262B-Lecture 5

Date created: 2021.02.03
N. of Pages: 15

Lipschitz Continuity of gradient:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x-y\| \qquad \forall x, y$$

$$\xrightarrow{\text{Thm}} f(x+y) \leq f(x) + \nabla f(x)^T y + \frac{L}{2}\|y\|^2 \qquad \forall x, y$$

Constant stepsize: $\min_{x \in \mathbb{R}^n} f(x)$

$x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$ where $\{\Delta x^{(k)}\}$ is gradient related. Assume Lipschitz continuity of gradient holds. Pick $\varepsilon \in (0, 1]$ and define: $\bar{\alpha}^{(k)} = \dfrac{|\nabla f(x^{(k)})^T \Delta x^{(k)}|}{L \|\Delta x^{(k)}\|^2}$.

$\Rightarrow$ every limit point of $\{x^{(k)}\}$ is a stationary point if $\alpha^{(k)} \in [\varepsilon, (2-\varepsilon)\bar{\alpha}^{(k)}]$.

$(0, 2\bar{\alpha}^{(1)}) \leftarrow$ pick arbitrary stepsize

$(0, 2\bar{\alpha}^{(2)}) \leftarrow \cdots$

$\vdots$

$(0, 2\bar{\alpha}^{(k)}) \leftarrow \cdots$

$\vdots$

Proof:

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

(previous thm)

$$f(x^{(k)}) + \nabla f(x^{(k)})^T (\alpha^{(k)} \Delta x^{(k)}) + \frac{1}{2} L \| \alpha^{(k)} \Delta x^{(k)} \|^2$$

(negative, descent direction)

$$\leq f(x^{(k)}) - \alpha^{(k)} | \nabla f(x^{(k)})^T \Delta x^{(k)} | + \frac{L}{2} \| \alpha^{(k)} \Delta x^{(k)} \|^2$$

$$= f(x^{(k)}) - \alpha^{(k)} \left(| \nabla f(x^{(k)})^T \Delta x^{(k)} | - \frac{L}{2} \alpha^{(k)} \| \Delta x^{(k)} \|^2 \right)$$

$\alpha^{(k)} \geq \varepsilon$

$$\geq \frac{1}{2} \varepsilon | \nabla f(x^{(k)})^T \Delta x^{(k)} |$$

$$\Rightarrow f(x^{(k+1)}) - f(x^{(k)}) \leq \frac{-\varepsilon^2}{2} | \nabla f(x^{(k)})^T \Delta x^{(k)} |$$

σ

Armijo rule, enough improvement

Gradient algorithm:

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

$$\Rightarrow \alpha^{(k)} = \frac{|\nabla f(x^{(k)})^T \Delta x^{(k)}|}{L \times \|\Delta x^{(k)}\|^2} = \frac{1}{L}$$

$$\Rightarrow \alpha^{(k)} \in [\varepsilon, \frac{2-\varepsilon}{L}], \quad \varepsilon > 0$$

$$\longrightarrow (0, \frac{2}{L})$$

Thm: If stepsize $\alpha^{(k)} = \alpha = \text{constant}$, then gradient algorithm works if

$$\alpha \in (0, \frac{2}{L})$$

Another version: diminishing stepsize.

If 1) Lipschitz continuity

2) $\alpha^{(k)} \longrightarrow 0$ as $k \to \infty$

3) $\sum_{k=1}^{\infty} \alpha^{(k)} \longrightarrow \infty \Longrightarrow$ every limit point is a stationary point.

what if L doesn't exist ?

