



262B-Lecture 8

Date created: 2021.02.11
N. of Pages: 12

$$\min f(x) \rightarrow \min \frac{1}{2} x^T Q x \quad \text{where } Q > 0$$

Gradient alg: or scaled gradient alg:

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

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



Convergence rate $\approx \frac{\text{c.d.}(Q)-1}{\text{c.d.}(Q)+1}$

Convergence rate $\approx \frac{\text{c.d.}(D^{\frac{1}{2}} Q D^{\frac{1}{2}})-1}{\text{c.d.}(D^{\frac{1}{2}} Q D^{\frac{1}{2}})+1}$

Best convergence rate \Rightarrow smallest condition number

$$\Rightarrow \text{c.d.}(D^{\frac{1}{2}} Q D^{\frac{1}{2}}) = 1 \Rightarrow D^{\frac{1}{2}} Q D^{\frac{1}{2}} = I$$

$$\Rightarrow D = Q^{-1} = \nabla^2 f(x)^{-1} \Rightarrow \text{converge in one iteration}$$

This inspires Newton's method:

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \underbrace{\left(\nabla^2 f(x^{(k)}) \right)^{-1}}_{D^{(k)}} \nabla f(x^{(k)}) \rightarrow \Delta x^{(k)}$$

Is this a descent direction? yes if $\nabla^2 f(x^{(k)}) > 0$

$$\nabla f(x^{(k)})^T \Delta x^{(k)} = - \nabla f(x^{(k)})^T \underbrace{\left(\nabla^2 f(x^{(k)}) \right)^{-1}}_{D^{(k)}} \nabla f(x^{(k)}) \leq 0$$

Newton-like methods:

Thm (Convergence) Consider the algorithm

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

Assume $\{x^{(k)}\} \rightarrow x_*$ where $\nabla f(x_*) = 0 \rightarrow$ FOC

$\nabla^2 f(x_*) \succ 0 \rightarrow$ SOC
sufficient

Assume that $\nabla f(x^{(k)}) \neq 0 \quad \forall k$ and

$$\lim_{k \rightarrow \infty} \frac{\|\Delta x^{(k)} - \nabla^2 f(x_*)^{-1} \nabla f(x^{(k)})\|}{\|\nabla f(x^{(k)})\|} = 0$$

Armijo: $\alpha = 1, 0 < \beta < 1, 0 < \sigma < \frac{1}{2}$

Then: 1. $\lim_{k \rightarrow \infty} \frac{\|x^{(k+1)} - x_*\|}{\|x^{(k)} - x_*\|} = 0 \rightarrow$ superlinear convergence

