262B-Lecture 4

Date created: 2021.01.28
N. of Pages: 12

$\min f(x)$: Algorithm

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

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

{ special : gradient

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

find ? → $g^{(k)}$: approximate gradient

$g^{(k)} = \nabla f(x^{(k)}) + \underbrace{e^{(k)}}_{\text{error}}$

$e^{(k)}$ is small relative to gradient

$\|e^{(k)}\| \leq \|\nabla f(x^{(k)})\| \quad \forall k$

downside: as we proceed, gradient gets smaller & so the error must be small.

$\Rightarrow$ Algorithm work!

$$\text{gradient} \cdot \text{direction} = \nabla f(x^{(k)})^T (-g^{(k)})$$

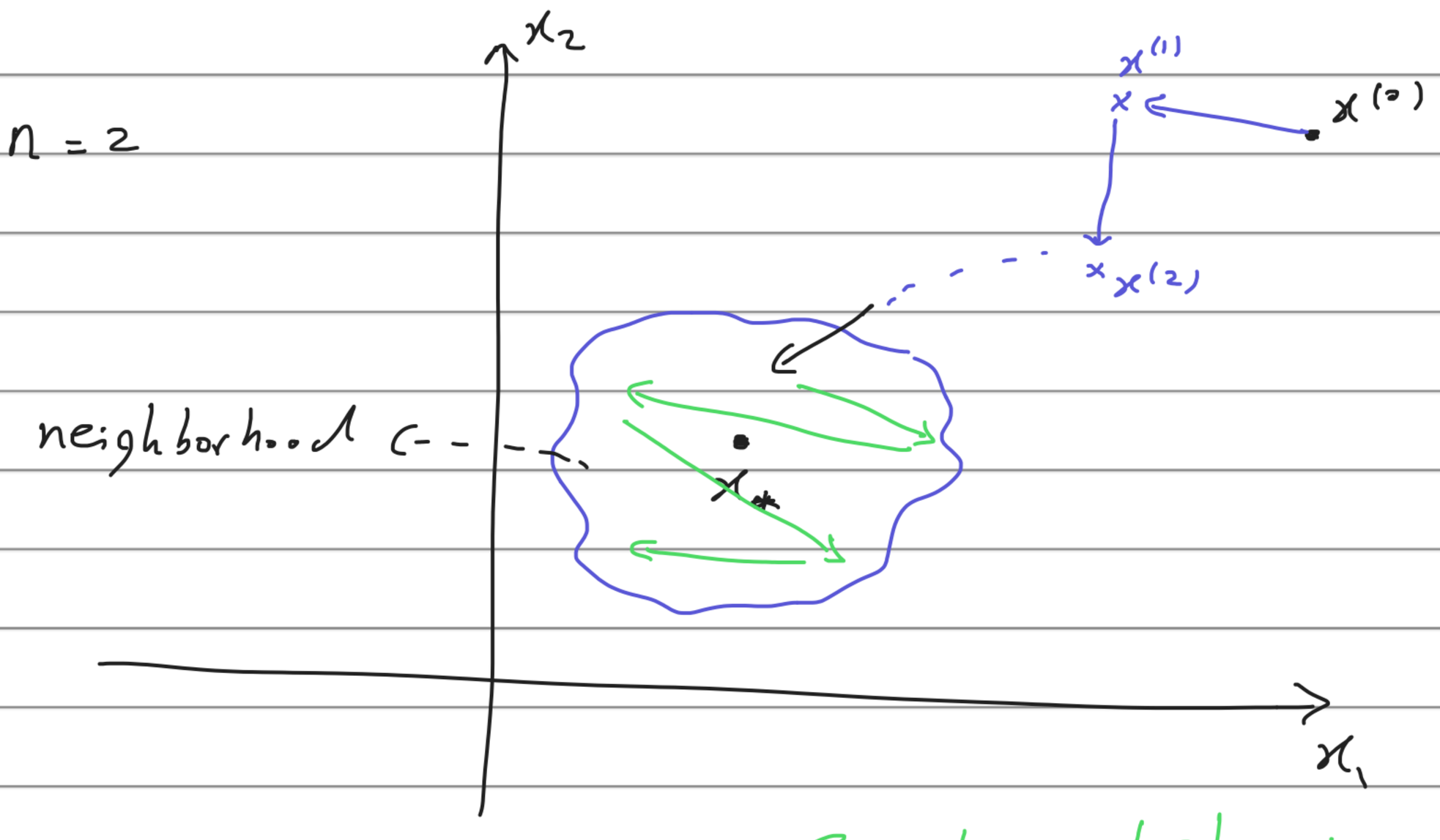
$$= -\| \nabla f(x^{(k)}) \|^2 - \nabla f(x^{(k)})^T e^{(k)}$$