$f(x) = x^4 \longrightarrow$ grows faster than a quadratic function

So far, we defined L globally.

what if we define it locally ?

previously : $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|$

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

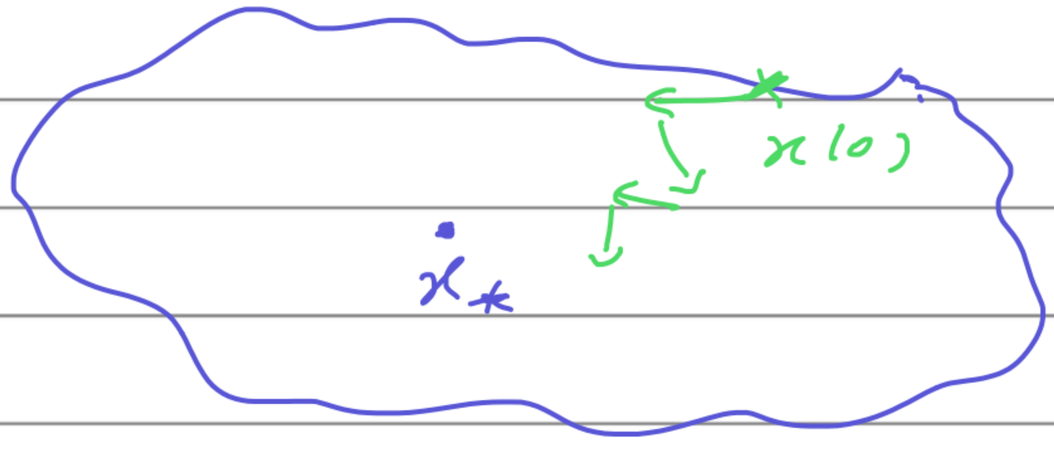
Now : $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\| \quad \forall\, x, y \in D$

D = a neighborhood of a stationary point x_*.

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

Property (*)

