GRADIENT DESCENT: $X_{t+1} \leftarrow X_t - \eta \nabla f(X_t)$ — (1)

[f - convex function, L-Lipshitz] $\|\nabla f(x)\| \leq L$

Thm: After t iterations

$$\frac{1}{t}\sum_{i=1}^{t} f(x_i) - \underbrace{f(x^*)}_{\min_x f(x)} \leq \boxed{O\left(\frac{RL}{\sqrt{t}}\right)}$$

where $R^2 = \|x_1 - x_*\|^2$

Proof: $f(x_t) - f(x_*) \leq \langle \nabla f(x_t), x_t - x_* \rangle$

$$\leq \left\langle \boxed{\frac{X_t - X_{t+1}}{\eta}}, x_t - x_* \right\rangle \longrightarrow (1)$$

$\left(\frac{X_t - Y_{t+1}}{\eta}\right)$

(law of Cosine)

$$\leq \frac{1}{2\eta}\left[\|X_t - X_{t+1}\|^2 + \|X_t - X_*\|^2 - \|X_{t+1} - X_*\|^2\right.$$

$$\underbrace{\|X_t - Y_{t+1}\|^2}_{-\nabla f(x) \leq L} + \|X_t - X_*\|^2 - \|Y_{t+1} - X_*\|^2$$

$$f(x_t) - f(x_*) \leq \frac{1}{2\eta}\left[\|x_{t+1} - x_t\|^2 + \|x_t - x_*\|^2 - \|x_{t+1} - x_*\|^2 \right]$$

$$D_t \overset{\shortparallel}{-} D_{t+1}$$

$$\boxed{\sum_{i=1}^{t} \left(f(x_i) - f(x_*)\right)} \leq \frac{1}{2\eta} \sum_{i=1}^{t} \|x_{i+1} - x_i\|^2 + \frac{1}{2\eta} \|x_1 - x_*\|^2 - \|x_{t+1} - x_*\|^2$$

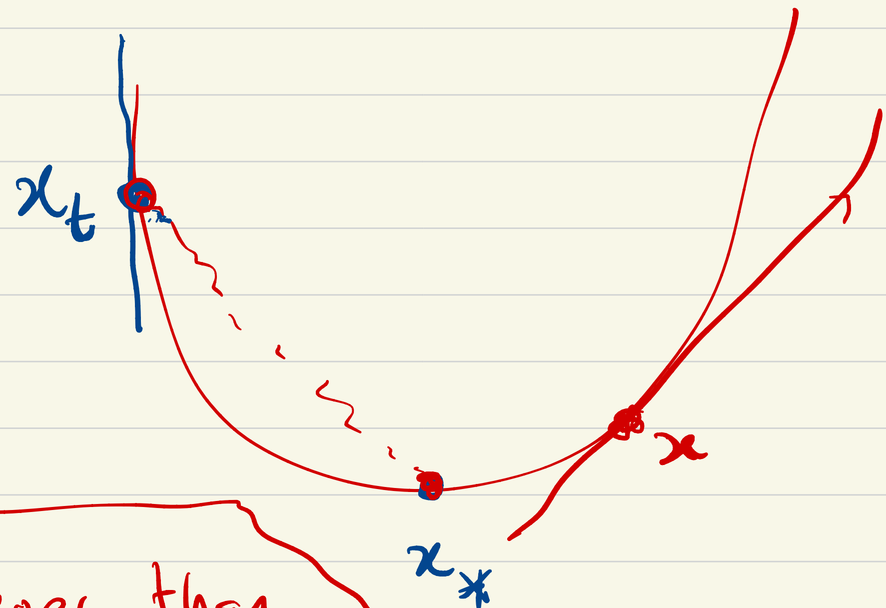
$$\boxed{\|x_{i+1} - x_i\|^2 = \|\eta \nabla f(x_i)\|^2 = \eta^2 L^2}$$

$$\leq \boxed{\frac{\eta L^2 t}{2} + \frac{R^2}{2\eta}} \quad \left(\text{Step}: \frac{R}{L\sqrt{t}}\right)$$

$$\frac{1}{t}\left(\sum f(x_i)\right) - f(x_*) \leq \boxed{\frac{\eta L^2}{2}} + \frac{R^2}{2\eta t}$$

Convex Fn:

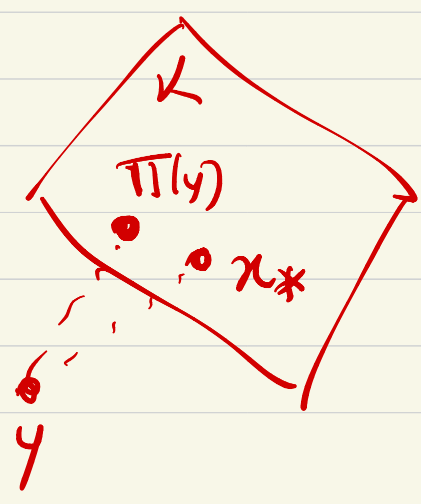
$$F(y) \geq \boxed{F(x) + \langle \nabla F(x), y - x \rangle}$$



