262B-Lecture 2

Date created: 2021.01.21
N. of Pages: 15

$\min f(x) \longrightarrow x_*$: local min $\longrightarrow \nabla f(x_*) = 0$

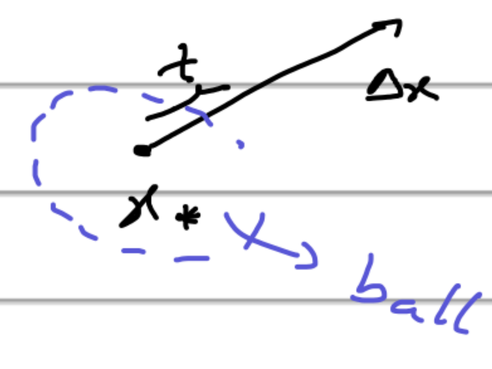
FOC

Proof 1 $\longrightarrow$ Proof 2 $\longrightarrow$ Proof 3

$g(t) = f(x_* + t\Delta x)$

$g'(0) = \lim_{t \to 0} \frac{f(x_* + t\Delta x) - f(x_*)}{t}$ ($f(x_* + t\Delta x) - f(x_*) \geq 0$, $t > 0$) ≥ 0

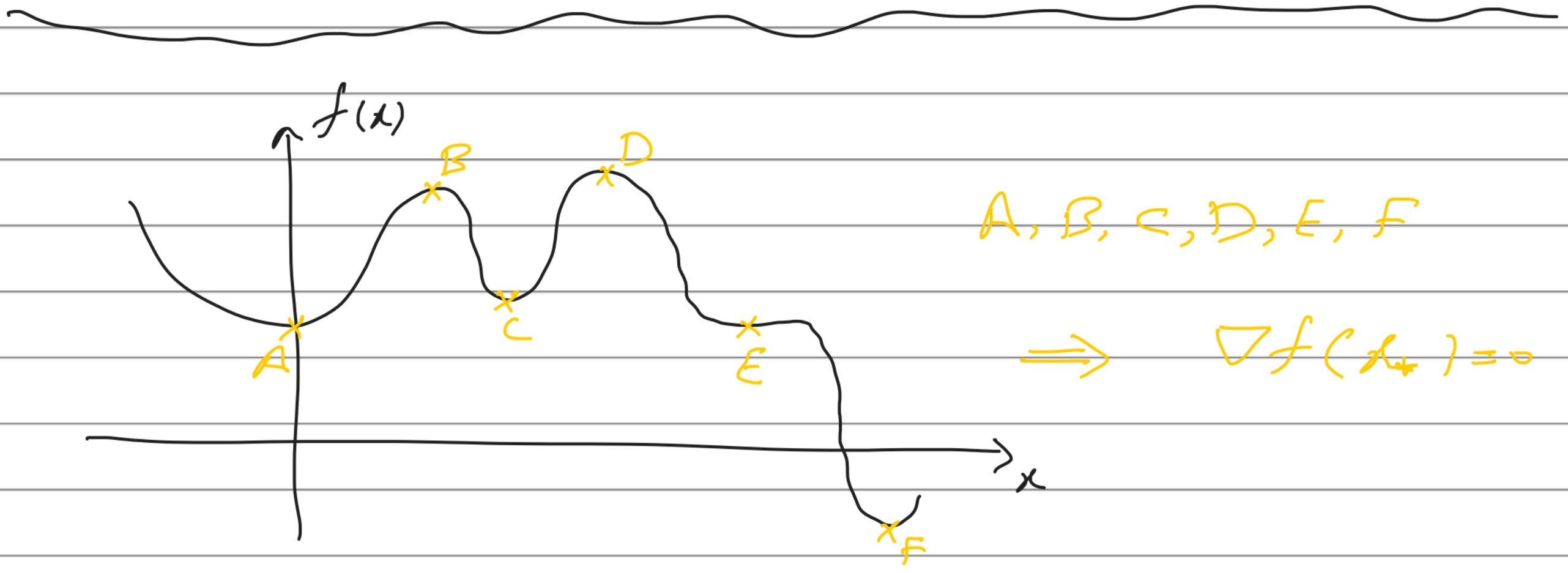
$t \downarrow 0$



$\Rightarrow \quad 0 \leq g'(0) = \nabla f(x_*)^T \Delta x \quad \forall \Delta x$

$\Delta x = -\nabla f(x_*) \Rightarrow 0 \leq -\|\nabla f(x_*)\|^2$

$\Rightarrow \|\nabla f(x_*)\| = 0 \Rightarrow \nabla f(x_*) = 0$



A, B, C, D, E, F

$\Rightarrow \nabla f(x_*) = 0$

SOC (necessary): second-order necessary condition

Thm: If x_* is a local min $\Rightarrow \nabla^2 f(x_*) \succeq 0$ (PSD)

$A \succeq 0$ PSD (positive semi-definite)
$A \succ 0$ PD
$A \preceq 0$ NSD
$A \prec 0$ ND

Proof: (by contradiction)

If $\nabla^2 f(x_*) \not\succeq 0 \Rightarrow$

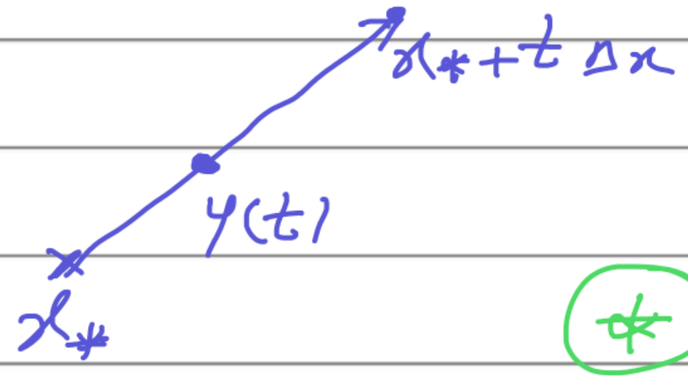
$\exists \Delta x : \Delta x^T \nabla^2 f(x_*) \Delta x < 0$

$$f(x_* + t\Delta x) = f(x_*) + t \nabla f(x_*)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(y(t)) \Delta x \quad (*)$$

