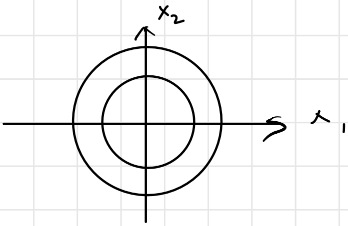
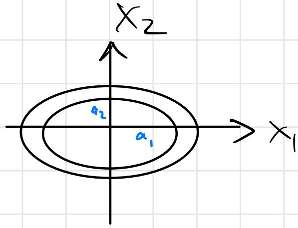
Lecture 5

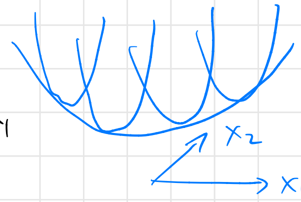
• Circle :



$$x_1^2 + x_2^2 = const$$

Ellipse :





$$\frac{x_1^2}{a_1^2} + \frac{x^2}{a_2^2} = const$$

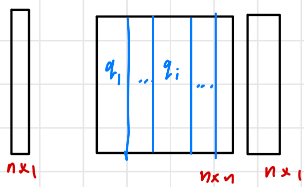
$$f(x) = \lambda_1 x_1^2 + \lambda_2 x_2^2$$

• Change of view :

noun verb

$$\begin{tikzcd} X_{n\times 1} \arrow[r, "H"] & X' \\ Y_{n\times 1} \arrow[u, "Q_{n\times n}"] \arrow[r, "\Lambda"] & Y' \arrow[u, "Q"'] \end{tikzcd}$$

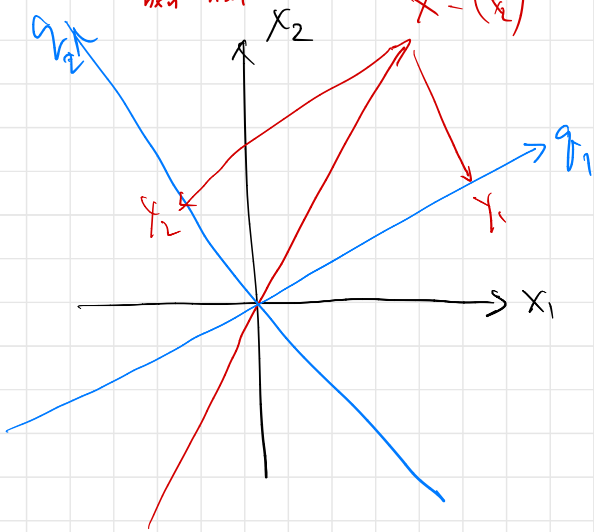
$$X = Q\,Y$$



$$X = (q_1, \dots q_i, \dots q_n) \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}$$

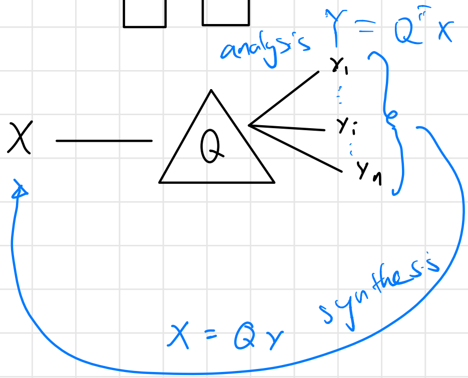
$$\boxed{x} = \boxed{q_1}\, y_1 + \dots \boxed{q_i}\, y_i + \dots \boxed{q_n}\, y_n$$

So $\langle q_i, q_j \rangle = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$



y_i coordinate

$$y_i = \langle \boxed{x_i}, \boxed{q_i} \rangle = q_i^T x \qquad Y = \begin{pmatrix} y_i \end{pmatrix} = \begin{pmatrix} q_i^T x \end{pmatrix} = \begin{pmatrix} q_i^T \end{pmatrix} x = Q^T x$$



