3. Line Search.

In practice, you don't know L. (largest eigenvalue of Hessian)

How to choose η^k s.t. $\{x^k\} \to$ stationary points.

3.1 Consider a general "descent method"

$$\boxed{\begin{array}{l} \text{For } k=0,1,\cdots \\ \quad p^k = (B^k)\nabla f(x^k) \\ \quad x^{k+1} = x^k - \eta^k p^k \end{array}} \longrightarrow \text{fine } \eta^k \text{ (line search)}$$

$\|Xw-y\|^2$

$= w^T X^T X w -$

$(Xw-y)^T(Xw-y)$

$= w^T X^T X w - 2y^T X w + y^T y$

$B^k = I \Rightarrow$ gradient descent

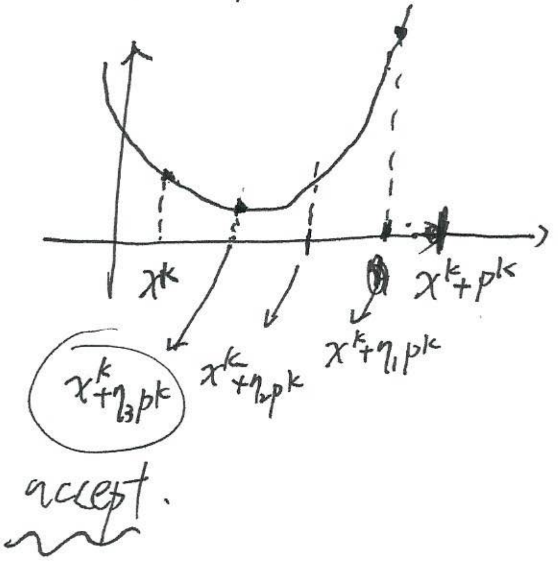
$B^k = \nabla^2 f(x)^{-1} \Rightarrow$ Newton's method.

The general "line search" strategy

Generate a sequence

$\eta_1, \eta_2, \eta_3, \cdots \in \mathbb{R}$

Stop when η_i satisfy some condition.



① What's the sequence?

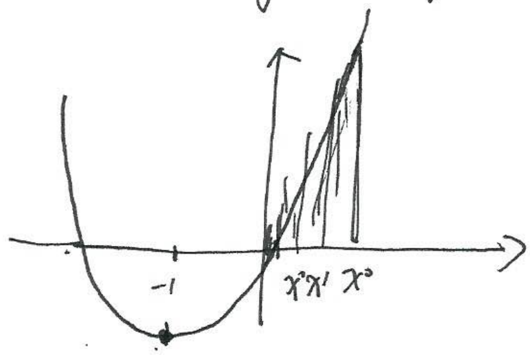
$\Rightarrow$ ② ~~Mean to~~ The condition for accepting a step size.

- A naive but wrong approach.

Condition: ~~$f(x^k+\eta p^k)$~~

$\underbrace{f(x^k+\eta p^k) < f(x^k)}$, then accept.

- No convergence guarantee:



$f(x)=(x+1)^2-1$

$\underbrace{\{x^k\}} = \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \cdots \frac{1}{2^k} \cdots\}$

$x^k \to 0$

Reason: step size is too small

$x^{k+1} \leftarrow x^k + \eta p^k$

Def: p is a descent direction iff $p^T \nabla f(x) < 0$

Thm: If p is a descent direction, then for any $c_1 \in (0,1)$

$$\exists \eta \text{ st } f(x+\eta p) \le f(x) + \underbrace{c_1 \eta \cdot \nabla f(x)^T p}_{<0}.$$

pf: $f(x+\eta p) = f(x) + \eta \cdot \nabla f(x)^T p + O(\eta^2) \leftarrow \eta^2 p^T \nabla^2 f(z) p$

$$= f(x) + c_1 \cdot \eta \cdot \nabla f(x)^T p + \underbrace{(1-c_1)\cdot \eta \cdot \nabla f(x)^T p + O(\eta^2)}$$

When η is small enough,

$$(1-c_1)\cdot \eta \cdot \underbrace{\nabla f(x)^T p}_{<0} < \underbrace{O(\eta^2)}$$

$$\le f(x) + c_1 \eta \cdot \nabla f(x)^T p$$

#

Thm: If $B \succ 0 I$ (symmetric), then $-B \nabla f(x)$ is a descent direction.

