# FYS-STK3155/4155 lecture, September 23, 2024

# Lecture FYS-STK3155/4155, September 23

Newton's Method

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

Taylor-expand around $\hat{\beta}$

$$C(\hat{\beta}) = C(\beta^{(n)}) + (\hat{\beta} - \beta^{(n)})^T g(\beta^{(n)}) + \frac{1}{2}(\hat{\beta} - \beta^{(n)})^T H(\beta^{(n)})(\hat{\beta} - \beta^{(n)}) + O((\hat{\beta} - \beta^{(n)})^3)$$

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

Linear Regression

$$g(\beta^{(n)}) = \left.\frac{\partial C}{\partial \beta}\right|_{\beta=\beta^{(n)}}$$

$$\underset{\text{Hessian matrix}}{H(\beta^{(n)})} = \frac{2}{n} \underbrace{X^T X}_{\text{independent of } \beta}$$

Logistic regression

$$\frac{\partial^2 C}{\partial \beta \partial \beta^T} = X^T W X$$

$$w_{ii} = p_i(\beta^{(n)})(1 - p_i(\beta^{(n)}))$$

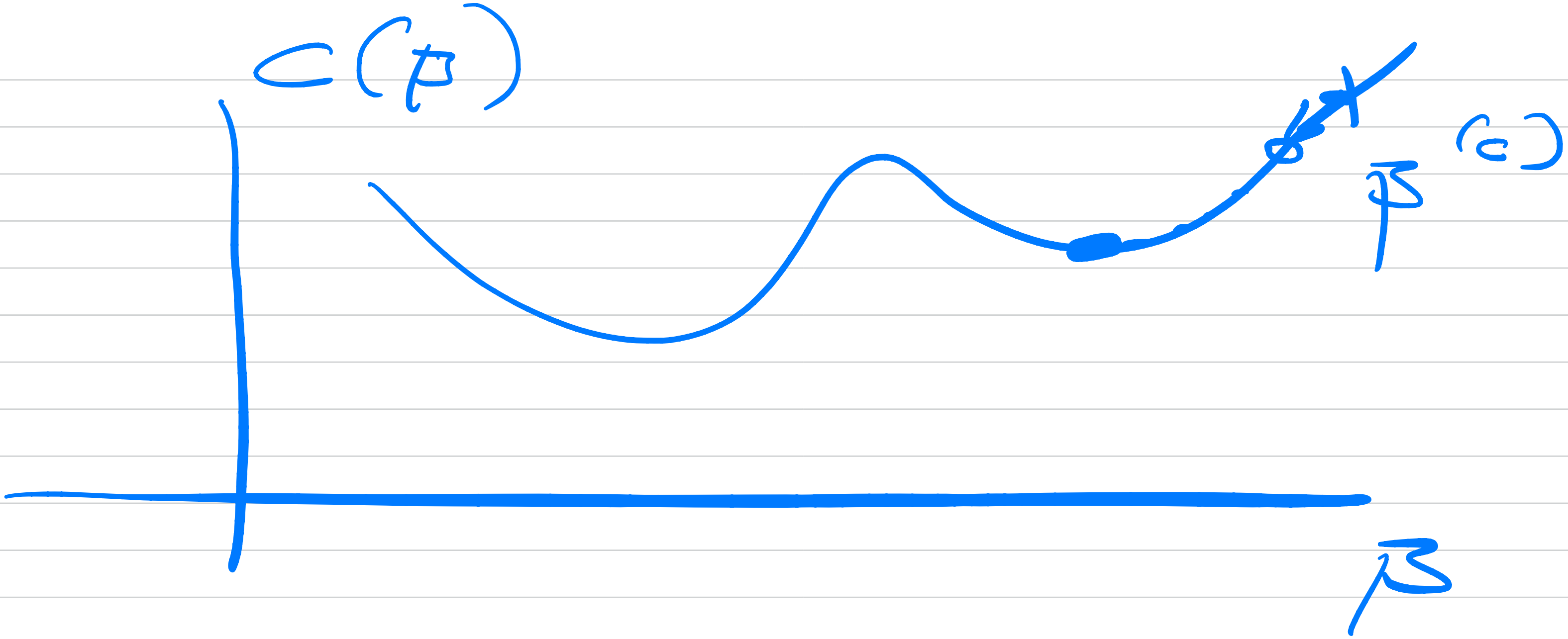
$x_i, y_i$

$$w_{ij} = 0 \quad \text{if } i \neq j$$

$\Rightarrow$

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

Newton-Raphson




Last week

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

Taylor expand $C(\beta)$ around $\beta^{(n)} - \gamma^{(n)} g^{(n)}$

