Chapter 4. Newton method.

4.1 Algorithm:

> **Newton method**
>
> - Initial $x = x^0$
> - For $k = 0, 1, 2, \cdots$
>
> $$\Delta x_{nt}^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$$
>
> $$x^{k+1} = x^k + \eta^k \cdot \Delta x_{nt}^k$$

GD: $\Delta x_{gd} = -\nabla f(x)$

Δx_{nt}^k : Newton direction.

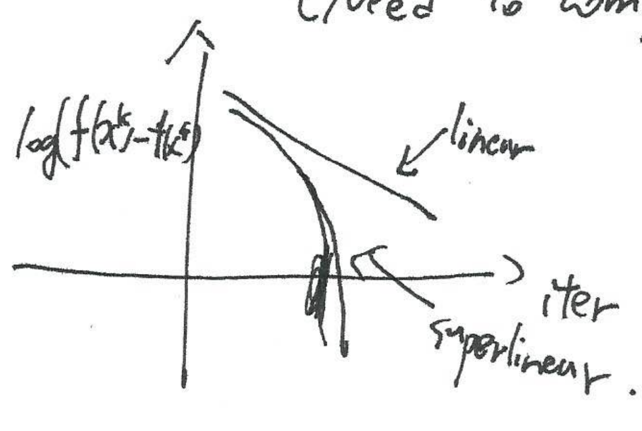
η^k : step size (usually chosen by backtracking line search + sufficient decrease condition)

Newton method: for strongly convex function:

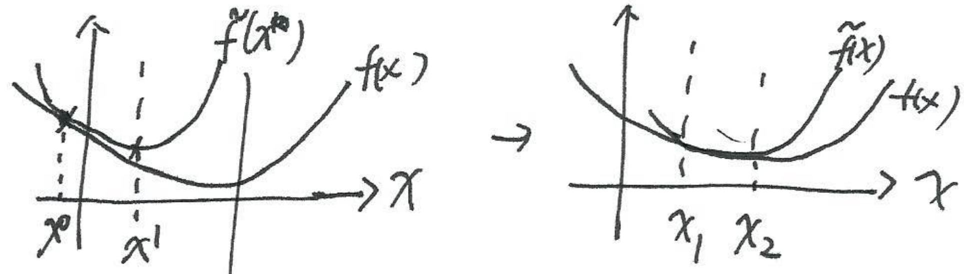
$$\nabla^2 f(x) \succeq mI, \quad m > 0$$

Properties: ① Faster convergence (Superlinear)

② Slower computation (Need to compute $\nabla^2 f(x)^{-1} \nabla f(x)$)



Intuitions: Another way to explain Newton method:



At each iteration, form the "best" quadratic approximation around x^k.

$$\tilde{f}(x^k+p) = f(x^k) + \nabla f(x^k)^T p + \frac{1}{2} p^T \nabla^2 f(x^k) p$$
$$\approx f(x^k+p)$$

Find the minimizer of $p^* = \arg\min \tilde{f}(x^k+p)$

$$\nabla f(x^k) + \nabla^2 f(x^k) p^* = 0$$
$$p^* = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$$

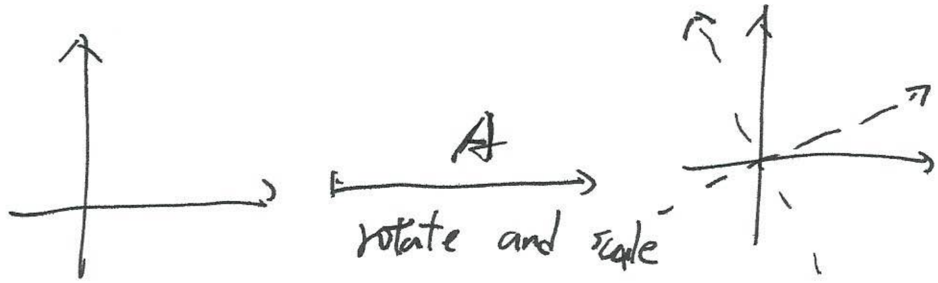
Comparing to gradient descent:

$$\tilde{f}(x^k+p) = f(x^k) + \nabla f(x^k)^T p + \frac{1}{2} p^T \boxed{I} p$$
$$p^* = -\nabla f(x^k)$$

4.2 Affine invariant:

Consider two optimization problems

$$\min_x f(x) \xmapsto{\;A\;} \min_y g(y) = f(A^{-1}y)$$



$$x \longmapsto y \; (= Ax)$$

$$x^1, x^2, \cdots \qquad f(x) \longmapsto g(y) \; (= f(A^{-1}y)) \qquad y^1, y^2, \cdots$$

$$x^0 \longmapsto y^0 \; (= Ax^0)$$

Def: The algorithm is "affine invariant"

if $\forall\, k = 1, 2, \cdots$, $y^k = Ax^k$

This means the algorithm will not change if we change the basis.

