262B-Lecture 6

Date created: 2021.02.04
N. of Pages: 13

$\min \frac{1}{2} x^T \underbrace{Q}_{>0} x \rightarrow$ Gradient Algorithm:

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

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

$$\Rightarrow \frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \underbrace{\Lambda(\alpha^{(k)})}_{\text{Minimize over step size}}$$

$$\Rightarrow e(x^{(k)}) \leq q \beta^k \quad \text{where}$$

$$\beta = \frac{c.d.(Q) - 1}{c.d.(Q) + 1}$$

- If $\alpha^{(k)} = \alpha = \text{Constant}$, then above β is the best rate possible.

- This upper bound is tight. If $x^{(0)}$ is an eigenvector for $\lambda_{min}(Q)$ or $\lambda_{max}(Q)$, then the inequality turns into an equality.

Ex: 1- $\frac{\lambda_{max}(Q)}{\lambda_{min}(Q)} = \frac{100}{1} \rightarrow \beta$ close to $1 \rightarrow$ slow

2- $c.d. = 1.1/1 \rightarrow \beta$ close to $0 \rightarrow$ fast

Method 2 : Constant stepsize / minimize $\mathcal{L}(\alpha)$

↓

Exact Line search

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

$$\alpha^{(k)} = \left(\underset{\alpha \geq 0}{\operatorname{argmin}} \; f(x^{(k)} - \alpha \nabla f(x^{(k)})) \right)$$

optimal point ← Given

$$\frac{\partial \cdots}{\partial \alpha} = 0$$

$$\Rightarrow 0 = \frac{\partial f(x^{(k)} - \alpha \underbrace{\nabla f(x^{(k)})}_{g^{(k)}})}{\partial \alpha} \quad (*)$$

$\nabla f(x^{(k)}) = Q x^{(k)} \longrightarrow$ call it $g^{(k)}$

$(*) \Rightarrow -(g^{(k)})^T \underbrace{\nabla f(x^{(k)} - \alpha g^{(k)})}_{Q(x^{(k)} - \alpha g^{(k)})} = 0$

$$\Rightarrow \alpha^{(k)} = \frac{(g^{(k)})^T g^{(k)}}{(g^{(k)})^T Q g^{(k)}}$$

$$\Rightarrow f(x^{(k+1)}) = \left(1 - \frac{\left((g^{(k)})^T (g^{(k)})\right)^2}{((g^{(k)})^T Q g^{(k)})((g^{(k)})^T Q^{-1} g^{(k)})}\right) f(x^{(k)})$$

$$\leq \left(\frac{c.d.(Q)-1}{c.d.(Q)+1}\right)^2 f(x^{(k)})$$

Define: $e(x) = f(x) - f(x_*)^0$

$$\Rightarrow \frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \left(\frac{c.d.(Q)-1}{c.d.(Q)+1}\right)^2 \rightarrow \beta$$

$\Rightarrow$ β is different for different stepsize methods, but they all depend on $c.d.(Q)$.

$\min \frac{1}{2} x^T Q x \xrightarrow{\text{generalize}} \min f(x)$

- Assume: $\exists\, m, M > 0$ s.t.

$$mI \preceq \nabla^2 f(x) \preceq MI \quad \forall x$$

strong convexity; PSD

$A \preceq B$

$B - A$: PSD

what if such m, M don't exist for $\forall x$?

Need to consider sub-level set

$$\{x \mid f(x) \leq f(x^{(0)})\} \quad \text{and}$$

then define m, M over this set

- since $mI \preccurlyeq \nabla^2 f(x_*) \implies$

soc sufficient ✓

- $\frac{M}{m}$ is an upper bound on c.d. $\nabla^2 f(x)$ over $\mathbb{R}^n$.

- previously: $f(x) - f(x_*) \leq \frac{\|\nabla f(x)\|^2}{2m}$

(m says when gradient is small, we are close to a solution).

