FYS-STK 3155/4155, Sept 30

$$\beta^{(n+1)} = \beta^{(n)} - \gamma^{(n)} g(\beta^{(n)})$$

$$C(\beta^{(n+1)}) = C(\beta^{(n)}) + (\beta^{(n+1)} - \beta^{(n)})^T g(\beta^{(n)}) + \frac{1}{2}(\beta^{(n+1)} - \beta^{(n)})^T H(\beta^{(n)}) \times (\beta^{(n+1)} - \beta^{(n)}) + \cdots$$

$$C(\beta) = \frac{1}{n}(y - X\beta)^T(y - X\beta)$$

$$H = \frac{2}{n} X^T X$$

$$b = \beta^{(n+1)} - \beta^{(n)}$$

$$C(\beta^{(n+1)}) = C_0 + b^T g(\beta^{(n)}) + \frac{1}{2} b^T H b$$

$$f(x) = \frac{1}{2} x^T A x - b^T x$$

$$\frac{\partial f(x)}{\partial x} = 0 = Ax - b$$

$$Ax = b$$

$$H = X^T X \in \mathbb{R}^{p \times p}$$

square & symmetric
positive definite matrix

$$\beta^{(n+1)} = \beta^{(n)} - \underbrace{H^{-1}(\beta^{(n)})}_{\gamma^{(n)}} g(\beta^{(n)})$$

- learning rate updates
  - linear update
  - constant
  - exponential update
  - momentum based

$$\beta^{(m+1)} = \beta^{(m)} + \gamma\, g(\beta^{(m)}) + \delta(\beta^{(m)} - \beta^{(m-1)})$$

- Adagrad (convex function)
- RMS prop (non-convex)
- Adam
- Full gradient calculation
- stochastic GD

Steepest descent

$$f(x) = \frac{1}{2} x^T A x - b^T x$$

$$\frac{\partial f}{\partial x} = 0 = b - Ax = -g$$

$$\boxed{x_{k+1} = x_k + \alpha_k r_k}$$

r_{k+1} = residual

$$r_{k+1} = b - A x_{k+1}$$

$$r_k = b - Ax_k$$

$$r_0 = b - Ax_0$$

$$r_{k+1}^T r_k = 0$$

$$r_{k+1} = b - Ax_{k+1}$$

$$= \underbrace{b - (Ax_k}_{r_k} + A\alpha_k r_k)$$

$$r_k^T r_{k+1} = 0 = r_k^T r_k + \alpha_k r_k^T A r_k$$

$$\Rightarrow \boxed{\alpha_k = \frac{r_k^T r_k}{r_k^T A r_k}}$$

$$r_k = b - Ax_k = -g_k$$

$$x_{k+1} = x_k + \alpha_k r_k$$

$$= x_k - \underbrace{\alpha_k}_{\substack{\parallel \\ \gamma_k}} g_k$$

$$\gamma_k = \frac{g_k^T g_k}{g_k^T H g_k}$$

$r_k = -g_k \quad A = H$

if $\quad H g_k = \lambda_k g_k$

$$\gamma_k = \frac{1}{\lambda_k}$$

AdaGrad $\quad \gamma_k \sim \frac{1}{g_k^T g_k}$

- Schedulers for γ_k
  - constant γ_0
  - linear

$$\gamma_k = (1-\alpha)\gamma_0 + \alpha \gamma_\tau$$

$$\gamma_\tau \sim \frac{1}{100}\gamma_0$$

$$\alpha = \frac{k}{\tau}$$

k — # iterations

τ — example in notebook

- $$\gamma_k = \frac{\gamma_0}{1 + k\gamma_\tau}$$

- exponential decay

$$\gamma_k = \gamma_0 \exp(-k\gamma_\tau)$$

- Adagrad

algorithm:
require initial γ_0
— " — β_0
adagrad with SGD
batches # epochs
$D = \{(x_0, y_0), (x_1, y_1) \ldots (x_{n-1}, y_{n-1})\}$

while stopping criterion not met

- compute gradient g_k

- Define $G = \sum_{i=1}^{k} g_i g_i^T$

- Define $\dfrac{\gamma_0}{\delta + \sqrt{G_{ii}}} = \gamma_K$

simple approach:
only diagonal elements
in $\sqrt{G_{ii}}$

- update

$$\beta_{K+1} = \beta_K - \frac{\gamma_0}{\underbrace{\delta}_{10^{-8} \sim 10^{-6}} + \sqrt{G_{ii}}} \odot g_K$$

$$x \odot y = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \odot \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 y_1 \\ x_2 y_2 \end{bmatrix}$$

end update,

RMS prop

require γ_0, β_0
decay rate ρ, $\delta \sim 10^{-8}$
while stopping criterion
not met

compute g_K

$$G_K = \rho G_{K-1} + (1-\rho) \sum_{i=0}^{K} g_i g_i^T$$

$$\gamma_K = \frac{\gamma_0}{\delta + \sqrt{G_K}}$$

↑ diagonal terms only

$$\beta_{K+1} = \beta_K - \gamma_K \odot g$$