$\nabla f(x_t)$ is steeper than $\boxed{\dfrac{f(x_t) - f(x_*)}{x_t - x_*}}$

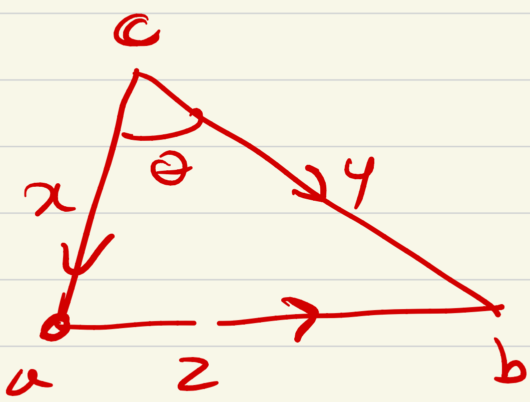
PROJECTIONS: $\forall x^* \in K$
$\forall y$

$$\|\Pi(y) - x_*\| \leq \|y - x_*\|$$



"Law of Cosine"

$$2\langle a-c, b-c\rangle = \|a-c\|^2 + \|b-c\|^2 - \|a-b\|^2$$



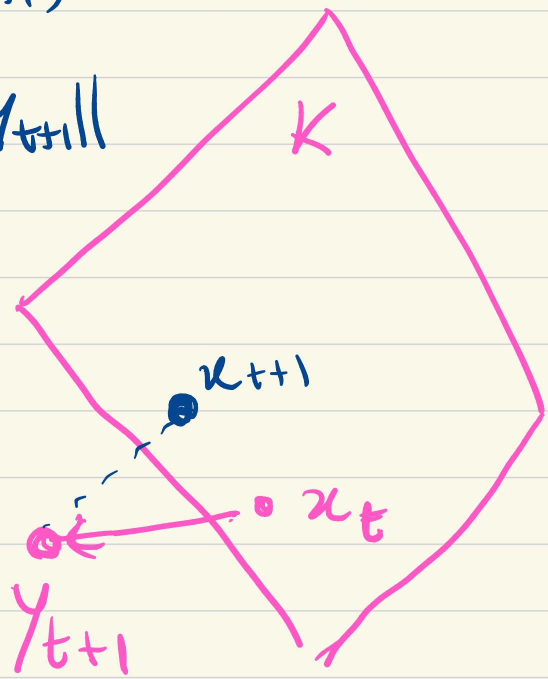
$$z^2 = x^2 + y^2 - 2xy\cos\theta$$
$$= \|x\|^2 + \|y\|^2 - 2\langle x, y\rangle$$

PROJECTED GRADIENT DESCENT

$$Y_{t+1} \leftarrow X_t - \eta \nabla f(X_t)$$

$$X_{t+1} \leftarrow \Pi_K (Y_{t+1}) = \min_{z \in K} \|z - Y_{t+1}\|$$

$$\min_{x \in K} f(x)$$



Comments:

1) # of iterations = independent of dimension

2) $O\left(\frac{1}{\sqrt{t}}\right)$ error in t iterations

$\downarrow$

optimal for Lipshitz functions (non smooth)

[Pseudopolynomial time]

for error ε take $O\left(1/\varepsilon^2\right)$ iterations

2^{-b} error $\qquad \downarrow$

$O(2^{2b})$

3) $\nabla f \leftarrow$ Gradient oracle

$$\nabla f(x) = \frac{1}{\sqrt{Var(\delta)}} \mathbb{E}_{\delta}\left[f(x+\delta)\, \vec{\delta} \right]$$

$$= \frac{f(x+\delta) - f(x-\delta)}{2\delta}$$

→ Stochastic gradient descent

$$E[\Theta(x)] = \nabla f(x)$$

→ Smooth & Strongly convex function:

After t iterations,

[Polynomial time algorithm]$^{\exp\left(-\frac{t}{(\beta/\alpha)}\right)}$

where β = smoothness parameter

α = strongly convexity parameter

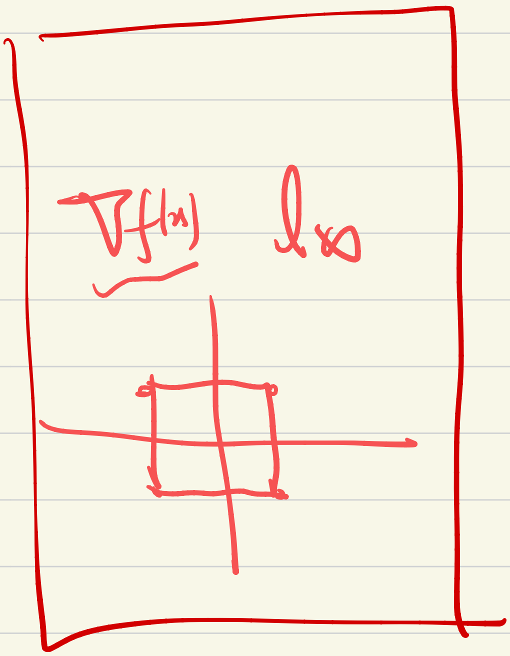
$$\boxed{\begin{aligned} \lambda_{min}(H_f(x)) &\geq \alpha \\ \lambda_{max}(H_f(x)) &\leq \beta \end{aligned}}$$

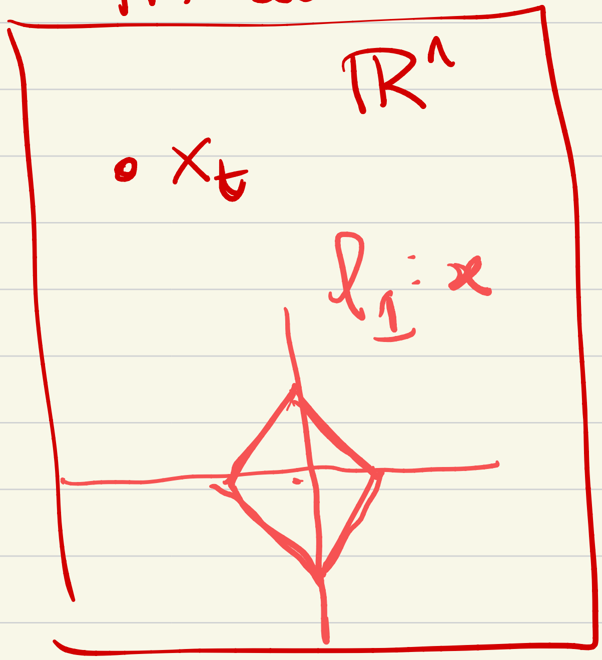
Mirror Descent

$$x_{t+1} \leftarrow \boxed{x_t - \eta \nabla f(x_t)}$$

Gradients (dual space)

Primal





$l_q = (\mathbb{R}^n, \|\cdot\|_q)$

$\|x\|_q = \left(\sum x_i^q\right)^{1/q}$

$l_p = \mathbb{R}^n,$

$\|x\|_p = \left(\sum x_i^p\right)^{1/p}$

$$\boxed{\frac{1}{q} + \frac{1}{p} = 1}$$

$x \in$ vector $\mathbb{R}^n$

$$\boxed{\langle \nabla f(x), y \rangle} = \nabla f(x)(y)$$

$$\boxed{\nabla f(x) : [\text{vector}] \to \mathbb{R}}$$

∇f : [vector] $\to \mathbb{R}$

$\nabla f \in$ linear function

$$\boxed{\|\nabla f\| = \sup_{\|x\|_p \le 1} |\underbrace{\nabla f(x)}|}$$

$$\|\nabla f\|_q = \sup_{\|x\|_p \le 1} |\langle \nabla f, x \rangle|$$

$$\boxed{\frac{1}{p} + \frac{1}{q} = 1}$$

Proximal Gradient Descent:

$$D_h(x_t \| x)$$

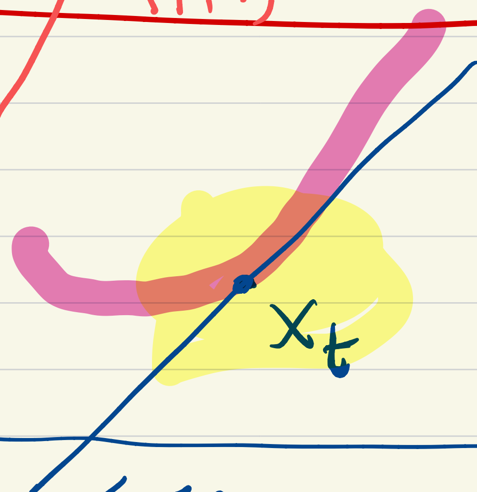
$$X_1 \leftarrow \text{starting point}$$

$$\boxed{X_{t+1} \leftarrow \arg\min_{x} \left\{ \eta \langle \nabla f(x_t), x \rangle + \frac{1}{2}\|x - x_t\|^2 \right\}}$$

$$\parallel$$

$$H(x)$$

drop terms don't



$$\boxed{H(x) = f(x_t) + \langle \nabla f(x_t), x - x_t \rangle}$$

To minimize: $\langle \eta \nabla f(x_t), x \rangle + \frac{1}{2}\|x - x_t\|^2$

$$\parallel$$

$$G(x)$$

$$\nabla_x G(x) = 0 \iff \eta \nabla f(x_t) + \frac{1}{2}\left[2(x - x_t)\right]$$

