Lecture FYS-STK3155/4155, September 28, 2023

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \; \underbrace{-\log P(D|\beta)}_{c(\beta)}$$

$$\left.\frac{\partial c}{\partial \beta}\right| = g$$

$$\frac{\partial^2 c}{\partial \beta \, \partial \beta^T} = H$$

perform Taylor expansion around $\hat{\beta}$

$\hat{\beta} - \beta^{(n)}$

$$C(\hat{\beta}) = C(\beta^{(n)}) + (g^{(n)})^T(\hat{\beta} - \beta^{(n)})$$

$$+ \frac{1}{2}(\hat{\beta} - \beta^{(n)})^T H^{(n)} (\hat{\beta} - \beta^{(n)})$$

$$+ \cdots$$

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

$$C(\hat{\beta}) \simeq C(\beta^{(n)}) + (g^{(n)})^T b$$

$$+ \frac{1}{2} b^T H^{(n)} b$$

$$\frac{\partial C}{\partial b} = 0 = g^{(n)} + H^{(n)} \cdot b$$

$$\hat{\beta} = \beta^{(n+1)} = \beta^{(n)} - \left[H(\beta^{(n)})\right]^{-1} g^{(n)}(\beta^{(n)})$$

in more general terms

$$C(\beta) \Rightarrow f(x) = c + g^T x + \frac{1}{2} x^T A x$$

$$\frac{\partial f}{\partial x} = 0 \quad Ax + g \Rightarrow$$

$$x = -A^{-1} g$$

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

learning rate

Taylor expand $C(\hat{\beta})$ again

$$C\left(\beta^{(n)} - \gamma g^{(n)}\right) \simeq C\left(\beta^{(n)}\right) - \gamma \left(g^{(n)}\right)^T g^{(n)} + \frac{1}{2}\gamma^2 \left(g^{(n)}\right)^T H^{(n)} g^{(n)} + \dots$$

