FYS 4411/9411 FEBRUARY 24, 2022

Gradient methods

$$\mathbb{E}[E_L(\vec{\alpha})] = \int_{\vec{R}\in D} d\vec{R}\; P(\vec{R};\vec{\alpha})\, E_L(\vec{R};\vec{\alpha})$$

$$\mathbb{E}\left[\frac{d\psi_T}{d\vec{\alpha}}\Big/\psi_T\right] = \int_{\vec{R}\in D} d\vec{R}\; P(\vec{R};\vec{\alpha})\,\frac{d\ln\psi_T}{d\vec{\alpha}}$$

$$\mathbb{E}\left[\frac{d\ln\psi_T}{d\vec{\alpha}}\, E_L(\vec{R};\vec{\alpha})\right]$$

$$\vec{\alpha}^{\,opt} = \underset{\vec{\alpha}\in\mathbb{R}^M}{\operatorname{argmin}}\; \mathbb{E}[E_L(\vec{\alpha})]$$

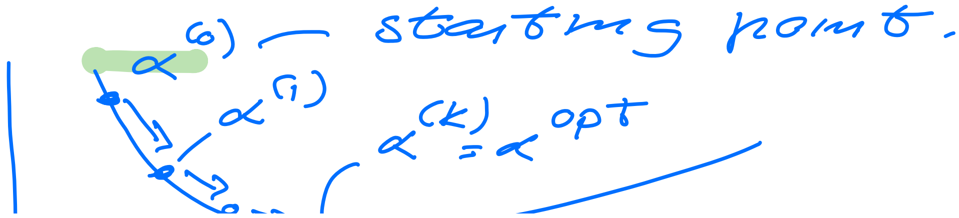
Simplest possible implementation:

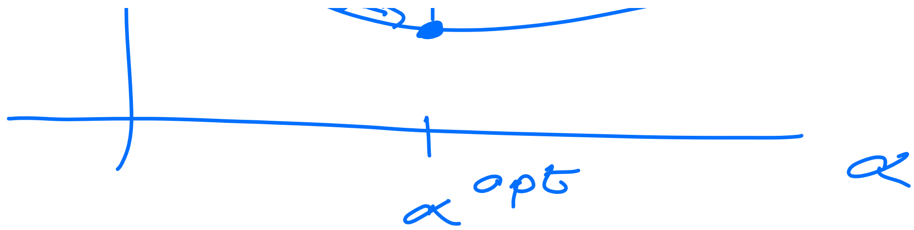
$$\vec{\alpha}^{(k+1)} = \vec{\alpha}^{(k)} - \gamma\, \vec{\nabla}_{\alpha}\, \mathbb{E}[E_L]_{\vec{\alpha}^{(k)}}$$

iteration

parameter (learning rate)

1-dim in $\vec{\alpha}$





stop when

$$\| \alpha^{(k+1)} - \alpha^{(k)} \|_2 \leq \delta$$

$$\delta \sim 10^{-8} - 10^{-10}$$

Gradient algorithms

- Steepest descent
- Newton-Raphson
- Conjugate gradient descent
- Gradient descent
- Approx to Newton-Raphson (Broyden et al)
- stochastic gradient descent

Belong to the class of convex optimization problems

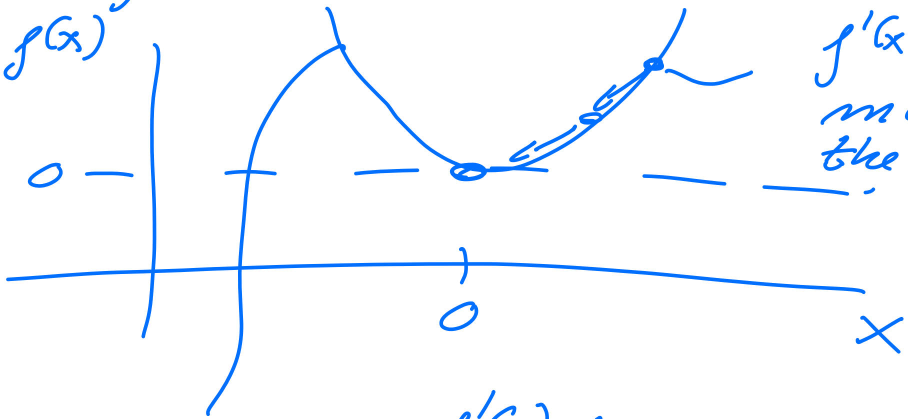
$$\mathbb{E}[E_L(\vec{\alpha})] \longrightarrow f(\vec{x})$$