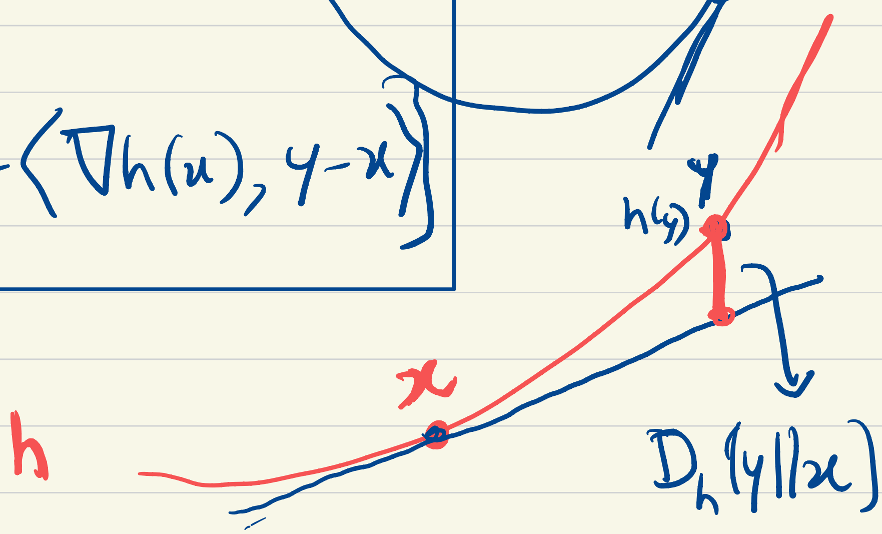
$$\iff \boxed{x = x_t - \eta \nabla f(x_t)}$$

Bregman Divergence:

Let h: strictly convex function

$h: \mathbb{R}^n \to \mathbb{R}$

Def: $$\boxed{D_h(y\|x) = h(y) - \left[h(x) + \langle \nabla h(x), y-x \rangle\right]}$$



Ex: $\boxed{h(x) = \|x\|^2}$

$$D_h(y\|x) = \|y-x\|^2$$

Ex: $\boxed{h(x) = \sum_{i=1}^{n} x_i \ln x_i - x_i}$ $(\in \mathbb{R}^n_+)$

$$D_h(y\|x) = \mathrm{KL\,Div}(y\|x) = \sum_i y_i \ln \frac{y_i}{x_i}$$

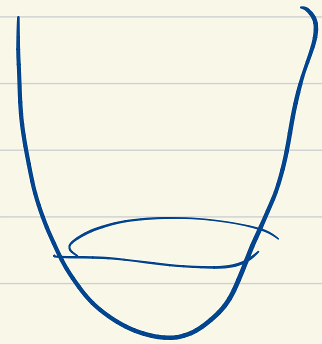
$h(x) = \|x\|^2 = \langle x, x \rangle$ ← Gradient Descent

$\sum x_i^2$

$$\boxed{h(x) = \langle x, Ax \rangle} \qquad \boxed{A \succeq 0}$$

$$\|$$

$$\boxed{\sum w_i x_i^2} \qquad w_i \geq 0$$

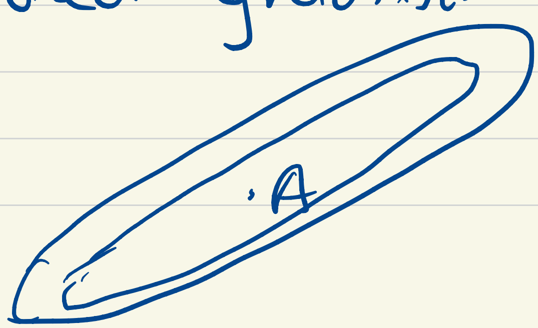


Excercise $\quad h(x) = x^T A x, \quad D_h =$

$$X_{t+1} \leftarrow \boxed{X_t - \eta A^{-1} \nabla f(x_t)}$$

$\downarrow$

Pre conditioned gradient descent:



Newton method

$$X_{t+1} \in \underset{x}{\operatorname{argmin}} \left\{ \langle \nabla f(x_t), x \rangle + \frac{1}{2} (x-x_t)^T \underbrace{H_f(x_t)}_{\downarrow} (x-x_t) \right\}$$

$H_f(x_t)$ = local Hessian of function $\succ 0$

$$\boxed{(x-x_t)^T H_f(x_t)(x-x_t)}$$

→ quadratic approximation

$$\boxed{\varepsilon\text{-accuracy} \rightarrow \log\log(1/\varepsilon)}$$