pf: $p^T \nabla f(x) = -\nabla f(x)^T B \nabla f(x) < 0$

#

3.2 Sufficient Decrease Condition.

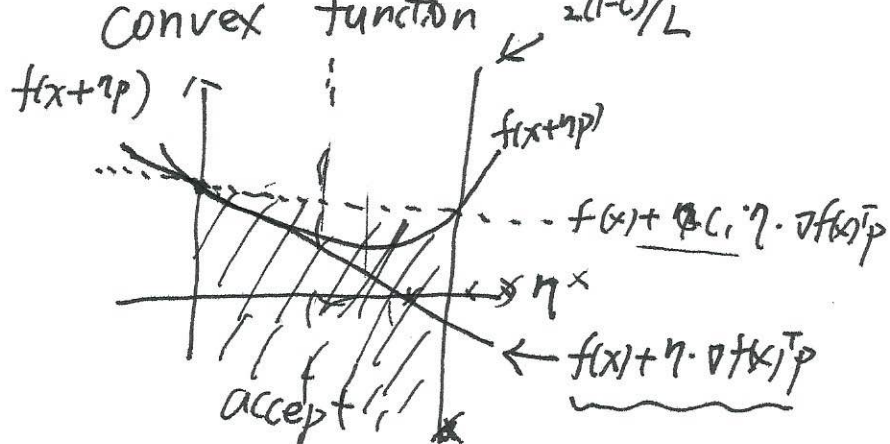
The step size η satisfy "sufficient decrease condition" if

$$f(x+\eta p) \le f(x) + c_1 \cdot \eta \cdot \nabla f(x)^T p$$

for some $c_1 \in (0,1)$

Idea:

convex function



noncovex func

Usually use $c_1 = 10^{-1}, 10^{-2}, 10^{-4}, \ldots$

sufficient decrease condition is satisfied $\forall \eta$ small enough.

☆ Sufficient decrease + Backtracking line search

$\Rightarrow$ convergence to stationary points

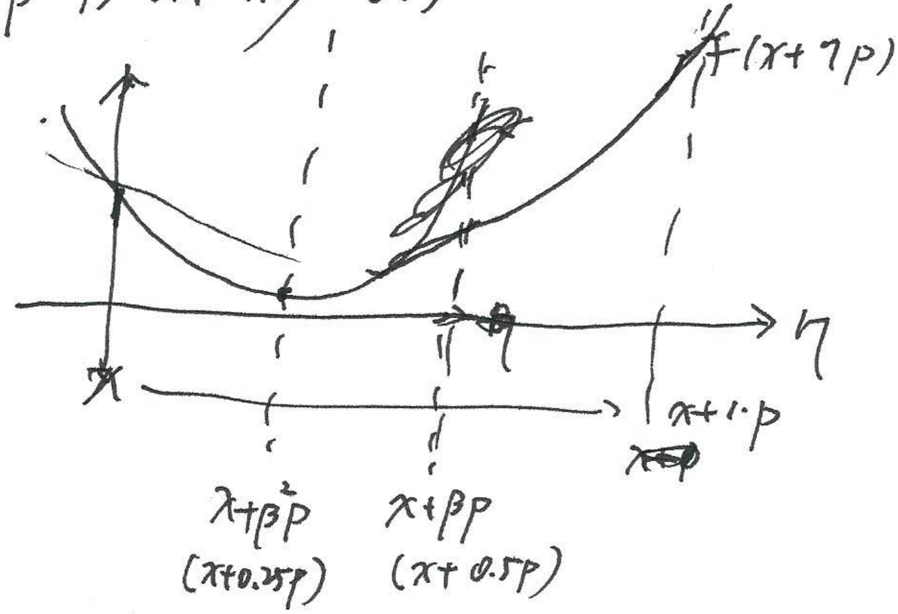
Algo: Backtracking line search, give p, x

- Initial $\eta = \eta^{init}$ (usually $\eta^{init} = 1$)
- While $f(x + \eta p) > f(x) + c_1 \cdot \eta \cdot \nabla f(x)^T p$

  $\eta = \beta \eta \quad (\beta < 1)$

  end
- Output step size η

- β is usually 0.5



Another way to present Backtracking line search:

Choose the largest number in $\{\eta^{init}, \beta\eta^{init}, \beta^2\eta^{init}, \dots\}$ that satisfy the sufficient decrease condition.

Thm: If f is L-Lipchitz and we apply GD with backtracking line search + sufficient decrease condition. Then any limit point is a stationary point.

pf: $f(x+\eta p) \leq f(x) - \eta(1-\frac{\eta L}{2}) \|\nabla f(x)\|^2$

When $\boxed{\eta \leq \frac{2(1-c)}{L}}$, $(1-\frac{\eta}{2}L) \geq c$

$\Rightarrow \eta$ satisfy sufficient decrease condition

$$\left(f(x+\eta p) \leq f(x) + c\cdot\eta\cdot\nabla f(x)^T p = f(x) - \eta\cdot c\cdot\|\nabla f(x)\|^2 \right)$$

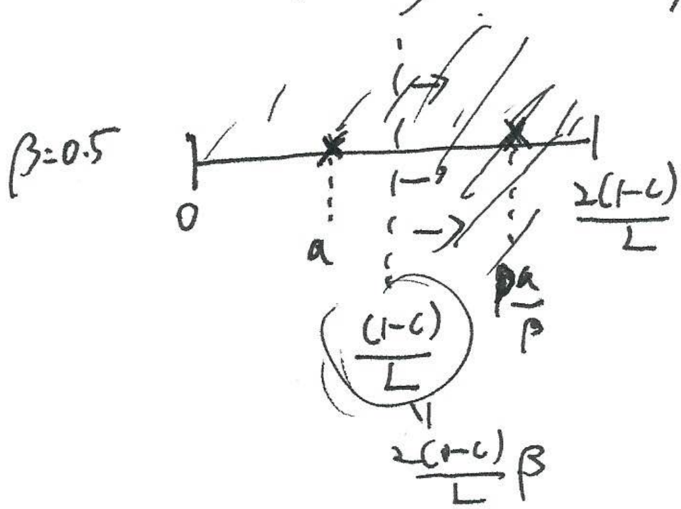
With Backtracking line search:

We'll choose $\eta \geq \frac{2(1-c)}{L}\beta = a$

$$f(x+\eta p) \le f(x) - a\,c\cdot\|\nabla f(x)\|^2$$

$$f(x^{k+1}) \le f(x^k) - a\,c\cdot\|\nabla f(x^k)\|^2$$

$\Rightarrow$ any limit point of $\{x^k\}$ is a stationary point. #



3.3 Wolfe condition.

① sufficient decrease: $f(x+\eta p) \le f(x) + c_1\cdot\eta\cdot\nabla f(x)^T p$

② avoid small stepsize: $\nabla f(x+\eta p)^T p \ge c_2\cdot\underbrace{\nabla f(x)^T p}_{<0}$

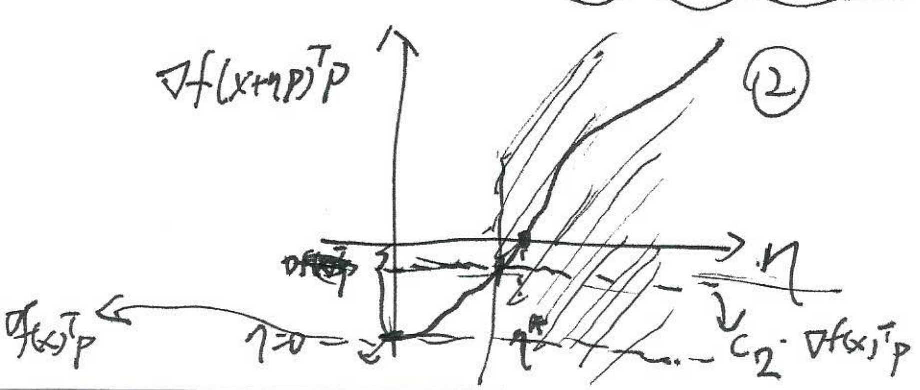
$$0 < c_1 < c_2 < 1$$

If we choose the "optimal" step size.

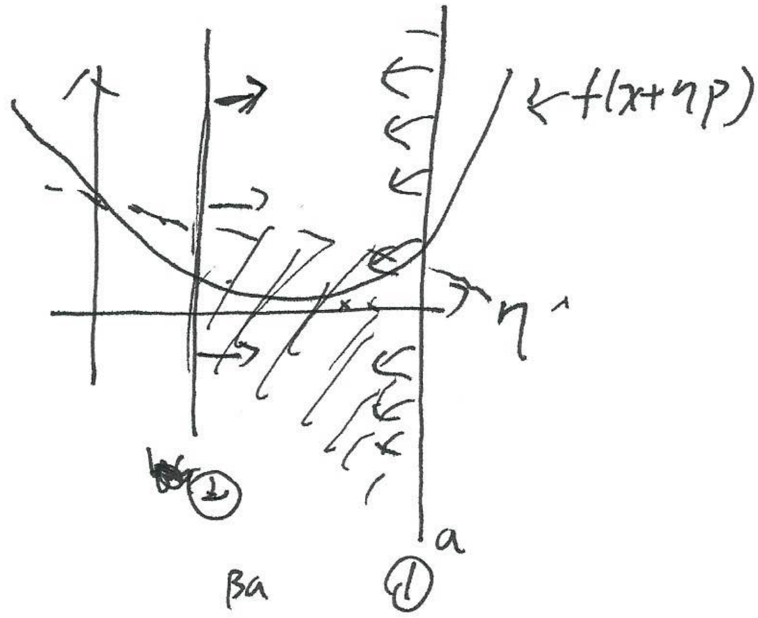
$$\eta^* = \arg\min_\eta f(x+\eta p)$$

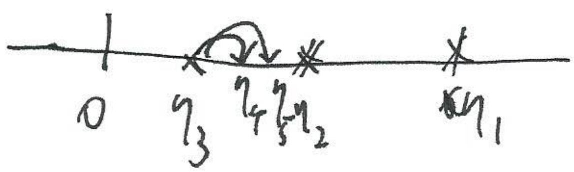
$$\cdot\ \frac{\partial}{\partial\eta} f(x+\eta^* p) = \nabla f(x+\eta^* p)^T p = 0$$

When $\eta = 0 \Rightarrow \nabla f(x)^T p < 0$



$\Rightarrow$ avoid small η





Thm: If f is continuously differentiable and bounded below, and L-Lipchitz, If we apply GD with step size satisfying Wolfe condition, then any limit point is a stationary point.