2. $\exists \bar{k} \geq 0$ s.t. $\alpha^{(k)} = 1 \quad \forall k \geq \bar{k}$

↓

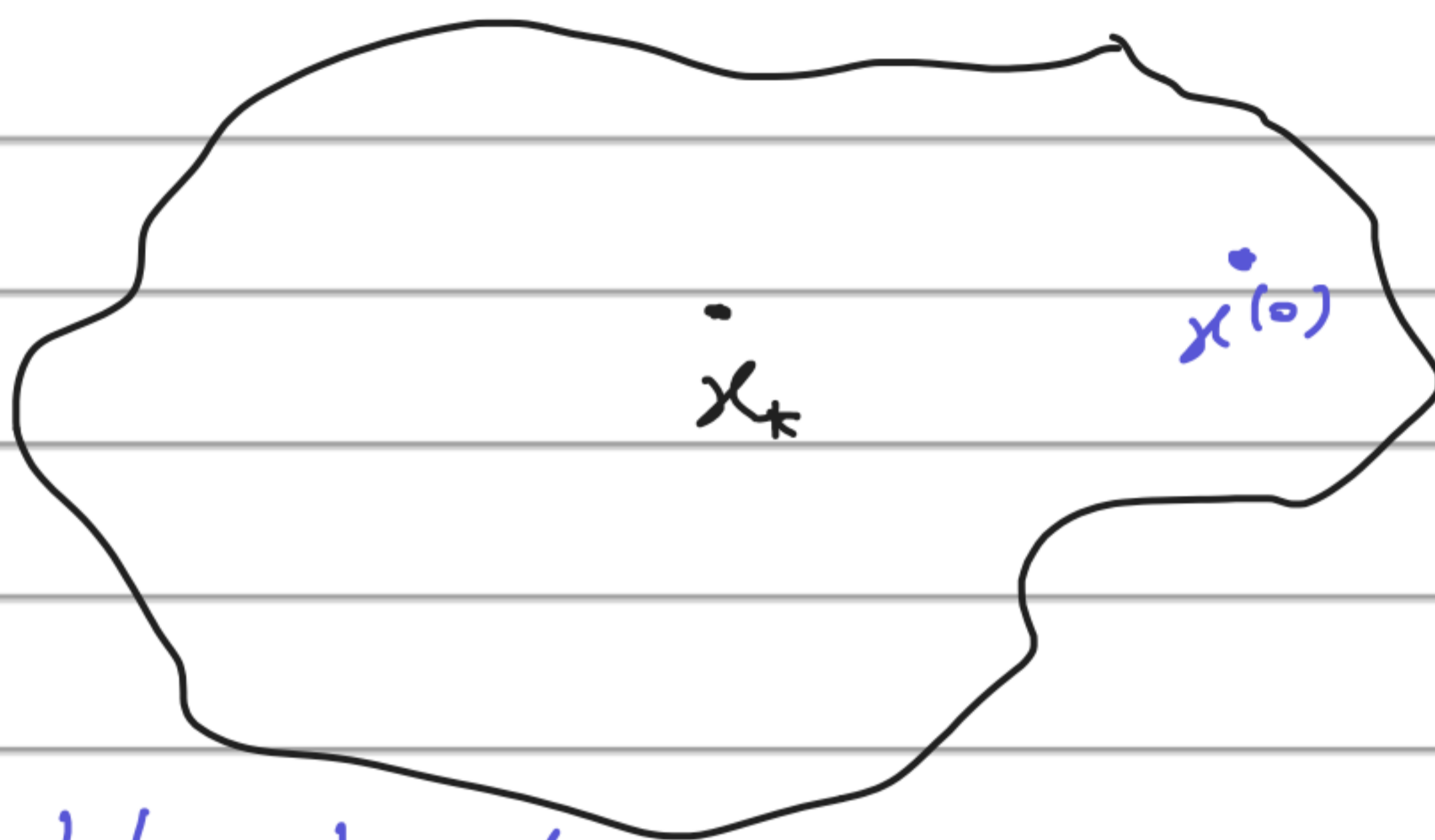
no reduction by Armijo after some time.

$\alpha^{(k)}$
⋮
 $\alpha = 1$
↓
 $\alpha \beta$
↓
 $\alpha \beta^2$
⋮

By-product: (this theorem & capture theorem)

pick x_* that satisfies FOC, SOC sufficient.

Then (capture) :



If $x^{(0)}$ is in a neighborhood

of x_* , then $\{x^{(k)}\} \rightarrow x_*$ if

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} \underbrace{\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})}_{\text{gradient related}}$$

gradient related

$$\frac{\| \Delta x^{(k)} - \nabla^2 f(x_*)^{-1} \nabla f(x^{(k)}) \|}{\| \nabla f(x^{(k)}) \|} \leq \| \nabla^2 f(x_*)^{-1} - \nabla^2 f(x^{(k)}) \|$$

$\rightarrow 0$ as $k \rightarrow \infty$

\Rightarrow Assumption is satisfied

\Rightarrow Newton's method has a 'super linear' convergence

if $x^{(0)}$ is close to x_* .