Property: Gradient Descent is "not" affine invariant,

Why? Original space $\xmapsto{\;A\;}$ y-space

$$y^1 \leftarrow y^0 - \nabla g(y^0)$$

$$x^1 \leftarrow x^0 - \nabla f(x^0) \qquad = y^0 - \frac{\partial}{\partial y} f(A^{-1}y)\Big|_{y=y^0}$$

$$= y^0 - A^{-T} \frac{\partial}{\partial x} f(x)\Big|_{x = x^0 = A^{-1}y}$$

$$\downarrow \qquad = y^0 - A^{-T} \nabla_x f(x^0)$$

$$y^1 \neq Ax^1 \neq y^1 = Ax^0 - A^{-T}\nabla_x f(x^0)$$

$$\| \quad Ax^0 - A\nabla f(x^0)$$

Thm: Newton method is affine invariant.

pf: $\Delta x_{nt}^{k} = -\nabla^2 f(x^k)^{-1}\cdot \nabla f(x^k)$

$$x^{k+1} = x^k - \underbrace{\eta^k}_{=1}\cdot \Delta x_{nt}^{k}$$

Original space $\longmapsto$ y-spase.

$$\cancel{x^{k+1} = x^k - \nabla_x f(x^k)}$$

$$x^{k+1} = x^k - \nabla^2 f(x^k)^{-1}\cdot \nabla f(x^k)$$

$$\cancel{y^{k+1} = y^k - \nabla_y g(y^k)}$$

$$\cancel{y^{k+1} = y^k - \nabla g(y^k}$$

$$
\begin{aligned}
y^{k+1} &= y^k - \nabla^2 g(y^k)^{-1}\nabla g(y^k)\\
&= y^k - \left(\frac{\partial^2}{\partial y^2} f(A^{-1}y)\Big|_{y=y^k}\right)^{-1}\cdot\left(\frac{\partial}{\partial y} f(A^{-1}y)\Big|_{y=y^k}\right)\\
&= y^k - \left(A^{-T}\frac{\partial^2}{\partial x^2} f(x^k)A^{-1}\right)^{-1}\cdot\left(A^{-T}\cdot\frac{\partial}{\partial x}f(x^k)\right)\\
&= y^k - A\cdot\nabla^2 f(x^k)^{-1}\cdot \cancel{A^{T}}\cdot\cancel{A^{-T}}\cdot\nabla f(x^k)\\
&= y^k - A\cdot\nabla^2 f(x^k)^{-1}\nabla f(x^k)\\
&= Ax^k - A\cdot\nabla^2 f(x^k)^{-1}\nabla f(x^k).
\end{aligned}
$$

$\left(g(y) = f(A^{-1}y)\right)$

$x = A^{-1}y$

$x^k = A^{-1}y^k$

$$Ax^{k+1} = y^{k+1}$$

#

4.3 Newton Decrement.

Newton Decrement:

$$\lambda(x) = \left(\nabla f(x)^T \cdot \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2}$$

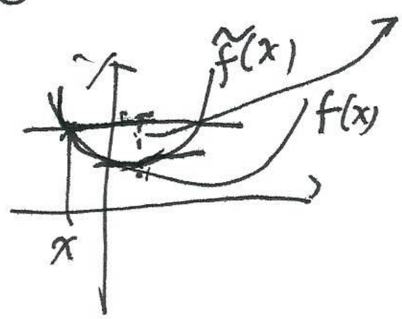
① In the line search (sufficient decrease condition).

