262B-Lecture 14

Date created: 2021.03.09
N. of Pages: 13

Convergence analysis of gradient projection methods:

Special case: $\min \frac{1}{2} x^T Q x - b^T x$, $Q \succ 0$

$$\text{s.t.} \quad x \in X$$

Assume: $\alpha^{(1)} = 1$, $s^{(k)} = s$

$$\Rightarrow \quad x^{(k+1)} = P_X \left(x^{(k)} - s \nabla f(x^{(k)}) \right)$$

$$\| x^{(k+1)} - x_* \| = \| P_X (x^{(k)} - s \nabla f(x^{(k)})) - P_X (x_* - s \nabla f(x_*)) \| \le$$

property 3 (non-expansive property)

$$\le \| (x^{(k)} - s \nabla f(x^{(k)})) - (x_* - s \nabla f(x_*)) \| = \| \underbrace{(I - sQ)} (x^{(k)} - x_*) \|$$

$$\le \max \{ |1 - s\lambda_{min}(Q)| , |1 - s\lambda_{max}(Q)| \} \| x^{(k)} - x_* \|$$

Optimize over s: $\dfrac{e^{(k+1)}}{e^{(k)}} \le \dfrac{c.d(Q) - 1}{c.d.(Q) + 1}$

Same rate as unconstrained optimization

(we drop the projection operator due to Property 3)

Difference: Unconstrained case $\rightarrow$ compute gradient

Constrained case $\longrightarrow$ optimization subproblem at every iteration.

What if c.d (Q) is large?

Change of variables:

$$x^T Q x \longrightarrow (x Q^{1/2})^T \underbrace{(Q^{1/2} x)}$$

$$y^{(k)} = Q^{1/2} x \implies \min_y h(y) = \min_x f(x)$$

$$h(y) = f(\underbrace{Q^{-1/2} y})$$

$$\begin{cases} y^{(k+1)} = y^{(k)} + \alpha^{(k)} (\bar{y}^{(k)} - y^{(k)}) \\ \bar{y}^{(k)} = P_Y (y^{(k)} - s^{(k)} \nabla h(y^{(k)})) \end{cases}$$

where $Y = \{ y \mid Q^{-1/2} y \in X \}$

Use the equations: $x^{(k)} = Q^{-1/2} y^{(k)}$, $\bar{x}^{(k)} = Q^{-1/2} \bar{y}^{(k)}$

$$\begin{cases} x^{(k+1)} = x^{(k)} + \alpha^{(k)} (\bar{x}^{(k)} - x^{(k)}) \\ \bar{x}^{(k)} = \arg\min_{x \in X} \left(\nabla f(x^{(k)})^T (x - x^{(k)}) + \frac{1}{2 s^{(k)}} (x - x^{(k)})^T Q (x - x^{(k)}) \right) \end{cases}$$

New condition number $= 1 \implies$ Convergence in one iteration

General case: Constrained Newton's method

$$\min f(x) \quad \text{s.t.} \quad x \in X$$

$$\implies \begin{cases} x^{(k+1)} = x^{(k)} + \alpha^{(k)} \left(\bar{x}^{(k)} - x^{(k)} \right) \\ \bar{x}^{(k)} = \underset{x \in X}{\arg\min} \left(\nabla f(x^{(k)})^T (x - x^{(k)}) + \frac{1}{2 s^{(k)}} (x - x^{(k)})^T \nabla^2 f(x^{(k)}) (x - x^{(k)}) \right) \end{cases}$$

Special case: $\alpha^{(k)} = s^{(k)} = 1$

$$\implies x^{(k+1)} = \underset{x \in X}{\arg\min} \left(\nabla f(x^{(k)})^T (x - x^{(k)}) + \frac{1}{2} (x - x^{(k)})^T \nabla^2 f(x^{(k)}) (x - x^{(k)}) \right)$$

$$x^{(k)} \longrightarrow x^{(k+1)} = x^{(k)} + \underbrace{\Delta x^{(k)}}_{x - x^{(k)}}$$