$$\leq -\| \nabla f(x^{(k)}) \|^2 + \| \nabla f(x^{(k)}) \| \cdot \| e^{(k)} \|$$

$$-\| \nabla f(x^{(k)}) \| \left(+ \underbrace{\| \nabla f(x^{(k)}) \| - \| e^{(k)} \|}_{>0 \text{ by assumption}} \right)$$

$\leq 0 \Rightarrow$ descent direction

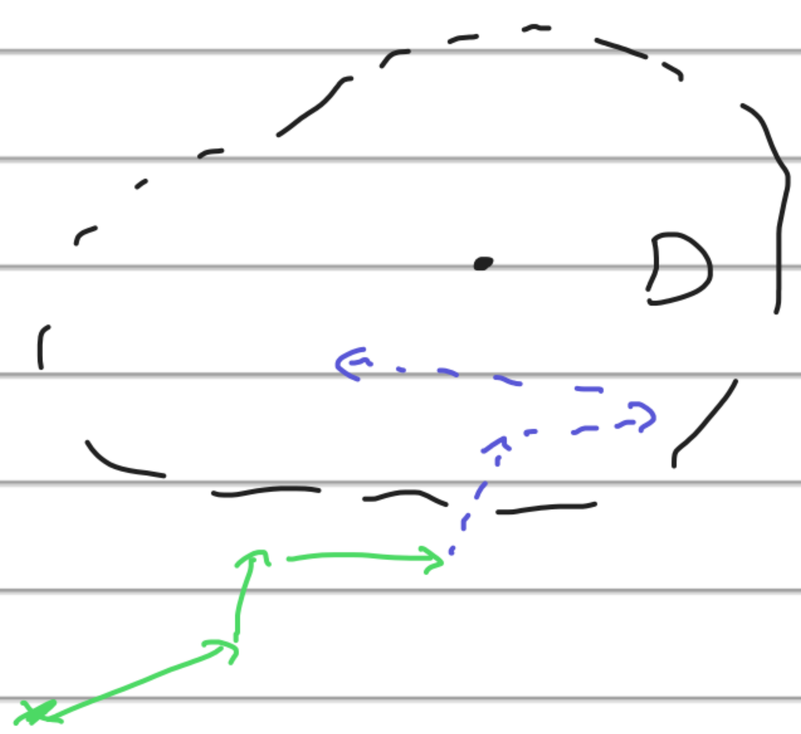
2. $e^{(k)}$ is bounded $\Rightarrow \underbrace{\| e^{(k)} \| \leq \delta \quad \forall k}$



erratic behavior

in the neighborhood

neighborhood: $\{x \mid \|\nabla f(x)\| \leq \delta\} = D$



outside of D:

descent algorithm

Assume: $x^{(k)} \notin D$

gradient . direction $= -\nabla f(x^{(k)})^T \underbrace{g^{(k)}}_{\nabla f(x^{(k)}) + e^{(k)}}$

$\leq -\|\nabla f(x^{(k)})\|^2 + \|\nabla f(x^{(k)})\| \times \delta$

$= -\|\nabla f(x^{(k)})\| \left(+ \underbrace{\|\nabla f(x^{(k)})\| - \delta}_{>0} \right)$

$\leq 0 \Rightarrow$ descent direction

$e^{(0)} \rightarrow e^{(1)} \rightarrow e^{(2)} \rightarrow \cdots$

special scenario: $e^{(k)} = \delta$

gradient alg.

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

perturbed gradient alg.

