262B-Lecture 16

Date created: 2021.03.16
N. of Pages: 17

$\min \; f(x) + h(x)$

s.t. $x \in X$

, assume:

f: convex & differentiable

h: convex but maybe non-smooth

Proximal gradient algorithm:

$$\begin{cases} z^{(k)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)}) & \longrightarrow \text{gradient alg. on } f(\cdot) \\ x^{(k+1)} = \underbrace{\text{prox}_{\alpha^{(k)}, h}(z^{(k)})}_{\text{argmin}_{x \in X} \left(h(x) + \frac{1}{2\alpha^{(k)}} \| x - z^{(k)} \|^2 \right)} & \longrightarrow \text{proximal on } h(\cdot) \text{ and } X \end{cases}$$

Two special cases:

$1. \; f(x) = 0 \Rightarrow z^{(k)} = x^{(k)}$

$\Rightarrow x^{(k+1)} = \text{prox}_{\alpha^{(k)}, h}(x^{(k)})$

$\Rightarrow$ proximal algorithm for $\min h(x)$ s.t. $x \in X$

$2. \; z - h(x) = 0$

$\text{prox}_{\alpha^{(k)}, h}(z) = \underset{x \in X}{\text{argmin}} \left(0 + \frac{1}{2\alpha^{(k)}} \| x - z \|^2 \right)$

$= \underset{x \in X}{\text{argmin}} \| x - z \| = P_X(z)$

$$\Rightarrow \quad x^{(k+1)} = \text{prox}_{\alpha^{(k)}, h}(z^{(k)}) = P_X(z^{(k)})$$

$$= P_X(x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)}))$$

$\Rightarrow$ Proximal gradient method is a generalization of projected gradient method.

Ex: $\min_{x \in \mathbb{R}^n} f(x) + |x|_1$ $\longrightarrow$ sparse solution in machine learning

special case: Lasso

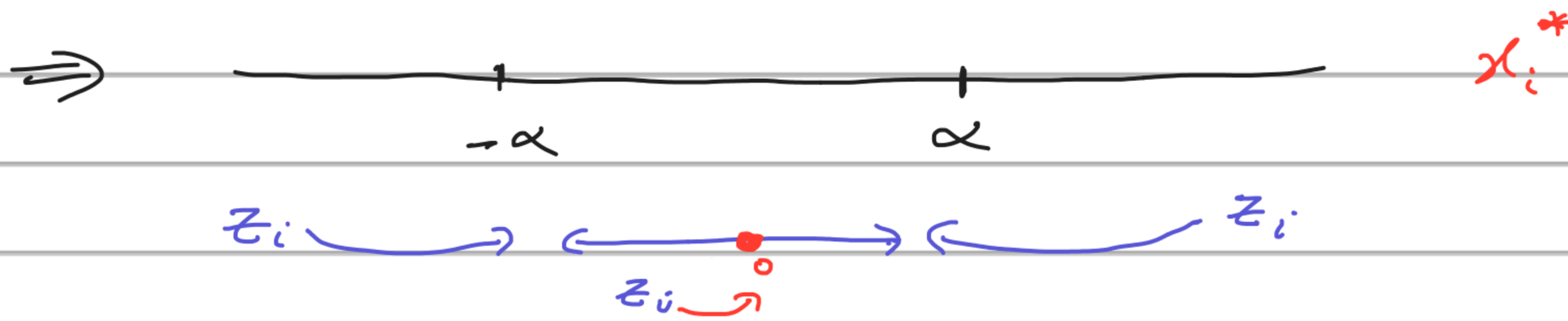
$$\min \|Ax - b\|^2 + \lambda |x|_1 \longrightarrow \min \frac{1}{\lambda} \|Ax - b\|^2 + |x|_1$$

