Choice of the norm

$f : \mathbb{R}^n \to \mathbb{R}$ convex

and $\quad\iff\quad \forall h \in \mathbb{R}^n$

$\|f'(y) - f'(x)\|_* \le L\|y-x\|. \qquad 0 \le \langle f''(x)h, h\rangle \le L\|h\|^2$

- Adam: $\approx$ gradient method with $\|\cdot\|_\infty$.
- Muon: $X \to \|X\|_{op}$. For NN layers.
- $\|\cdot\|_1 \Rightarrow$ greedy coordinate descent.

...

Euclidean norms $\quad \|\cdot\| = \|\cdot\|_2$

$$x_{k+1} = x_k - \frac{1}{L} f'(x_k), \quad k \ge 0.$$

$$\frac{1}{2}\|x_{k+1} - x^*\|^2 = \frac{1}{2}\|x_k - x^* - \frac{1}{L} f'(x_k)\|^2 =$$

$$= \frac{1}{2}\|x_k - x^*\|^2 + \frac{1}{L}\Big[\underbrace{\frac{1}{2L}\|f'(x_k)\|^2 + \langle f'(x_k), x^* - x_k\rangle}_{\le f(x^*) - f(x_k) \le 0}\Big]$$

$$\le \frac{1}{2}\|x_k - x^*\|^2.$$

Proposition $\forall k \ge 0$: $\quad \|x_k - x^*\| \le \|x_0 - x^*\| \equiv R$.

Rate for the Gradient Method.

1. $f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|f'(x_k)\|^2$

2. convexity: $f(x_k) - f^* \leq \langle f'(x_k), x_k - x^* \rangle \leq \|f'(x_k)\| \cdot R.$

Denote $F_k = f(x_k) - f^* > 0$

$$\boxed{F_k - F_{k+1} \geq \frac{1}{2L}\|f'(x_k)\|^2 \geq \frac{1}{2LR^2} F_k^2 .}$$

$F_{k+1} < F_k$.

$F_0 = f(x_0) - f^* \leq \frac{L}{2}\|x_0 - x^*\|^2 = \frac{LR^2}{2}$.

For smooth funct:
$f(y) \leq f(x) + \underline{\langle f'(x), y-x \rangle} + \frac{L}{2}\|y-x\|^2$
$f'(x^*) = 0$.

$$\frac{1}{F_{k+1}} - \frac{1}{F_k} = \frac{F_k - F_{k+1}}{F_k \cdot F_{k+1}} \geq \frac{1}{2LR^2} \cdot \frac{F_k^2}{F_k \cdot F_{k+1}} \geq \frac{1}{2LR^2}.$$

<u>Telescope it</u>: $\frac{1}{F_k} \geq \frac{1}{F_0} + \frac{k}{2LR^2} \geq \frac{k+4}{2LR^2}$.

<u>Theorem</u>. On smooth convex functions, the rate of GM is

$$f(x_k) - f^* \leq \frac{2LR^2}{k+4}.$$

<u>Therefore</u>, To find $f(x_k) - f^* \leq \varepsilon$ it's enough to do

$$k = \left\lfloor \frac{2LR^2}{\varepsilon} \right\rfloor + 1 \text{ iterations.}$$

The Gradient Norm Minimization.

Non-convex:

$$\min_{0 \le i \le k-1} \|f'(x_i)\|^2 \le \frac{2L(f(x_0) - f^*)}{k}.$$

$\|f'(\bar{x})\| \le \varepsilon$

$O\left(\frac{LF_0}{\varepsilon^2}\right)$

Compare with

$$f(x_k) - f^* \le \frac{2LR^2}{k+4}.$$

$R = \|x_0 - x^*\|$

$O\left(\frac{LR^2}{\varepsilon}\right)$

<u>Improved rate</u>. Run for $2K$ iterations, $K \ge 1$.

$$\min_{K \le i \le 2K-1} \|f'(x_i)\|^2 \le \frac{2L(f(x_K) - f^*)}{K} \le \frac{4L^2R^2}{K^2}.$$

$$\Rightarrow \|f'(\bar{x})\| \le \varepsilon : O\left(\frac{LR}{\varepsilon}\right)$$

<u>Distance</u> $\|x_k - x^*\| \overset{?}{\to} 0$ <u>impossible</u> to guarantee for convex.

Strongly Convex Smooth Functions.