sublevel set : $S = \{x \mid f(x) \le f(x^{(0)})\}$

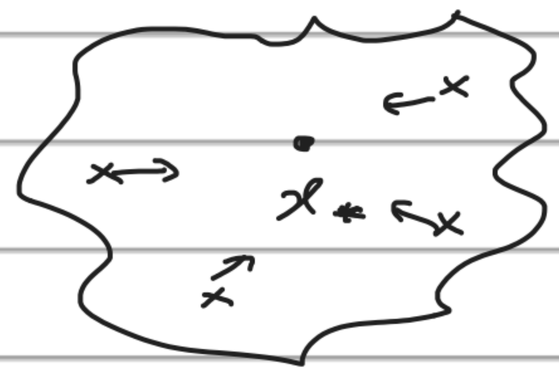


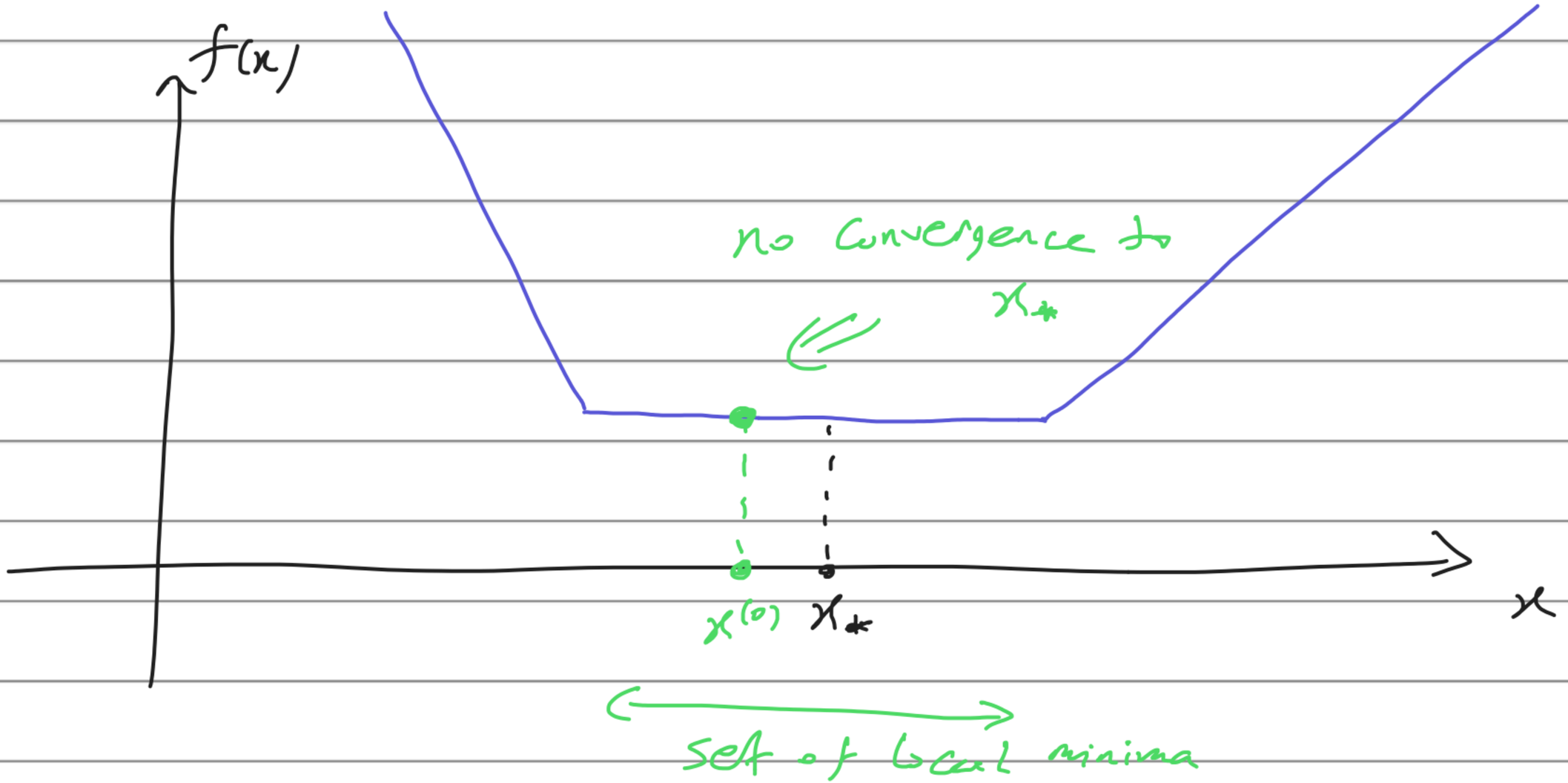
$\Rightarrow D : \subseteq S$

(a little bit bigger than S)

Capture theorem : If $x^{(0)}$ is close enough to an isolated local min x_*, then

$\{x^{(k)}\} \longrightarrow x_*$.





Convergence rate

Alg : $x^{(0)} \longrightarrow x^{(1)} \longrightarrow x^{(2)} \longrightarrow \cdots \longrightarrow x_*$

a bunch of points in $\mathbb{R}^n$

Assume: $\{x^{(k)}\} \longrightarrow$ unique limit point = x_*

Measure of speed: $\longrightarrow$ scalar

error: 1 - $e(x) = \|x - x_*\|$

2 - $e(x) = |f(x) - f(x_*)|$

- Asymptotic convergence rate

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

beginning unimportant tail matters.

$\{e^{(k)}\}$: Alg $\underline{\underline{1}}$ versus Alg $\underline{\underline{2}}$

$\underline{\underline{1}}$: 1 0.9 0.8 0.2 10^{-2} ....

$\underline{\underline{2}}$: 1 0.7 0.6 0.3 10^{-3} ....

need a baseline for comparison.