(x_*: nominal, $t\Delta x$: perturbation)

$y(t)$: new point $\rightarrow$ depends on t

$t \to 0 : y(t) \to x_* \quad \preceq 0 \quad \nabla^2 f(y(t)) \to \nabla^2 f(x_*)$



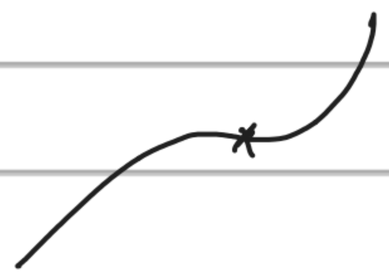
$\Rightarrow \Delta x^T \nabla^2 f(y(t)) \Delta x \to \Delta x^T \nabla^2 f(x_*) \Delta x < 0$

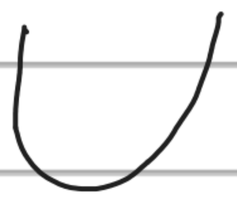
$(*) \Rightarrow f(x_* + t\Delta x) < f(x_*)$

new point; local min

$\times$

$\min x^2$, $\min x^3$, $\min x^4$





$f'(0)=0$ $f'(0)=0$ $f'(0)=0$

$f''(0)\geq 0$ $f''(0)\geq 0$ $f''(0)\geq 0$

$x_* = 0$ satisfies FOC, SOC

SOC (sufficient) : second-order sufficient condition

Thm: If x_* s.t. $\nabla f(x_*) = 0$, $\nabla^2 f(x_*) > 0$

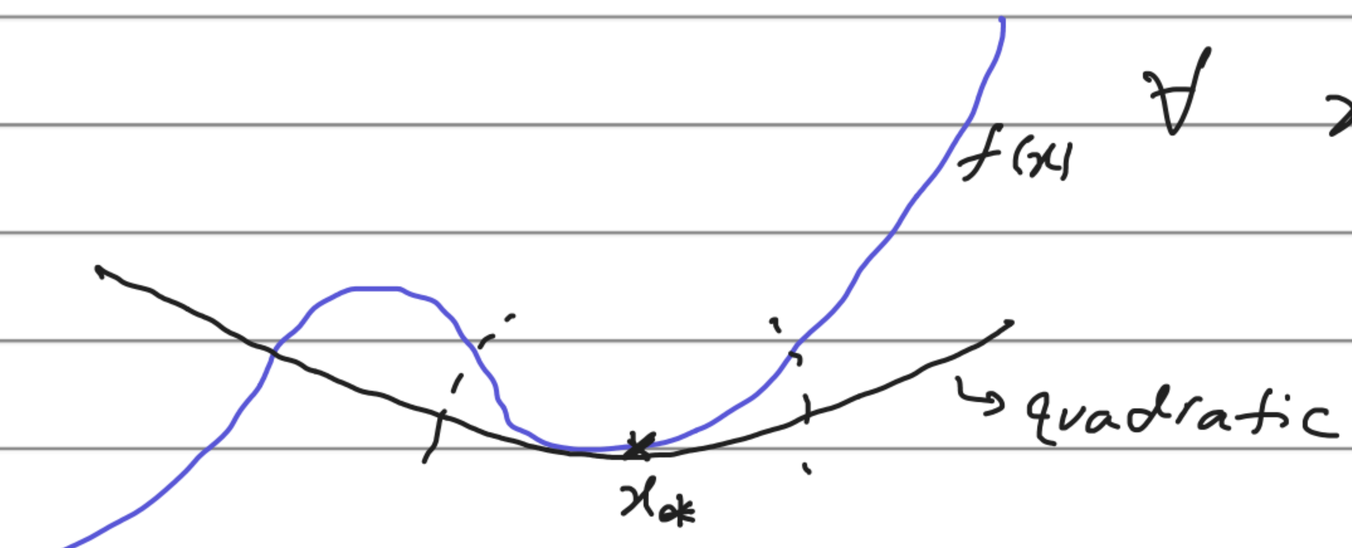
then: 1- x_* is a strict local min

2- $\exists\ \varepsilon > 0, \mu > 0$ s.t.

$$f(x) \geq f(x_*) + \mu \|x - x_*\|^2$$

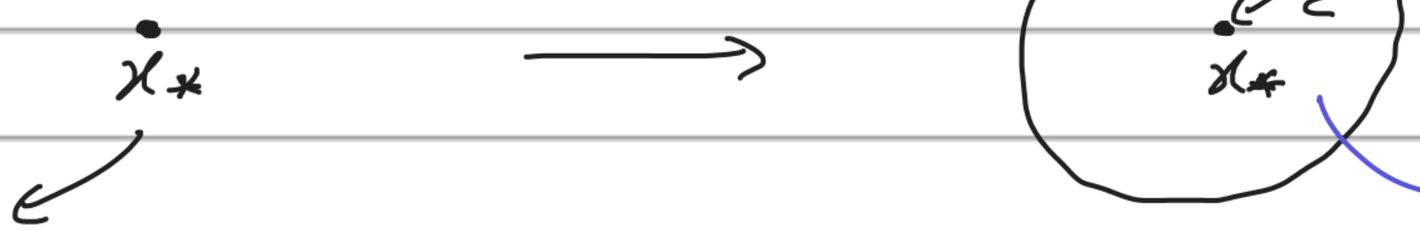
quadratic in x

$\forall\ x \in B$ → centered at x_* of radius ε



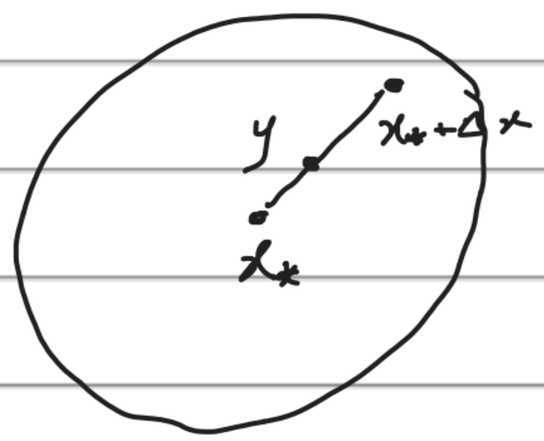
Proof: $\nabla^2 f(x_*) > 0 \implies \exists\ \epsilon, \lambda > 0$

s.t. $\min \text{eig}\left(\nabla^2 f(x)\right) \geq \lambda \quad \forall x : \|x - x_*\| \leq \epsilon$