$f: \mathbb{R}^n \to \mathbb{R}$ strongly convex and smooth if

$$\mu I \preceq f''(x) \preceq LI \qquad \forall x \in \mathbb{R}^n \qquad (1)$$

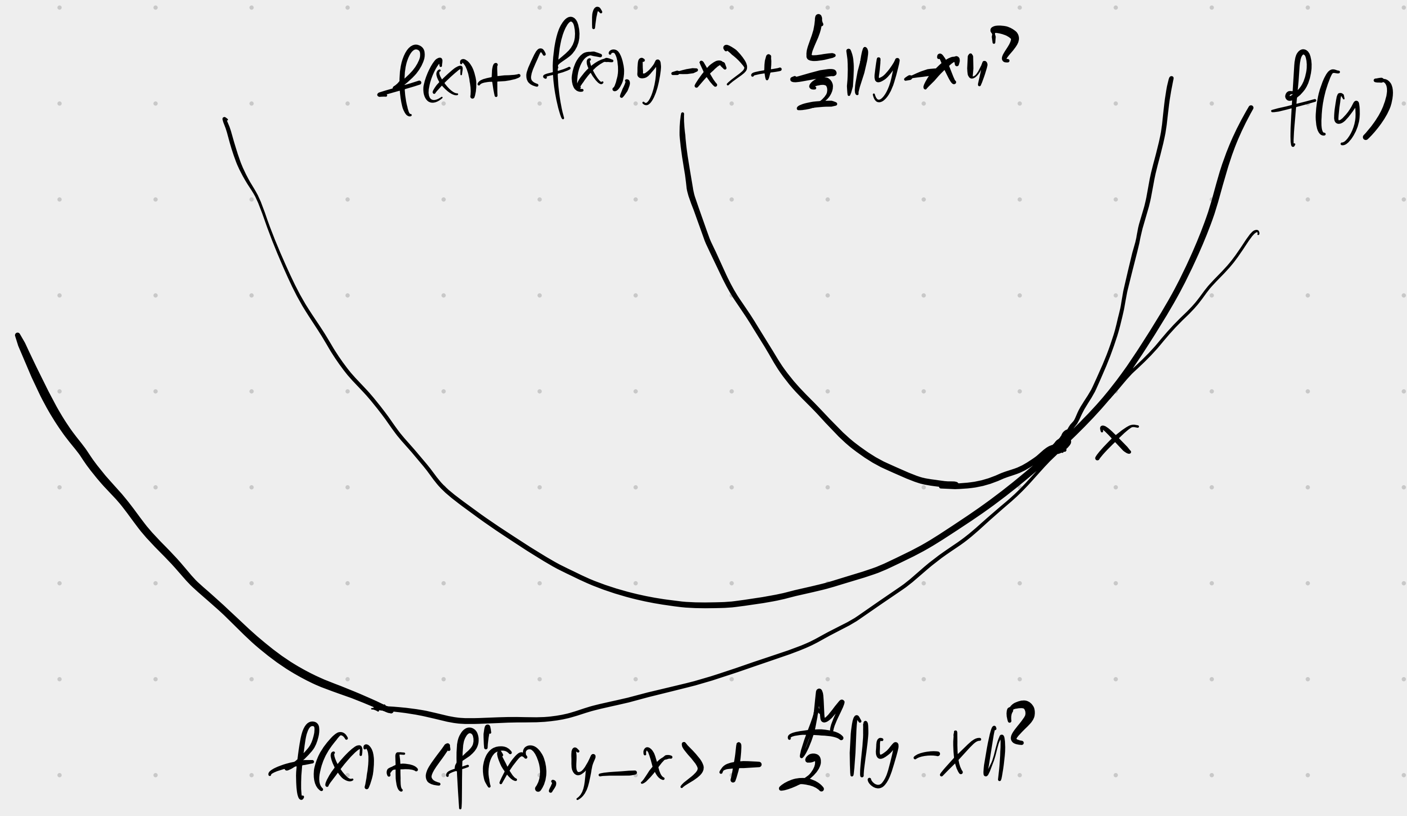
$0 < \mu \leq L$.

when $\mu = 0 \Rightarrow$ convex functions.

<u>Theorem</u> Condition (1) is equivalent to

(2) $\mu\|y-x\|^2 \leq \langle f'(y) - f'(x), y-x \rangle \leq L\|y-x\|^2$

(3) $\frac{\mu}{2}\|y-x\|^2 \leq f(y) - f(x) - \langle f'(x), y-x \rangle \leq \frac{L}{2}\|y-x\|^2$.