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

bias

$\min \quad f(x)$

$\min \quad f(x) + \underbrace{\delta^T x}_{\text{Linear}}$

3- SGD : stochastic gradient descent

$$g^{(k)} = \nabla f(x^{(k)}) + \underbrace{e^{(k)}}_{\text{random}}$$

$\min \quad f(x) \longrightarrow \mathbb{E}_w \{ F(x, w) \}$

uncertainty

(stochastic or simulation optimization)

$$\nabla f(x) = \mathbb{E}_w \{ \nabla_x F(x, w) \}$$

hard to compute in real world

$\longrightarrow$ simulate

Basic approach:

gradient : $\nabla f(x^{(k)}) \xrightarrow{\text{approximate}} \nabla_x F(x^{(k)}, \underbrace{w^{(k)}}_{\text{one sample}})$

$$\underbrace{e^{(k)}} = \nabla_x F(x^{(k)}, w^{(k)}) - \mathcal{E}_w(\nabla F(x^{(k)}, w))$$

Error : random variable

$\mathcal{E}(e^{(k)}) = 0 \Rightarrow$

$\nabla f(x^{(k)})(-g^{(k)}) < 0$ on average

$\Rightarrow$ SGD : $e^{(k)}$: i.i.d., zero mean, bounded

Thm: $\alpha^{(k)} \to 0$, $\sum \alpha^{(k)} = \infty$, $\sum (\alpha^{(k)})^2 < \infty$

$\Rightarrow$ SGD : Converge to a stationary point.

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

Assume $\{\Delta x^{(k)}\}_{k=1}^{\infty}$ is gradient related.

And $\alpha^{(k)}$ → exact line search
→ Limited line search
→ Armijo rule (3)

$\Rightarrow$ Every Limit point of $\{x^{(k)}\}_{k=1}^{\infty}$ is a stationary point.

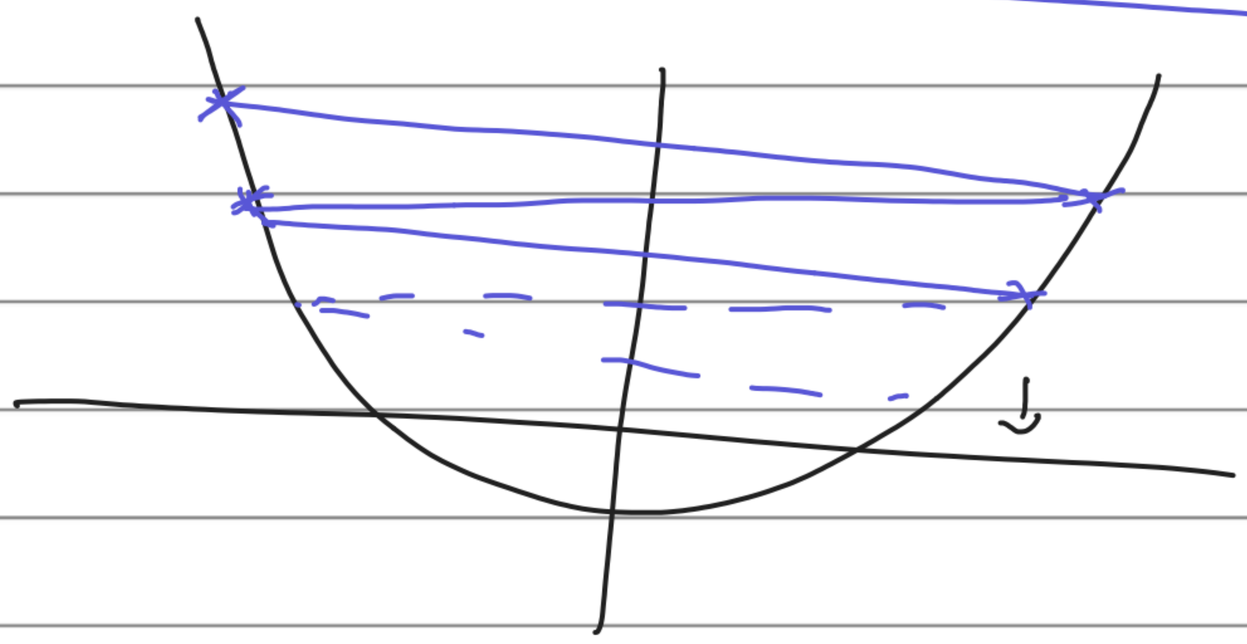
no convergence
two limit points: $+1, -1$

Proof of (3): (by contradiction)

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

$f(x^{(0)}) > f(x^{(1)}) > f(x^{(2)}) > \cdots$ $\Longrightarrow$ a sequence in $\mathbb{R}$ that is monotonic

