262B-Lecture 7

Date created: 2021.02.09
N. of Pages: 12

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

($m > 0$)

$\Rightarrow$ Linear convergence

- $\min f(x)$ if $m = 0$ $\Rightarrow$ sublinear convergence

More precisely: $O(1/k)$

Assumptions: $\nabla f(x^{(k)})^T \Delta x^{(k)} \leq -c \|\nabla f(x^{(k)})\|^2$ (1)

($c > 0$)

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

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

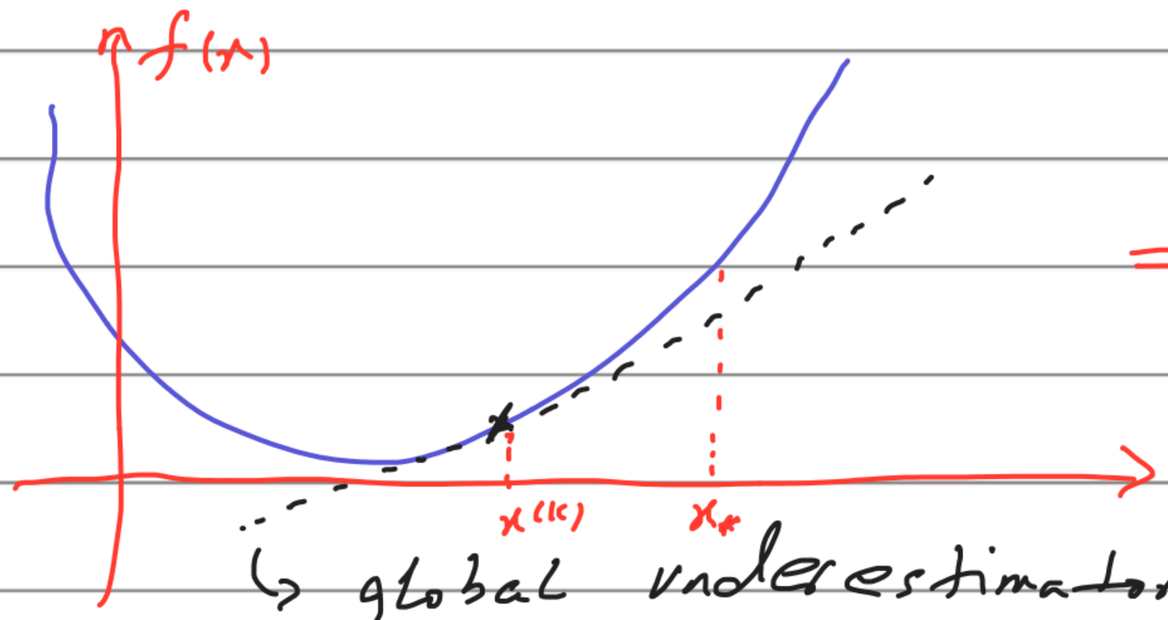
Define: X_* = set of all global minima

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

Proof: Based on a previous theorem, we know

that $\{e(x^{(k)})\} \to 0$ as $k \to \infty$

Now: $\nabla^2 f(x) \succeq 0 \quad \forall x \Rightarrow$ Conven



↳ global underestimator

$\Rightarrow f(x^{(k)}) - f(x_*) \leq \nabla f(x^{(k)})^T (x^{(k)} - x_*)$ (*)

$\leq \|\nabla f(x^{(k)})\| \cdot \|x^{(k)} - x_*\|$

minimize both sides of (*) with respect to x_* :

$$f_* = \min f(x)$$

$$\Rightarrow f(x^{(k)}) - f_* \leq \|\nabla f(x^{(k)})\| \times d(x^{(k)}, X_*) \quad (2)$$

(1), (2) $\Rightarrow$ (3) if error: $f(x) - f_*$

$$\text{then, } e(x^{(k+1)}) \leq e(x^{(k)}) - \frac{c\varepsilon^2 e(x^{(k)})^2}{2d(x^{(k)}, X_*)} \quad \forall k$$

---

proof for constant stepsize : (3)