Hessian is PD

Hessian at each point inside the ball is PD.



$$f(\underbrace{x_* + \Delta x}_{x}) = f(x_*) + \nabla f(x_*)\,\Delta x + \frac{1}{2}\Delta x^T \nabla^2 f(y)\,\Delta x$$

($\nabla f(x_*) = 0$; $\nabla^2 f(y)$: min eig $\geq \lambda$.)

$$\geq f(x_*) + \frac{\lambda}{2}\|\Delta x\|^2$$

$\Delta x^T A\, \Delta x \geq \min \text{eig}(A)\, \|\Delta x\|^2$

$$\implies f(x) \geq f(x_*) + \underbrace{\frac{\lambda}{2}}_{\mu}\|x - x_*\|^2$$

$\min f(x) \longrightarrow \underbrace{\nabla f(x_*) = 0}_{\text{FOC}}, \underbrace{\nabla^2 f(x_*) \geq 0}_{\text{SOC}}$

$\Longrightarrow$ How to find such point x_* ?

Descent methods : can't find x_* in one shot.

$\Longrightarrow$ Find it gradually

$x^{(0)} \longrightarrow x^{(1)} \longrightarrow x^{(2)} \longrightarrow \cdots \longrightarrow x^{(k)} \longrightarrow x^{(k+1)} \longrightarrow \cdots$

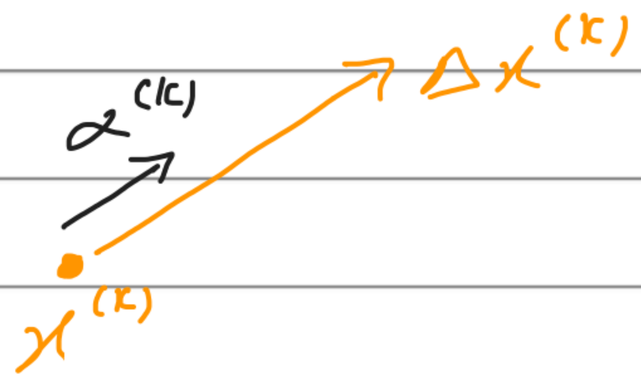
(generate a sequence of point)

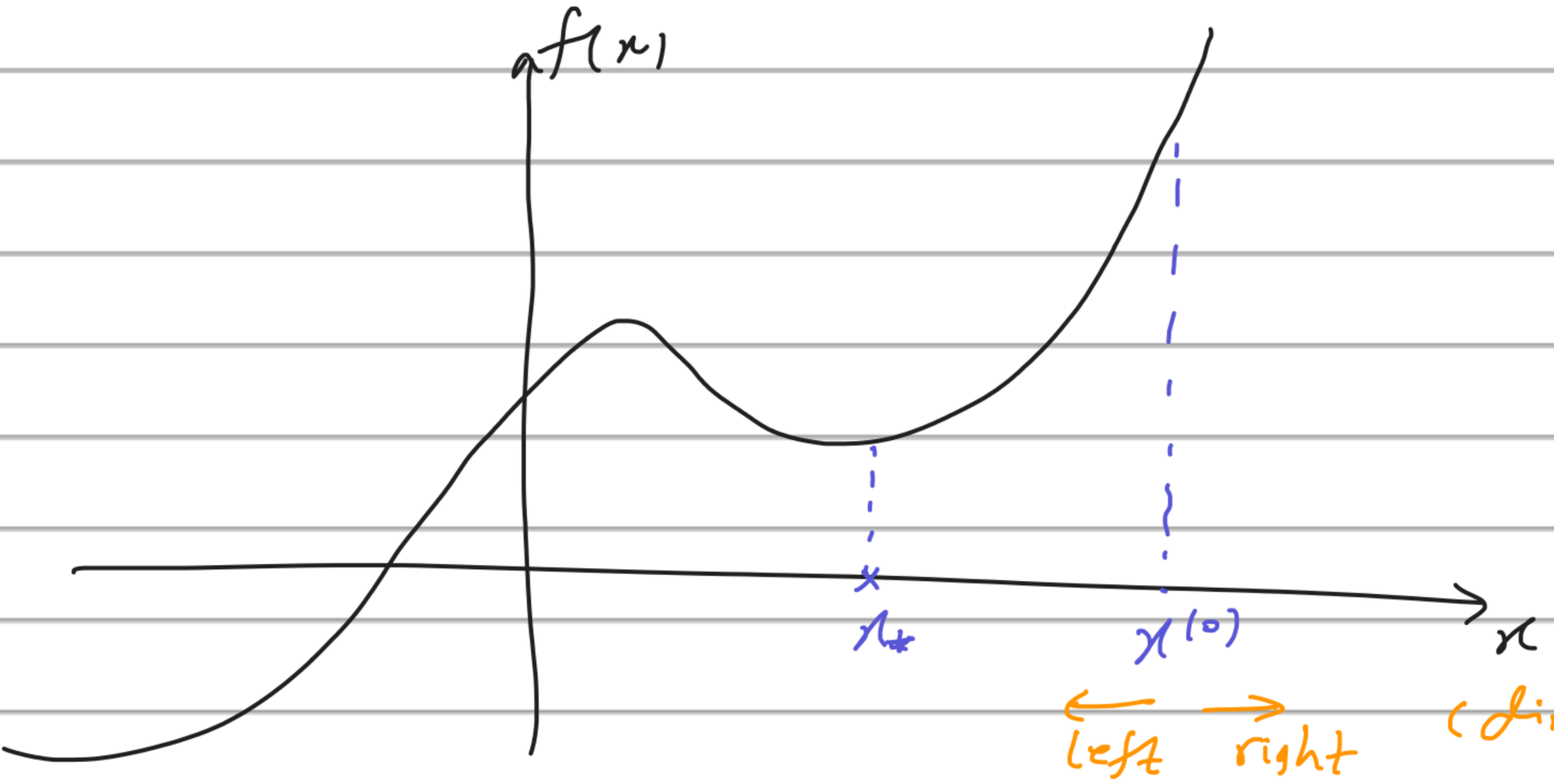
improvement at each iteration:

$f(x^{(0)}) > f(x^{(1)}) > f(x^{(2)}) > \cdots$

$x^{(k)} \longrightarrow x^{(k+1)}$ updating rule:

$$\underbrace{x^{(k+1)}}_{\text{new point}} = \underbrace{x^{(k)}}_{\text{old point}} + \underbrace{\alpha^{(k)}}_{\text{stepsize}} \underbrace{\Delta x^{(k)}}_{\text{direction}}$$



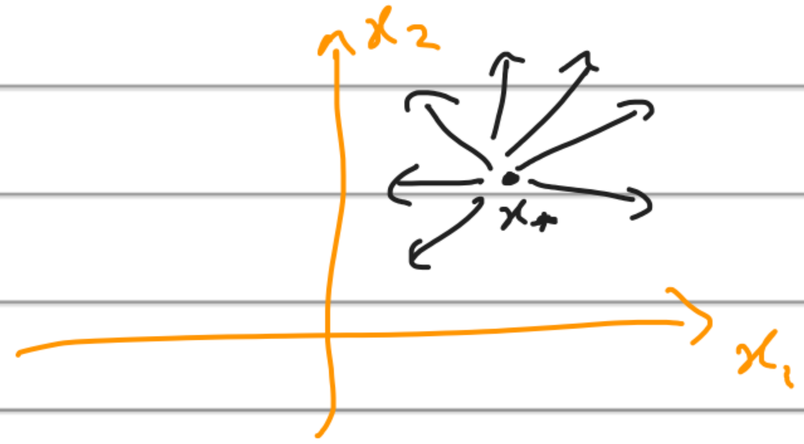


$n = 1$

← → (direction)

left right

← ← how much (stepsize)



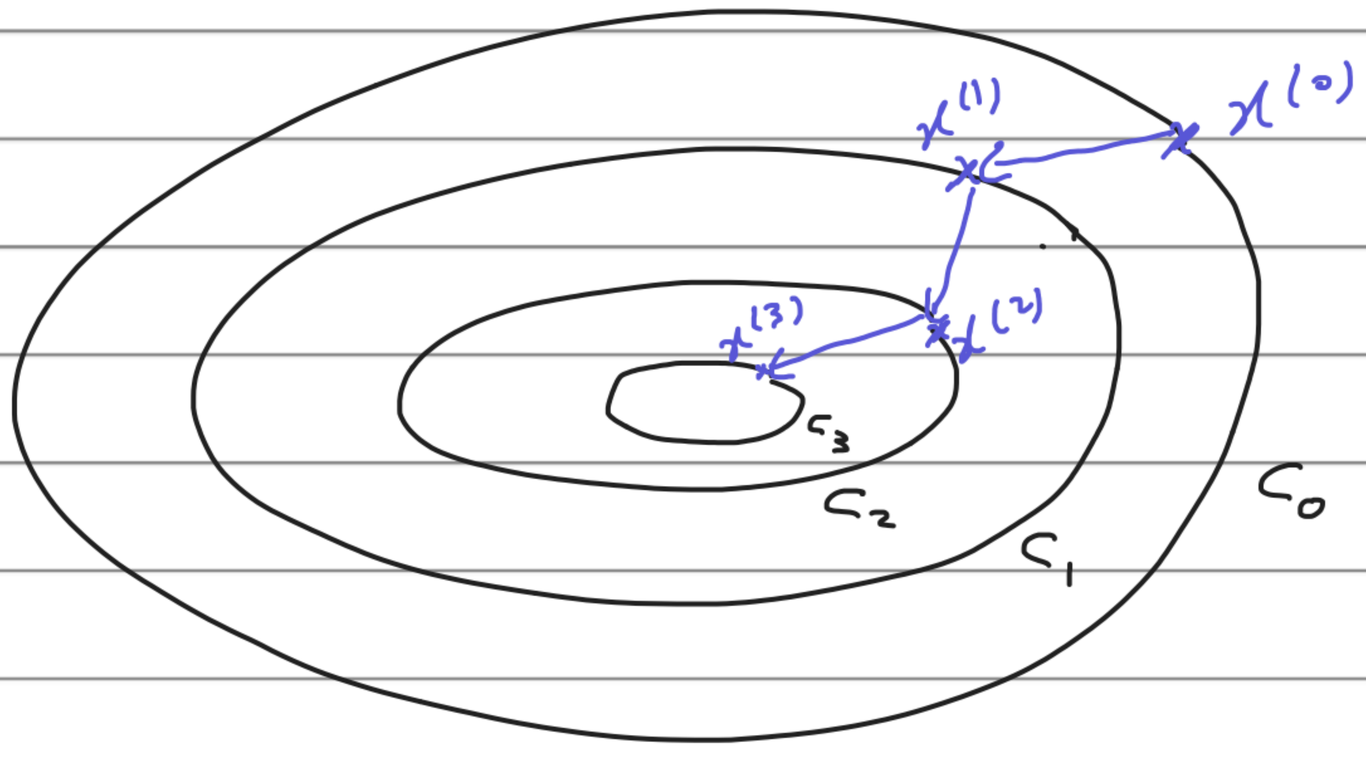
$n = 2$

Level set (c) $= \{ x \mid f(x) = c \}$

$\downarrow$

$\in \mathbb{R}$

$n = 2$



$$\underbrace{x^{(k+1)}}_{\text{new}} = \underbrace{x^{(k)}}_{\text{old}} + \underbrace{\alpha^{(k)} \Delta x^{(k)}}_{\text{perturbation}}$$

dominates

first term → rest

$$f(x^{(k+1)}) = f(x^{(k)}) + \alpha^{(k)} \nabla f(x^{(k)})^T \Delta x^{(k)} + \dots$$

improvement : $f(x^{(k+1)}) < f(x^{(k)})$

Thm: 1- If $\nabla f(x^{(k)})^T \Delta x^{(k)} > 0$, then

$\exists \; I > 0$ s.t.

$$f(x^{(k+1)}) > f(x^{(k)}) \quad \forall \alpha^{(k)} \in [0, I]$$

2- If $\nabla f(x^{(k)})^T \Delta x^{(k)} < 0$, then

$\exists \; I > 0$ s.t.