<u>Consequence 1</u> Plug $x := x^*$: then

$$\frac{\mu}{2}\|y-x^*\|^2 \leq f(y) - f(x^*) \leq \frac{L}{2}\|y-x^*\|^2 \quad \forall y$$

$\Rightarrow$ the solution is unique.

<u>Consequence 2</u> $f(y) \geq f(x) + \langle f'(x), y-x \rangle + \frac{\mu}{2}\|y-x\|^2$

Minimize in y both LHS and RHS:

$$f^* \geq f(x) - \frac{1}{2\mu}\|f'(x)\|^2 .$$

Rearranging:

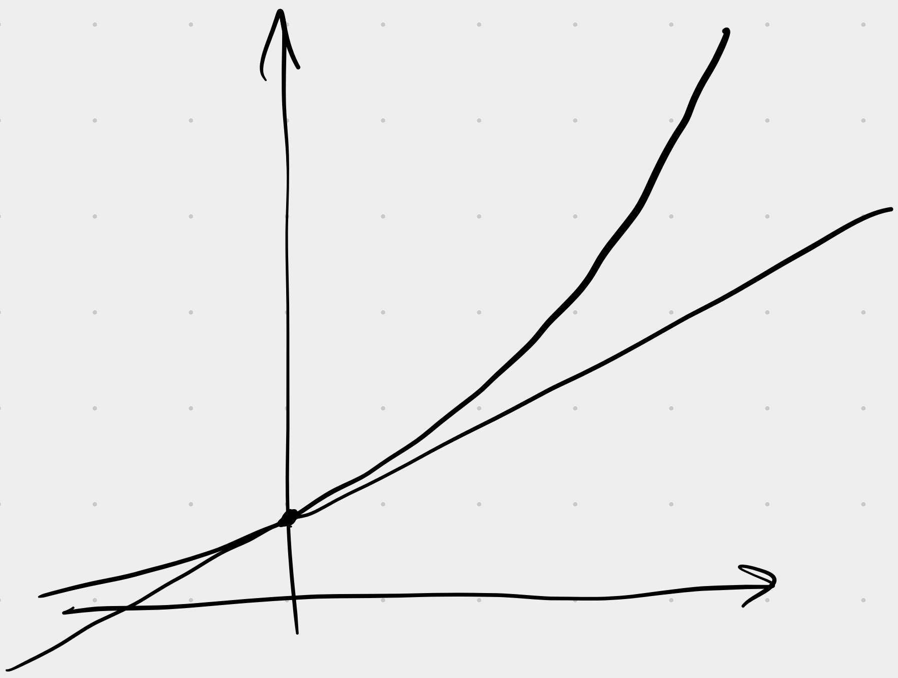
$$\frac{1}{2\mu}\|f'(x)\|^2 \geq f(x) - f^* \overset{?}{\geq} \frac{1}{2L}\|f'(x)\|^2 .$$

Convergence of GM on strongly convex smooth functions.

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|f'(x_k)\|^2 \geq \frac{\mu}{L}\left(f(x_k) - f^*\right)$$

For $F_k = f(x_k) - f^*$, we get: ...... $(f(x_{k+1}) - f^*)$?

$$F_{k+1} \leq \left(1 - \frac{\mu}{L}\right) F_k \leq \exp\left(-\frac{\mu}{L}\right) F_k$$



Theorem For GM on strongly convex and smooth funct., we have:

$$f(x_k) - f^* \leq \exp\left(-k \cdot \frac{\mu}{L}\right)\left(f(x_0) - f^*\right)$$

$$f(x_k) - f^* \leq \min \frac{2LR^2}{k+4} .$$

To get: $f(x_k) - f^* \le \varepsilon$ it's enough to perform

$$K = \frac{L}{\mu} \log \frac{f(x_0) - f^*}{\varepsilon}.$$

The condition number of the problem $\frac{L}{\mu} \ge 1$.

Minimizing Quadratic Functions.

