ML ERASMUS, NOV 21, 2022

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

- update of $\gamma^{(n)}$
  - keep fixed for all iteration -n-
  - linear change

$$\gamma^{(k)} = (1-\alpha)\gamma_0 + \alpha\gamma_\tau$$

$$\alpha = \frac{k}{\tau} \qquad \tau = \text{parameter}$$

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

$$\gamma_0 \in [10^{-5}, 10^{-4}, 10^{-3}, \dots 10^{-1}]$$

- $$\gamma^{(k)} = \frac{\gamma_0}{1+k\gamma_\tau}$$

- exponential

$$\gamma^{(k)} = \gamma_0 \exp(-k\gamma_\tau)$$

- Approximation of Hessian matrix (info on gradients)

- Adagrad
- RMSprop
- ADAM

- Efficient calculation of gradients
  - momentum based gradients
  - stochastic gradient descent.

- Momentum based gradients

$$m \frac{d^2x}{dt^2} + \underbrace{\mu}_{\text{friction/drag}} \frac{dx}{dt} = F(x) = -\underbrace{\nabla V(x)}_{\text{gradient}}$$

$$\frac{d^2x}{dt^2} \simeq \frac{x_{t+\Delta t} - 2x_t + x_{t-\Delta t}}{(\Delta t)^2}$$

$$\frac{dx}{dt} \simeq \frac{x_{t+\Delta t} - x_t}{\Delta t}$$

$$m\,\frac{\left(x_{t+\Delta t} - 2x_t + x_{t-\Delta t}\right)}{(\Delta t)^2} +$$

$$\mu\left(\frac{x_{t+\Delta t} - \Delta t}{\Delta t}\right) = -\nabla V(x)$$

Define

$$\Delta x_{t+\Delta t} = x_{t+\Delta t} - x_t$$

$$\Delta x_t = x_t - x_{t-\Delta t}$$

$$\Delta x_{t+\Delta t} = -\frac{(\Delta t)^2}{m+\mu\Delta t}\nabla V(x) + \frac{m}{m+\mu\Delta t}\Delta x_t$$

$$\delta = \frac{m}{m+\mu\Delta t}$$

$$\gamma = \frac{(\Delta t)^2}{m+\mu\Delta t}$$

$$\Delta x_{t+\Delta t} = -\gamma\nabla V(x) + \delta\,\Delta x_t$$

memory parameter

$$\Delta x_t = x_t - x_{t-\Delta t}$$

$$\left(\beta^{(n+1)} = \beta^{(n)} - \gamma \nabla C(\beta^{(n)}) \right.$$

$$\Delta \beta^{(i+1)} = \beta^{(i+1)} - \beta^{(i)}$$

$$\beta^{(i+1)} = \beta^{(i)} - \gamma^{(i)} \nabla C(\beta^{(i)}) + \delta(\beta^{(i)} - \beta^{(i+1)})$$

Rewritten as

$$v^{(i)} = \delta(\beta^{(i)} - \beta^{(i-1)}) - \gamma g(\beta^{(i)})$$

$$\beta^{(i+1)} = \beta^{(i)} + v^{(i)}$$

$$\delta \in [0, 1]$$

Algorithm

Require : learning rate γ
momentum parameter δ

require : initial β and v

$$v^{(i)} = \delta(\beta^{(i)} - \beta^{(i-1)}) - \gamma g(\beta^{(i)})$$

$$\beta^{(i+1)} = \beta^{(i)} + v^{(i)}$$

while stopping criterion not met

- compute gradient (g)
- compute v-update

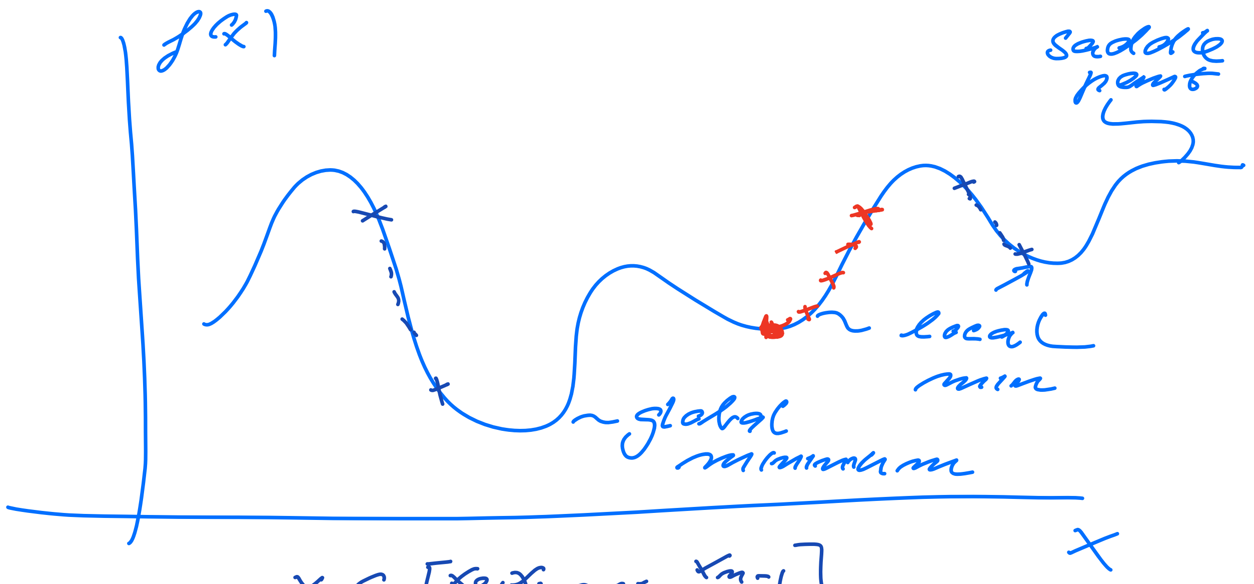
$$v^{(i)} = \delta(\beta^{(i)} - \beta^{(i-1)}) - \gamma g(\beta^{(i)})$$

- apply update

$$\beta^{(i+1)} = \beta^{(i)} + v^{(i)}$$

end while.

Stochastic gradient descent for the calculation of g



$x \in [x_0, x_1, \ldots, x_{n-1}]$

- split n (total # of x values) in so-called batches with a given number of data points – m –
- Total of M-batches
- $n = 100$ $M = 10$, then in each mini-batches we have 10 points.

Steepest Descent

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

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

$$Ax = b \qquad \left(x = \underset{H^{-1}}{\underset{\uparrow}{A^{-1}}} \underset{g(\beta^{(n)})}{\underset{\uparrow}{b}} \right)$$

Define $r = b - Ax$
we have the solution when
residual $r = 0$

start with a guess x_0

$$r_0 = -Ax_0 + b$$

in general

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

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

$$r_{k+1} = b - A(x_k + \alpha_k r_k)$$

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

$$r_{k+1} = r_k - \alpha_k A r_k$$

we want $r_{k+1} = 0$

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

multiply with r_k^T

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

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

learning rate

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

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

$$x_{k+1} = x_k - \alpha_k g(x_k)$$

interpret as learning rate $\gamma^{(k)}$

$$A = \text{Hessian matrix} = \frac{\partial^2 C}{\partial \beta \partial \beta^T}$$