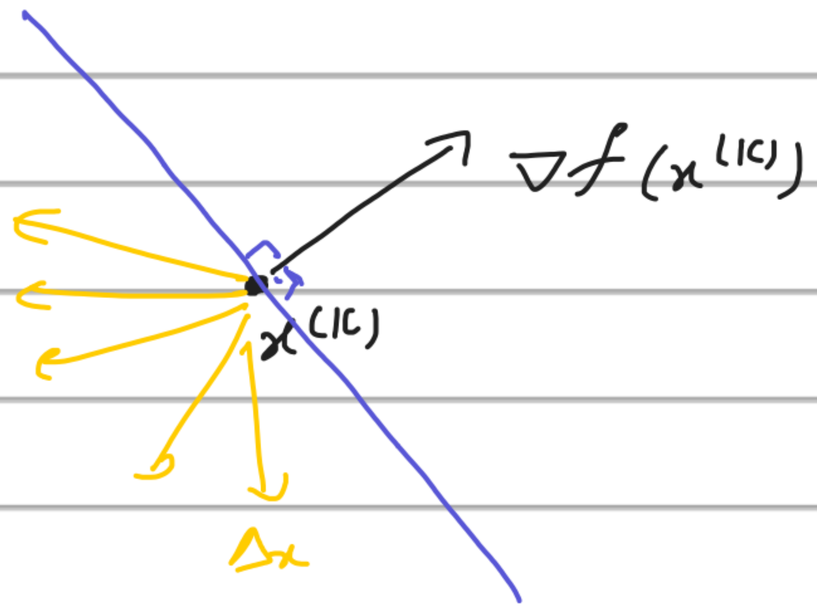
$$f(x^{(k+1)}) < f(x^{(k)}) \quad \forall \alpha^{(k)} \in [0, I]$$

