FYS-STK3155/4155 week 37, September 8-12, 2025

FYS-STK3155/4155 week 37, Sept. 8-12

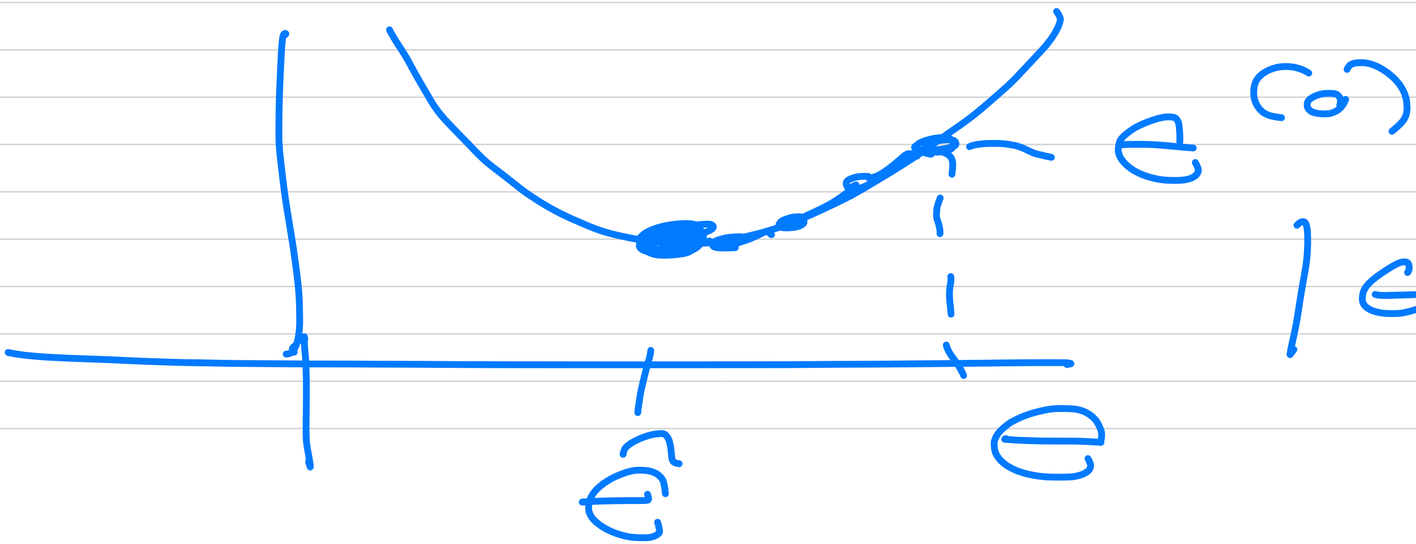
Gradient descent methods

- Plain GD
- Momentum GD
- other simple updates of learning rates
- ADAgrad, RMSprop, ADAM
- Stochastic GD
- Examples

What did we obtain last week?

$$\theta^{(n+1)} = \theta^{(n)} - \left(H(\theta^{(n)})\right)^{-1} \nabla_\theta C(\theta^{(n)})$$

- start with a guess $\theta^{(0)}$



$$|\theta^{(n+1)} - \theta^{(n)}| \le \varepsilon \approx 10^{-5}$$

$$\left(H^{(n)}\right) = \frac{\partial^2 C\left(\theta^{(n)}\right)}{\partial^2 \theta} \to$$

$$\eta^{(n)}$$

learning rate

$$\theta^{(n+1)} = \theta^{(n)} - \eta \nabla_\theta C\left(\theta^{(n)}\right)$$

plain/simple Gradient descent (GD)

Gradients

OLS

$$\nabla_\theta C = \frac{2}{n}\left(X^T X\theta - X^T y\right)$$

$X \in \mathbb{R}^{n \times p}$ $\qquad \theta \in \mathbb{R}^p$

$y \in \mathbb{R}^n$

Ridge

$$\nabla_\theta C = \frac{2}{n}\left(X^T X\theta - X^T y\right) + \lambda \cdot 2\theta$$

LASSO

$$\frac{2}{n}\left(X^T X\theta - X^T y\right) + \lambda\,\mathrm{sgn}(\theta)$$

$$\theta^{(n+1)} = \theta^{(n)} - \eta\, g^{(n)}$$

$\nabla_\theta C(\theta^{(n)})$

Taylor-expand around

keep only terms to 2nd derivative

$$C(\epsilon^{(n+1)}) = C(\epsilon^{?})$$

$$= C(\epsilon^{(n)}) + g^{(n)^T}(\epsilon^{(n)} - \eta g^{(n)})$$

$$+ \frac{1}{2}(\epsilon^{(n)} - \eta g^{(n)})^T H^{(n)}$$

$$\times (\epsilon^{(n)} - \eta g^{(n)})$$

optimal η ?