$\min\limits_{\Delta x}$ quadratic approximation of $f(x^{(k)} + \Delta x)$

s.t. $\Delta x = x - x^{(k)}$, $x \in X$

Recall general case:

$$\begin{cases} x^{(k+1)} = x^{(k)} + \alpha^{(k)} \left(\bar{x}^{(k)} - x^{(k)} \right) \\ \bar{x}^{(k)} = P_X \left(x^{(k)} - s^{(k)} \nabla f(x^{(k)}) \right) \end{cases}$$

Thm: If we use Armijo rule or Limited Line search for $\alpha^{(k)}$ and pick $s^{(k)}$ = constant & arbitrary positive $\Longrightarrow$ every limit point of $\{x^{(k)}\}$ is a stationary (FOC) point

---

Thm: If we pick $\alpha^{(k)} = 1$ and $s^{(k)}$ based on Armijo rule $\Longrightarrow$ every limit point of $\{x^{(k)}\}$ is a stationary point.

Thm (Projected gradient method):

Use the algorithm $x^{(k+1)} = P_X\left(x^{(k)} - s\nabla f(x^{(k)})\right)$

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

$\forall x, y \in X$

$-\quad 0 < s < \dfrac{2}{L} \rightarrow$ (Like the unconstrained case,

$\Rightarrow$ every Limit point of $\{x^{(k)}\}$ is stationary.

proof: $x^{(k+1)} = P_X\left(x^{(k)} - s\nabla f(x^{(k)})\right)$

$$\Rightarrow \left(\left(x^{(k)} - s\nabla f(x^{(k)})\right) - \underbrace{x^{(k+1)}}\right)^T\left(x - \underbrace{x^{(k+1)}}\right) \le 0$$

$\forall x \in X$

Property of projection

Pick $x = x^{(k)}$

$$\Rightarrow \left(\nabla f(x^{(k)})\right)^T\left(x^{(k+1)} - x^{(k)}\right) \le \frac{-1}{s}\|x^{(k+1)} - x^{(k)}\|^2 \quad (*)$$

Also,

$$f(x^{(k+1)}) \le f(x^{(k)}) + \nabla f(x^{(k)})^T\left(x^{(k+1)} - x^{(k)}\right) + \frac{L}{2}\|x^{(k+1)} - x^{(k)}\|^2 \quad (**)$$

(*), (**) $\Longrightarrow$ (***)

$$f(x^{(k+1)}) - f(x^{(k)}) \leq \underbrace{\left(\frac{-1}{s} + \frac{L}{2}\right)}_{<0 \text{ since } s < \frac{2}{L}} \|x^{(k+1)} - x^{(k)}\|^2$$

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

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

Consider subsequence corresponding to a Limit point of $\{x^{(k)}\}_1^\infty \longrightarrow \{x^{(k)}\}_{k \in K} \longrightarrow x_*$ Limit point

As $k \to \infty$, Left side of (***) $\longrightarrow 0$

right side of (***) ≤ 0

$\Longrightarrow$ right side of (***) $\longrightarrow 0$ as $k \to \infty$

$$\Longrightarrow \bar{x}^{(k)} - x^{(k)} \longrightarrow 0$$

$$\Longrightarrow P_X(x^{(k)} - s\nabla f(x^{(k)})) - x^{(k)} \longrightarrow 0$$

Take the Limit:

$$P_X(x_* - s\nabla f(x_*)) = x_* \implies x_* = \text{foc point}$$

Convergence rate of projected gradient method:

Thm: Assume: $-\ \nabla^2 f(x) \succeq m I$, $m>0$, $\forall x \in X$

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

$$\implies \|x^{(k+1)} - x_*\| \le \underbrace{\sqrt{1-\frac{m}{L}}}_{\text{Linear Convergence}} \|x^{(k)} - x_*\|$$

if $s = \frac{1}{L}$

Proof: $x^{(k+1)} = P_X\left(x^{(k)} - \frac{1}{L}\nabla f(x^{(k)})\right)$

$$\implies \left(\left(x^{(k)} - \frac{1}{L}\nabla f(x^{(k)})\right) - x^{(k+1)}\right)^T\left(x - x^{(k)}\right) \le 0 \quad \forall x \in X$$