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

$$\Rightarrow e(x^{(k+1)}) \leq e(x^{(k)}) \left(1 - \underbrace{\frac{c\varepsilon^2 e(x^{(k)})}{2d(x^{(k)}, X_*)^2}}_{a^{(k)}}\right)$$

since $e(x) \geq 0 \Rightarrow 1 - a^{(k)} \geq 0$

$$(1 - a^{(k)})^{-1} = 1 + a^{(k)} + (a^{(k)})^2 + (a^{(k)})^3 + \cdots \geq 1 + a^{(k)}$$

$$\Rightarrow e(x^{(k+1)})^{-1} \geq e(x^{(k)})^{-1}\left(1 - a^{(k)}\right)^{-1}$$

$$\geq e(x^{(k)})^{-1}\left(1 + a^{(k)}\right)$$

$$= e(x^{(k)})^{-1} + \frac{c\varepsilon^2}{2d(x^{(k)}, X_*)^2}$$

$\Rightarrow$ add them up, then take the inverse:

$$e(x^{(k)}) \leq \left(\frac{1}{e(x^{(0)})} + \frac{c\varepsilon^2}{2}\sum_{i=0}^{k-1}\frac{1}{d(x^{(i)}, X_*)^2}\right)^{-1}$$

$$\Rightarrow k\,e(x^{(k)}) \leq \left(\frac{1}{k\,e(x^{(0)})} + \frac{c\varepsilon^2}{2k}\sum_{i=0}^{k-1}\frac{1}{d(x^{(i)}, X_*)^2}\right)^{-1}$$

A

$k \to \infty$: $e(x^{(k)}) \to 0 \Rightarrow d(x^{(i)}, X_*) \to 0$

$$\Rightarrow \frac{1}{d(x^{(i)}, X_*)^2} \Rightarrow \infty \Rightarrow \frac{1}{k}\sum_{i=0}^{k-1}\frac{1}{d(x^{(i)}, X_*)^2} \Rightarrow \infty$$

$$\Rightarrow A = (\infty)^{-1} = 0$$

$$\Rightarrow \lim_{k\to\infty} k\,e(x^{(k)}) = 0 \Rightarrow \boxed{e(x^{(k)}) = O\left(\frac{1}{k}\right)}$$

Summary :

	Strongly Convex ($m>0$)	Convex ($m\geq 0$)
	$O\left(\log\left(\frac{1}{\varepsilon}\right)\right)$	$O\left(\frac{1}{\varepsilon}\right)$
	heavy ball	Nestrov ← Can we improve it ?

Acceleration :

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

$$\Downarrow$$

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)}) + \beta^{(k)} \underbrace{(x^{(k)} - x^{(k-1)})}_{\text{momentum}}$$

If $\alpha^{(k)} = \alpha$, $\beta^{(k)} = \beta$

$\Rightarrow$ Heavy-ball method

Case study : $\min \frac{1}{2} x^T \underbrace{Q}_{>0} x$ $(m>0)$

Gradient alg : $\frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \frac{c.d.(Q)-1}{c.d.(Q)+1}$

accelerated gradient : $\frac{e(x^{(k+1)})}{e(x^{(k)})} \leq \frac{\sqrt{c.d.(Q)}-1}{\sqrt{c.d.(Q)}+1}$ for some choices of α and β

Proof: Homework 2

Ex: $c.d.(Q) = 10^4 \rightarrow$ ill-conditioned

$\Rightarrow$ extremely slow convergence for gradient alg.

But $\sqrt{c.d.(Q)} = 10^2 \rightarrow$ much faster

$\Rightarrow$ This sort of acceleration is good when $\nabla^2 f(x_*) > 0$.

previously: $\left(\frac{c.d.(Q)-1}{c.d.(Q)+1}\right)^k \leq \varepsilon$

Now: $\left(\frac{\sqrt{c.d.}-1}{\sqrt{c.d.}+1}\right)^k \leq \varepsilon$

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

acceleration improves the constant of O.

Ex: $1000 \times \log \frac{1}{\varepsilon} \Rightarrow 20 \times \log \frac{1}{\varepsilon}$

improvement

Nesterov's acceleration method:

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

$$\Downarrow$$

$$\begin{cases} y^{(k)} = x^{(k)} + \beta^{(k)} (x^{(k)} - x^{(k-1)}) & : \text{intermediate parameter} \\ x^{(k+1)} = y^{(k)} - \alpha \nabla f(y^{(k)}) & : \text{apply gradient alg. to } y^{(k)} \end{cases}$$

Assume: $\beta^{(k)} \to 1$ as $k \to \infty$

Original proof: $\beta^{(k)} = \frac{k-1}{k+2}$

special: $f(x^{(k+1)}) \not\leq f(x^{(k)})$

$\Rightarrow$ Not a descent algorithm

- Assume: $0 \preceq \nabla^2 f(x) \preceq M I \Rightarrow m = 0$

If $\alpha \leq \frac{1}{M} \Rightarrow$

$$f(x^{(k)}) - f_* \leq \frac{2 d(x^{(0)}, X_*)^2}{\alpha (k+1)^2}$$

$\Rightarrow$ Convergence rate $= O(\frac{1}{k^2})$

If we want $f(x^{(k)}) - f_* \leq \varepsilon$

$$\Rightarrow \frac{2 d(x^{(0)}, X_*)^2}{\alpha (k+1)^2} \leq \varepsilon \Rightarrow .$$