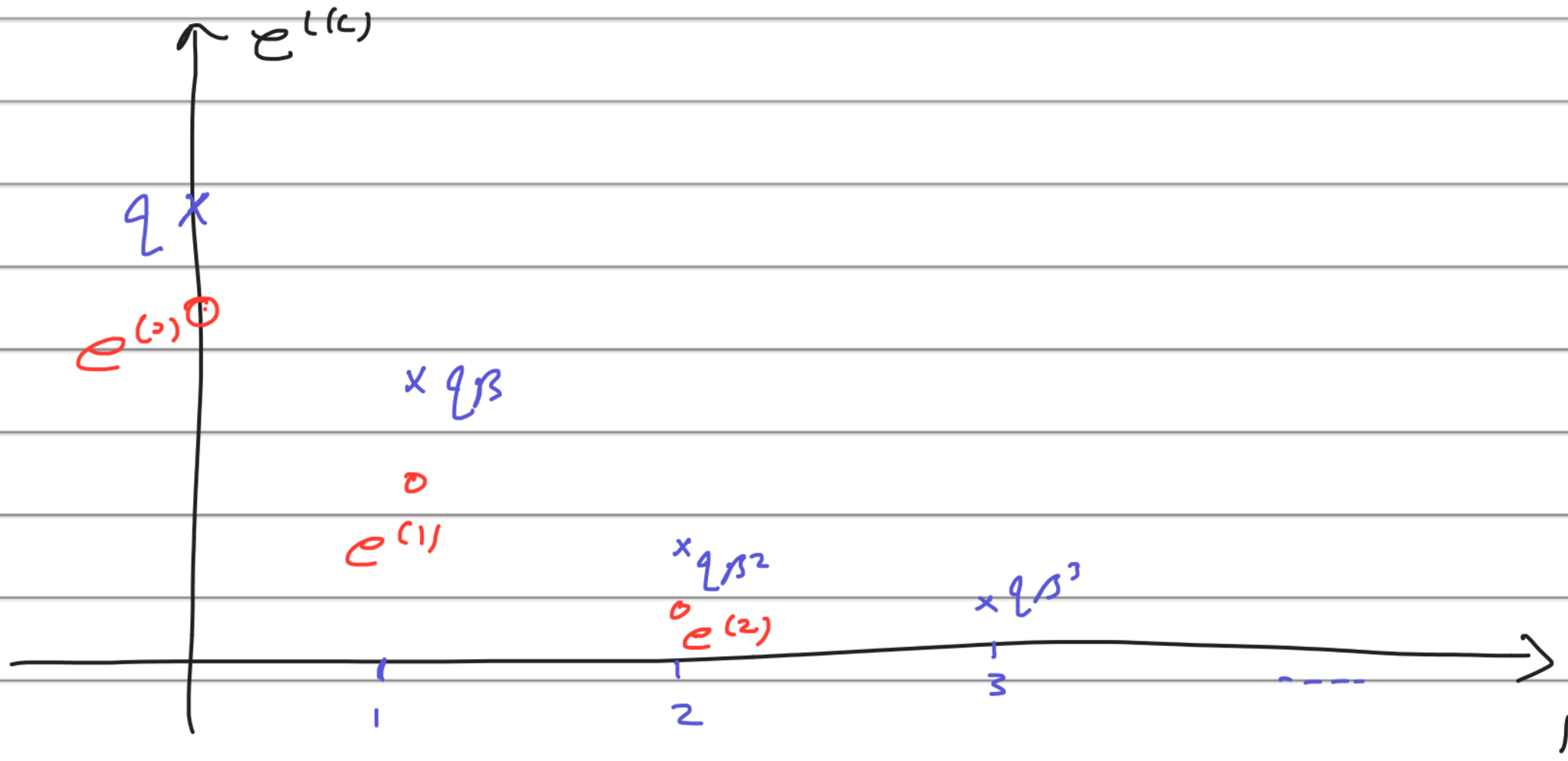
Baseline: pick $\beta \in (0,1)$

generate a sequence: $1 \to \beta \to \beta^2 \to \beta^3 \to \cdots$

geometric sequence

Rescale $\{e^{(k)}\} \longrightarrow$ pick $q > 0$.

Compare $\{e^{(k)}\}_{k=1}^{\infty}$ against $\{q\beta^k\}_{k=1}^{\infty}$



$e^{(k)} = e(x^{(k)}) \leq q\beta^k \quad (\forall\ k)$ ← ①

error sequence — baseline

$e(x^{(k)}) \leq \tilde{q}\beta^k$

$\forall\ k \geqslant \bar{k}$ ← ②

large enough

If that happens $\longrightarrow$ $\{e(x^{(k)})\}$

converges linearly or geometrically

with factor β.

---

special case: $e(x^{(k)}) = q\beta^{k}$

$$\Rightarrow \frac{e(x^{(k+1)})}{e(x^{(k)})} = \beta \Rightarrow$$

error reduces by factor β at

each iteration.

