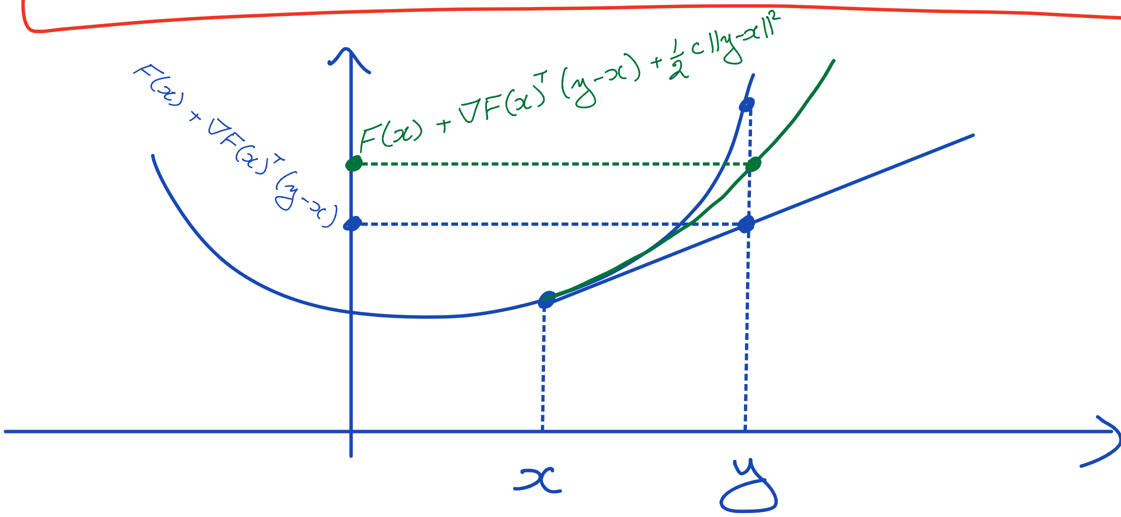
"Review" & Insights into GD and other 1^{st} order methods

One more definition : Strong convexity

A function F is stongly convex if $\exists c > 0$ such that $\forall x, y \in \mathbb{R}^d$

$$F(y) \geq F(x) + \nabla F(x)^T (y-x) + \frac{1}{2} c \|y-x\|^2$$



and it also guarantees

$$\boxed{\left(F(x) - F(x^*) \right) \leq \frac{\|\nabla F(x)\|^2}{2c}}$$

optimum

This is a fact. You'll prove it in HW

Want to minimize $F : \mathbb{R}^d \to \mathbb{R}$

Gradient Descent

$$x^{(t+1)} = x^{(t)} - \alpha_t \nabla F(x^{(t)})$$

A simple analysis of GD: Assume to start that

$$\|\nabla^2 F\| \le L$$

$$F(x^{(t+1)}) = F(x^{(t)} - \underbrace{\alpha_t \nabla F(x^{(t)})}_{h})$$

$$= F(x^{(t)}) - h^T \nabla F(x^{(t)}) + \frac{1}{2} \underbrace{h^T \nabla^2 F(\xi) h}$$

Taylor's theorem

$$= \langle h, \nabla^2 F(\xi) h \rangle$$

Cauchy $\quad \le \|h\| \, \|\nabla^2 F(\xi) h\|$

$\|Ax\| \le \|A\| \, \|x\|$ $\quad \le \|h\|^2 \, \|\nabla^2 F(\xi)\|$

$$\le L \|h\|^2$$

$$\le F(x^{(t)}) - \alpha_t \|\nabla F(x^{(t)})\|^2 + \frac{L\alpha_t^2}{2} \|\nabla F(x^{(t)})\|^2$$

$\|\nabla^2 F\| \le L$

$$\Rightarrow F(x^{(t+1)}) \leq F(x^{(t)}) - \alpha_t \underbrace{\left(1 - \frac{L\alpha_t}{2}\right)}_{\text{pick } \alpha < \frac{1}{L}} \|\nabla F(x^{(t)})\|^2$$

$$\Rightarrow \boxed{F(x^{(t+1)}) \leq F(x^{(t)}) - \frac{\alpha}{2} \|\nabla F(x^{(t)})\|^2}$$