$$\frac{dC}{d\eta} = 0 \implies$$

$$-g^{T(n)} g^{(n)} + \eta\, g^{T(n)} H^{(n)} g^{(n)} = 0 \quad \Rightarrow$$

$$\eta^{(n)} = \frac{g^{T(n)} g^{(n)}}{g^{T(n)} \underbrace{H^{(n)} g^{(n)}}}$$

$$H^{(n)} g^{(n)} = \lambda g^{(n)}$$

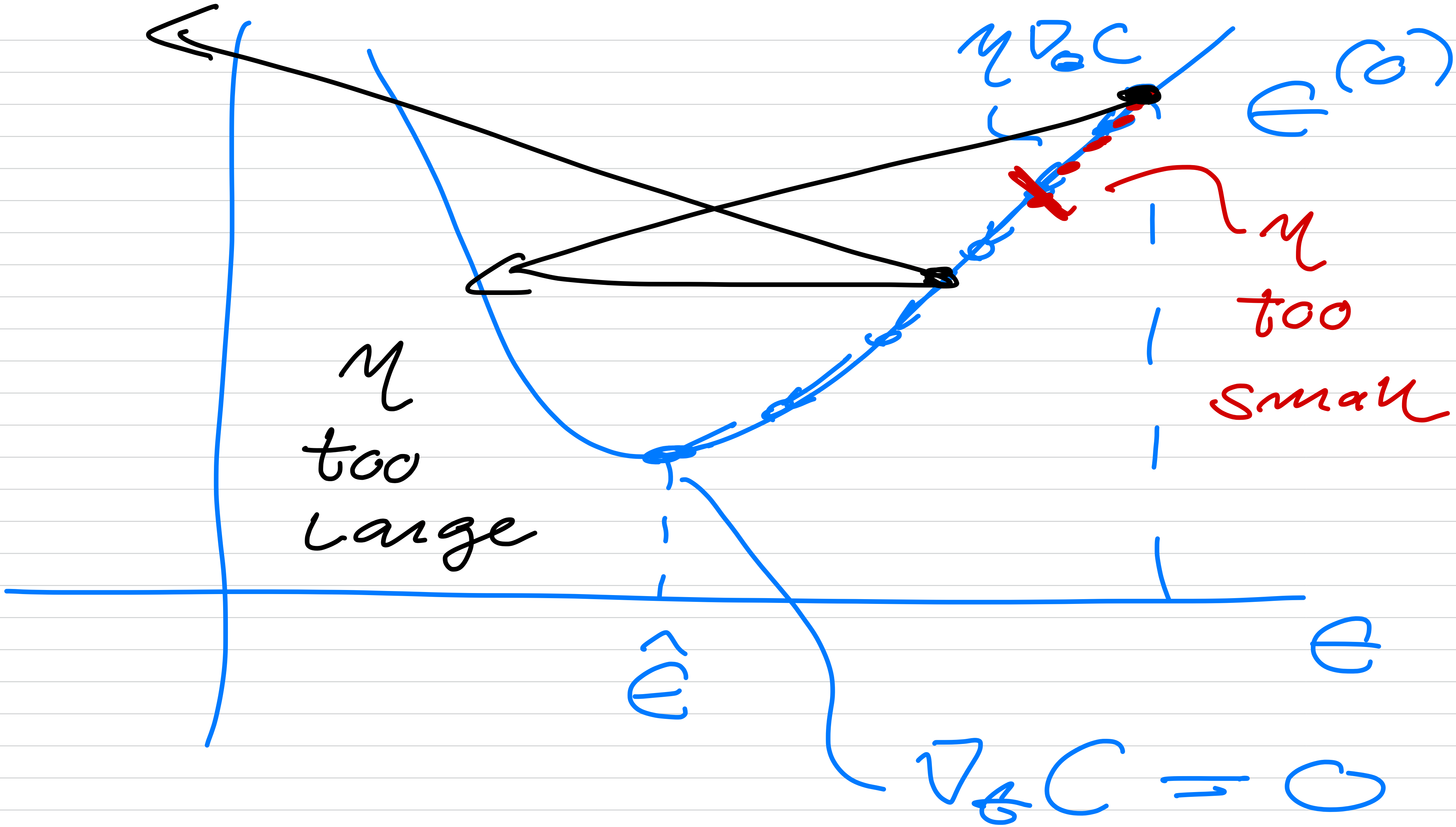
$$\eta^{(n)} = \frac{g^T g}{\lambda g^T g} = \frac{1}{\lambda}$$

η - requirement

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

Largest eigenvalue

of $H^{(n)}$



$\eta \nabla_\theta C$
$\theta^{(0)}$
η too small
η too large
$\hat{\theta}$
θ
$\nabla_\theta C = 0$

GD with momentum

Newton's eq. of motion

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

Discretize

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

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

Define

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

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

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

$$\Delta x_{t+\Delta t} = -\vec{g}\,\underbrace{\frac{\Delta t^2}{m+\mu\Delta t}}_{\eta} + \underbrace{\frac{m}{m+\mu\Delta t}}_{\delta}\Delta x_t$$