Thm: $\{e^{(k)}\}$ converges linearly if

$$\lim_{k\to\infty} \frac{e^{(k+1)}}{e^{(k)}} < 1$$

with what factor ? $\beta = \lim_{k\to\infty} \frac{e(x^{(k+1)})}{e(x^{(k)})}$

Alg 1: $\longrightarrow$ $\beta = 0.9$ slow

Alg 2: $\longrightarrow$ $\beta = 0.1$ fast

Thm: As β decreases, algorithm gets faster.

what if $\beta = \lim_{k \to \infty} \frac{e(x^{(k+1)})}{e(x^{(k)})} = 0$?

That means the sequence $\{e^{(k)}\}$ is faster than $\{\beta^k\}$ for every $0 < \beta < 1$.

$\Longrightarrow$ converges super Linearly

Def: $\{e(x^{(k)})\}$ converges superLinearly with order $\underline{p}$ if

$p > 1$

$$e(x^{(k)}) \leq q \times \beta^{(p^k)}$$

for some $q > 0$ &

($\forall$ k or Large enough k) $0 < \beta < 1$

$$\beta = 0.5 \quad , \quad q = 1 \quad , \quad K = 10$$

Linear

$$e(x^{(10)}) \leq (0.5)^{10}$$

small

super Linear $p = 2$

$$e(x^{(10)}) \leq (0.5)^{(2^{10})}$$

$2^{10} = 1024$

0.5^{1024}

Extremely small

---

special case:

$$e(x^{(K)}) = q \times \beta^{(p^K)} \qquad \forall K$$

$$\Rightarrow \underbrace{e(x^{(K+1)})}_{\text{new error}} = q \times \beta^{(p^{K+1})} = q \times \left(\beta^{(p^K)}\right)^p$$

$$= q \times \left(\frac{e(x^{(K)})}{q}\right)^p$$

$$= \underbrace{q^{(1-p)}}_{\text{any number}} \times \underbrace{e(x^{(K)})^p}_{\text{old error}}$$

Thm: $\{e^{(k)}\}$ converges superlinearly with order p if

$$\lim_{k\to\infty} \frac{e(x^{(k+1)})}{e(x^{(k)})^p} < \infty$$

note:
since $x^{(k)} \to x_*$,
we have
$e(x^{(k)}) \to 0$

$p=2$: Quadratic convergence

gradient alg $\longrightarrow$ Linear convergence

newton's alg $\longrightarrow$ quadratic convergence

Role of condition number:

$Q \succ 0 \longrightarrow$ condition number (c.d.)

$$= \frac{\max \text{eig } Q}{\min \text{eig } Q} = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}$$

$$\geq 1$$

$f : \mathbb{R}^n \longrightarrow \mathbb{R}$, x_* : satisfies SOC (sufficient)

$\Longrightarrow$ strict local min

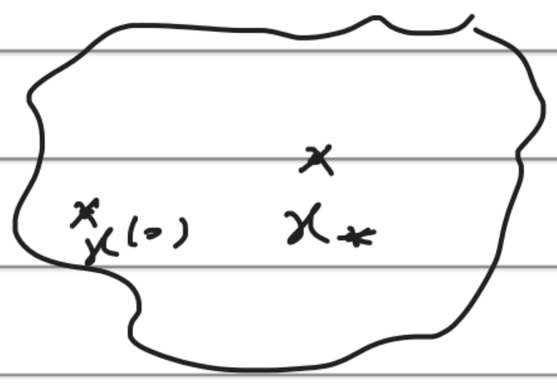
isolated

$\Longrightarrow$ $f(x) = f(x_*) + \nabla f(x_*)^T (x - x_*)$

$+ \frac{1}{2} (x - x_*)^T \nabla^2 f(x_*) (x - x_*)$ (*)

$+ O(\|x - x_*\|^3)$

Now, if x is close to x_*,



$\nabla^2 f(x_*) \longrightarrow Q$

$x - x_* \longrightarrow x$

Capture thm $\Longrightarrow$ right side (*) $\sim$ $\frac{1}{2} x^T Q x$;