- Convergence rate: 1- exact line search

2- Back tracking

- Let's start with case 1.

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

$$= \min_{\alpha \geq 0} f(\underbrace{x^{(k)}}_{\text{nominal}} - \underbrace{\alpha \nabla f(x^{(k)})}_{\text{perturbation}})$$

$$= \min_{\alpha \geq 0} \Big(f(x^{(k)}) + \nabla f(x^{(k)})^T (-\alpha \nabla f(x^{(k)})) + \frac{1}{2} (-\alpha \nabla f(x^{(k)}))^T (\nabla^2 f(\underbrace{z^{(k)}}_{\text{new point}}) (-\alpha \nabla f(x^{(k)})) \Big)$$

$$\preceq M I$$

$$\leq \min_{\alpha \geq 0} \Big(f(x^{(k)}) + (-\alpha + \frac{M\alpha^2}{2}) \| \nabla f(x^{(k)}) \|_2^2 \Big)$$

$$\frac{\partial \cdots}{\partial \alpha} = 0 \Rightarrow \alpha = \frac{1}{M}$$

$$\Rightarrow f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{1}{2M} \underbrace{\| \nabla f(x^{(k)}) \|^2}_{\geq 2m (f(x^{(k)}) - f(x_*))}$$

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

$$\Rightarrow \boxed{e(x^{(k+1)}) \leq (1 - \frac{m}{M}) e(x^{(k)})}$$

β

$\Rightarrow$ Linear converge rate with

$$\beta = \left(1 - \frac{m}{M}\right) = 1 - \frac{1}{c.d.}$$

$\min f(x) \longrightarrow x^{(0)} \rightarrow x^{(1)} \rightarrow \cdots \rightarrow x^{(k)}$

$\vdots$

stop at some point

stop if $f(x^{(k)}) - f(x_*) \leq \underbrace{\varepsilon}_{\text{user defined}}$

$$\Rightarrow e(x^{(k)}) \leq \left(1 - \frac{m}{M}\right)^k e(x^{(0)}) \leq \varepsilon$$

$$\Rightarrow k \geq \frac{\log\left(\frac{f(x^{(0)}) - f(x_*)}{\varepsilon}\right)}{\underbrace{\log\left(1 - \frac{m}{M}\right)^{-1}}_{\rightarrow \text{Constant depending on Condition number}}}$$

$\Rightarrow$ Number of iterations $= O\left(\log \frac{1}{\varepsilon}\right)$

Also, number of iterations is proportional to log of suboptimality gap $f(x^{(0)}) - f(x_*)$

Redo convergence analysis for case $\subseteq$:

Backtracking

α
$\downarrow$
$\alpha\beta$
$\downarrow$
$\alpha\beta^2$
$\vdots$
$\downarrow$
$\alpha\beta^{i_k}$

$\alpha^{(k)}$: $\longrightarrow$ if

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

$$\sigma \nabla f(x^{(k)})^T (+\alpha^{(k)} \underbrace{\Delta x^{(k)}}_{-\nabla f(x^{(k)})}) = -\sigma \alpha^{(k)} \|\nabla f(x^{(k)})\|^2$$

Armijo Rule $(0 < \sigma < 1)$ (1)

Pick an arbitrary non-negative number j:

$$f(x^{(k)} + \alpha\beta^j \Delta x^{(k)}) - f(x^{(k)}) =$$

$$\nabla f(x^{(k)})^T (\alpha\beta^j \underbrace{\Delta x^{(k)}}_{-\nabla f(x^{(k)})}) + \frac{1}{2} (\alpha\beta^j \Delta x^{(k)})^T \underbrace{\nabla^2 f(z^{(k)})}_{\leq MI} \times (\alpha\beta^j \Delta x^{(k)})$$

$$\leq -\left(1 - \frac{M\alpha\beta^j}{2}\right)(\alpha\beta^j) \|\nabla f(x^{(k)})\|^2 \quad (2)$$