$$f(x+\eta \Delta x_{nt}) \le f(x) + c\nabla f(x)^T \eta \Delta x_{nt}$$

$$= f(x) + c\eta \cdot \nabla f(x)^T \left(\nabla^2 f(x)^{-1} \nabla f(x)\right)$$

$$= f(x) + c \cdot \eta \cdot \lambda(x)^2$$

② Thm: $f(x) - \inf_y \tilde{f}(y) = \frac{1}{2}\lambda(x)^2$.



why? $\tilde{f}(y) = f(x) + \nabla f(x)^T (y - \overset{\nabla^2 f(x)^{-1}\nabla f(x)}{x}) + \frac{1}{2}(y-x)^T \nabla^2 f(x) \cdot (y-x)$

$$x + \Delta x^*_{nt} = \arg\min_y \tilde{f}(y)$$

$$\tilde{f}(x+\Delta x^*_{nt}) = f(x) - \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) + \frac{1}{2}\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla^2 f(x) \cdot \nabla^2 f(x)^{-1} \cdot \nabla f(x)$$

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

$$= f(x) - \frac{1}{2}\underbrace{\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)}_{\lambda(x)^2}$$

③ $\lambda(x)$ can be used as stopping condition.

Newton' method { with line search & stopping condition

Initial x^0

For $k = 0, 1, 2, \cdots$

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

$\lambda^2 = \nabla f(x^k)^T \nabla^2 f(x^k)^{-1} \nabla f(x) = -\nabla f(x^k)^T \Delta x_{nt}^k$

② Stop if $\lambda^2/2 \leq \varepsilon$ (or $\lambda^2/2 \leq \varepsilon \cdot \lambda_0^2/2$)

③ For $\eta = 1, \beta, \beta^2, \cdots$

Quit if $f(x^k + \eta \Delta x_{nt}^k) \leq f(x^k) + c \cdot \eta \cdot \lambda^2$

④ Update $x^{k+1} = x^k + \eta \cdot \Delta x_{nt}^k$

GD : $0 < c < 1$

Newton : $0 < c < \frac{1}{2}$

4.4 Convergence.

Assumptions:

① f is strongly convex, twice differentiable.

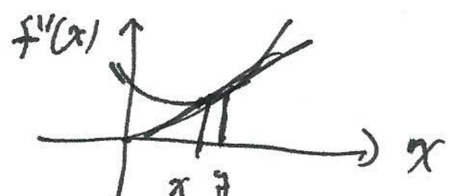
M- Lipchitz :

$$mI \preceq \nabla^2 f(x) \preceq MI \quad (m, M > 0)$$

② $\nabla^2 f(x)$ is Lipchitz continuous:

$$\| \nabla^2 f(x) - \nabla^2 f(y) \|_2 \leq L \|x - y\|_2 \quad \forall x, y.$$

$\Downarrow$



~~if f is~~

if $\nabla^3 f(x)$

$\Rightarrow \| \nabla^3 f(x) \| \leq L$

Convergence properties : the iterations fall into "two" phases

- Phase I : the "global" or "damped." phase.
- Phase II : the "local" or "quadratic" phase.

There are numbers $0 < \alpha \leq m^2/M$ and $\gamma > 0$, st.

- Phase I: if $\|\nabla f(x^k)\|_2 \geq \alpha$, then

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

- Phase II : if $\|\nabla f(x^k)\|_2 < \alpha$, then

① The backtracking line search give you $\eta = 1$

quadratic $\xrightarrow{\text{convergence}}$ ② $\frac{M}{2m^2} \|\nabla f(x^{k+1})\|_2 \leq \left(\frac{M}{2m^2} \|\nabla f(x^k)\|_2 \right)^2$

Lipchitz continuous. (Clarify the definition)

Def : If a function f is Lipchitz continuous

iff $$\|f(x) - f(y)\| \le L \cdot \|x - y\| \quad \forall x, y$$

L : Lipchitz constant , $L > 0$.