Min $f(x)$ where $f(x) = \frac{1}{2} x^T Q x$

$\Longrightarrow \nabla f(x) = Qx$, $\nabla^2 f(x) = Q > 0$

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

$$\Rightarrow x^{(k+1)} = x^{(k)} - \alpha^{(k)} Q x^{(k)}$$

$$= \underbrace{(I - \alpha^{(k)} Q)}\, x^{(k)} \quad (*)$$

Error : $e(x^{(k)}) = \| x^{(k)} - \underbrace{x^*}_{\to 0} \| = \| x^{(k)} \|$

$(*) \Rightarrow \| x^{(k+1)} \|^2 = (x^{(k)})^T \underbrace{(I - \alpha^{(k)} Q)^2}_{\geq 0} x^{(k)}$

$$\Rightarrow \underbrace{\| x^{(k+1)} \|^2}_{e(x^{(k+1)})^2} \leq \lambda_{max} (I - \alpha^{(k)} Q)^2 \underbrace{\| x^{(k)} \|^2}_{e(x^{(k)})^2}$$

$x^T \underbrace{A}_{\geq 0} x \leq \lambda_{max}(A) \| x \|$

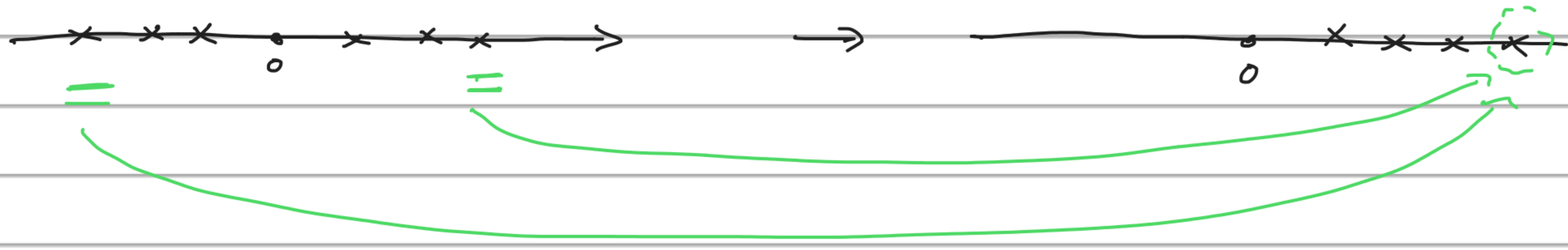
$$\Rightarrow \frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \sqrt{\lambda_{max} (I - \alpha^{(k)} Q)^2}$$

eigs of Q : $\lambda_1, \lambda_2, \ldots, \lambda_n$

eigs of $I - \alpha^{(k)} Q$: $1 - \alpha^{(k)} \lambda_1, 1 - \alpha^{(k)} \lambda_2, \ldots, 1 - \alpha^{(k)} \lambda_n$

eigs of $(I - \alpha^{(k)} Q)^2$: $(1 - \alpha^{(k)} \lambda_1)^2, (1 - \alpha^{(k)} \lambda_2)^2, \ldots$

eigs of $I - \alpha^{(k)} Q$ $\longrightarrow$ eigs of $(I - \alpha^{(k)} Q)^2$



min eig of Q : m (little)

max eig of Q : M (captical)