Under extra assumptions: quadratic convergence

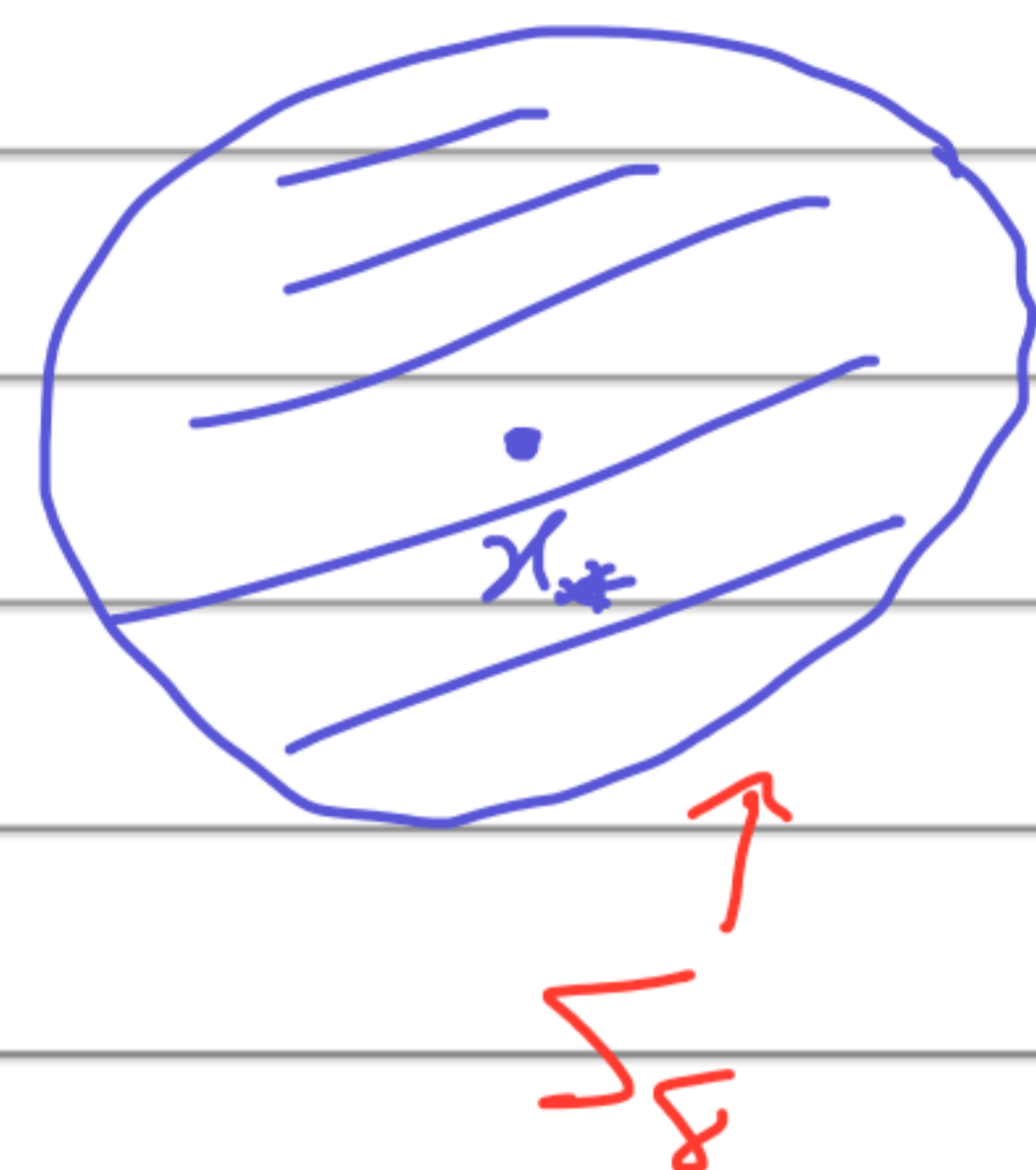
Thm: Assume $\exists L > 0, m > 0, \delta > 0$ s.t.