If f is differentiable :

f is L-Lipchitz continuous iff $\|\nabla f(x)\| \le L$ $\forall x$

When we say f is L-Lipchitz

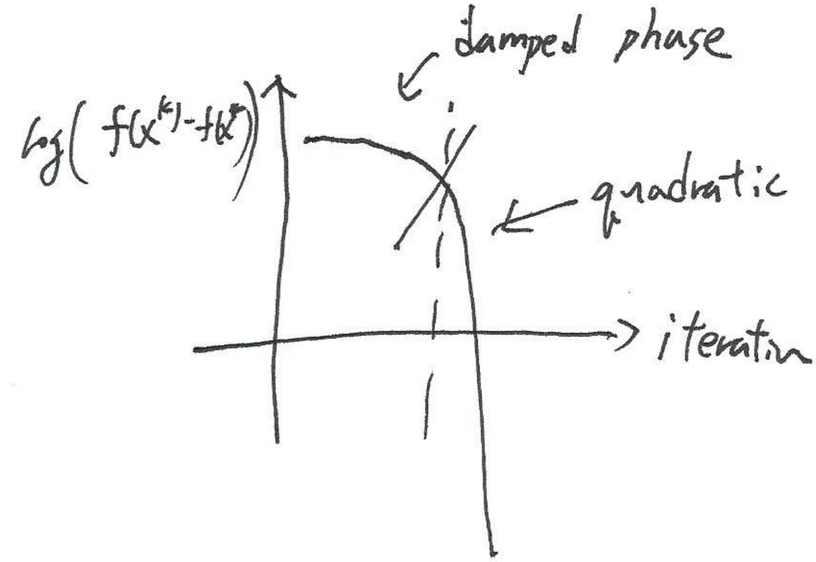
$\iff$ ∇f is L-Lipchitz continuous

✓ $\iff \|\nabla f(x) - \nabla f(y)\| \le L \|x - y\| \quad \forall x, y$

In Newton

we want $\nabla^2 f$ is L-Lipchitz continuous

✓ $\iff \|\nabla^2 f(x) - \nabla^2 f(y)\| \le L \|x - y\|$



How many iterations do we need to get "ε - accurate solution" ?

$\Downarrow$

$$f(x^k) - f(x^*) \leq \varepsilon$$

Phase I : $f(x^{k+1}) - f(x^k) \leq -\gamma$

l iterations in phase I: $l \leq \dfrac{f(x^0)-f(x^*)}{\gamma}$

Phase II : $\|\nabla f(x^k)\| < \alpha \leq \dfrac{m^2}{M}$ & $\dfrac{M}{2m^2}\|\nabla f(x^{k+1})\|_2 \leq \left(\dfrac{M}{2m^2}\|\nabla f(x^k)\|\right)^2$

$\Downarrow$

a iterations

① After a iterations :

$$\frac{M}{2m^2}\|\nabla f(x^{l+a})\| \leq \left(\frac{M}{2m^2}\|\nabla f(x^{l})\|\right)^{2^a}$$

$$\leq \left(\frac{1}{2}\right)^{2^a}$$

$$\|\nabla f(x^{l+a})\| \leq \frac{2m^2}{M}\left(\frac{1}{2}\right)^{2^a}$$

② Connect $\|\nabla f(x^{l+a})\|$ to $f(x^{l+a}) - f(x^*)$

$$\forall y, x \qquad f(y) \ge \left\{ f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}\|y-x\|^2 \right\} := g(y)$$

$$\tilde{y} = \operatorname*{argmin}_y g(y)$$

$$\nabla f(x) + m(\tilde{y} - x) = 0$$

$$\tilde{y} = x - \frac{1}{m}\nabla f(x)$$

$$f(x^*) \ge g(x^*) \ge g(\tilde{y}) = f(x) - \nabla f(x)^T \left(\frac{1}{m}\nabla f(x)\right) + \frac{m}{2}\left\|\frac{1}{m}\nabla f(x)\right\|^2$$

$$= f(x) - \frac{1}{m}\|\nabla f(x)\|^2 + \frac{1}{2m}\|\nabla f(x)\|^2$$

$$= f(x) - \frac{1}{2m}\|\nabla f(x)\|^2$$

$$f(x^{l+a}) - f(x^*) \le \frac{1}{2m}\|\nabla f(x^{l+a})\|^2$$

$$\le \frac{1}{2m}\left(\frac{2m^2}{M}\left(\frac{1}{2}\right)^{2^a}\right)^2$$

$$= \frac{2m^3}{M^2}\cdot\left(\frac{1}{2}\right)^{2^{a+1}} := \varepsilon_0 \cdot \left(\frac{1}{2}\right)^{2^a}$$