- Note for any symmetric matrix H, we have

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

Consider $x' = Hx$

$$Qy' = HQY$$

$$\begin{aligned} Y' &= Q^{-1} H Q Y \\ &= Q^T H Q Y \\ &= Q^T Q \Lambda Q^T Q Y \\ &= \Lambda Y \end{aligned}$$

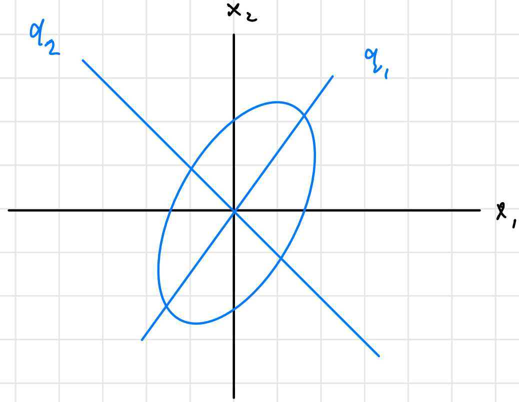
Note:

$$\begin{aligned} \langle x, x' \rangle &= |x||x'| \cos\theta \\ &= |Y||Y'| \cos\theta \end{aligned}$$

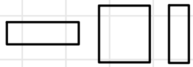
Rotation doesn't change length or angle

$$x^T x' = x^T H x = Y^T Y' = Y^T \Lambda Y$$

- Quadratic Form



$$f(x) = x^T H x$$



$$= y^T \Lambda y$$

$$= \sum_{i=1}^{n} \lambda_i y_i^2, \quad \lambda_i > 0$$

- Taylor Expansion:

$F(x)$

$x = x_0 + \vec{u} t$

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

small o

$$= f(0) + f'(0) t + f''(0) t^2 + o(t^2)$$

$$= F(x_0) + \langle \underbrace{F'(x_0)}_{g}, x - x_0 \rangle + \frac{1}{2} (x - x_0)^T \underbrace{F''(x_0)}_{H} (x - x_0) + o(|x - x_0|^2)$$

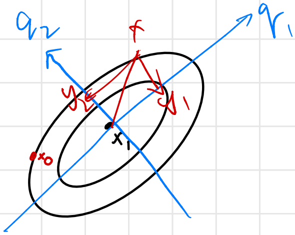
$$= F(x_0) + \langle g, x - x_0 \rangle + \frac{1}{2} (x - x_0)^T H (x - x_0) + o(|x - x_0|^2)$$

$$= \text{const} + \frac{1}{2} (x - x_0 + H^{-1} g)^T H (x - x_0 + H^{-1} g)$$ complete the square

$$= \text{const} + \frac{1}{2} (x - x_1)^T H (x - x_1)$$

where $x_1 = x_0 - H^{-1} g = x_0 - F''(x_0)^{-1} F'(x_0)$

- So the shape is given by



$y = Q^T(x - x_1)$, x_1 is a minimum since we have $\sum \lambda_i y_i^2$, $\lambda_i > 0$

- Newton - Raphson

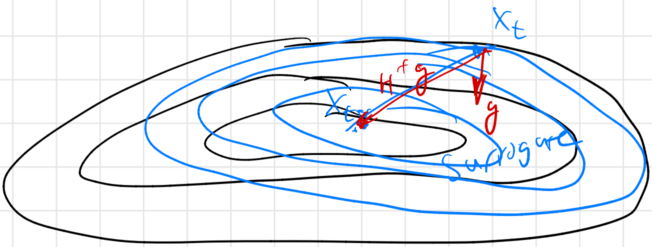
$$\text{approx } \underbrace{F(x)}_{\text{objective}} \doteq \underbrace{F(x_t) + \langle F'(x_t), x - x_t \rangle + \frac{1}{2}(x - x_t)^T F''(x_t)(x - x_t)}_{\text{surrogate}}$$