$$1 - \underbrace{\|\nabla^2 f(x) - \nabla^2 f(y)\|}_{\text{matrix}} \underbrace{\| \cdot \|_2}_{\substack{\text{2-norm} \\ \downarrow}} \leq L \underbrace{\|x-y\|_2}_{\substack{\text{vector} \\ \text{length}}} \quad \forall x, y \in S_\delta$$

$$2 - \nabla^2 f(x) \geq m I \quad \forall x \in S_\delta$$

$$3 - \frac{L\delta}{2m} < 1$$

where $S_\delta = \{x \mid \|x - x_*\| \leq \delta\}$



Then, $x^{(0)} \in S_\delta \implies$

$$1 - x^{(k)} \in S_\delta \quad \forall k \quad (\text{invariant set})$$

$$2 - \|x^{(k+1)} - x_*\| \leq \frac{L}{2m} \|x^{(k)} - x_*\|^2 \quad (\text{quadratic convergence})$$

Understand condition 1: about third derivatives

Condition 3: \rightarrow pick $\delta > 0$ to satisfy it

(if third derivatives don't exist \implies super linear but not quadratic)

proof by induction: base: $x^{(0)} \in \Sigma_g$

Assume $x^{(k)} \in \Sigma_g \xrightarrow{?} x^{(k+1)} \in \Sigma_g$

$$\nabla f(x^{(k)}) = \int_0^1 \frac{d \nabla f(x_* + t(x^{(k)} - x_*))}{dt} dt \quad \boxed{\nabla f(x_*) = 0}$$

$$= \left(\int_0^1 \nabla^2 f(x_* + t(x^{(k)} - x_*)) dt \right) \cdot (x^{(k)} - x_*) \quad (*)$$

(related gradient at the current point to Hessian over a line)

$$\|x^{(k+1)} - x_*\| = \|x^{(k)} - \underbrace{\nabla^2 f(x^{(k)})^{-1}} \nabla f(x^{(k)}) - x_*\|$$

$$= \underbrace{\| \nabla^2 f(x^{(k)})^{-1} \|}_{\text{matrix}} \underbrace{\| \nabla^2 f(x^{(k)}) (x^{(k)} - x_*) - \nabla f(x^{(k)}) \|}_{\text{vector}}$$

$$\leq \| \nabla^2 f(x^{(k)})^{-1} \|_2 \times \| \text{vector} \| \leq \frac{1}{m} \times \| \text{vector} \|$$

$$\underbrace{\text{max eig } \nabla^2 f(x^{(k)})^{-1}}_{\leq m^{-1}} \quad (**)$$

(use $x^{(k)} \in \Sigma_g$)

\circledast , \circledast \Rightarrow

$$\|x^{(k+1)} - x_*\| \leq \frac{1}{m} x$$

$$\left\| \int_0^1 (\nabla^2 f(x^{(k)}) - \nabla^2 f(x_* + t(x^{(k)} - x_*))) dt \cdot (x^{(k)} - x_*) \right\|_2$$

$$\leq \frac{1}{m} \int_0^1 \|\nabla^2 f(x^{(k)}) - \nabla^2 f(x_* + t(x^{(k)} - x_*))\|_2 dt \cdot \|x^{(k)} - x_*\|$$

$$\leq \frac{1}{m} \int_0^1 L \underbrace{\|x^{(k)} - (x_* + t(x^{(k)} - x_*))\|}_{(1-t)\|x^{(k)} - x_*\|} dt \cdot \|x^{(k)} - x_*\|$$

$$= \frac{\|x^{(k)} - x_*\|^2}{m} \times L \left(\int_0^1 (1-t) dt \right)^{1/2}$$

$$\Rightarrow \|x^{(k+1)} - x_*\| \leq \frac{L}{2m} \|x^{(k)} - x_*\|^2 \quad (\text{part } \underline{2})$$

$$\leq \frac{L}{2m} \|x^{(k)} - x_*\| \times \|x^{(k)} - x_*\|$$

$\leq \delta$ (by

assumption)

$$\leq \left(\frac{L\delta}{2m} \right) \|x^{(k)} - x_*\| \leq \|x^{(k)} - x_*\|$$