① (assumed only $\|\nabla^2 F\| \leq L$, $\alpha < \frac{1}{L}$)

From here, two possibilities:

(A) Can't / Don't assume strong convexity. Then, rearranging ① & taking a telescoping sum.

$$\frac{\alpha}{2} \sum_{t=0}^{T-1} \|\nabla F(x^{(t)})\|^2 \leq \sum_{t=0}^{T-1} \left[F(x^{(t)}) - F(x^{(t+1)})\right]$$

$$= F(x^{(0)}) - F(x^{(T)})$$

Telescoping

$$\leq \underbrace{F(x^{(0)}) - F(x^*)}_{\text{constant}}$$

Thus as $T \to \infty$ $\|\nabla F(x^{(T)})\|^2 \to 0$

(Otherwise the LHS can't be smaller than the RHS as $T \to \infty$)

Moreover $\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F(x^{(t)})\|^2 \leq \frac{2}{\alpha T}\left[F(x^{(0)}) - F(x^*)\right]$.

Average $\|\nabla F\|^2 \longrightarrow 0$ so we "find" the minimizer.

(B) Can assume that F is strongly convex.

Recall: strong convexity with constant c

$\Rightarrow \quad F(x+h) \geq F(x) + h^T \nabla F(x) + \frac{c}{2}\|h\|^2$

& strong convexity also gives us:

$$(2) \quad \boxed{\|\nabla F(z)\|^2 \geq 2c\left[F(z) - F(x^*)\right]}$$

(HW)

Recall (1)

$$F(x^{(t+1)}) \leq F(x^{(t)}) - \frac{\alpha}{2}\|\nabla F(x^{(t)})\|^2$$

$$\Rightarrow \quad F(x^{(t+1)}) - F(x^*) \leqslant F(x^{(t)}) - F(x^*) - \frac{\alpha}{2} \|\nabla F(x^{(t)})\|^2$$

$$\leqslant F(x^{(t)}) - F(x^*) - \frac{\alpha}{2} \cdot 2c \cdot [F(x^{(t)}) - F(x^*)]$$

by ②

$$\leqslant \left(1 - \frac{c}{L}\right)[F(x^{(t)}) - F(x^*)]$$

pick $\alpha = 1/L$

$$\leqslant \left(1 - \frac{c}{L}\right)^{t+1}[F(x^{(0)}) - F(x^*)]$$

So: GD w/ strong convexity
w/ $\|\nabla^2 F\| \leqslant L$
w/ $\alpha = \frac{1}{L}$

guarantees

$$\boxed{F(x^{(t)}) - F(x^*) \leqslant \left(1 - \frac{c}{L}\right)^t [F(x^{(0)}) - F(x^*)]}$$

Fact: $1 - \frac{c}{L} \leqslant e^{-c/L}$ (why?)