$$\gamma^{(n)} = \frac{g^T g}{g^T H(\beta^{(n)}) g} \qquad g = g(\beta^{(n)})$$

steepest descent

starting point ; gradient descent

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

learning rate

Different approaches

1) keep $\gamma^{(n)} = \gamma$ fixed for every iteration

choose $\gamma = [10^{-5}, 10^{-4}, \ldots 10^{-1}]$

2) Gradient descent (GD) with momentum (memory)

3) Adagrad ($\gamma$ has info $g(\beta^{(n)})$)

4) RMS prop ( —— " —— )

5) ADAM ( —— " —— )

optimal value

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

eigvalue of $H(P^{(n)})$

$$f(x_1, x_2) = \frac{1}{2} \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 20 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 5 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\left( \frac{1}{2} x^T A x - b^T x \right)$$

$$= x_1^2 + x_1 x_2 + 10 x_2^2 - 5x_1 - 3x_2$$

$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 + x_2 - 5 = 0$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = 0 = x_1 + 20x_2 - 3$$

$$x_1 = 97/39 \approx 2,47$$

$$x_2 = 1/39 \approx 0,026$$

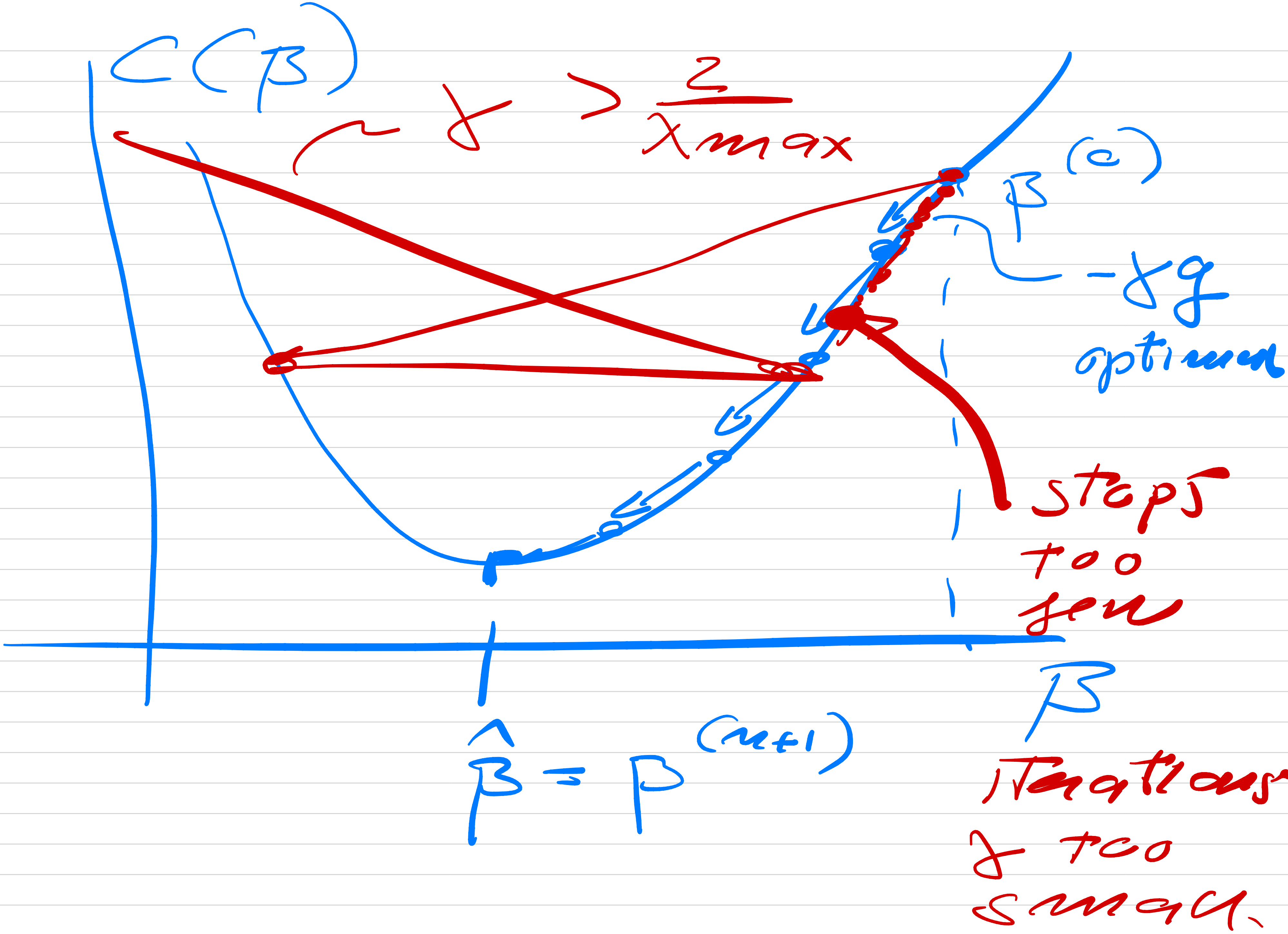
$$x^{(0)} = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$$