property of Projection

Pick $x = x_*$

$$\implies \nabla f(x^{(k)})^T(x^{(k+1)} - x_*) \le L(x^{(k)} - x^{(k+1)})^T(x^{(k+1)} - x_*)$$

$$0 \le f(x^{(k+1)}) - f(\underbrace{x_*}_{\substack{\text{global} \\ \text{min} \\ \text{(convex)}}}) = \underbrace{\boxed{f(x^{(k+1)}) - f(x^{(k)})}}_{A}$$

$$+ \underbrace{\boxed{f(x^{(k)}) - f(x_*)}}_{B} \le \underbrace{\nabla f(x^{(k)})^T (x^{(k+1)} - x^{(k)}) + \frac{L}{2} \| x^{(k+1)} - x^{(k)} \|^2}_{A'}$$

$$+ \underbrace{\nabla f(x^{(k)})^T (x^{(k)} - x_*) - \frac{m}{2} \| x^{(k)} - x_* \|^2}_{B'}$$

A' = quadratic over-estimator

B' = quadratic under-estimator

$$\Rightarrow \left(0 \le \nabla f(x^{(k)})^T (x^{(k+1)} - x_*) + \frac{L}{2} \| x^{(k+1)} - x^{(k)} \|^2 - \frac{m}{2} \| x^{(k)} - x_* \|^2 \right) \quad (2)$$

$$(1), (2) \Rightarrow \left(\frac{L}{2} \| x^{(k+1)} - x^{(k)} \|^2 - \frac{m}{2} \| x^{(k)} - x_* \|^2 + L (x^{(k)} - x^{(k+1)})^T (x^{(k+1)} - x_*) \ge 0 \right) \quad (3)$$

Recall: $e^{(k+1)} = x^{(k+1)} - x_*$, $e^{(k)} = x^{(k)} - x_*$

③ $\Longrightarrow \quad 0 \leq \frac{L}{2} \| e^{(k+1)} - e^{(k)} \|^2 - \frac{m}{2} \| e^{(k)} \|^2$

$+ L (e^{(k)} - e^{(k+1)})^T e^{(k+1)}$

$= (\frac{L}{2} - \frac{m}{2}) \| e^{(k)} \|^2 - \frac{L}{2} \| e^{(k+1)} \|^2$

$\Longrightarrow \quad \| e^{(k+1)} \| \leq \sqrt{1 - \frac{m}{L}} \; \| e^{(k)} \|$

Summary

Conditional gradient	projected gradient
f: Convex $\Longrightarrow O(\frac{1}{\varepsilon})$	f: Convex $\Longrightarrow O(\frac{1}{\varepsilon})$
f: strongly Convex $\Longrightarrow O(\frac{1}{\varepsilon})$	f: strongly Convex $\Longrightarrow \underbrace{O(\log \frac{1}{\varepsilon})}_{\text{Linear Convergence}}$
$\begin{cases} f: \text{Convex} \\ X: \text{strongly convex} \end{cases} \Longrightarrow \underbrace{O(\log \frac{1}{\varepsilon})}_{\text{Linear Convergence}}$	

Can we do better than Linear Convergence ?

Thm (Constrained Newton's method):