$$\vec{\nabla}_{\vec{x}} f(\vec{\alpha}) = \vec{\nabla} f(\vec{\alpha}) = 0$$

want. the roots of $\vec{\nabla} f(\vec{x})$

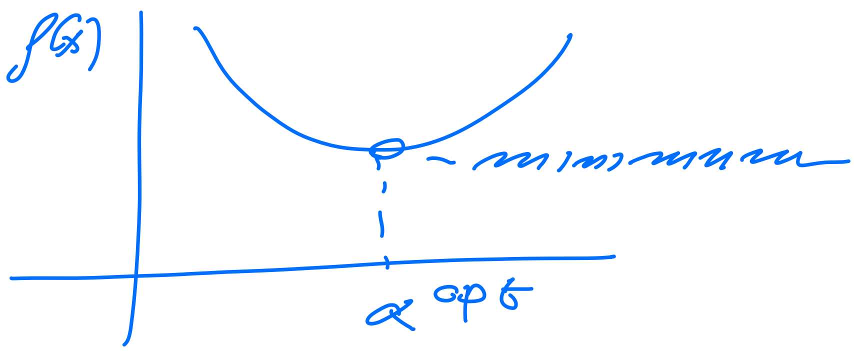
$$f(x) = 1/2\, x^2$$

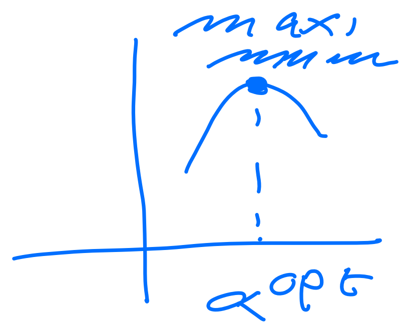
$f'(x) = x$

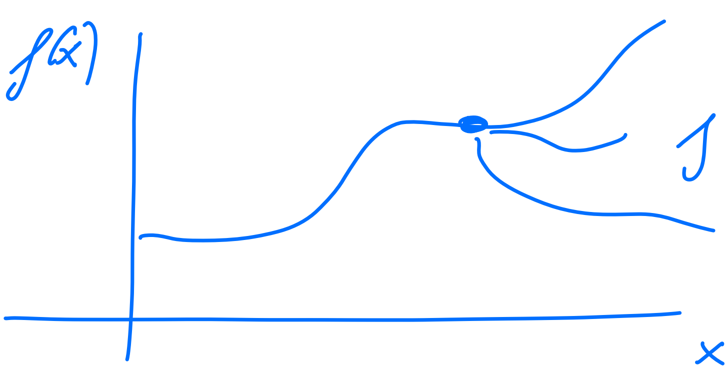


$x > 0$
$f'(x) > 0$
move to the left

$x < 0$; $f'(x) < 0$
$f(x)$ is decreased by moving to the right

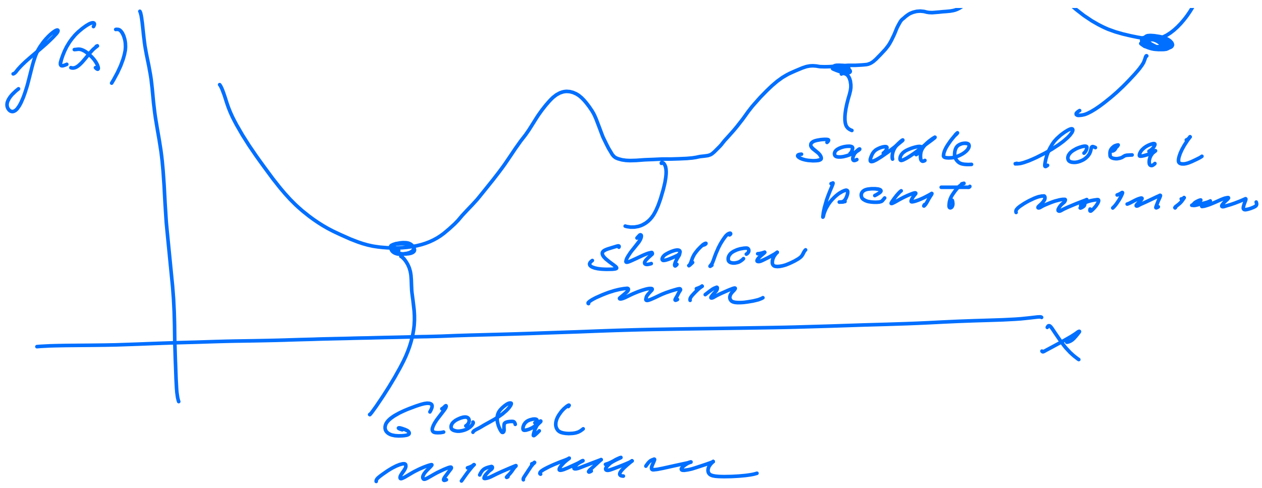






$f'(x) = 0$
saddle point

General case



$$\frac{\partial^2 f}{\partial x_i \partial x_j} = H_{ij}$$ matrix elements of Hessian matrix

In most cases H is positive definite, all eigenvalues are larger than zero $\Rightarrow$ convex optimization problem.

Taylor expansion of $f(\vec{x})$

$$f(\vec{x}) \simeq f(\vec{x}_0) + (\vec{x} - \vec{x}_0)^T \cdot \vec{g}$$

$$\left(\vec{\nabla} f(\vec{x}) = \vec{g} \right)$$

$$+ \frac{1}{2} (\vec{x} - \vec{x}_0)^T H (\vec{x} - \vec{x}_0)$$

$$\vec{x} - \vec{x}_0 \simeq -\gamma \vec{g} \qquad \gamma = \text{parameter.}$$

$$f(\vec{x}) = f(\vec{x}_0) - \gamma \vec{g}^T \vec{g} + \frac{1}{2} \gamma^2 \vec{g}^T H \vec{g}$$

optimal γ

$$\frac{d f(\vec{x})}{d \gamma} = 0 \Rightarrow$$

$$\vec{g}^T \vec{g} = \gamma \vec{g}^T H \vec{g} \Rightarrow$$

$$\gamma^{opt} = \frac{\vec{g}^T \vec{g}}{\vec{g}^T H \vec{g}} \quad \text{(steepest descent)}$$

Newton-Raphson:

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \frac{1}{H(f(x^{(k)}))} \times \vec{\nabla} f(x^{(k)})$$

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \gamma \cdot \vec{\nabla} f(x^{(k)})$$

Gradient descent assumes γ can be seen as a tuning parameter

$\gamma \in [10^{-5}, 10^{-4}, 10^{-3}, \ldots 10^{-1}]$

- in first approximation use GD with a parameter γ chosen between $10^{-4} - 10^{-1}$

- next step
  - stochastic gradient descent
  - Broyden et al (Newton-Raphson based)

- Steepest descent & conjugate gradient methods
- Essence: being able to evaluate H.

$H \to A$ (matrix $\in \mathbb{R}^{n \times n}$)

$Ax = b$ $\quad\quad$ $x^T A x$ is > 0

A is a positive definite matrix

The problem of solving $Ax = b$ is equivalent to the problem of optimizing the quadratic function

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

$$\frac{dq}{dx} = 0 \Rightarrow Ax = b$$

$$Ax^{(k+1)} = b$$

$x^{(0)}$ is a starting guess

if A is a positive definite matrix, then starting with a random $x^{(0)}$ and iterating will always lead, after k-iterations, to the "exact" value x which solves $Ax = b$

$$\| x^{(k+1)} - x^{(k)} \|_2 \leq \delta$$

Steepest descent

Steepest descent

$q(x)$ minimized

$Ax = b$

Define a residual

$r = b - Ax$

$Ax^{(0)} = b$

$r^{(0)} = b - Ax^{(0)}$

$r^{(k+1)} = b - Ax^{(k+1)}$

to be continued

optimal residuals involve

$$\frac{r^{T(k)} r^{(k)}}{r^{T(k)} A r^{(k)}} = \gamma^{(k)}$$