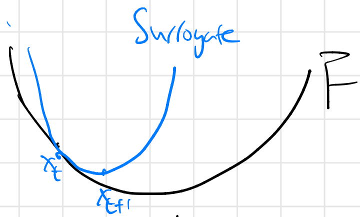
$\max_x F(x)$

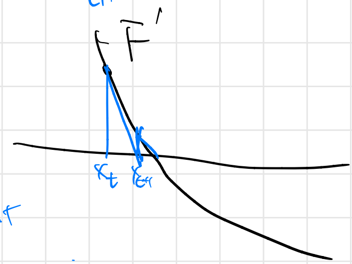
max

$$x_0 \longrightarrow \cdots \longrightarrow x_t \longrightarrow x_{t+1} \longrightarrow \cdots$$

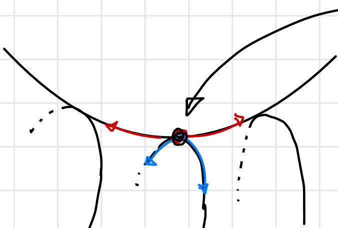
$$x_{t+1} = x_t - \underbrace{F''(x_t)^{-1}}_{H^{-1}} \underbrace{F'(x_t)}_{g}$$





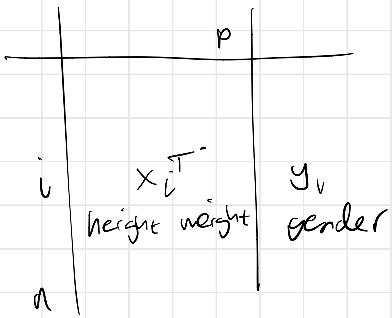


-



saddle point can be problematic

- logistic regression:



$$L(\beta) = \sum_{i=1}^{n} \left(y_i s_i - \log(1 + e^{s_i}) \right)$$

$$s_i = x_i^T \beta$$

$$L'(\beta) = \sum_{i=1}^{n} \left(x_i y_i - x_i \frac{e^{s_i}}{1 + e^{s_i}} \right)$$

$$= \sum_{i=1}^{n} x_i (y_i - p_i)$$

$$L''(\beta) = -\sum_{i=1}^{n} x_i x_i^T \underbrace{p_i(1 - p_i)}_{w_i}$$

$$= -\sum_{i=1}^{n} w_i x_i x_i^T$$

$$\beta_{t+1} = \beta_t + \left(\sum_{i=1}^{n} w_i x_i x_i^T \right)^{-1} \left(\sum_{i=1}^{n} x_i e_i \right)$$ IRLS

Newton - Raphson

- Gradient Descent / Ascent

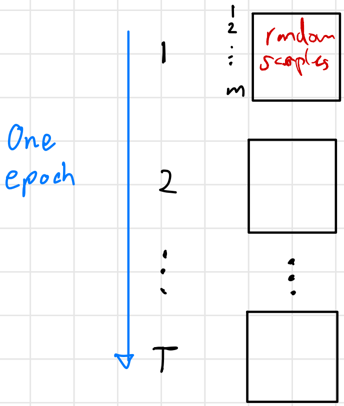
θ : unknown parameter

$Loss(\theta)$

$$\theta_{t+1} = \theta_t - \underset{\uparrow}{\eta_t}\, Loss'(\theta_t)$$

step-size learning rate

- Mini-batch :



$$Loss_t(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i s_i - \log\left(1+e^{s_i}\right)\right)$$

batch t

$$g_t = Loss_t'(\theta_t)$$

$$= -\frac{1}{m}\sum_{i=1}^{m} x_i\,(\overbrace{y_i - \underset{\uparrow\,\theta_t}{p_i}}^{e_i})$$