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

Strongly Convex ($m > 0$)	Convex ($m \geq 0$)
Gradient: $O(\log \frac{1}{\varepsilon})$	Gradient $O(\frac{1}{\varepsilon})$
$\downarrow$ heavy ball	$\downarrow$ Nestrov
$O(\log \frac{1}{\varepsilon})$	$O(\frac{1}{\sqrt{\varepsilon}})$
but with a	
better constant	

Ex: $\varepsilon = 10^{-4} \Rightarrow$ Compare $10^4 = \frac{1}{\varepsilon}$ to $10^2 = \frac{1}{\sqrt{\varepsilon}}$

Most of our analysis : $\nabla^2 f(x) \geq 0$.

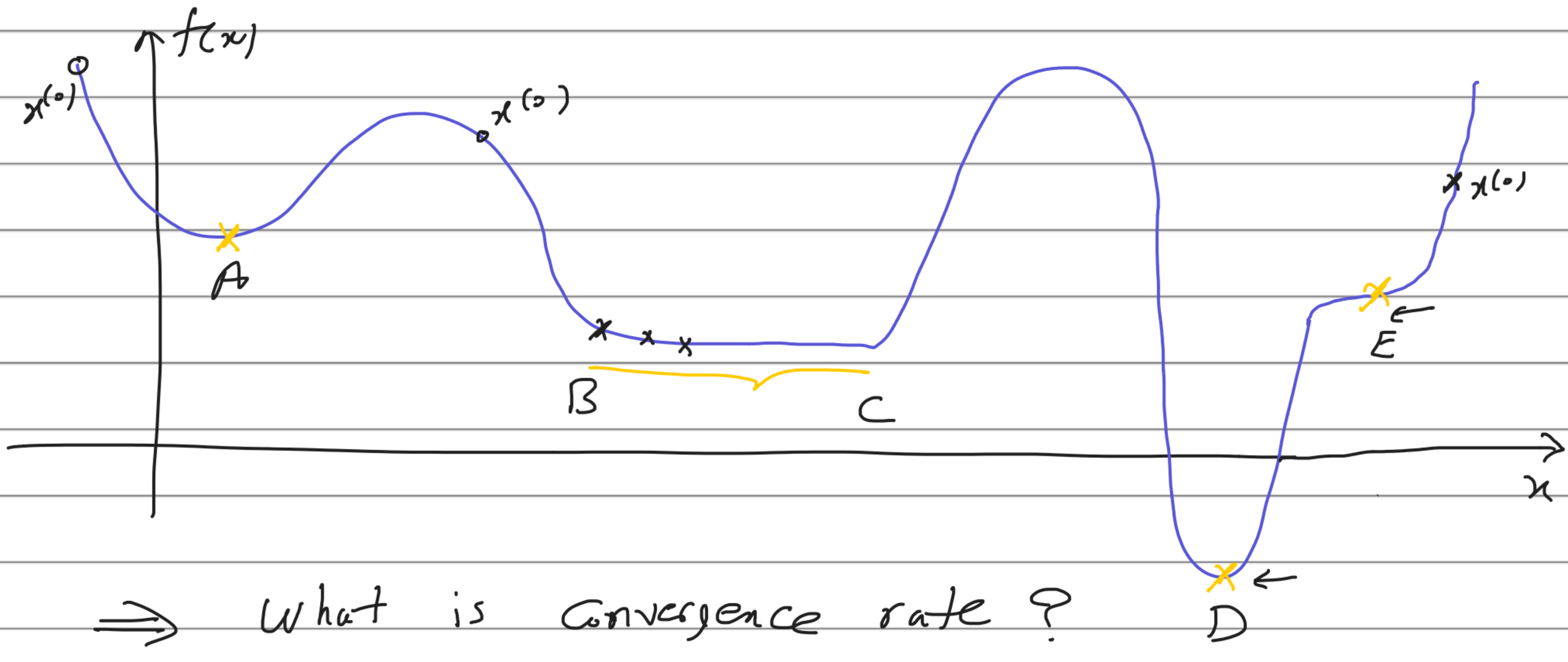
$\forall x$

$\Downarrow$

Convex case

Generalize to the non-convex case using the idea of sub-level set if $x^{(0)}$ is close to a solution x_*.