$1 - f(x^{(k)}) \longrightarrow -\infty$



$2 - f(x^{(k)}) \longrightarrow f_*$

$x^{(k)} \longrightarrow \{-1, +1\}$

$f(x^{(k)}) \longrightarrow 0$

$\sigma \in (0,1)$

$$\underbrace{f(x^{(k+1)}) - f(x^{(k)})}_{\text{improvement}} \leq \sigma \nabla f(x^{(k)})^T (\alpha^{(k)} \Delta x^{(k)}) \quad (*)$$

$$\leq 0$$

$k \to \infty$: $f(x^{(k)}) \longrightarrow f_* \longrightarrow$ Left side $(*) \to 0$

Limit point : pick an arbitrary Limit point $\bar{x}$

$$\{x^{(k)}\}_{k=1}^{\infty} \xrightarrow{\text{sub-sequence}} \{x^{(k)}\}_{k \in K} \xrightarrow{\text{Converge}} \bar{x}$$

not a Limit point

$(*) \Rightarrow$ $\limsup_{\substack{k \to \infty \\ k \in K}} \alpha^{(k)} \nabla f(x^{(k)})^T \Delta x^{(k)} = 0$ (***)

gradient related : $\limsup_{\substack{k \to \infty \\ k \in K}} \nabla f(x^{(k)})^T \Delta x^{(k)} < 0$ (****)

(***), (****) $\Rightarrow$ $\limsup_{\substack{k \to \infty \\ k \in K}} \alpha^{(k)} = 0$

$\Rightarrow$ $\alpha^{(k)} \longrightarrow 0$ as $k \longrightarrow \infty$, $k \in K$

$$\lim_{\substack{k\to\infty \\ k\in K}} \Delta x^{(k)} = \overline{\Delta x}$$

$\{\Delta x^{(k)}\}_k$ is bounded,

so it has a convergent

subsequence with index set

$\overline{K} \subseteq K$. assume $\overline{k} = k$

to simplify proof.

Backtracking:

α

$\alpha\beta^2$

$\Rightarrow \alpha^{(k)} \to 0$

k : Large

$\Rightarrow \exists\ \overline{k}$ s.t. $\forall\ k \geq \overline{k}$: $\alpha^{(k)} < \alpha$

Armijo: $f(x^{(k)} + \alpha^{(k)} \Delta x^{(k)}) - f(x^{(k)}) \leq \sigma \nabla f(x^{(k)})^T (\alpha^{(k)} \Delta x^{(k)})$

① but

$$f\left(x^{(k)} + \frac{\alpha^{(k)}}{\beta} \Delta x^{(k)}\right) - f(x^{(k)}) > \sigma \nabla f(x^{(k)})^T \left(\frac{\alpha^{(k)}}{\beta} \Delta x^{(k)}\right)$$