vector

$$x^{(i+1)} = x^{(i)} - \alpha \nabla f(x^{(i)})$$

$$|x^{(i+1)} - x^{(i)}| \leq \varepsilon \approx 10^{-5}$$

$$(10^{-8})$$


$C(\beta)$
$\gamma > \frac{2}{\lambda_{max}}$
$\beta^{(0)}$
$-\gamma g$
optimum
stops too few iterations
$\gamma$ too small.
$\beta$
$\hat{\beta} = \beta^{(n+1)}$

# GD with momentum

particle moving with friction and with force $F(x) = -\vec{\nabla} V(x)$

Newton's eq. of motion!

$$m \frac{d^2 x}{dt^2} + \underbrace{\mu \frac{dx}{dt}}_{\text{Friction}} = \underbrace{-\nabla V(x)}_{\text{think of as our gradient}}$$

Discretize 1st & 2nd derivatives

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

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

Define

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

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

$$\Delta x_{t+\Delta t} = \underbrace{\frac{-(\Delta t)^2}{m+\mu\Delta t}}_{\sim\gamma} \vec{\nabla} V(x) + \underbrace{\frac{m}{m+\mu\Delta t}}_{\sim\delta} \Delta x_t$$

$$\lim_{\mu\to 0} \delta = 1 \quad \wedge \quad \lim_{\mu\to\infty} \delta = 0$$

$$m, \mu, \Delta t > 0 \implies \delta \in [0,1]$$

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

$$x_t \Rightarrow \beta^{(i)}$$

$$x_{t+\Delta t} \Rightarrow \beta^{(i+1)}$$

$$x_{t-\Delta t} \Rightarrow \beta^{(i-1)}$$

$$\vec{\nabla} V \Rightarrow g(\beta^{(i)})$$

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

Algorithm

fix initial guess for $\beta^{(0)}$

fix momentum $\delta$

fix learning rate $\gamma$

initialize $v^{(i)}$

while stopping criterion ~~net~~ met

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

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

end while