Distributed Optimization:

$$\min_{x \in \mathcal{X}} \quad f(x) = \frac{1}{n}\sum_i f_i(x)$$

graph: $G = (V, \mathcal{E})$

ea. node i has their x_i and access to $f_i(x)$ and infor from $\mathcal{N}(i)$

Motivating Example:

i : separate set of training data

x : parameters

linear $\quad f_i(x) = \| z_i - H_i x \|^2$

gen. $\quad f_i(x) = \| h_i(z_i, x) \|^2$

Properties of $f_i(x)$:

$f_i(x)$: convex "bowl shaped"

$g_i \in \partial f_i(x)$ subgradient

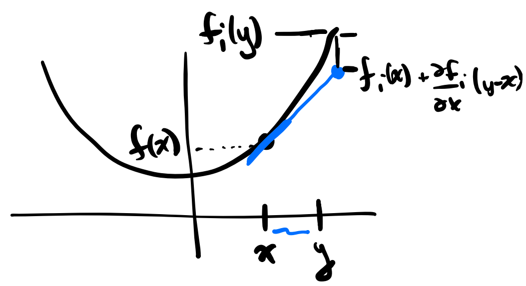
$g_i = \frac{\partial f_i}{\partial x}$

Convexity: x, y lower bound

$$f_i(y) \geq f_i(x) + \frac{\partial f_i}{\partial x}(y - x) + \varepsilon \|y - x\|^2$$

convexity

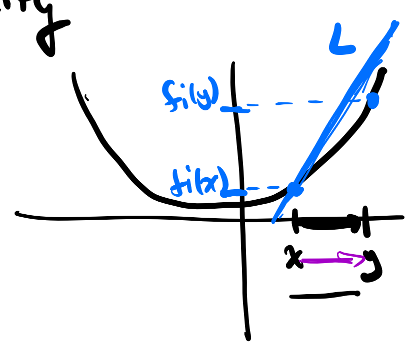
strong convexity



L-Lipschitz continuity:

$$\| f_i(y) - f_i(x) \| \leq L \| x - y \|$$

for any $x, y \in \mathcal{X}$



for an L-Lipschitz function: f_i with $g_i = \frac{\partial f_i}{\partial x}$

$\|g_i\|_* \le L$

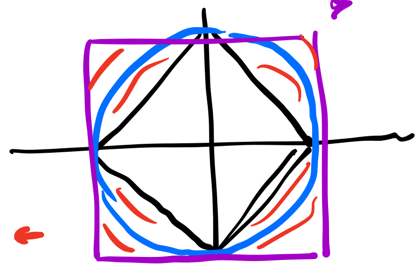
Norms:

$$\|x\|_1 = \sum_i |x_i|$$

$$\|x\|_2 = \Big(\sum_i |x_i|^2\Big)^{1/2}$$

$$\|x\|_p = \Big(\sum_i |x_i|^p\Big)^{1/p} \qquad 1 \le p \le \infty$$

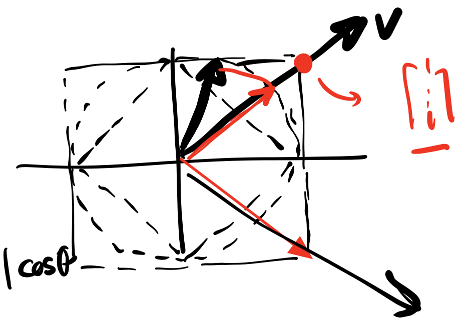
$$\|x\|_\infty = \max_i |x_i|$$



dual norm:

$$\|v\|_{p,*} = \sup_{\|u\|_p = 1} \langle v, u \rangle$$

$\|v\|\|u\|\cos\theta$



$$\boxed{\|v\|_{2,*} = \|v\|_2}$$

$$\|v\|_{\infty,*} = \|v\|_1$$

$$\|v\|_{1,*} = \|v\|_\infty$$

Proximal Function $\varphi(x)$ used for projection $\Pi_\chi^\varphi(z,\alpha)$

$\varphi(x)$: strongly convex

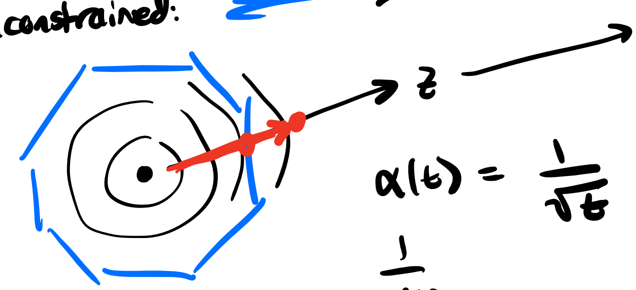
$$\varphi(x) = \frac{1}{2}\|x\|_2^2$$

$$x = \Pi_\chi^\varphi(z,\alpha) = \underset{x\in\chi}{\operatorname{argmin}} \Big\{ -\langle z, x\rangle + \frac{1}{\alpha}\varphi(x) \Big\}$$