Question : what if $f(x)$ is highly non-convex and $x^{(0)}$ is random (far away from x_*) ?



$\Rightarrow$ What is convergence rate ?

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

x_* : stationary point

Convergence : get to a neighborhood of a stationary point ?

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

Assume : Constant stepsize $\alpha \in [\underbrace{\bar{\varepsilon}}_{>0}, \frac{2-\bar{\varepsilon}}{L}]$

$\Rightarrow$ stopping criterion : $\| \nabla f(x^{(k)}) \| \leq \varepsilon$

previous theorem : $\{x^{(k)}\}$: Limit points are stationary points.

$\Downarrow$ Proof

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

$\Downarrow$ add this up over different values of k.

$$\Rightarrow (f(x^{(k+1)}) - f(x^{(0)})) \leq -\frac{\bar{\varepsilon}^2}{2} \sum_{i=0}^{k} \| \nabla f(x^{(i)}) \|^2 \quad (1)$$

$$\geq \underbrace{\min f(x)}_{\text{global min}} - f(x^{(0)})$$

$$\frac{\varepsilon^2}{2} \leq \sum_{i=0}^{K} \|\nabla f(x^{(i)})\|^2 \geq \frac{\varepsilon^2}{2} \times (K+1)$$

$$\times \min_{0 \leq i \leq K} \|\nabla f(x^{(i)})\|^2 \qquad (2)$$

If we want to make sure that

$$\min_{0 \leq i \leq K} \|\nabla f(x^{(i)})\|^2 \leq \varepsilon^2 \qquad (3)$$

(1), (2), (3) $\Rightarrow$

$$\left(\min_{1 \leq i \leq K} \|\nabla f(x^{(i)})\|^2 \right) \leq \frac{f(x^{(0)}) - \min f(x)}{\frac{\varepsilon^2 (K+1)}{2}}$$

global min

$$\leq \varepsilon^2$$

$\Rightarrow$ enough to have:

# of iterations : $K = O\left(\frac{1}{\varepsilon^2}\right)$

$$\frac{f(x^{(0)}) - \min f(x)}{\varepsilon^2 (K+1) \times \frac{1}{2}} \leq \varepsilon^2 \Rightarrow K+1 \sim \frac{1}{\varepsilon^2}$$

strongly convex ($m>0$)	convex ($m\geq 0$)	non-convex
Gradient: $O(\log \frac{1}{\varepsilon})$	Gradient: $O(\frac{1}{\varepsilon})$	Gradient: $O(\frac{1}{\varepsilon^2})$
↓ heavy ball	↓ Nestrov	↓
$O(\log \frac{1}{\varepsilon})$	$O(\frac{1}{\sqrt{\varepsilon}})$	?
better constant		(state-of-the-art)

first-order methods: (we use gradient to generate the points)

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

↓

scaled gradient: $x^{(k+1)} = x^{(k)} - \alpha^{(k)} \underbrace{D^{(k)}}_{>0} \nabla f(x^{(k)})$

Assume $D^{(k)} = D$, $f(x) = \frac{1}{2} x^T Q x$

$\Rightarrow x^{(k+1)} = x^{(k)} - \alpha^{(k)} D Q x^{(k)} \times D^{-1/2}$

Define: $y^{(k)} = D^{-1/2} x^{(k)}$

$\Rightarrow y^{(k+1)} = y^{(k)} - \alpha^{(k)} D^{-1/2} D Q D^{1/2} y^{(k)}$

$\Rightarrow y^{(k+1)} = y^{(k)} - \alpha^{(k)} (D^{1/2} Q D^{1/2}) y^{(k)}$

$\Downarrow$

gradient alg for $\min \frac{1}{2} y^T (D^{1/2} Q D^{1/2}) y$

$\Rightarrow \frac{\|y^{(k+1)}\|}{\|y^{(k)}\|} \leq \frac{c.d.(D^{1/2} Q D^{1/2}) - 1}{c.d.(D^{1/2} Q D^{1/2}) + 1}$ (optimal stepsize)

$\Rightarrow \frac{\sqrt{(x^{(k+1)})^T D^{-1} x^{(k+1)}}}{\sqrt{(x^{(k)})^T D x^{(k)}}} \leq \frac{c.d.(D^{1/2} Q D^{1/2}) - 1}{c.d.(D^{1/2} Q D^{1/2}) + 1}$

error in weighted norm