$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\sigma \alpha^{(k)} \| \nabla f(x^{(k)}) \|^2 \quad (1)$$

$$f(x^{(k)} - \alpha\beta^j \Delta x^{(k)}) - f(x^{(k)}) \leq -\left(1 - \frac{M\alpha\beta^j}{2}\right)(\alpha\beta^j) \times \| \nabla f(x^{(k)}) \|^2 \quad (2)$$

function $1 - \frac{M\alpha\beta^j}{2}$ in terms of j

As $j \to \infty$, this function goes to 1.

Also, $0 < \sigma < 1 \Rightarrow \left(1 - \frac{M\alpha\beta^j}{2} \geq \sigma \right.$ for large j $\left.\right)$ $(0 < \beta < 1)$

Let μ denote the smallest nonnegative integer s.t. $1 - \frac{M\alpha\beta^\mu}{2} \geq \sigma$

$\Rightarrow$ Two scenarios

1. $1 - \mu = 0 \Rightarrow \alpha$: small
2. $\mu > 0 \Rightarrow 1 - \frac{M\alpha\beta^{\mu-1}}{2} < \sigma$

$\downarrow$

α, $\alpha\beta$, $\vdots$ $\rightarrow i_k$

$\alpha\beta^\mu$, $\alpha\beta^{\mu+1}$, $\vdots$ $\rightarrow$ Armijo rule due to (2) $\Rightarrow i_k \leq \mu$

$\Rightarrow \alpha^{(k)} \geq \alpha\beta^\mu$

① $\Rightarrow f(x^{(k+1)}) - f(x^{(k)}) \leq$

$-\sigma \alpha^{(k)} \| \nabla f(x^{(k)}) \|^2 \leq$

$-\sigma \alpha \beta^{\mu} \| \nabla f(x^{(k)}) \|^2$,

$\geq 2m \, (f(x^{(k)}) - f(x_*))$

$e(x) = f(x) - f(x_*)$ (3)

$\Rightarrow e(x^{(k+1)}) \leq e(x^{(k)}) \times (1 - \sigma \alpha \beta^{\mu} (2m))$

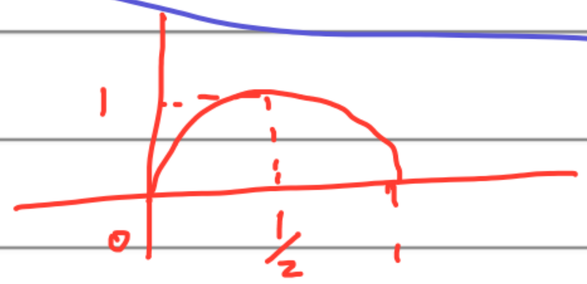
Assume: Scenario ② happens.

$1 - \frac{m \alpha \beta^{\mu-1}}{2} < \sigma \Rightarrow -\alpha \beta^{\mu} < \frac{2(1-\sigma)\beta}{m}$ (4)

③, ④ $\Rightarrow$

$\frac{e(x^{(k+1)})}{e(x^{(k)})} \leq 1 - (4\sigma(1-\sigma)) \times \beta \times \left(\frac{m}{M}\right)$,

< 1 < 1 cd.



$\Rightarrow$ Convergence rate $\sim$ Condition number

So far: $\exists\ m > 0$ s.t. $\nabla^2 f(x) \succ mI$

what if that m doesn't exist ?

- Let's focus on the case $\nabla^2 f(x) \succeq 0$.

$\forall x$

$\Rightarrow$ Convergence could be pretty slow.

And goes from linear to sublinear.