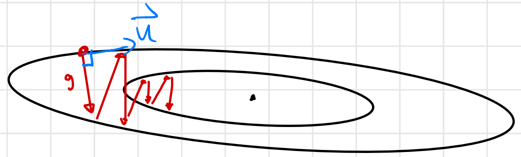
Stochastic Gradient Descent

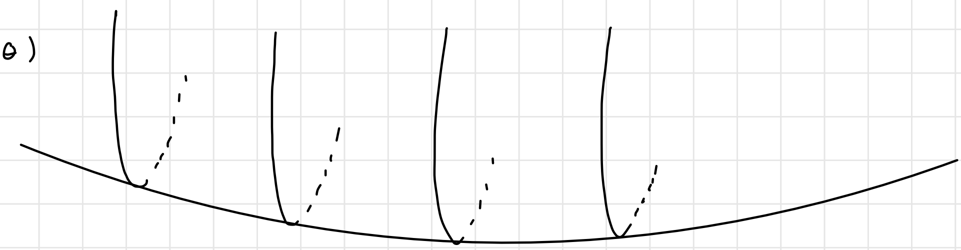
$$\theta_{t+1} = \theta_t - \eta_t\, g_t$$

- Issues with SGD:

- consider the following loss function

$f'(0) = \langle g, \vec{u} \rangle$





- Think Newton - Raphson

$$\begin{array}{ccc} g & \xrightarrow{H^{-1}} & H^{-1}g \\ {\scriptstyle Q}\uparrow & & \uparrow{\scriptstyle Q} \\ \tilde{g} & \xrightarrow{\Lambda^{-1}} & \Lambda^{-1}\tilde{g} \end{array} = \begin{pmatrix} \frac{\tilde{g}_1}{\lambda_1} \\ \vdots \\ \frac{\tilde{g}_n}{\lambda_n} \end{pmatrix}$$

GD: $\Theta_{t+1} = \Theta_t - \eta_t L'(\Theta_t)$ $\quad H = \frac{1}{\eta_t} I$

Newton: $\Theta_{t+1} = \Theta_t - H^{-1} L'(\Theta_t)$

GD : $\theta_{t+1} = \theta_t - \eta_t L'(\theta_t)$ $\quad H = \frac{1}{\eta_t} I$

$$\underbrace{L(\theta)}_{\text{objective}} \doteq \underbrace{L(\theta_t) + \langle \theta - \theta_t, \overset{g}{L'(\theta_t)} \rangle + \frac{1}{2}(\theta - \theta_t)^T \frac{1}{\eta_t} I (\theta - \theta_t)}_{\text{surrogate}}$$

$\Downarrow$

Poor approximation $\longrightarrow$ $\frac{1}{2\eta_t} |\theta - \theta_t|^2$

no curvature info

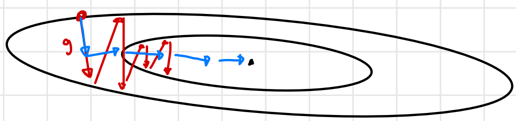
Newton: $\theta_{t+1} = \theta_t - H^{-1} L'(\theta_t)$

- Momentum:

friction $\cdot g$

$v_t = \gamma v_{t-1} + \eta_t g_t$ $\quad$ heavy ball

$\theta_{t+1} = \theta_t - v_t$ $\quad$ (cancelling oscillation /variance)

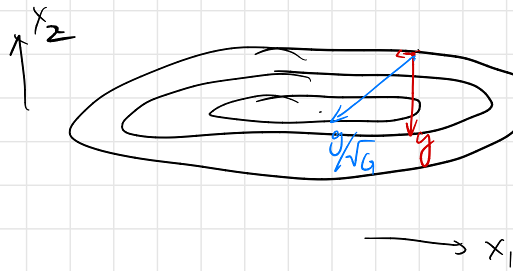


- Adaptive Gradient:

$$G_t = G_{t-1} + g_t^2$$

$$\theta_{t+1} = \theta_t - \eta_t \frac{g_t}{\sqrt{G_t + \varepsilon}}$$

element-wise operation



- Recall $g_t = -\frac{1}{m} \sum_{i=1}^{m} x_i(y_i - p_i)$

  if $x_{ij} = 0$ for most i (sparse feature)
  then $L'(\beta_j)$ small

  → then once we see $x_{ij} \neq 0$ we take a large step.