3- If $\nabla f(x^{(k)})^T \Delta x^{(k)} = 0 \Rightarrow$

Look at sign $(\Delta x^{(k)})^T \nabla^2 f(x^{(k)}) \Delta x^{(k)}$

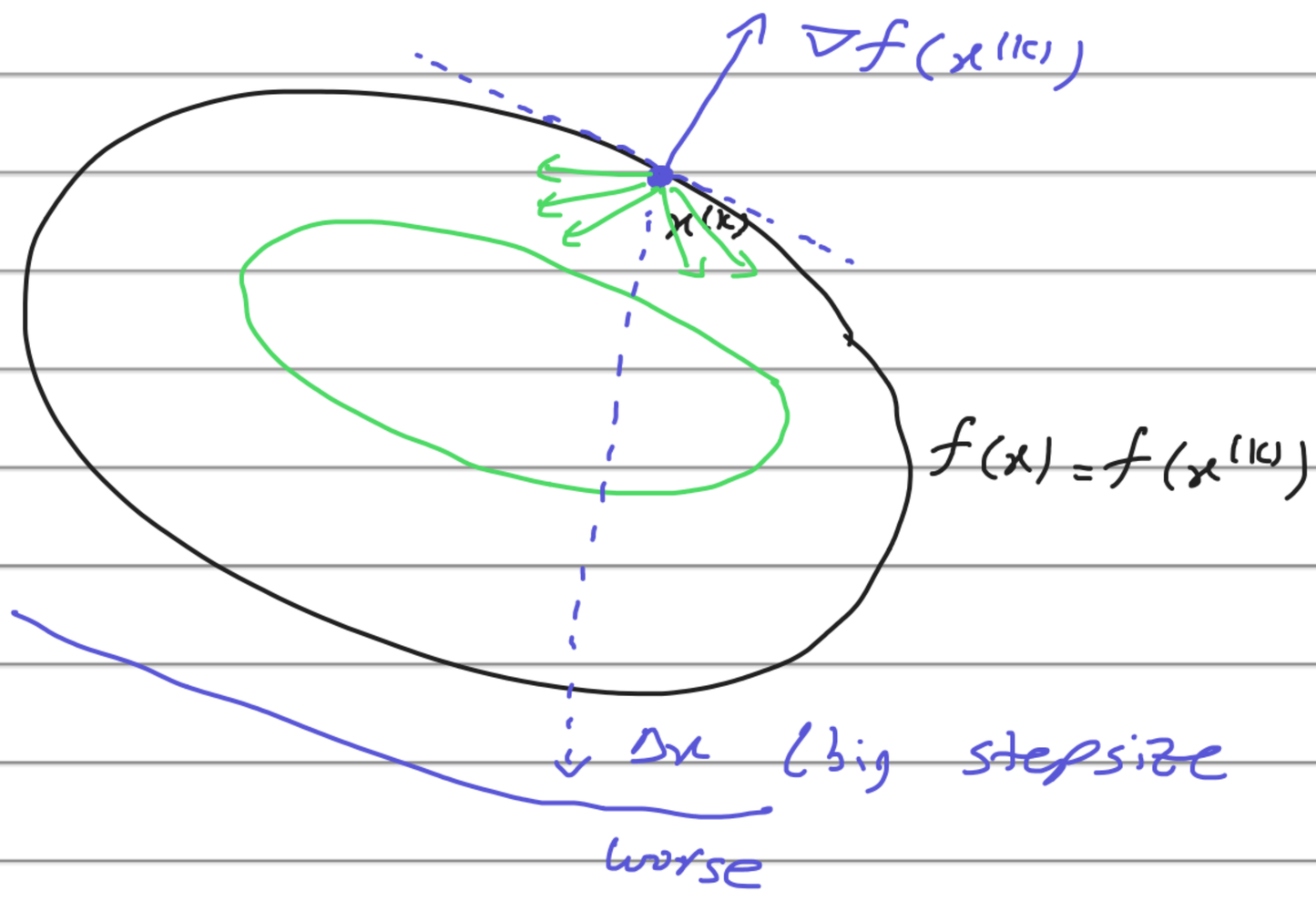
Descent direction: Δx is called descent

if $\nabla f(x^{(k)})^T \Delta x < 0$



angle between direction & gradient

$\Rightarrow$ There are infinitely many descent directions



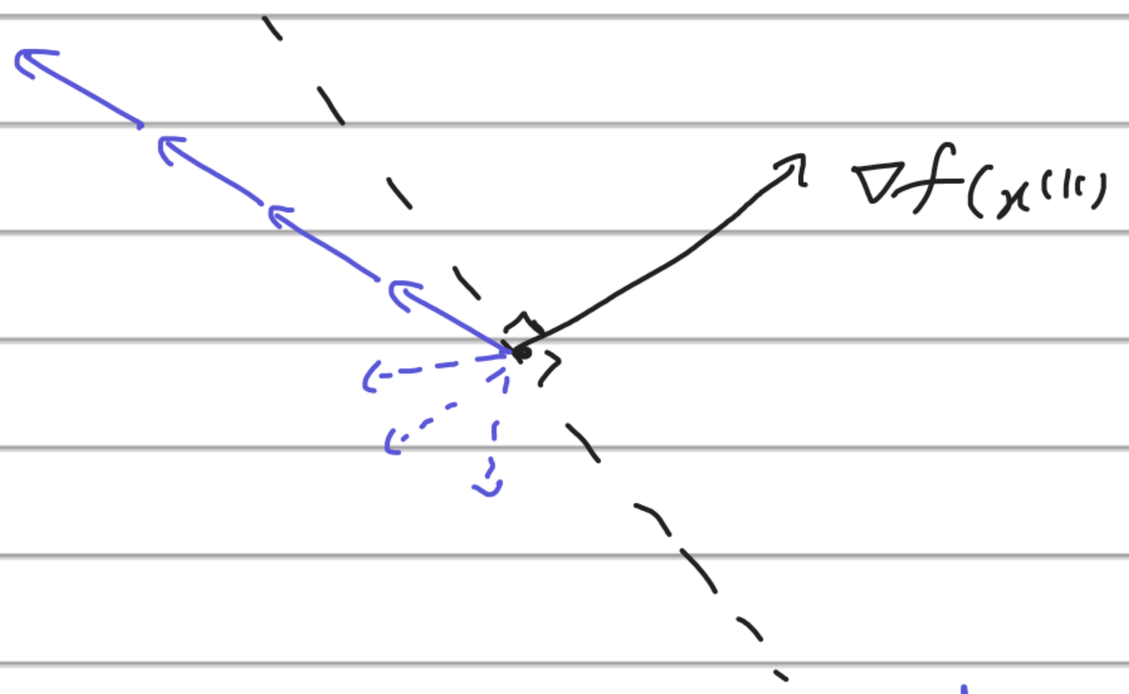
Design the best direction ?

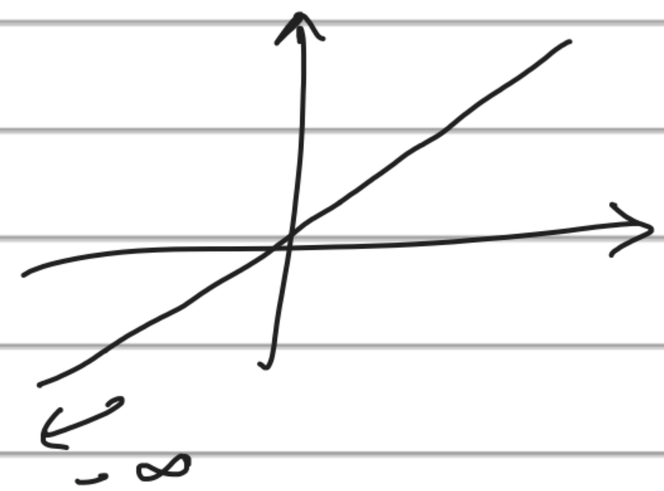
$$f(x^{(k+1)}) = f(x^{(k)}) + \alpha^{(k)} \nabla f(x^{(k)})^T \Delta x^{(k)} + \cdots$$