Two sequences:

① $1 \quad \beta \quad \beta^2 \quad \beta^3 \quad \cdots \quad \beta^k \quad \cdots \longrightarrow$ linear convergence

② $1 \quad \frac{1}{2} \quad \frac{1}{3} \quad \cdots \quad \frac{1}{k} \longrightarrow$ extremely slow

$\beta = 0.2$, $k = 100$: $(0.2)^{100}$ versus $\frac{1}{100}$

① $\longrightarrow$ under assumption $\exists\ m > 0$

② $\longrightarrow$ if $\nexists\ m > 0$

- without (local) strong convexity assumption,

$\{e^{(k)}\}$ is just a bit faster than $\{\frac{1}{k}\}$

i.e $\lim_{k \to \infty} \frac{e^{(k)}}{1/k} = 0$

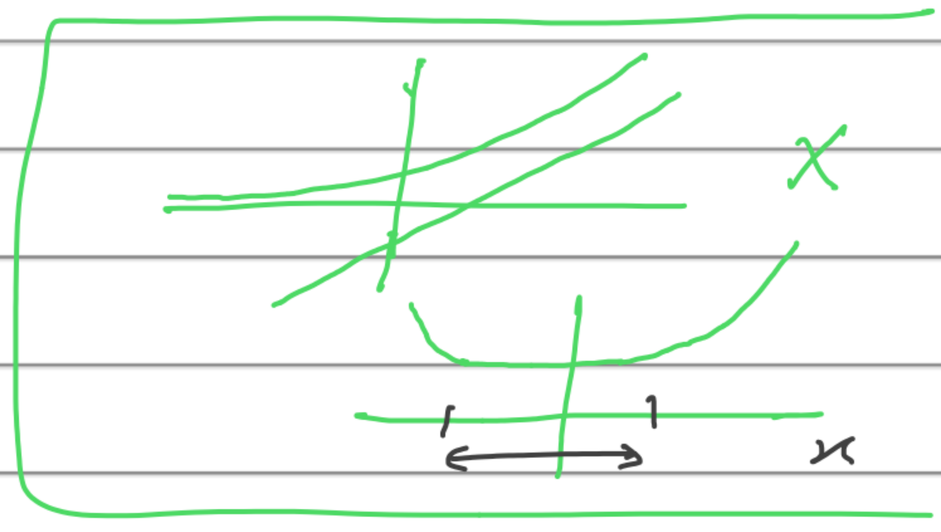
$\Rightarrow e^{(k)} = o(1/k)$

small o

Thm: - Assume that $\nabla^2 f(x) \geq 0$ &

set of global solutions, denoted as X_*, is non-empty & bounded

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



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

where $\exists\ c > 0$ s.t. $\nabla f(x^{(k)})^T \Delta x^{(k)} \leq -c \| \nabla f(x^{(k)}) \|^2$ (special case of gradient related)

- $\alpha^{(k)} \in [\varepsilon, (2-\varepsilon)\bar{\alpha}^{(k)}]$

where $\bar{\alpha}^{(k)} = \frac{|\nabla f(x^{(k)})^T \Delta x^{(k)}|}{L \| \Delta x^{(k)} \|^2}$