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \qquad A \in \mathbb{R}^{n\times n},\ b \in \mathbb{R}^n$$

$\min_{x \in \mathbb{R}^n} f(x)$

$$A = A^T \succ 0.$$

$f'(x) = Ax - b$ — one matrix-vector product, affine

$f'(x^*) = Ax^* - b = 0$

$f''(x) = A \qquad 0 < \mu = \lambda_{min}(A) \le \lambda_{max}(A) = L.$

Proposition

$$f(y) - f(x) - \langle f'(x), y - x\rangle \equiv \frac{1}{2}\langle f'(y) - f'(x), y - x\rangle$$

$$\equiv \frac{1}{2}\langle f''(x)(y-x), y - x\rangle$$

$$\equiv \frac{1}{2}\langle A(y-x), y-x\rangle = \frac{1}{2}\|y-x\|_A^2.$$

$\Rightarrow \quad f(y) - f(x) \equiv \frac{1}{2}\langle f'(y) + f'(x), y - x\rangle.$

GM has complexity $O\left(\frac{L}{\mu}\cdot \ln\frac{1}{\varepsilon}\right)$. $\Rightarrow$ Conjugate Gradient Method

Polyak's Heavy-Ball Method.

Iterate $k \geq 0$:

$$x_{k+1} = x_k - \alpha f'(x_k) + \beta (x_k - x_{k-1})$$

$0 \leq \beta \leq 1$ is the momentum parameter.

Gradient Method with Momentum

Init: $x_0 \in \mathbb{R}^n$, $s_0 = 0$.

Iterate $k \geq 0$:

$$s_{k+1} = \beta s_k + f'(x_k)$$

$$x_{k+1} = x_k - \frac{1}{L} s_{k+1}.$$

Analysis $\quad \beta = 1.$

$$s_{k+1} = \sum_{i=0}^{k-1} f'(x_i)$$

$$\frac{1}{k} s_{k+1} = \frac{1}{k} \sum_{i=0}^{k-1} f'(x_i) = f'(\bar{x}_k), \quad \bar{x}_k = \frac{1}{k} \sum_{i=0}^{k-1} x_i.$$

$$f(x_{k+1}) \leq f(x_k) + \langle f'(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$= f(x_k) - \frac{1}{L}\langle \underbrace{f'(x_k)}, s_{k+1} \rangle + \frac{1}{2L}\|s_{k+1}\|^2$$

$$= f(x_k) - \frac{1}{L}\langle s_{k+1} - s_k, s_{k+1} \rangle + \frac{1}{2L}\|s_{k+1}\|^2$$

$$= f(x_k) - \frac{1}{2L}\|s_{k+1}\|^2 + \frac{1}{L}\langle s_k, s_{k+1} \rangle$$

<u>Proposition</u>

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|s_{k+1}\|^2 - \frac{1}{L}\langle s_k, s_{k+1} \rangle$$

<u>Proposition</u>

$$f(x_k) - f(x_{k+1}) = \frac{1}{2}\langle \underbrace{f'(x_k) + f'(x_{k+1})}_{s_{k+2} - s_k}, \overbrace{x_k - x_{k+1}}^{\frac{1}{L}s_{k+1}} \rangle$$

$$= \frac{1}{2L}\langle s_{k+2}, s_{k+1} \rangle - \frac{1}{2L}\langle s_k, s_{k+1} \rangle$$

<u>Rearranging</u>

$$f(x_k) - f(x_{k+1}) + \frac{1}{2L}\langle s_k, s_{k+1} \rangle = \frac{1}{2L}\langle s_{k+1}, s_{k+2} \rangle$$

<u>Telescoping</u>

$$f(x_0) - f^* \geq f(x_0) - f(x_k) = \frac{1}{2L}\langle s_k, s_{k+1} \rangle.$$

Combine them together:

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|s_{k+1}\|^2 - 2(f(x_0) - f^*)$$

Telescope it:

$$f(x_0) - f^* \geq f(x_0) - f(x_k)$$

$$\geq \frac{1}{2L}\sum_{i=1}^{k}\|s_i\|^2 - k \cdot 2(f(x_0) - f^*).$$

Theorem

$$2L(1+2k)(f(x_0) - f^*) \geq \sum_{i=1}^{k}\|s_i\|^2 = \sum_{i=1}^{k} i^2\|f'(\bar{x}_i)\|^2$$

where $\bar{x}_i = \sum_{j=0}^{i-1} x_j$.

Denote $g_k := \min\{\|f'(\bar{x}_0)\|, \dots, \|f'(\bar{x}_{k-1})\|\}$

$$g_k^2 \cdot \underbrace{\sum_{i=1}^{k} i^2}_{\approx k^3} \leq 2L(1+2k)(f(x_0) - f^*)$$

We finally get the rate:

$$g_k^2 \leq \frac{12L(f(x_0) - f^*)}{k \cdot (k+1)}$$

$\square$