Assume $\nabla^2 f(x) \succ 0 \quad \forall x \in X$

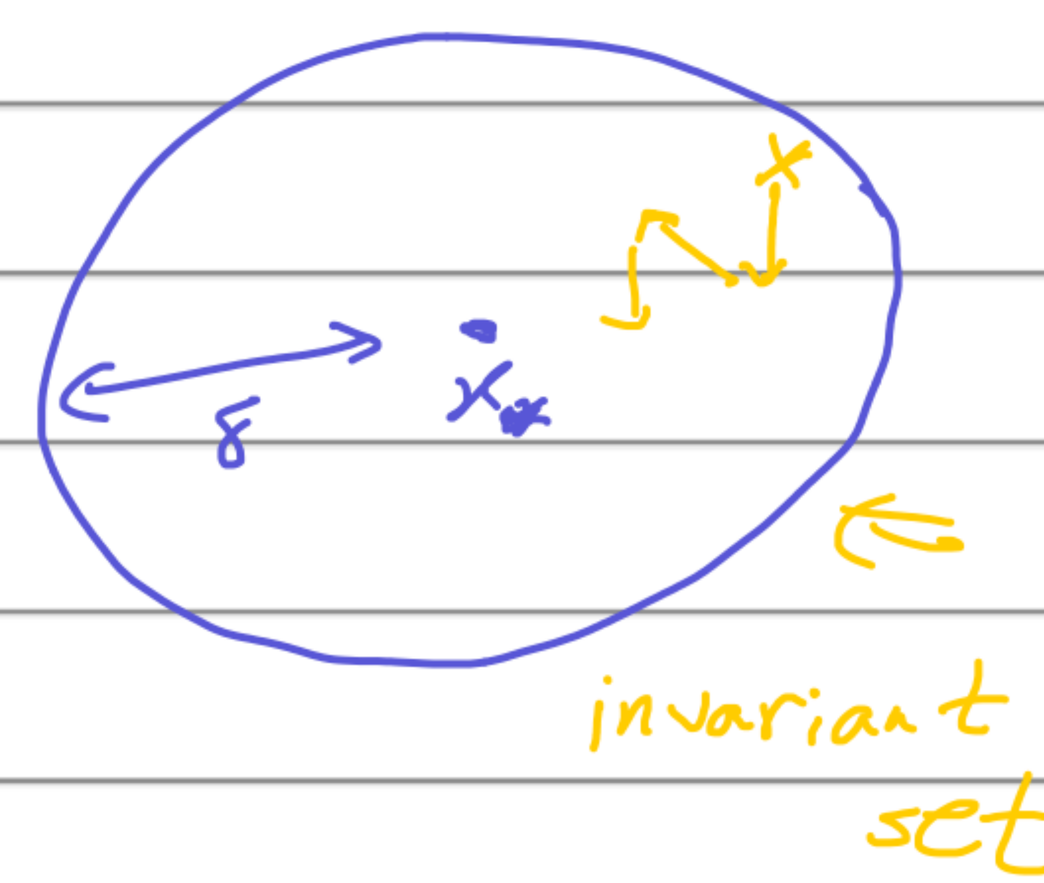
Let x_* be the global min of $\min\limits_{x \in X} f(x)$

Then $\exists \, \delta > 0$ s.t. $\|x^{(0)} - x_*\| < \delta$,

then $\{x^{(k)}\}$ generated by constrained Newton's method with $\alpha^{(k)} = s^{(k)} = 1$:

$$x^{(k+1)} = \operatorname*{argmin}_{x \in X} \left(\nabla f(x^{(k)})^T (x - x^{(k)}) + \frac{1}{2} (x - x^{(k)})^T \nabla^2 f(x^{(k)}) (x - x^{(k)}) \right)$$

has three properties:



1- $\|x^{(k)} - x_*\| \leq \delta \quad \forall k$

2- $x^{(k)} \to x_*$

3- $\|x^{(k)} - x_*\| \to 0$ superlinearly.

(like before Lipschitz continuity of $\nabla^2 f(x)$ may give us quadratic convergence)

Proximal algorithms (ideal for non-smooth objectives)

$$\min \ f(x) \quad \text{s.t. } x \in X$$

, assume:

- $f(x)$: convex over $\mathbb{R}^n$
- X : convex

Proximal algorithm:

$$x^{(k+1)} = \underset{x \in X}{\operatorname{argmin}} \left(f(x) + \underbrace{\frac{1}{2\alpha^{(k)}} \| x - x^{(k)} \|^2}_{\text{proximal term}} \right)$$

plays the role of stepsize

Proximal term - regularizes x with respect to $x^{(k)}$

At every iteration, the sub-problem resembles the original problem $\min_{x \in X} f(x)$, what is the point?