↳ cond'n number of the problem $\frac{L}{c} = \kappa$

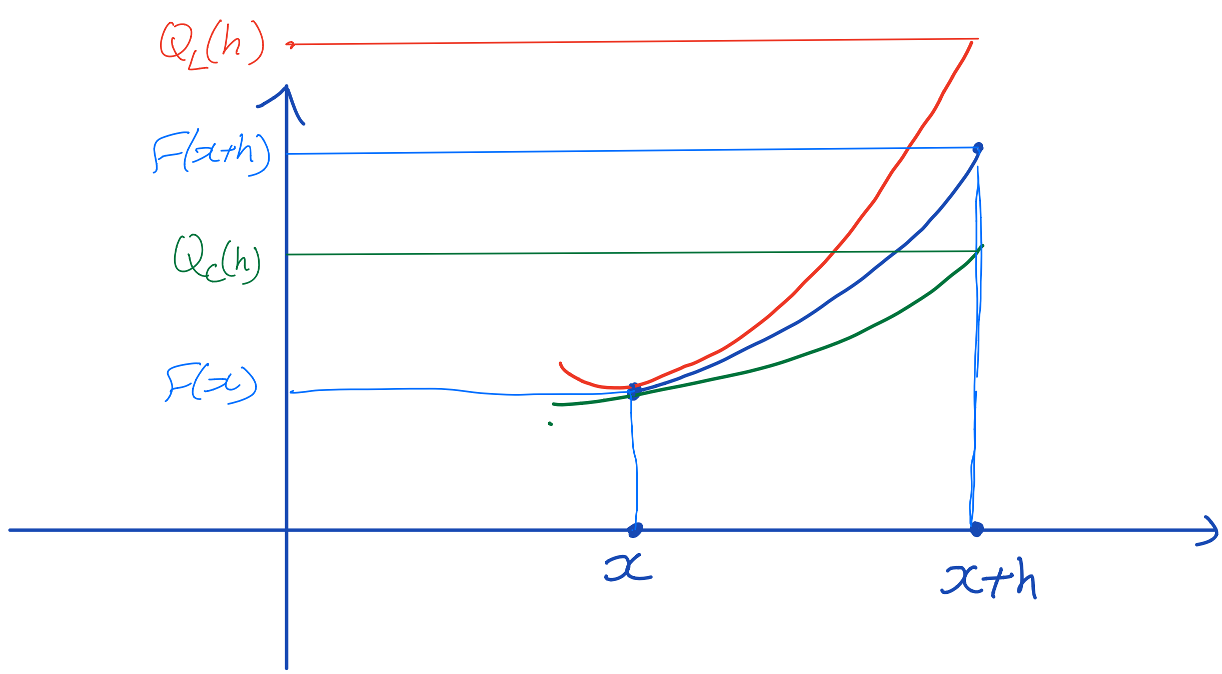
What is this condition number ?

Recall : $\|\nabla^2 F\| \leq L$

$$\Rightarrow F(x+h) \leq \underbrace{F(x) + h^T \nabla F(x) + \frac{L}{2}\|h\|^2}_{\substack{\text{Quadratic in } h \\ Q_L(h)}}$$

Also : Strong convexity

$$\Rightarrow F(x+h) \geq \underbrace{F(x) + h^T \nabla F(x) + \frac{c}{2}\|h\|^2}_{\substack{\text{Quadratic in } h \\ Q_c(h)}}$$



So bad condition number $\Rightarrow$ slower convergence.

Some solutions:

GD with momentum

$$x^{(t+1)} = x^{(t)} - \alpha \nabla F(x^{(t)}) + \beta(x^{(t)} - x^{(t-1)})$$

remembers the history

In the strongly convex case:

Convergence is like $\beta^t \|x^{(0)} - x^*\|$

"best" $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

better dependence on the cond'n number than GD $\Rightarrow$ can be much faster than GD

In non-strongly convex case, momentum can fail.

$\Rightarrow$ Nesterov Acceleration

$$x^{(t+1)} = x^{(t)} + \beta(x^{(t)} - x^{(t-1)}) - \alpha \nabla F\left(x^{(t)} + \beta(x^{(t)} - x^{(t-1)})\right)$$

Take a GD step not from $x^{(t)}$ but from $x^{(t)} + \beta(x^{(t)} - x^{(t-1)})$

Now, no longer only works for SC function, and also convergences is like

$$\left(\sqrt{\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa}}}\right)^t$$

Alternative, more "classical approach"

Conjugate Gradient:

Looks more complicated, simple to implement

Algorithm (FR):

Initialize $x^{(0)}$, set $F_0 = F(x^{(0)})$

$$\nabla F_0 = \nabla F(x^{(0)})$$

$$P_0 = -\nabla F_0$$

While $\nabla F_t \neq 0$, or ∇F_t too big

- Find α_t using line-search
- set $x^{(t+1)} = x^{(t)} + \alpha_t P_t$
- Evaluate $\nabla F_{t+1} = \nabla F(x^{(t+1)})$
- $\beta^{FR}_{t+1} = \dfrac{\nabla F_{t+1}^T \nabla F_{t+1}}{\nabla F_t^T \nabla F_t}$
- $P_{t+1} = -\nabla F_{t+1} + \beta^{FR}_{t+1} P_t$ — (*)

Converges in $\leq$ n-steps when F is quadratic with positive def'n A.