≤ 1

$$\leq \delta \Rightarrow x^{(k+1)} \in \Sigma_\delta$$

(part $\underline{1}$)

Analysis in the convex case:

Assumptions: 1 - $mI \preceq \nabla^2 f(x) \preceq MI \quad \forall x$
(Convex)

2 - $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L \|x - y\|$

Run Newton's method with Armijo rule s.t.

$$\alpha = 1, \quad 0 < \beta < 1, \quad 0 < \sigma < \frac{1}{2}$$

Define: $\eta = 3(1 - 2\sigma) \frac{m^2}{L}$

$$\gamma = \sigma \beta \eta^2 \frac{m}{m^2}$$

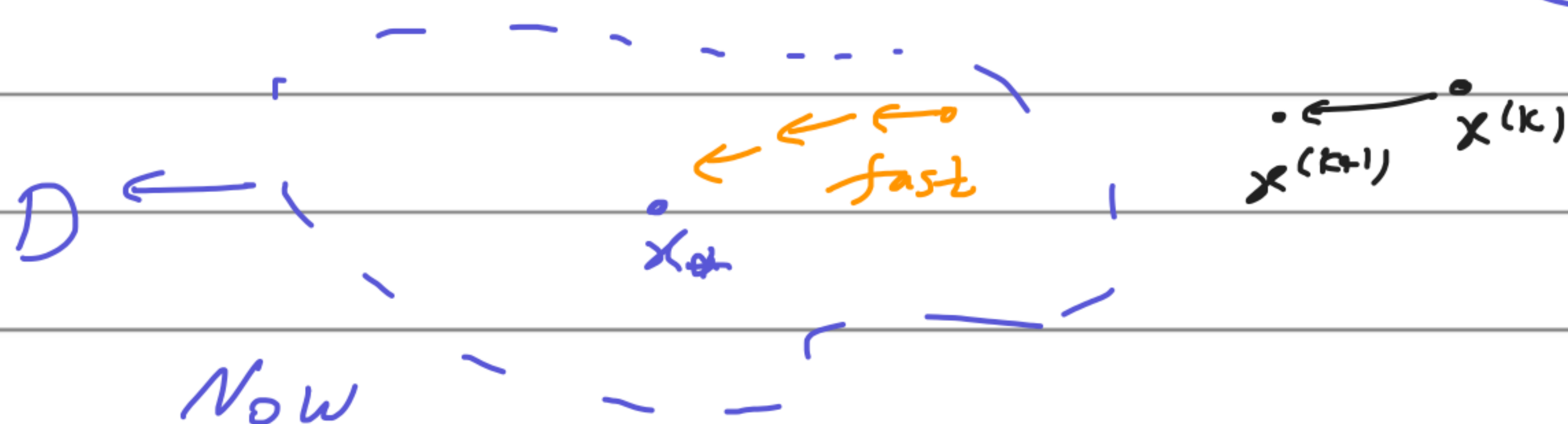
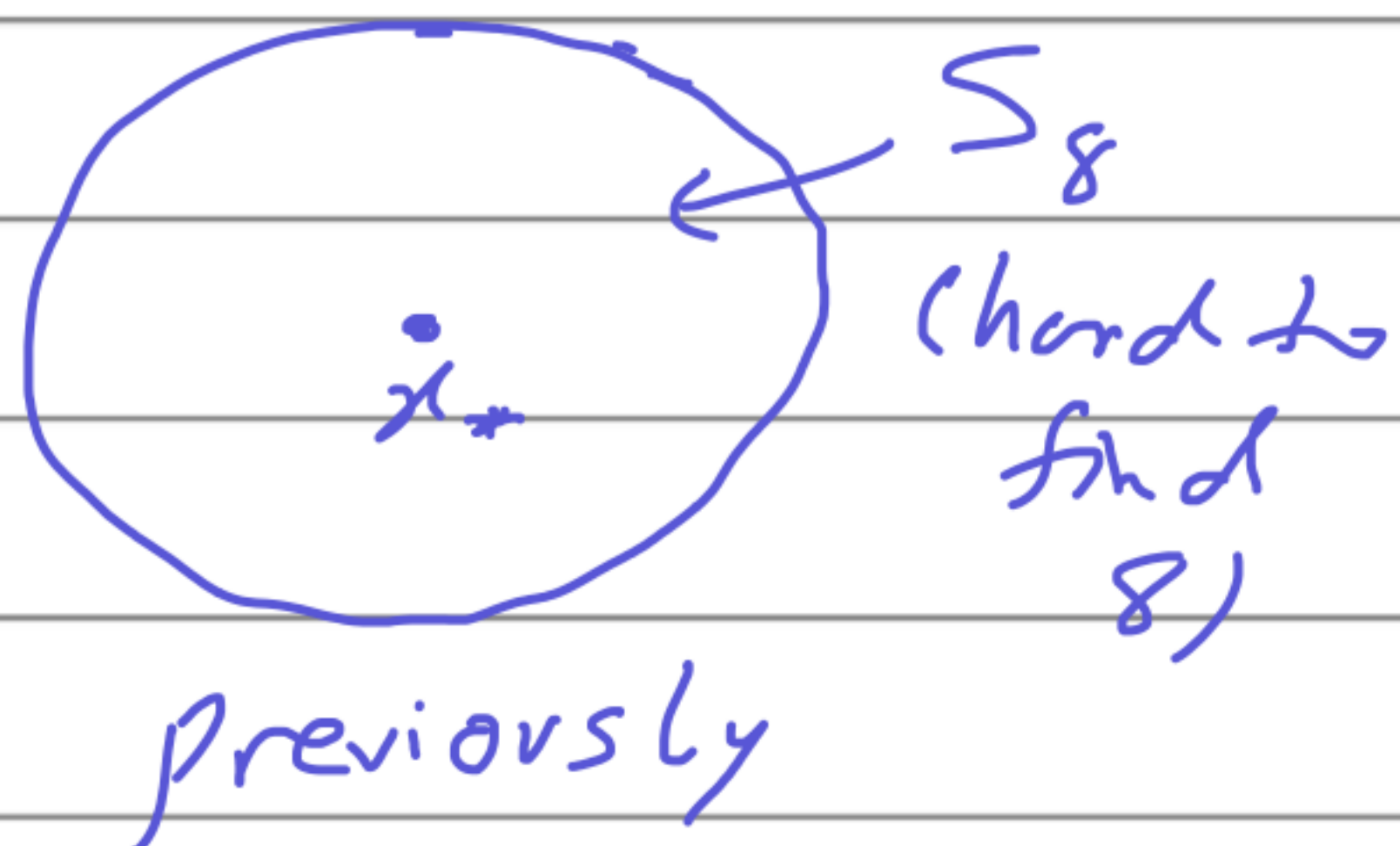
Define: $D = \{x \mid \|\nabla f(x)\| < \eta\}$

Then: 1 - If $x^{(k)} \notin D \Rightarrow f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$

2 - If $x^{(k)} \in D$ then $\alpha^{(k)} = 1$

Constant improvement

and have quadratic convergence.