unconstrained:

$$-z^T + \frac{1}{\alpha} x^T = 0$$

$$\Rightarrow \quad x = \alpha z$$

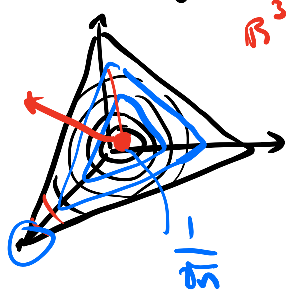


$\alpha(t) = \frac{1}{\sqrt{t}}$ ↓

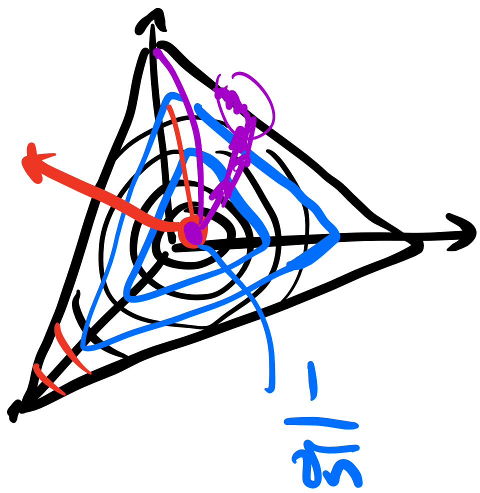
$\frac{1}{\alpha(t)}$ ↑

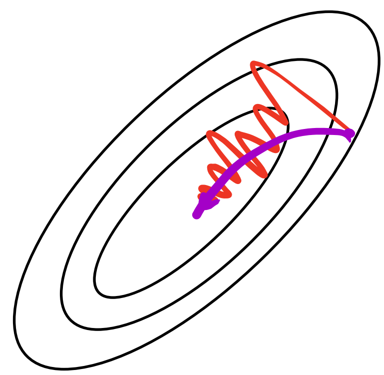
if x is on simplex "x is a discrete probability dist."

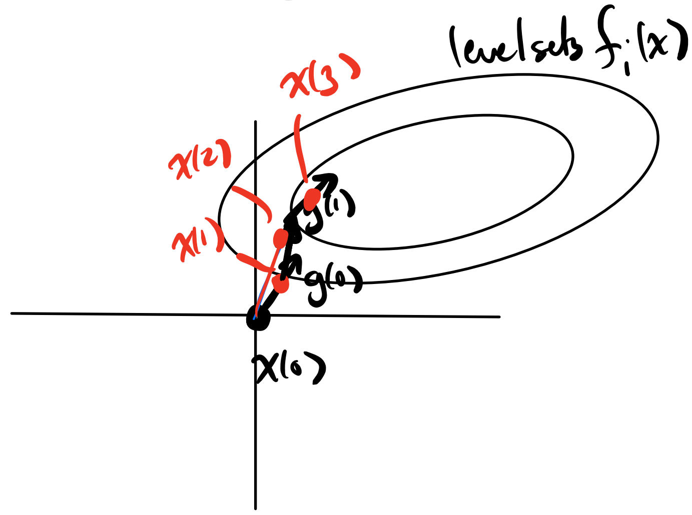
$\Delta_n = \{x \in \mathbb{R}^n \mid \mathbb{1}^T x = 1,\ x \geq 0\}$ $\quad x \in \mathbb{R}^3$



$\Psi(x) = \sum_i x_i \log(x_i) - x_i$ and $\|\cdot\|_1$







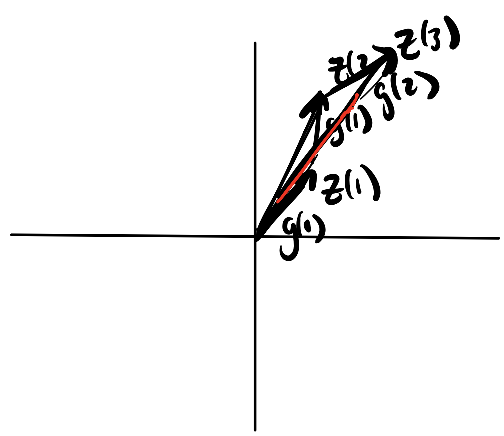
Dual Averaging:

$z(t+1) = z(t) - g(t)$

$g(t) \in \partial f_i(x_i(t))$

$g(t) = \frac{\partial f_i}{\partial x}(x_i(t))$

$x(t+1) = \Pi_X^{\varphi}(-z(t+1), \alpha(t))$



Traditional:

$$x(t+1) = \Pi\left(\underline{x(t) + \alpha g(t)}\right)$$

Linear Dynamics:

$$\underline{z(t+1)} = \underline{z(t) - g(t)} \Longrightarrow \underline{z(t+1) =} \underset{\downarrow}{\overset{I}{\underline{A}}}\ \underline{z(t) + \overset{-I}{\underline{B}}\, u(t)}$$

$g(t)$

$$z(0) = z_0$$

$$z(1) = Az(0) + Bu(0)$$

$$z(2) = A^2 z(0) + \underline{ABu(0)} + Bu(1)$$

$$z(3) = A^3 z(0) + \overset{Az(1)}{A^2 Bu(0)} + ABu(1) + Bu(2)$$

$$\vdots$$

$$z(t) = \underline{A}^t \cancel{z(0)}^{\,0} + \sum_{j=0}^{t-1} \underline{A}^{t-j-1} \underline{B}\, u(j)$$

$$\boxed{z(t) = \underbrace{A^t z(0)}_{\text{drift term}}} + \underbrace{\begin{bmatrix} A^{t-1}B & A^{t-2}B & \cdots & AB & B \end{bmatrix}}_{\text{Reachability Matrix}} \begin{bmatrix} u(0) \\ \vdots \\ u(t-1) \end{bmatrix}$$

Distributed Scheme

ea. node i $\{x_i(t), z_i(t)\}$

compute $g_i(t) \in \partial f_i(t)$ → local f_i

receive $z_j(t) \in j \in N(i)$

Communication matrix $P \in \mathbb{R}^{n \times n}$

P is doubly stochastic, symmetric

$P_{ij} > 0$ if and only if $j \in N(i)$

$$\sum_j P_{ij} = \sum_{j \in N(i)} P_{ij} = 1 \qquad P\mathbb{1} = \mathbb{1}$$

$$\sum_i P_{ij} = \sum_{i \in N(j)} P_{ij} = 1 \qquad \mathbb{1}^T P = \mathbb{1}^T$$

Before: $z(t+1) = z(t) - \underline{g(t)}$

Now:

$$z_i(t+1) = \sum_{j \in N(i)} P_{ij} z_j(t) - g_i(t) \qquad x_i(t+1) = \Pi_{\mathcal{X}}^{\varphi}(-z_i(t+1), \alpha(t))$$

vector case for z_i & x_i

for scalar: z_i, x_i, g_i

$$z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad g = \begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix}$$

$$z(t+1) = Pz(t) - g$$

$$z(t) = P^t \cancel{z(0)}^{\,0} - \sum_{s=0}^{t-1} P^{t-s-1} g(s)$$

Question:

ea. z_i is a descent direction ← want to agree on optimal descent direction

what is the average descent direction doing?

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

in scalar z_i case:

$$z$$

$$\frac{1}{n}\mathbb{1}^T \left[\cancel{z(t+1) = Pz(t) - g} \right]$$

$$\frac{1}{n}\mathbb{1}^T z(t+1) = \frac{1}{n}\mathbb{1}^T P z(t) - \frac{1}{n}\mathbb{1}^T g$$

$$\mathbb{1}^T$$

$$\bar{z}(t+1) = \bar{z}(t) - \frac{1}{n}\sum_i g_i(x_i)$$

average descent direction

will be overall gradient when all x_is agree.

$$\hat{x}_i(T) = \frac{1}{T}\sum_{t=1}^{T} x_i(t) \longrightarrow$$ time average of x_i

Thm 1: Basic Convergence

$$f(\hat{x}_i(T)) - f(x^*) \leq \frac{1}{T\alpha(T)}\psi(x^*) + \frac{L^2}{2T}\sum_{t=1}^{T}\alpha(t-1) + \boxed{\frac{3L}{T}\max_j \sum_{t=1}^{T}\alpha(t)\|\bar{z}(t) - z_j(t)\|_*}$$

AP ✓

converges to 0

Thm 2: Rates (spectral gap)

$$f(\hat{x}_i(T)) - f(x^*) \leq 8\frac{RL}{\sqrt{T}}\frac{\log(T\sqrt{n})}{\sqrt{1-\sigma_2(P)}}$$ for all $i \in V$

$1 - \sigma_2(P)$ = spectral gap.

$\bar{\omega}$