$\Rightarrow$ All limits points of $\{x^{(k)}\}$ are optimal, and

$$e(x^{(k)}) = f(x^{(k)}) - f_* = O\left(\frac{1}{k}\right)$$

$$\boxed{f_* = \min f(x)}$$

Implication: $e(x^{(k)}) \leq \frac{q}{k} \leq \varepsilon$

$\Rightarrow$ number of iterations $= O\left(\frac{1}{\varepsilon}\right)$

$m > 0$

$\#$ iterations $= O(\log \frac{1}{\varepsilon})$

L

Complexity = Linear in number of digits

$m = 0$

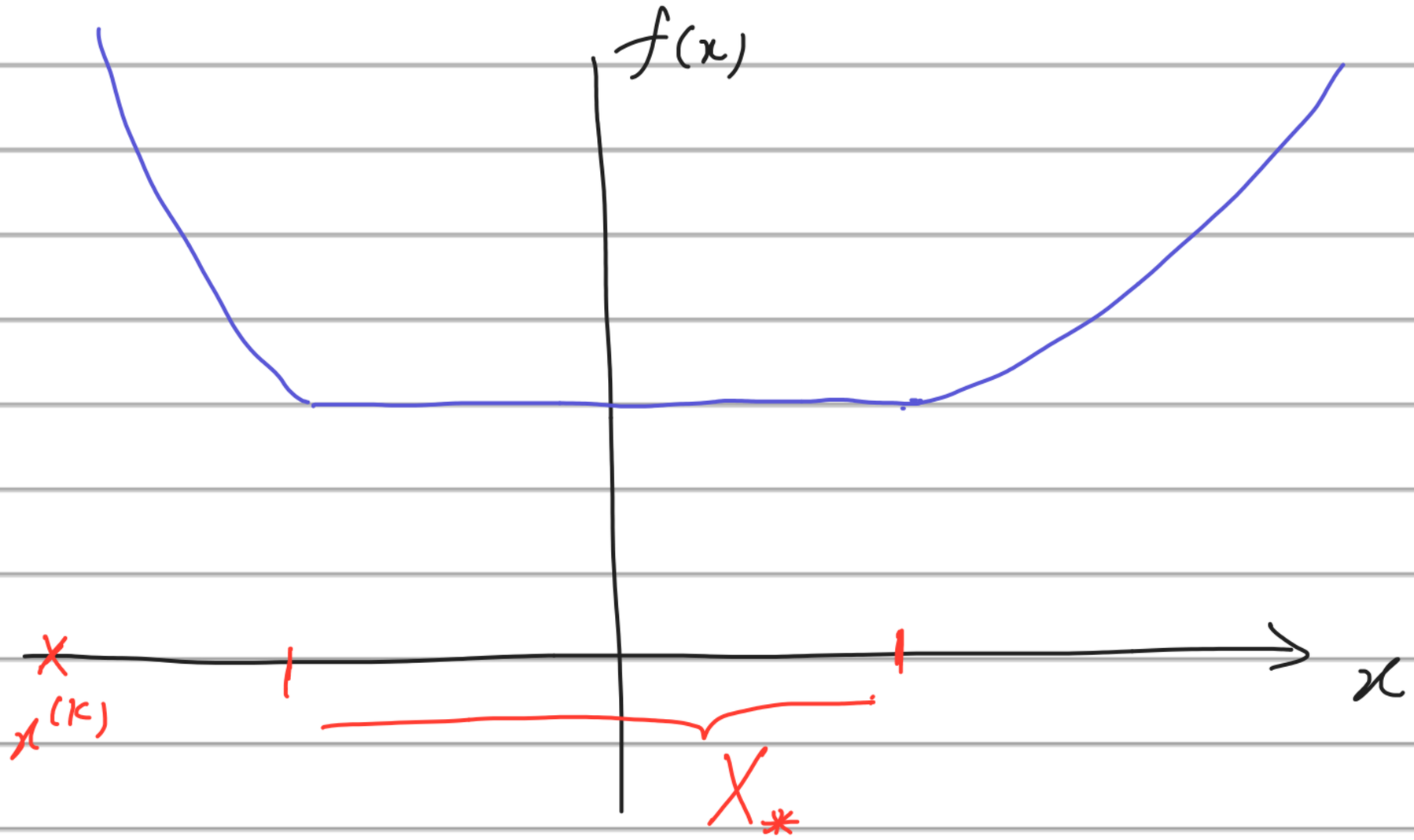
$\#$ iterations $= O(\frac{1}{\varepsilon})$

10^L

$\varepsilon = 10^{-L}$

L: digits

Complexity = exponential in number of digits



$e(x) = \|x - x_*\| \longrightarrow x_*$: not unique

↓

measures the distance between point x

& set X_*.

$$d(x, X_*) = \min_{x_* \in X_*} \|x - x_*\|$$