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

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

$$\Delta x_{t+\Delta t} = -\eta \vec{g} + \delta \Delta x_t$$

$$x_t \longrightarrow \theta^{(n)}$$

$$x_{t+\Delta t} \longrightarrow \theta^{(n+1)}$$

$$x_{t-\Delta t} \longrightarrow \theta^{(n-1)}$$

$$\theta^{(n+1)} = \theta^{(n)} - \eta\, g(\theta^{(n)}) + \delta\left[\theta^{(n)} - \theta^{(n-1)}\right]$$

momentum param

(memory)

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

algorithm:

fix initial guess $\theta^{(0)}$

fix $\eta^{(0)}$

fix momentum δ

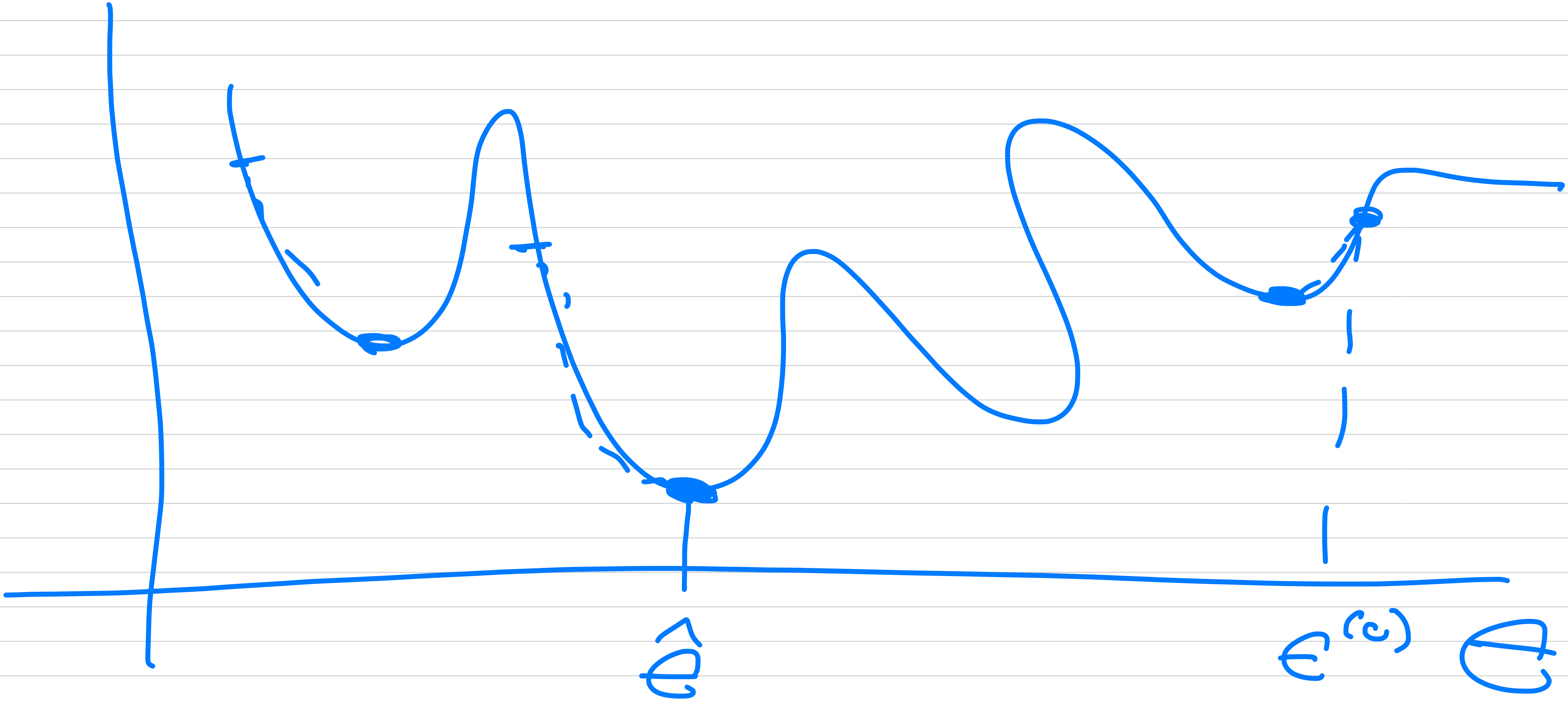
initialize vector $v^{(0)}$

while stopping criterion not met $\quad \theta^{(n)} + \alpha\,\theta^{(n-1)}$

— $v^{(n)} = \delta\left(\theta^{(n)} - \theta^{(n-1)}\right) - \eta\, g\left(\theta^{(n)}\right)$

— $\theta^{(n+1)} = \theta^{(n)} + v^{(n)}$

end while



θ̂
θ(0)
θ

cheap ways to update η

- η constant

- exponential decay

$$\eta^{(k)} = \eta^{(0)} \exp(-k\gamma_\eta)$$

$$\gamma_\eta \sim \eta^{(0)}/100 \text{ or similar.}$$

parameters

- linear

$$\eta^{(k)} = (1-\alpha)\,\eta^{(0)} + \alpha\,\delta_\tau$$

$$\delta_\tau \approx \eta^{(0)}/100$$

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