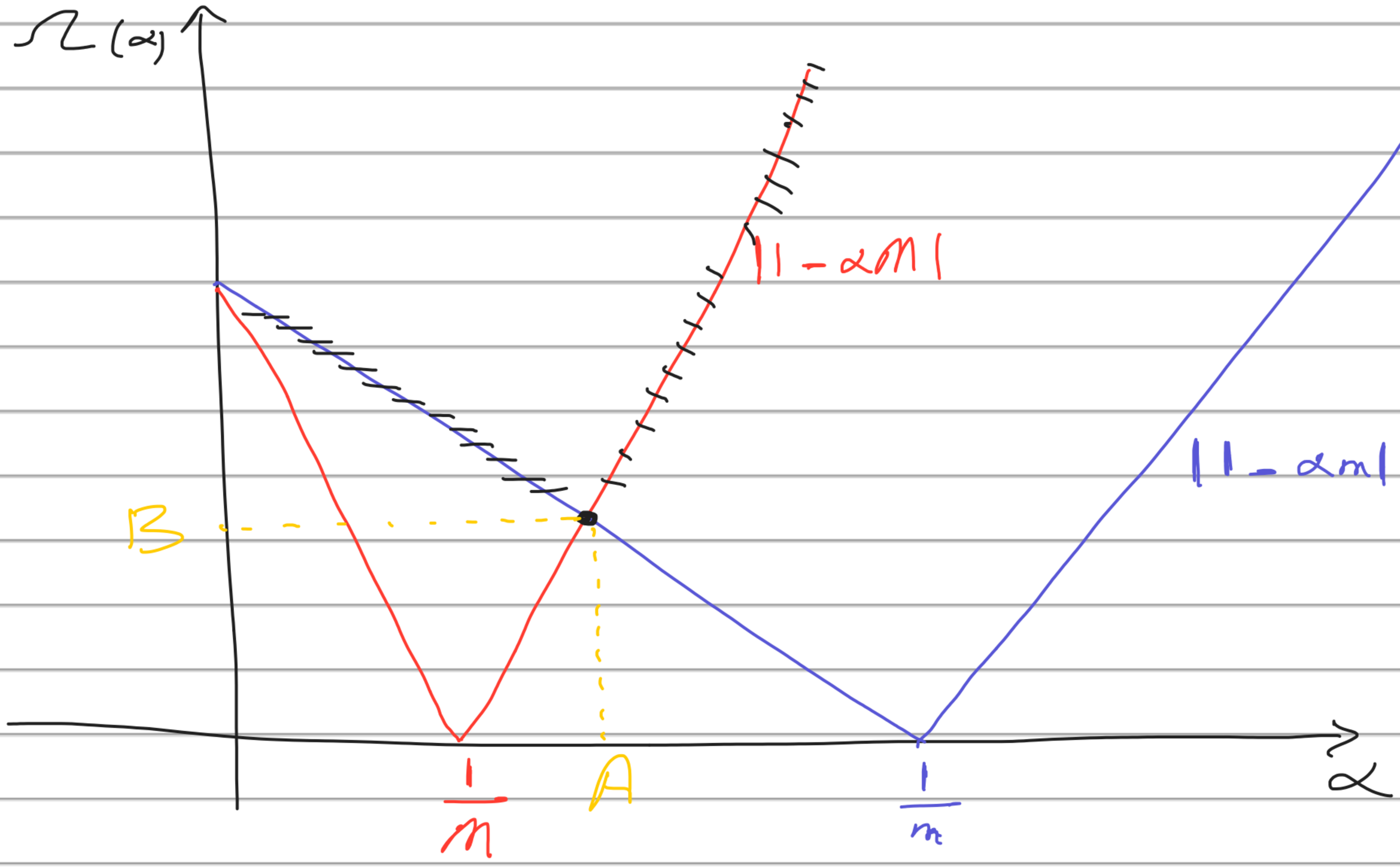
$$\Rightarrow \lambda_{max}(I - \alpha^{(k)} Q)^2 = \max\{(1-\alpha^{(k)} m)^2, (1-\alpha^{(k)} M)^2\}$$

$$\Rightarrow \frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \max\{|1-\alpha^{(k)} m|, |1-\alpha^{(k)} M|\}$$

Ω

How to pick $\alpha^{(k)}$: two methods

method 1 : Find $\alpha^{(k)}$ to minimize upper bound. $\longrightarrow$ $\min_{\alpha^{(k)}} \Omega$



$$A = \frac{2}{m+M} \quad , \quad B = \frac{M-m}{M+m} = \frac{c.d.(Q)-1}{c.d.(Q)+1}$$

Optimize upper bound to find stepsize:

$$\{e^{(k)}\}_{k=1}^{\infty} \longrightarrow \frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \frac{c.d.(Q)-1}{c.d.(Q)+1}$$

$0 < \beta < 1$

$\Rightarrow$ Gradient alg: Linear convergence with rate depending on c.d.