Stopping criterion : $\| \nabla f(x^{(k)}) \| \leq \epsilon$

Instead: $\underbrace{\frac{1}{2} \nabla f(x^{(k)})^T \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})}_{A} \leq \epsilon$

$$x^{(k)} \rightarrow x^{(k+1)} : f(x^{(k)} + \Delta x^{(k)}) = f(x^{(k)}) + \nabla f(x^{(k)})^T \Delta x^{(k)} + \frac{1}{2} (\Delta x^{(k)})^T \nabla^2 f(x^{(k)}) \Delta x^{(k)}$$

A : difference between $f(x)$ & its quadratic approximation

Thm: Number of iterations =

get to set D + Converge inside D

$$\leq \frac{f(x^{(0)}) - f(x^*)}{\gamma} + \log_2 \log_2 \frac{(2m^3/L^2)}{\epsilon}$$

$$= \underbrace{O(\log \log \frac{1}{\epsilon})}_{\text{almost constant}}$$

almost constant

What if $f(x)$ is non convex & $x^{(0)}$

is far away from x_* ?

Then $\nabla^2 f(x^{(k)}) \not> 0$

Method 1: Design $\Delta^{(k)}$ s.t.
Correction

$$\Delta^{(k)} + \nabla^2 f(x^{(k)}) > 0$$

$$\text{Then: } \Delta x^{(k)} = - \left(\Delta^{(k)} + \nabla^2 f(x^{(k)}) \right)^{-1} \nabla f(x^{(k)})$$

$$, \Delta^{(k)} \rightarrow 0 \text{ as } k \rightarrow \infty$$

Method 2: Trust region

$$f(x^{(k)} + \Delta x^{(k)}) \xrightarrow{\text{approximate}} \left(f(x^{(k)}) + \nabla f(x)^T \Delta x^{(k)} + \frac{1}{2} (\Delta x^{(k)})^T \nabla^2 f(x^{(k)}) \Delta x^{(k)} \right)$$

Quadratic approximation

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

$$\Rightarrow \Delta x_* = - \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$$

if Hessian > 0

If Hessian $\nabla^2 f(x^{(k)})$ is not positive definite, then just use the optimization directly:

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

Solution could be $-\infty$

So, we should do a restriction to a local region, named trust region

$$\{ \Delta x \mid \|\Delta x\| \leq \gamma^{(k)} \}$$

$\Delta x^{(k)}$: solution to the following optimization:

$$\begin{cases} \min_{\Delta x} f(x^{(k)}) + \nabla f(x^{(k)})^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x^{(k)}) \Delta x \\ \text{s.t.} \quad \|\Delta x\| \leq \gamma^{(k)} \end{cases} \implies (\Delta x)^T \Delta x \leq (\gamma^{(k)})^2$$

Zero duality gap (duality) (S-lemma) : $\lambda^{(k)}$ = Lagrange multiplier for constraint

Move the constraint up:

$$\min_{\Delta x} f(x^{(k)}) + \nabla f(x^{(k)})^T \Delta x + \frac{1}{2} \Delta x^T \left(\nabla^2 f(x^{(k)}) + \lambda^{(k)} \times 2 \times I \right) \Delta x$$

$$\Rightarrow \Delta x_* = - \left(\nabla^2 f(x^{(k)}) + 2\lambda^{(k)} I \right)^{-1} \nabla f(x^{(k)})$$

Method $\underline{=}$

Thm (trust region) : If $\gamma^{(k)}$ is small enough,

$$\text{then } f(x^{(k)} + \Delta x_*) < f(x^{(k)})$$

$n = \text{step size } (\alpha^{(k)} = 1)$

unless $x^{(k)}$: FOC, SOC necessary

Newton's method : second-order method

since $x^{(k)}$ depends on second derivatives.

second-order
methods

first-order
methods

fast
($\log \log \frac{1}{\epsilon}$)

slow
($\frac{1}{\epsilon}$ or $\log \frac{1}{\epsilon}$)