Proximal operator: $h(x) = |x|_1$

$$\text{prox}_{\alpha, \lambda}(z) \longrightarrow \min_x |x|_1 + \frac{1}{2\alpha} \|x - z\|^2$$

$$\longrightarrow \min_x \underbrace{\sum_{i=1}^{\hat{}} (|x_i| + \frac{1}{2\alpha}(x_i - z_i)^2)}_{\text{decomposable}}$$

$$\Rightarrow x_i^* = \begin{cases} z_i - \alpha & \text{if } z_i > \alpha \\ 0 & \text{if } z_i \in [-\alpha, \alpha] \\ z_i + \alpha & \text{if } z_i < -\alpha \end{cases}$$



Prox operator in this case is called soft thresholding or shrinkage operator.

$$\min_{x \in \mathbb{R}^n} f(x) + |x|_1$$

$$\Rightarrow x^{(k+1)} = \underbrace{ST_{\alpha^{(k)}}}_{\text{soft thresholding}}\left(x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)})\right)$$

Recall : - proximal gradient algorithm is a general case of projected gradient algorithm.

- Convergence proofs for projected gradient algorithms are based on :

$$\| P_X(x) - P_X(y) \| \leq \|x - y\| \Rightarrow \text{non-expansive property}$$

Thm: $\| prox_{\alpha,h}(x) - prox_{\alpha,h}(y) \| \leq \| x - y \| \quad \forall x, y \in \mathbb{R}^n$

proof in a special case: $h(\cdot)$ = differentiable

shorthand notation: $z(x) = prox_{\alpha,h}(x)$

$z(y) = prox_{\alpha,h}(y)$

$$\Rightarrow z(x) = \underset{w \in X}{\operatorname{argmin}} \underbrace{\left(h(w) + \frac{1}{2\alpha} \| w - x \|^2 \right)}_{\text{Convex}}$$

(FOC: gradient of objective =

$$\nabla h(z(x)) + \frac{z(x) - x}{\alpha}$$

$$\rightarrow \left(\nabla h(x(z)) + \frac{z(x) - x}{\alpha} \right)^T (w - z(x)) \geq 0$$

$\forall w \in X$

since $z(y) \in X$, pick $w = z(y)$

$$\Rightarrow \boxed{\left(\nabla h(x(z)) + \frac{z(x) - x}{\alpha} \right)^T (z(y) - z(x)) \geq 0.}$$

(1)

Redo for $z(y)$:

$$\left(\nabla h(z(y)) + \frac{z(y)-y}{\alpha}\right)^T \left(z(x) - z(y)\right) \geq 0 \qquad (2)$$

Also, Convexity of $h(\cdot)$:

$$\left(\nabla h(z(x)) - \nabla h(z(y))\right)^T \left(z(x) - z(y)\right) \geq 0 \qquad (3)$$

(1), (2), (3) :

$$\left((z(x)-x) - (z(y)-y)\right)^T \left(z(x) - z(y)\right) \leq 0$$

$$\Rightarrow \|z(x) - z(y)\|^2 \leq (x-y)^T (z(x) - z(y))$$

$$\leq \|x-y\| \times \|z(x) - z(y)\| \qquad (*)$$

1- If $\|z(x) - z(y)\| = 0 \Rightarrow \|z(x) - z(y)\| \leq \|x-y\|$ ($\|z(x) - z(y)\| = 0$)

2- If $\|z(x) - z(y)\| \neq 0$

$\Rightarrow (*)$: $\|z(x) - z(y)\| \leq \|x-y\|$

similar to projected gradient method:

1 - $\|\nabla f(x) - \nabla f(y)\| \leq L \|x-y\|$ $(\forall x,y \in \mathbb{R}^n)$

$\Rightarrow$ iteration complexity $= O(\frac{1}{\varepsilon})$

2 - $\|\nabla f(x) - \nabla f(y)\| \leq L \|x-y\|$ $\forall x,y \in \mathbb{R}^n$