②

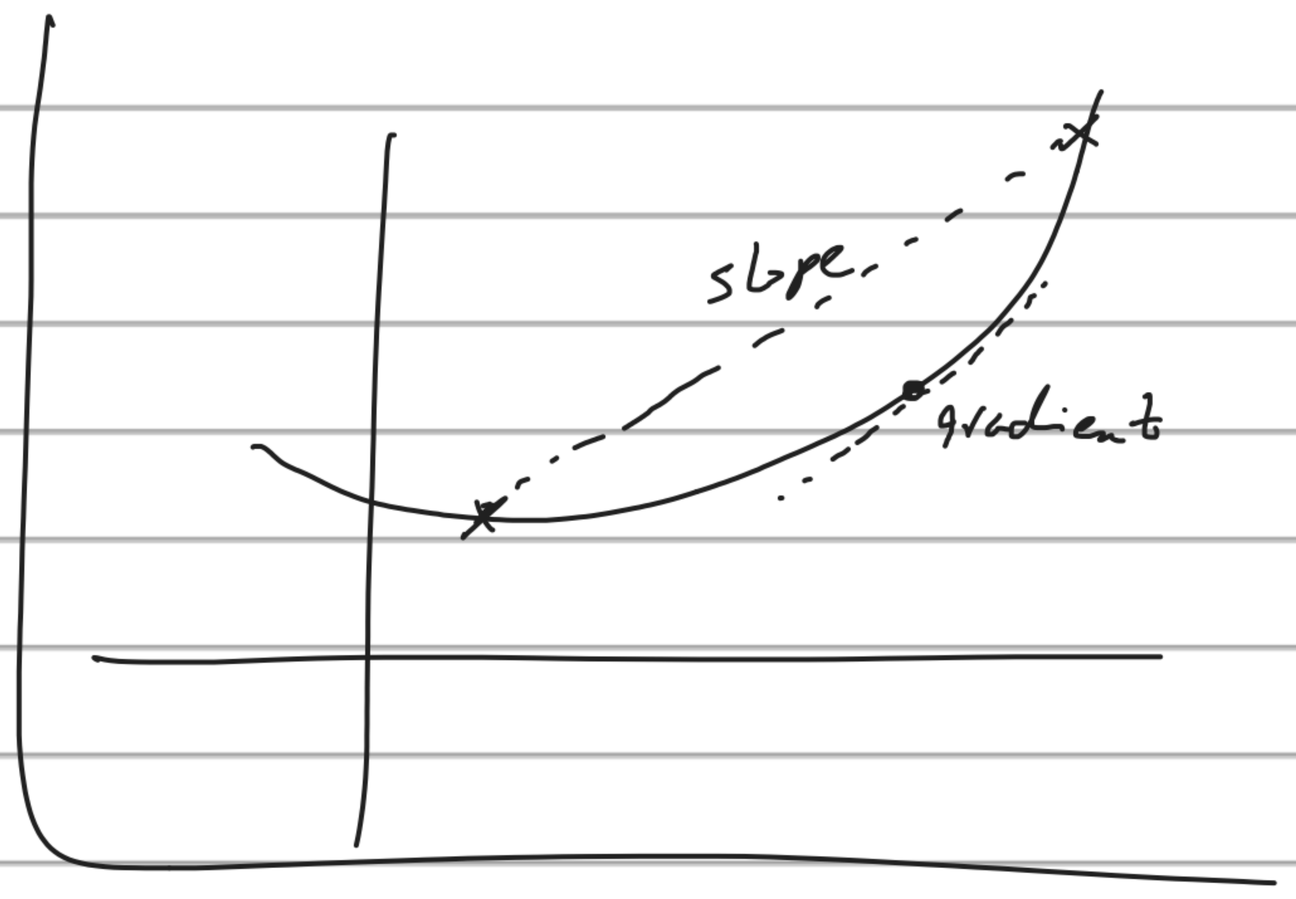
α

$\vdots$

$\times\ \alpha\beta^{t-1}$

$\checkmark\ \alpha\beta^t = \alpha^{(k)}$

$$\frac{f\left(x^{(k)} + \frac{\alpha^{(k)}}{\beta} \Delta x^{(k)}\right) - f(x^{(k)})}{\frac{\alpha^{(k)}}{\beta}} = \nabla f\left(x^{(k)} + \tilde{\alpha}^{(k)} \Delta x^{(k)}\right)^T \Delta x^{(k)}$$



$$\Rightarrow \nabla f\left(\bar{x}^{(k)} + \tilde{\alpha}^{(k)} \bar{\Delta x}^{(k)}\right)^T \bar{\Delta x}^{(k)} > \sigma \nabla f\left(\bar{x}^{(k)}\right)^T \bar{\Delta x}^{(k)}$$

$k \longrightarrow \infty \quad , \; k \in \mathbb{K}$

$$\Rightarrow \underbrace{(1-\sigma)}_{>0}, \nabla f(\bar{x})^T \bar{\Delta x} > 0 \Rightarrow \nabla f(\bar{x})^T \bar{\Delta x} > 0$$

$\bar{x}$: Limit point; $\bar{\Delta x}$: Limit of direction

$0 < \sigma < 1$

Also, $\nabla f(\bar{x})^T \bar{\Delta x} < 0$ (descent)

$\Rightarrow$ Contradiction

Lipschitz continuity of gradient:

$$\exists\, L > 0 \quad : \quad \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|$$

$$\forall\, x, y \in \mathbb{R}^n$$

Thm: Lipschitz continuity of gradient

implies:

$$\underline{f(x+y)} \leq \underline{f(x)} + y^T \nabla f(x) + \frac{L}{2} \| y \|^2$$

$$\forall\, x, y$$

$$\left(x = 0 \implies f(y) \leq \underbrace{f(0) + y^T \nabla f(0) + \frac{L}{2} \| y \|^2}_{\text{quadratic in } y} \right)$$