expensive

cheap

Gradient : Compute $\nabla f(x^{(k)}) \rightarrow$ cheap

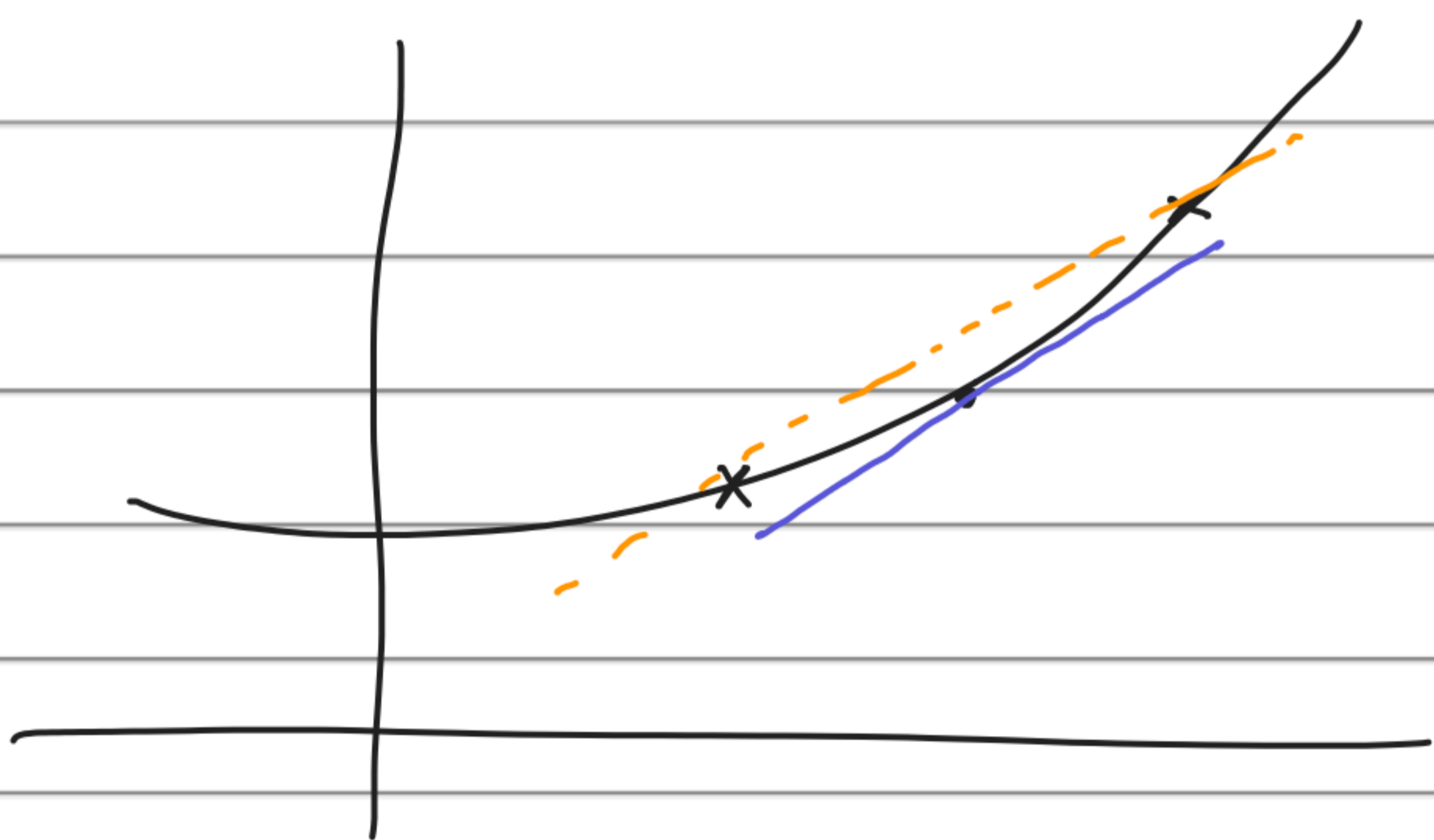
Newton : Compute $-\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$

\rightarrow expensive

Can we design a first-order method

that mimics the behavior of a second-

order method? quasi-newton



approximate gradient

by a slope obtained

from two points.

$$\Rightarrow \nabla^2 f(x^{(k+1)}) (x^{(k+1)} - x^{(k)}) \approx \frac{\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})}{x^{(k+1)} - x^{(k)}}$$