reduce — reduce

Best direction : $\min_{\Delta x} \nabla f(x^{(k)})^T \Delta x$

$\nabla f(x^{(k)})$ — linear





shouldn't confuse direction with stepsize

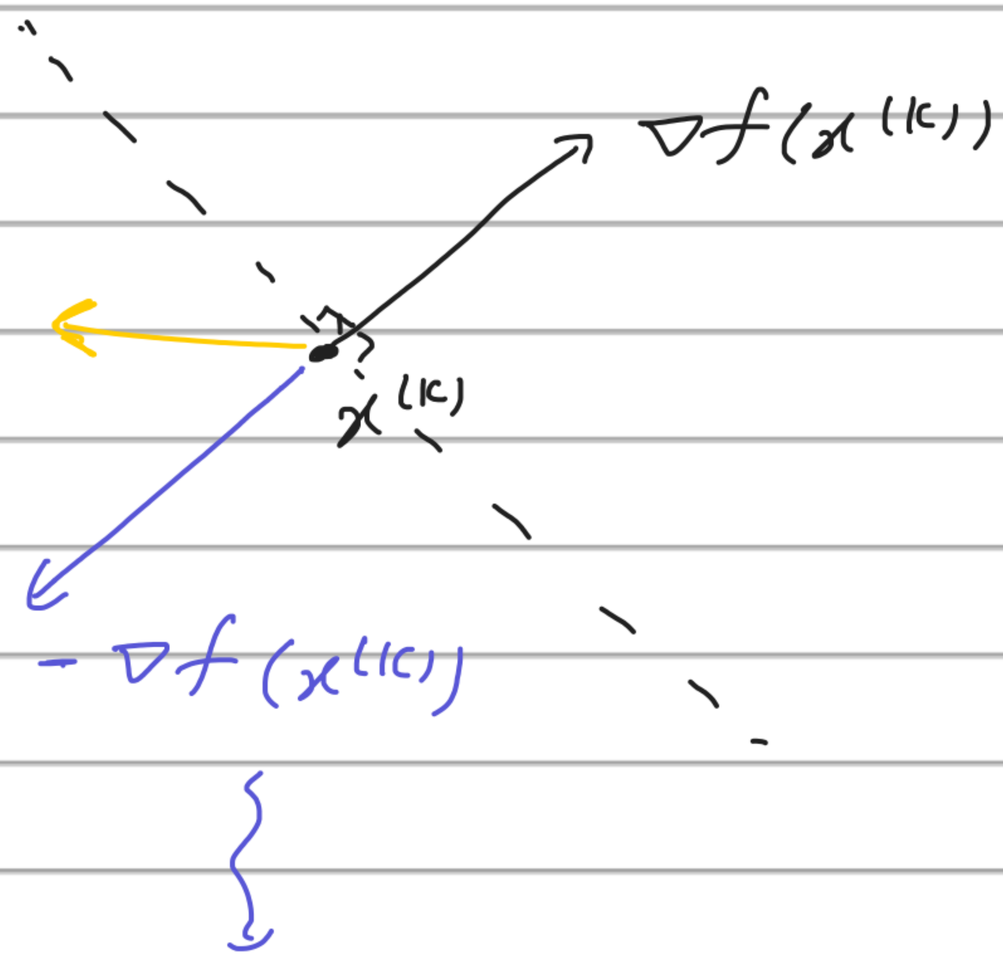
$\Longrightarrow$ ($\min_{\Delta x} \nabla f(x^{(k)})^T \Delta x$: s.t. $\|\Delta x\| \leq 1$)

$\|x\|_1 = \sum |x_i|$, $\|x\|_2 = (\sum x_i^2)^{1/2}$

$\|x\|_\infty = \max_i (|x_i|)$, ...

standard norm : length , $\|\Delta x\|_2$

$\Longrightarrow \Delta x = - \dfrac{\nabla f(x^{(k)})}{\|\nabla f(x^{(k)})\|} \xrightarrow{\text{proportional}} - \nabla f(x^{(k)})$



steepest descent (w.r.t. $\|\cdot\|_2$)

$$\Delta x^{(k)} = -\nabla f(x^{(k)})$$ (Gradient algorithm)

generalize

$$\Delta x^{(k)} = -\underbrace{D^{(k)}}_{PD} \nabla f(x^{(k)})$$

Thm: If $D^{(k)} > 0 \Rightarrow \Delta x^{(k)}$: descent direction

$$\nabla f(x^{(k)})^T \Delta x^{(k)} = -\underbrace{\nabla f(x^{(k)})^T}_{\text{vector}} \underbrace{D^{(k)}}_{\substack{\text{matrix} \\ > 0}} \underbrace{\nabla f(x^{(k)})}_{\text{vector}}$$

$$\leq 0$$

$$D^{(k)} = \nabla^2 f(x^{(k)})^{-1} \quad \text{if} \quad \nabla^2 f(x^{(k)}) > 0.$$

$$\Rightarrow \text{Newton's method}$$

old ✓ new ? old ✓ direction ?

$$x^{(k)} \longrightarrow x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$$

$$f(x^{(k+1)}) = f(x^{(k)}) + \nabla f(x^{(k)})\, \Delta x^{(k)} + \frac{1}{2} (\Delta x^{(k)})^T \nabla^2 f(x^{(k)})\, \Delta x^{(k)} + \cdots$$

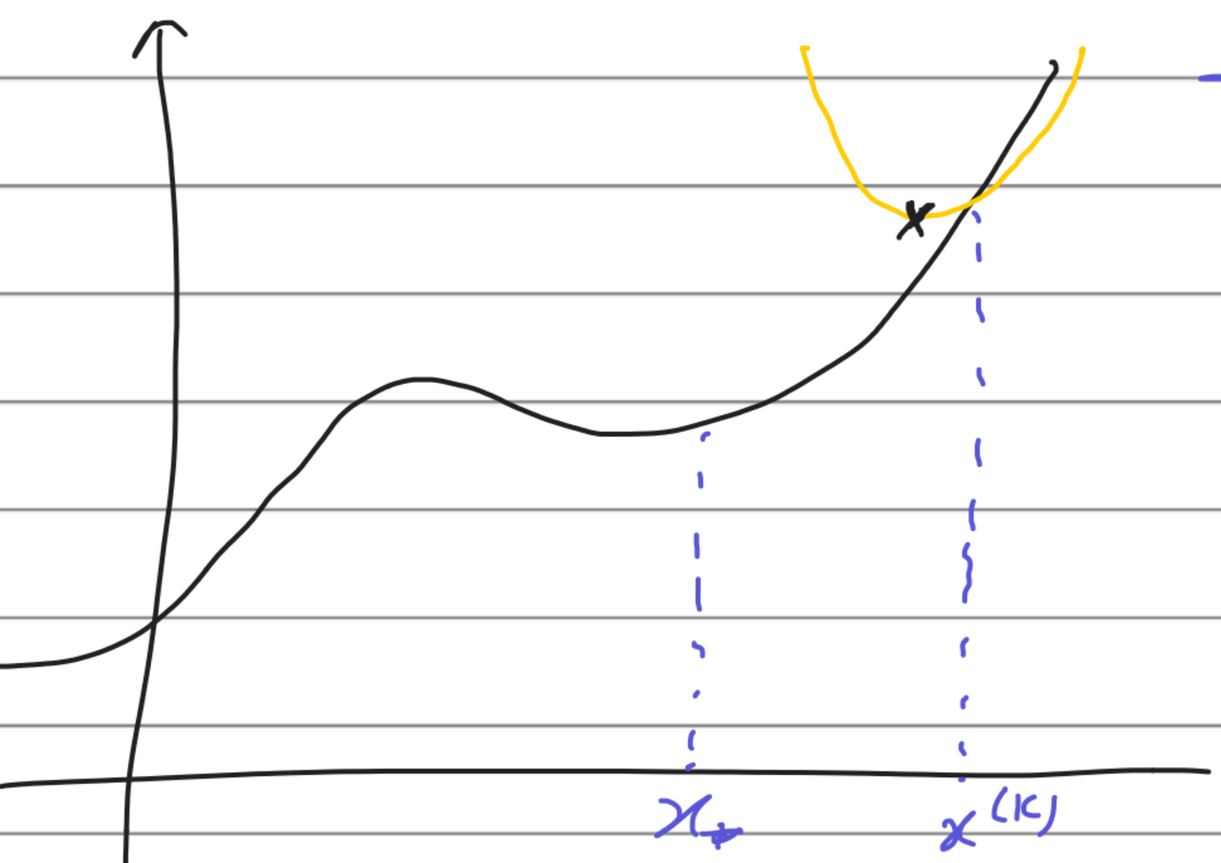
$$f(x^{(k+1)}) \overset{?}{<} f(x^{(k)})$$