1- If $f(x)$ is twice differentiable:

$$\nabla^2 \left(f(x) + \frac{1}{2\alpha^{(k)}} \| x - x^{(k)} \|^2 \right) \succeq$$

$$\boxed{\left(\min_{x \in X} \operatorname{eig}(\nabla^2 f(x)) + \frac{1}{\alpha^{(k)}} \right)} \times I \qquad \forall x \in X$$

$\Rightarrow \alpha^{(k)}$ makes the objective strongly convex even if $f(x) \neq$ strongly convex

$\left(\min_{x \in X} \operatorname{eig}(\nabla^2 f(x)) = 0 \right)$

Also, if $f(x) =$ strongly convex, $\alpha^{(k)}$ improves the coefficient of strongly convex (m)

2 - Assume $f(x)$ is not differentiable.

Define: $\underbrace{M_{\alpha,f}}_{\text{Moreau envelope}}(z) = \inf_{x \in X} \left(f(x) + \frac{1}{2\alpha} \|x - z\|^2 \right)$ $\quad \alpha > 0$

$M_{\alpha,f}(z)$: $z \in \mathbb{R}^n$ not just $z \in X$

properties:

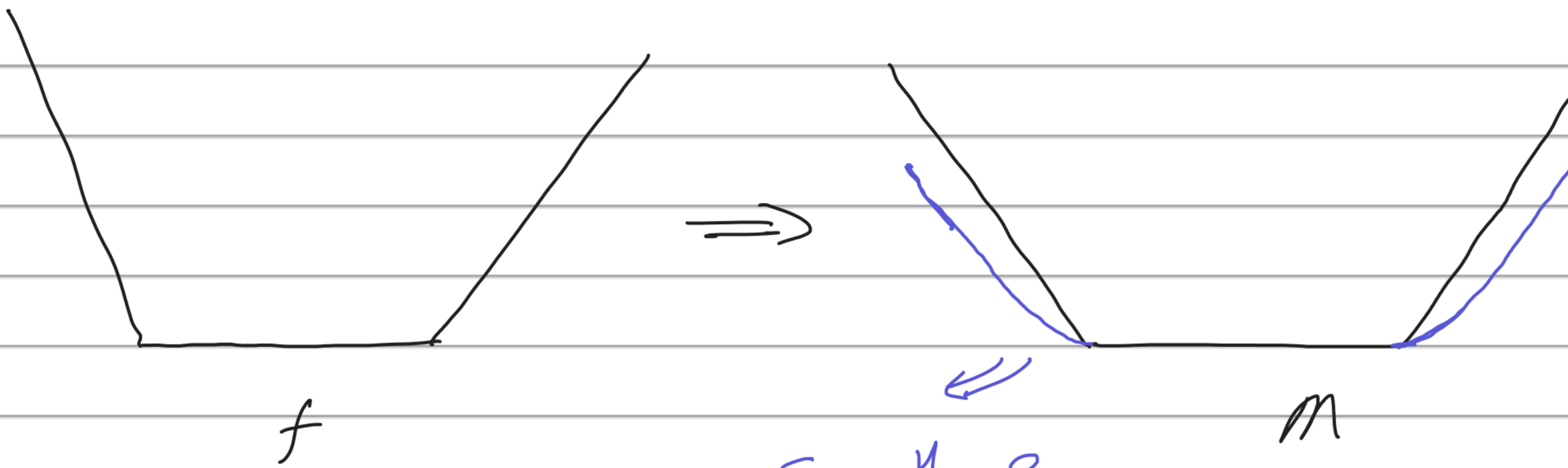
1 - $M_{\alpha,f}(z)$: convex & differentiable $\forall z \in \mathbb{R}^n$

Note: $g(x,z) = f(x) + \frac{1}{2\alpha} \|x - z\|^2$

: jointly convex in (x,z)

$$\inf_{x \in \underbrace{X}_{\text{Convex set}}} \underbrace{g(x,z)}_{\text{Convex function}} \longrightarrow \text{Convex function in } z$$

$$f(x) \longrightarrow M_{\alpha,f}(z) \xrightarrow{\text{relabel variable}} M_{\alpha,f}(x)$$



Smooth &

same set of global minima

$\boxed{\text{smoothing}}$ technique