THEOREM:

f: β-Smooth, convex, GD works for $\eta \leq 1/\beta$

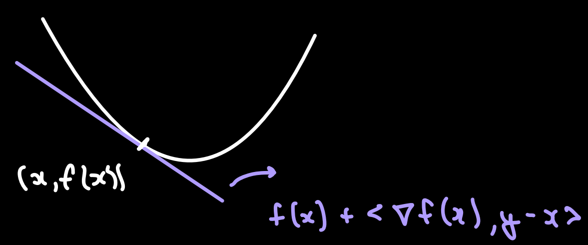
$$f(x_k) \leq f(x_*) + \frac{2\beta \|x_* - x_0\|}{k} \quad \left(\eta = \frac{1}{\beta}\right)$$

PROOF:

Recall:

CONVEXITY:

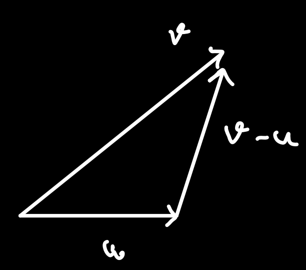
$$\forall x, y \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$



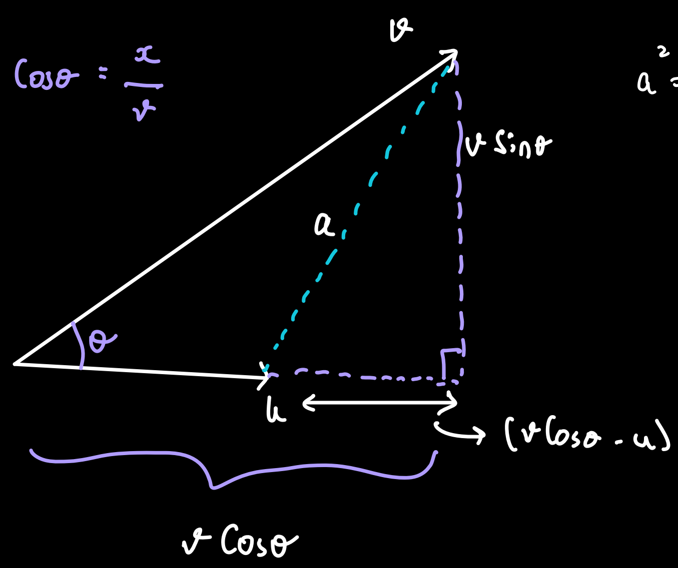
PROPERTIES OF VECTORS:

a) $\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2\langle u, v \rangle$

a') $2\langle u, v \rangle = \|u\|^2 + \|v\|^2 - \|u - v\|^2$



$\cos\theta = \dfrac{x}{v}$



$$a^2 = v^2\sin^2\theta + (v\cos\theta - u)^2$$
$$= v^2\sin^2\theta + v^2\cos^2\theta + u^2 - 2uv\cos\theta$$
$$= u^2 + v^2 - 2uv\cos\theta$$
$$\langle u, v\rangle$$

b) $|\langle u, v\rangle| \leq \|u\| \cdot \|v\|$ (CAUCHY - SCHWARZ)

SO ALL THE NEEDED EQUATIONS:

Convexity: $\forall x, y \quad f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$

β-Smooth: $f(x_{k+1}) \leq f(x_k) - \dfrac{1}{2\beta}\|\nabla f(x_k)\|_2^2$

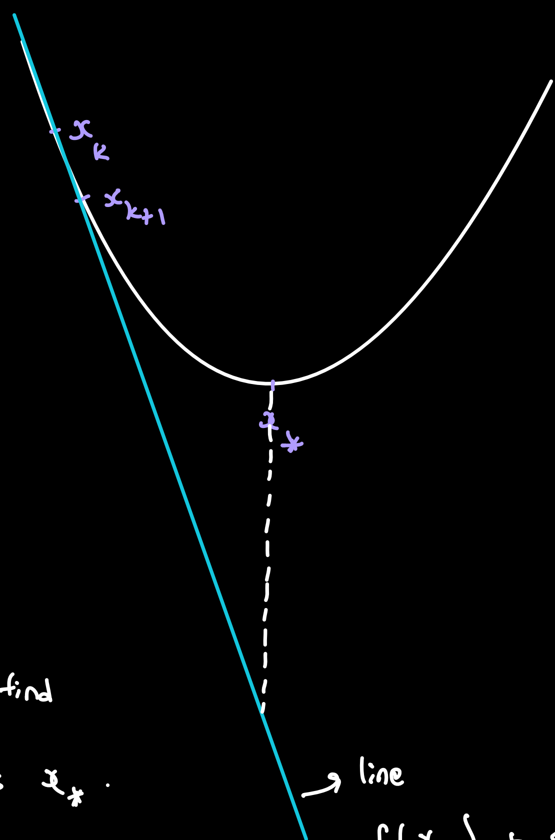
GD: $x_{k+1} = x_k - \eta \nabla f(x_k)$