Take derivative wrt γ

$(g^{(n)} \Rightarrow g \,;\; H^{(n)} \rightarrow H)$

$$\gamma = \frac{g^T g}{g^T H g}$$

$$Hg = \lambda g$$

$$\gamma = \frac{g^T g}{\lambda g^T g} = \frac{1}{\lambda}$$

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

Adagrad

RMSprop

ADAM

Momentum

Plain gradient descent

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

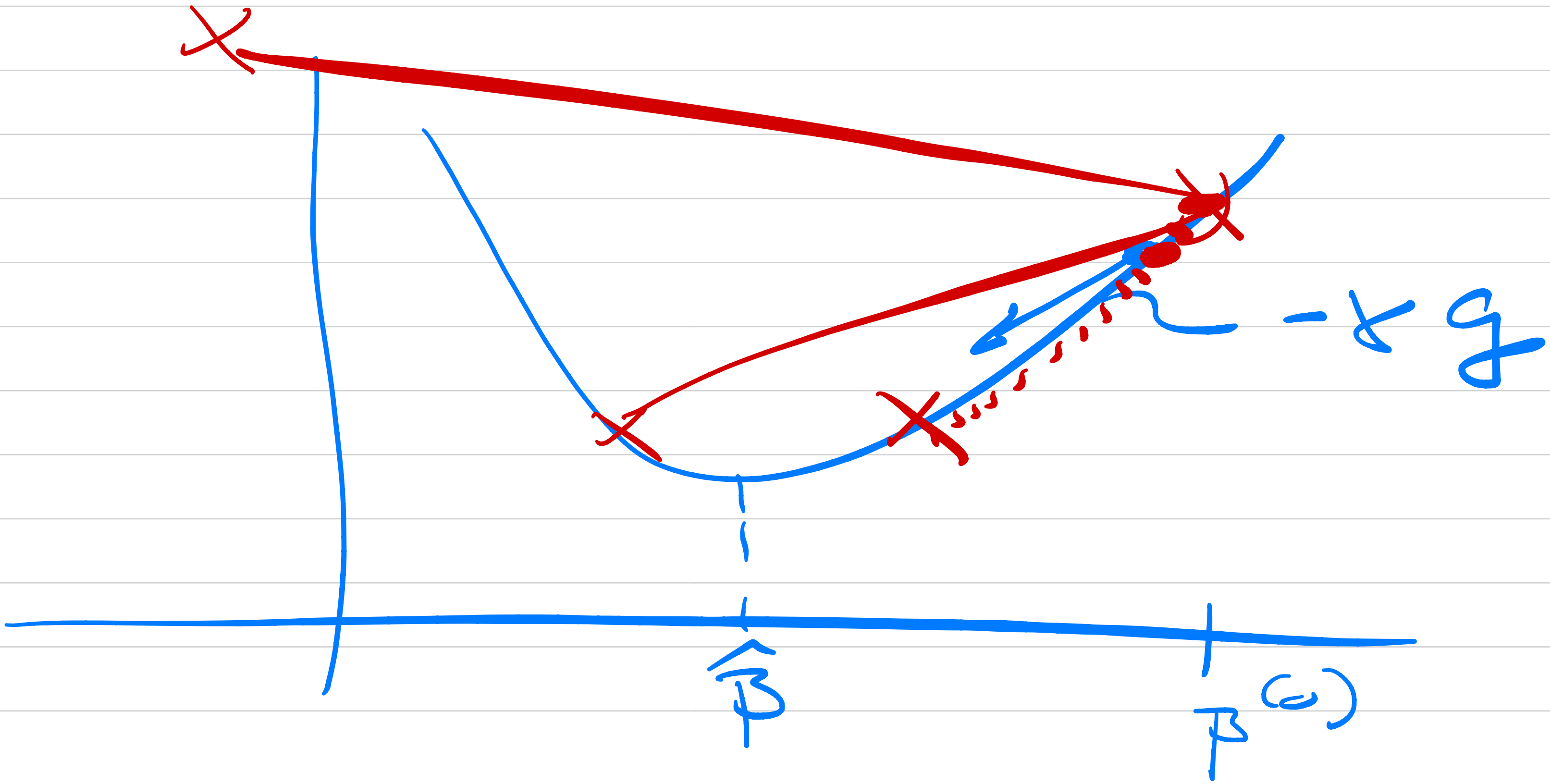
Convergence test

$$\| \beta^{(n+1)} - \beta^{(n)} \|_1 \leq \varepsilon \approx 10^{-8}$$

Explicit convergence criterion

$$\gamma < \frac{2}{\lambda_{max}}$$

λ_{max} largest eigenvalue of H



Gradient descent with momentum.

Newton's eq of motion

$$m \frac{d^2x}{dt^2} + \mu \frac{dx}{dt} = F(x) = -\nabla V(x)$$

discretize wrt t, Δt

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

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

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

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

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

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

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

$$+ \frac{m}{m+\mu\Delta t} \Delta x_t$$

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

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

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

$$\delta(x_t - \underbrace{x_{t-\Delta t}})$$

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

memory term

$$x_t \Rightarrow B_i \qquad x_{t+\Delta t} \Rightarrow B_{i+1}$$

$$\nabla V \Rightarrow g(B_i) \qquad \Delta B_{i+1} = B_{i+1} - B_i$$

$$\beta_{i+1} = \beta_i - \gamma_i\, g(\beta_i) + \delta(\beta_i - \beta_{i-1})$$

algorithm:

Require: learning rate γ
and δ

Require: initial β_i value
and v_i

$$v_i = \delta(\beta_i - \beta_{i-1}) - \gamma\, g(\beta_i)$$

$$\beta_{i+1} = \beta_i + v_i$$

while stopping criterion not met
- compute gradient

- compute

$$v_i = \delta(\beta_i - \beta_{i-1}) - \gamma g_i'$$

- update β_{i+1}

$$\beta_{i+1} = \beta_i + v_i'$$

end while.