minimize

$$\min_{\Delta x^{(k)}} f(\underbrace{x^{(k)} + \Delta x^{(k)}}_{\text{new point}}) = \min_{x} f(x)$$

approximate

$$\min_{\Delta x^{(k)}} \underbrace{f(x^{(k)})}_{r} + \underbrace{\nabla f(x^{(k)})}_{q^T} \Delta x^{(k)} + \frac{1}{2} (\Delta x^{(k)})^T \underbrace{\nabla^2 f(x^{(k)})}_{P} \Delta x^{(k)}$$



$$\min_{\Delta x} \; r + q^T \Delta x + \frac{1}{2} \Delta x^T P \Delta x$$

$$\downarrow \text{Gradient} = 0$$

$$\Delta x = -P^{-1} q$$

$$\Delta x^{(k)} = -P^{-1} g = -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$$

$$= -D^{(k)} \nabla f(x^{(k)})$$

$$D^{(k)} = I \quad \text{or} \quad \nabla^2 f(x^{(k)})^{-1}$$

Gradient — Newton

Gradient → slow (cheap)

between <- - - - ->

Newton → fast (expensive)

$A^{-1} \longrightarrow O(n^3)$

$$D^{(k)} = \text{diagonal} \quad \& \quad D^{(k)}_{ii} = \left(\frac{\partial^2 f(x^{(k)})}{\partial x_i \partial x_i} \right)^{-1}$$

↳ cheap

If $\nabla^2 f(x^{(k)}) > 0 \implies D^{(k)} > 0$

$D^{(0)}, D^{(1)}, \ldots D^{(p)}, D^{(p+1)}, D^{(p+2)}, \ldots$

$\nabla^2 f(x^{(0)})^{-1}$ $\qquad$ $\nabla^2 f(x^{(p)})^{-1}$

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$$

✓ ? ✓

How to find stepsize ?

1 - exact line search

2 - Limited Line search

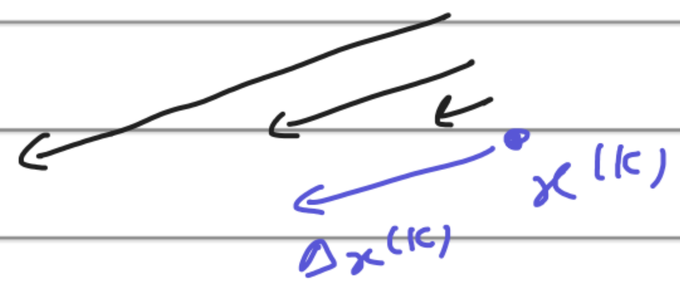
3 - backtracking

$f(x^{(k+1)}) < f(x^{(k)})$ ⟶ best improvement

reduce

$$\min_{\alpha} f(x^{(k)} + \alpha \Delta x^{(k)})$$

$$s.t. \quad \alpha \geq 0$$



⟶ $\alpha^{(k)}$ = optimal α

exact line search

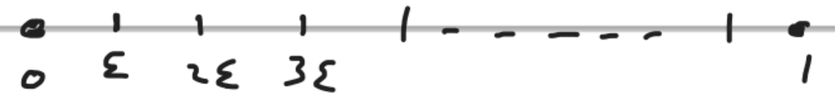
$$\min_{x} f(x) \longrightarrow x^{(0)} \longrightarrow x^{(1)} \longrightarrow x^{(2)} \longrightarrow \dots$$

$\min_{\alpha}$ $\min_{\alpha}$ $\min_{\alpha}$

$x \in \mathbb{R}^n$ ⟶ multi-variate opt

$\alpha \in \mathbb{R}$ ⟶ univariate opt

$$\min_{\alpha \in \mathbb{R}} g(\alpha) \quad \text{s.t.} \quad 0 \le \alpha \le 1$$



check $g(\cdot)$ at all grid points to find the best one. $\longrightarrow \frac{1}{\varepsilon}$ evaluations

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad 0 \le x_i \le 1$$

$\Longrightarrow$ gridding : $\left(\frac{1}{\varepsilon}\right)^n$ evaluations

$\varepsilon = 0.5$, $n = 400$ $\Longrightarrow$ # eval $\geq$

atoms in observable Universe

Limited line search:

$$\min f(x^{(k)} + \alpha \Delta x^{(k)}) \quad \text{s.t.} \quad 0 \le \alpha \le T$$

T ← threshold

Backtracking : No optimization

Pick $\alpha > 0$, $0 < \beta < 1$

α
$\alpha\beta$
$\alpha\beta^2$
$\alpha\beta^3$ ⟶ pick the first one
⋮ ↓
0 s.t. $f(\text{new point}) < f(\text{old point})$