VECTOR: $\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2\langle u, v\rangle$

$|\langle u, v\rangle| \leq \|u\| \cdot \|v\|$

β- SMOOTH IMPLIES:

$$f(x_{k+1}) \le f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2 \quad - (1)$$



We want to find progress towards x_*.

line
$f(x_k) + \langle \nabla f(x_k), y - x_k \rangle$

From convexity:

$$f(x_*) \ge f(x_k) + \langle \nabla f(x_k), x_* - x_k \rangle$$

$$f(x_k) \le f(x_*) - \langle \nabla f(x_k), x_* - x_k \rangle \quad - (2)$$

Combine ① and ②

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

$$f(x_{k+1}) - f(x_*) \leq \frac{1}{2\beta} \left[2\beta \langle \nabla f(x_k), x_k - x_* \rangle - \|\nabla f(x_k)\|^2 \right.$$

G.D:

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

$$\eta = \frac{1}{\beta}$$

$$x_{k+1} = x_k - \frac{1}{\beta} \nabla f(x_k)$$

$$\nabla f(x_k) = \beta . (x_k - x_{k+1})$$

$$f(x_{k+1}) - f(x_*) \leq \frac{1}{2\beta} \left[2\beta \langle \beta . (x_k - x_{k+1}), x_k - x_* \rangle - \beta^2 \|x_k - x_{k+1}\|^2 \right]$$

$$= \frac{\beta}{2} \left[2 . \langle x_k - x_{k+1}, x_k - x_* \rangle - \|x_k - x_{k+1}\|^2 \right]$$

Apply ⓐ'

$$2 \langle u, v \rangle = \|u\|^2 + \|v\|^2 - \|u - v\|^2$$

$$= \frac{\beta}{2}\Big[\ \|x_k - x_{k+1}\|^2 + \|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2$$
$$- \|x_k - x_{k+1}\|^2\Big]$$

Therefore,

$$f(x_{k+1}) - f(x_*) \leq \frac{\beta}{2}\left(\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2\right)$$

(for $k=0$)

$$f(x_1) - f(x_*) \leq \frac{\beta}{2}\left(\|x_0 - x_*\|^2 - \|x_1 - x_*\|^2\right)$$

(for $k=1$)

$$f(x_2) - f(x_*) \leq \frac{\beta}{2}\left(\|x_1 - x_*\|^2 - \|x_2 - x_*\|^2\right)$$

$$\vdots$$

$$f(x_{k+1}) - f(x_*) \leq \frac{\beta}{2}\left(\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2\right)$$

ADD UP ALL k INEQUALITIES

$$\sum_{i=1}^{k+1}\left(f(x_i) - f(x_*)\right) \leq \frac{\beta}{2}\left[\|x_0 - x_*\|^2 - \|x_{k+1} - x_*\|^2\right]$$

Now by the monotonicity of GD,

$$f(x_{k+1}) \leq f(x_i) \quad \forall\, i \leq k+1$$

$$f(x_{k+1}) - f(x_*) \leq f(x_i) - f(x_*)$$

So $$\sum_{i=1}^{k+1} \left(f(x_i) - f(x_*)\right) \geq (k+1)\left(f(x_{k+1}) - f(x_*)\right)$$

$$(k+1) \cdot \left(f(x_{k+1}) - f(x_*)\right) \leq \frac{\beta}{2}\left[\|x_0 - x_*\|^2 - \|x_{k+1} - x_*\|^2\right]$$

$$\leq \frac{\beta}{2}\|x_0 - x_*\|^2$$

$$f(x_{k+1}) - f(x_*) \leq \frac{\beta}{2(k+1)}\|x_0 - x_*\|^2.$$

SUMMARY:

If f is β-smooth, then

$$f(x_{k+1}) \leq f(x_*) + \frac{\beta}{2(k+1)}\|x_0 - x_*\|^2$$

1. What do you really need to run GD?

→ The only ability required is to compute gradients.

