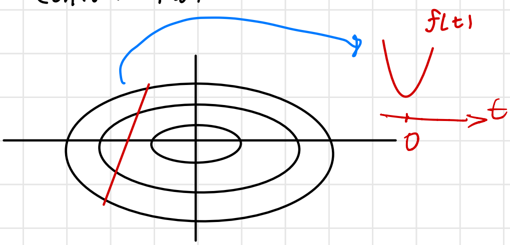
$$\frac{\partial^2 y}{\partial x\, \partial x^T} = A + A^T$$

• Taylor Expansion

$$\underset{1\times1}{y} = F(\underset{n\times1}{x}) = F(x_0) + \langle \underset{n\times1}{F'(x_0)}, \underset{n\times1}{x-x_0} \rangle + \frac{1}{2} \underset{1\times n}{(x-x_0)^T} \underset{n\times n}{F''(x_0)} \underset{n\times1}{(x-x_0)}$$

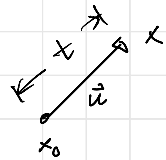



Contour plot:




$$x = x_0 + \vec{u}\, \underset{1\times1}{t}$$

$$|\vec{u}| = 1$$




$$f(t) = F(x) = F(x_0 + \vec{u} t)$$

$$\doteq f(0) + f'(0)\, t + \frac{1}{2} f''(0)\, t^2$$

note: $f'(t) = \frac{\partial y}{\partial t} = \frac{\partial y}{\partial x^T} \frac{\partial x}{\partial t} = F'(x)^T u$

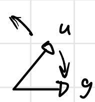
$= \langle F'(x), u \rangle$

$$f'(0) = \langle \underbrace{F'(x_0)}_{g}, u \rangle = \langle g, u \rangle$$

Implies $f'(0)$ is maximized when $u$ aligned with $g$

$g$ is the gradient

the steepest direction.




$= |g||u| \cos(\theta)$

$= |g| \cos(\theta)$

$\overset{\max}{\Longrightarrow} u \propto g$

• note: $f''(t) = \frac{\partial f'}{\partial t} = \frac{\partial}{\partial t} u^T F' = u^T \frac{\partial}{\partial t} F'$

$$= u^T \frac{\partial F'}{\partial x^T} \frac{\partial x}{\partial t} = u^T \frac{\partial^2 F}{\partial x \partial x^T} u$$

$$= u^T F''(x) u$$

$$f''(0) = u^T F''(x_0) u = u^T H u$$

Going back to $f(t) \doteq f(0) + f'(0)t + \frac{1}{2} f''(0) t^2$
so we have

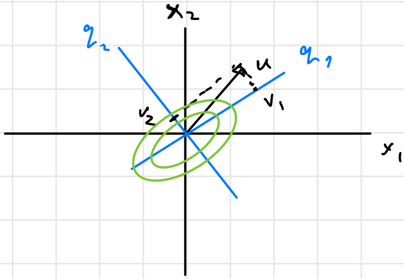
$$f(t) \doteq F(x_0) + \langle F'(x_0), x - x_0 \rangle + \frac{1}{2}(x - x_0)^T F''(x_0)(x - x_0)$$

$$H = Q \Lambda Q^T$$

$$u^T H u = \underbrace{u^T Q}_{v^T} \Lambda \underbrace{Q^T u}_{v} = v^T \Lambda v = \sum_{i=1}^{n} \lambda_i v_i^2$$

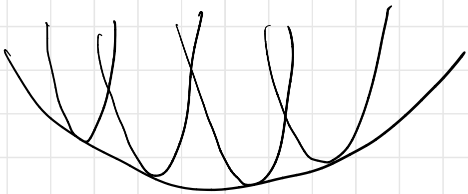
Change of Basis $\left\{ \begin{array}{l} v = Q^T u \\ u = Q v \end{array} \right.$

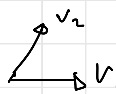
$$u = \begin{pmatrix} \vec{q}_1 & \cdots & \vec{q}_i & \cdots & \vec{q}_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ \vdots \\ v_n \end{pmatrix} = \vec{q}_1 v_1 + \cdots + \vec{q}_i v_i + \cdots + \vec{q}_n v_n$$




$\lambda_i \equiv$ curvature along $\vec{q}_i$

$$v_i = q_i^T u = \langle u, \vec{q}_i \rangle$$

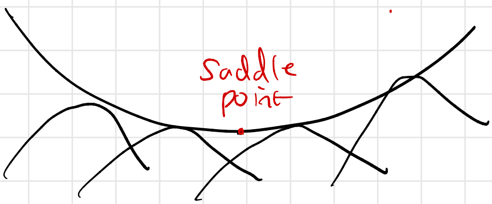


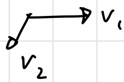



$$\lambda_1 v_1^2 + \lambda_2 v_2^2$$

$\lambda_1 > 0$ $\quad$ $\lambda_2 > 0$

$(\lambda_1 = 1)$ $\quad$ $(\lambda_2 = 10)$




$\lambda_1 > 0$ $\quad$ $\lambda_2 < 0$

$(\lambda_1 = 1)$ $\quad$ $(\lambda_2 = -5)$

curve up $\quad$ curve down