Proof: $\underline{g(t)} = \underline{f(x + ty)}$, $t \in \mathbb{R}$

$$f(x+y) - f(x) = g(1) - g(0) = \int_0^1 \frac{dg(t)}{dt}\, dt$$

$$= \int_0^1 y^T \underline{\nabla f(x + ty)}\, dt = \int_0^1 y^T \nabla f(x)\, dt$$

$$+ \int_0^1 y^T \left(\nabla f(x + ty) - \nabla f(x) \right) dt$$

$$f(x+y) - f(x) = \int_0^1 y^T \nabla f(x) \, dt +$$

$$\int_0^1 y^T \left(\nabla f(x+ty) - \nabla f(x) \right) dt \leq$$

$$\int_0^1 \underbrace{y^T \nabla f(x)}\, dt + \int_0^1 \|y\| \cdot \underbrace{\| \nabla f(x+ty) - \nabla f(x) \|}_{\leq L \| x+ty - x \| = t \|y\|} dt$$

$$= y^T \nabla f(x) + L \|y\|^2 \underbrace{\int_0^1 t \, dt}_{1/2}$$

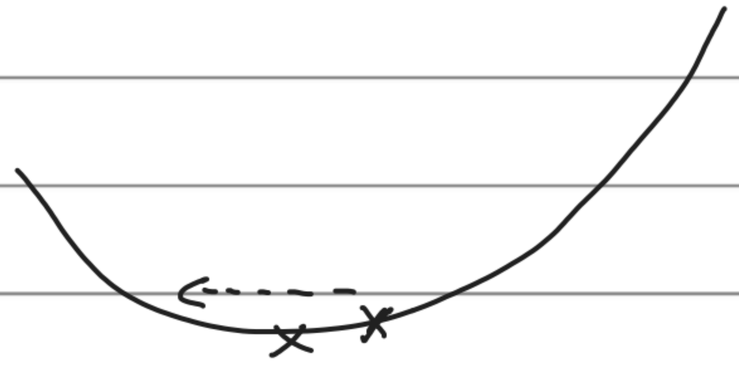
Rough idea : $x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)})$

implement : $\alpha^{(k)}$? $\longrightarrow$ $\alpha^{(k)} = \alpha$

$\alpha < \frac{2}{L} \Longrightarrow x^{(k)} \longrightarrow$ stationary point

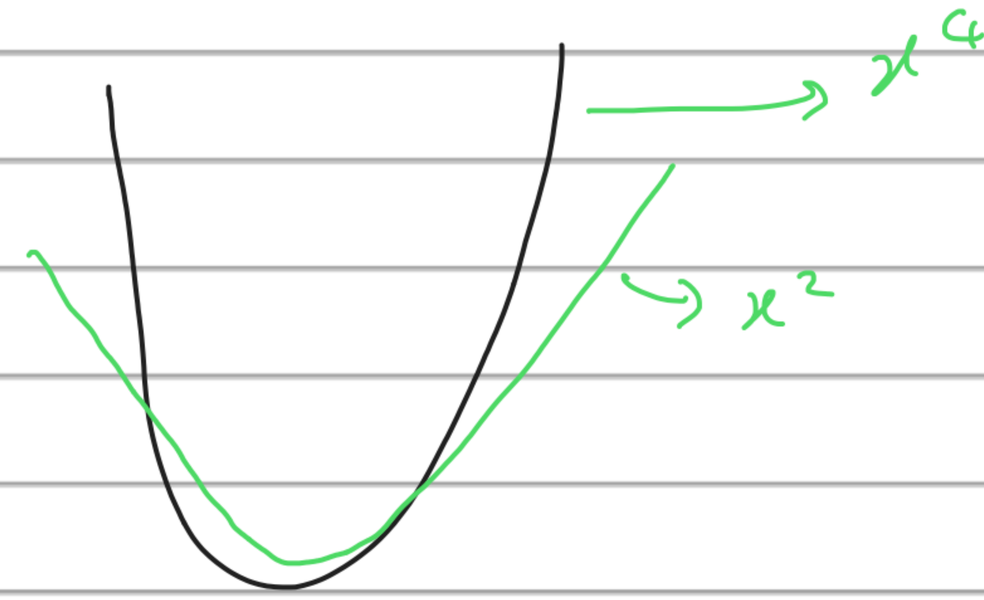
$x^{(k)} \longrightarrow$ close to x_* , $\nabla f(x^{(k)}) \longrightarrow 0$

$\underbrace{\alpha}_{\text{Constant}} \times \underbrace{\nabla f(x^{(k)})}_{\text{re-scaling}} \longrightarrow 0$



2 : $f(x) \leq$ quadratic function

$f(x) = x^4$:



$\Rightarrow$ local analysis: $f(x) \leq$ quadratic function in a neighborhood