FIRST - ORDER METHODS OF OPTIMIZATION:

We have a subroutine that computes $\nabla f(x)$ at any point x.

2. What is the best you can do with First-Order method?

NESTEROV'S ACCELERATED GRADIENT DESCENT (NAGD) 1983:

$$f(x_k) \leq f(x_*) + \frac{\|x_0 - x_*\|^2}{\beta \cdot k^2}$$

(k^2 not k)!

Remark: To get within ε of the optimum

GD takes → $1/\varepsilon$ iterations

NAGD takes → $1/\sqrt{\varepsilon}$ iterations.

NAGD:

Start with $x_0 = y_0 = z_0$ MOMENTUM UPDATE

For $i = 0, \dots$:

$$x_{i+1} = y_i - \eta \nabla f(y_i)$$

$$z_{i+1} = z_i - \eta_i \nabla f(y_i)$$

$$y_{i+1} = \alpha_i z_i + (1-\alpha_i) x_i$$

THEOREM:

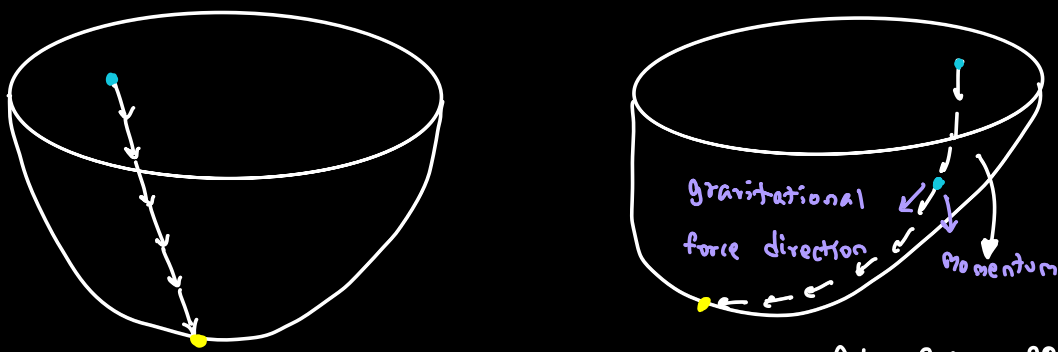
f is convex and smooth

$$\eta \leq 1/\beta$$

$$\eta_i = \frac{(i+1)\eta}{2}$$

$$\alpha_i = \frac{2}{i+3}$$

$$f(x_k) \leq f(x_*) + \frac{2\beta \|x_0 - x_*\|^2}{k^2}$$



→ At every point gravitational pull ≡ along gradient.

New velocity is a combination of current velocity and Force!

Intuitively : velocity ≡ change in position

"$x_{k+1} - x_k$" ≡ Some combination of "$x_k - x_{k-1}$" and "$-\nabla f(x_k)$".

THEOREM:

NAGD is the best you can do among all first-Order Methods!

[NAGD is not compulsorily monotone]

Demo:

Least Squares Regression: $(x_1, y_1), \cdots, (x_n, y_n)$

Parameter family: $h_w(x) = \langle w, x \rangle$

ERM: $L(w) = \frac{1}{n} \sum_{i=1}^{n} (\langle w, x_i \rangle - y_i)^2$

$$\nabla L(w) = \frac{1}{n} \sum_{i=1}^{n} 2(\langle w, x_i \rangle - y_i)\, x_i \,.$$

In matrix notation:

$$L(w) = \frac{1}{n} \left\| \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{X} \, w - y \right\|^2$$

$$= \frac{1}{n} \| Xw - y \|^2$$

$$\nabla L(w) = \frac{2}{n} X^T (Xw - y)$$