Want: $\varepsilon_0 \cdot \left(\frac{1}{2}\right)^{2^a} < \varepsilon$

$$\Rightarrow a \ge \log_2 \log_2 (\varepsilon_0 / \varepsilon)$$

Total number of iterations

$$l + a \le \frac{f(x^0) - f(x^*)}{\gamma} + \log_2\log_2(\varepsilon_0/\varepsilon)$$

#

Damped Newton Phase:

Thm: $\exists\ \alpha \le \frac{m^2}{M}$ such that
if $\|\nabla f(x^k)\| \ge \alpha$, then $f(x^{k+1}) - f(x^k) \le -\gamma$

pf: Let $x = x^k$, $x^+ = x^{k+1}$

$$x^+ = x - \eta \cdot \nabla^2 f(x)^{-1} \nabla f(x) = x + \eta\, \Delta x_{nt}$$

$$f(x^+) \le f(x) + \nabla f(x)^T (\eta \cdot \Delta x_{nt}) + \frac{M}{2} \|\eta \cdot \Delta x_{nt}\|^2$$

$$= f(x) - \eta \cdot \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$$

$$+ \frac{M}{2} \eta^2 \cdot \nabla f(x)^T \nabla^2 f(x)^{-1} \cdot \nabla^2 f(x)^{-1} \cdot \nabla f(x)$$

$$\le f(x) - \eta \cdot \lambda(x)^2 + \frac{\eta^2 M}{2m} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$$

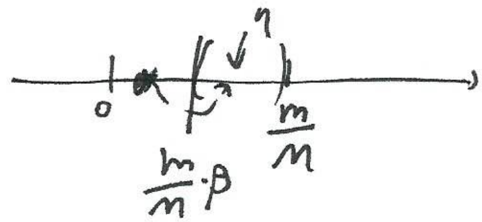
$$= f(x) - \eta \lambda(x)^2 + \frac{M\eta^2}{2m} \lambda(x)^2$$

When $\eta \le \frac{m}{M}$ $\Rightarrow f(x^+) - f(x) \le -\eta \lambda(x)^2 + \frac{\eta}{2} \lambda(x)^2$

$$= -\frac{\eta}{2} \lambda(x)^2$$

$(c < \frac{1}{2}) \rightarrow$

$$\le -c \cdot \eta \cdot \nabla f(x)^T \Delta x_{nt}$$



step size $\ge \frac{m}{M} \beta$.

$$\Rightarrow f(x^+) - f(x) \le -c \cdot \eta \cdot \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$$

$$\le -c \cdot \frac{m}{M} \beta \cdot \left(\frac{1}{M} \cdot \|\nabla f(x)\|^2\right)$$

$$\le \underbrace{-c \cdot \frac{m}{M^2} \beta \cdot \alpha^2}_{= \gamma}$$

#

Quadratic convergence phase:

① Prove $\eta = 1$ after $\|\nabla f(x)\| \le \alpha \le m^2/M$

(see page 490–491 "Convex Optimization")

② Quadratic convergence when $\|\nabla f(x)\| \le \alpha \le m^2/M$

pf : $\|\nabla f(x^+)\|_2 = \|(\nabla f(\overset{x^+}{x+\Delta x_{nt}}) - \nabla f(x)) - \underbrace{\nabla^2 f(x)\Delta x_{nt}}_{(-\nabla f(x))}\|_2$

$$= \left\| \int_0^1 \nabla^2 f(x+\alpha\Delta x_{nt})\Delta x_{nt}\, d\alpha - \int_0^1 \nabla^2 f(x)\Delta x_{nt}\, d\alpha \right\|_2$$

$$= \left\| \int_0^1 (\nabla^2 f(x+\alpha\Delta x_{nt}) - \nabla^2 f(x))\Delta x_{nt}\, d\alpha \right\|_2$$

L-Lipschitz of $\nabla^2 f$

$$\le \left\| \int_0^1 L\cdot\|\alpha\cdot\Delta x_{nt}\|\cdot\|\Delta x_{nt}\|\, d\alpha \right\|_2$$

$$= L\|\Delta x_{nt}\|^2 \cdot \int_0^1 \alpha\, d\alpha$$

$$= \frac{L}{2}\|\nabla^2 f(x)^{-1}\nabla f(x)\|^2$$

$$\le \frac{L}{2m^2}\|\nabla f(x)\|^2$$

#