$\nabla^2 f(x) \geq m I$, $m > 0$, $\forall x \in \mathbb{R}^n$

$\Rightarrow \|x^{(k+1)} - x_*\| \leq \sqrt{1 - \frac{m}{L}} \|x^{(k)} - x_*\|$

if $\alpha^{(k)} = \frac{1}{L}$

$\Rightarrow$ iteration complexity $= O(\log \frac{1}{\varepsilon})$

Note: number of iterations is like unconstrained case, but we shold take a proximal step at every iteration:

per-iteration complexity $\sim h(\cdot), X$

chapter 4 :

previously, we studied :

$$\begin{cases} \min \quad f(x) \longrightarrow \text{non-convex} \\ \text{s.t} \quad x \in X \longrightarrow \text{convex} \longrightarrow \text{geometric FOC, SOC} \end{cases}$$

what if X = non-convex ?

- optimization with equality constraints :

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

$$\text{s.t.} \quad \begin{cases} h_1(x) = 0 \\ h_2(x) = 0 \\ \vdots \\ h_m(x) = 0 \end{cases} \longrightarrow \underbrace{h(x) = 0}_{\text{vector}} \longrightarrow \text{nonLinear}$$

Definition: Cone of first-order feasible directions at $y \in \mathbb{R}^n$

$$\underbrace{V(y)}_{\text{Variation}} = \{ \Delta x \mid \nabla h_i(y)^T \Delta x = 0, \quad i = 1, \dots, m \}$$

Idea: pick a feasible y : $h(y)=0$

perturb it as $y+\Delta y$ s.t. : $h(y+\Delta y)=0$

$$h_i(y)=0 \quad , \quad 0 = h_i(y+\Delta y) = \cancel{h_i(y)} + \nabla h_i(y)^T \Delta y + O(\|\Delta y\|^2)$$

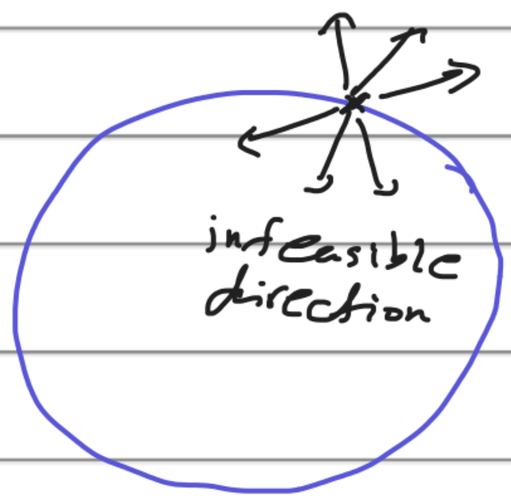
small

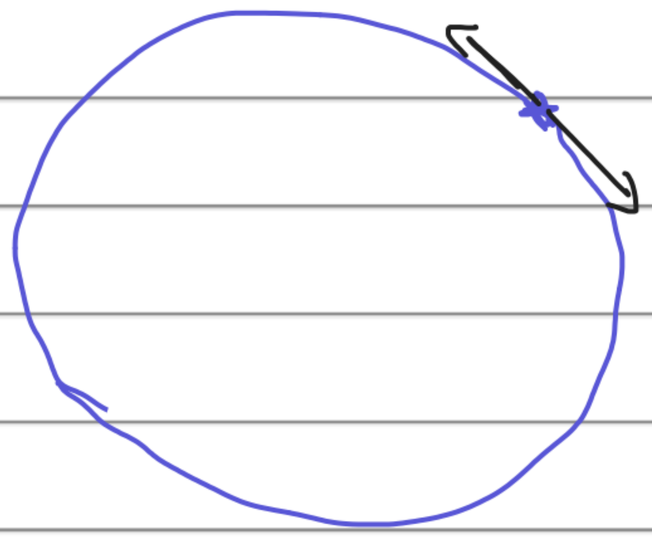
If Δy : small $\Rightarrow$ $\nabla h_i(y)^T \Delta y \simeq 0$

$\Rightarrow$ $\Delta x \in V(y)$: $h(\underbrace{y+\varepsilon \Delta x}_{\Downarrow}) \simeq 0$ if ε = small

almost feasible

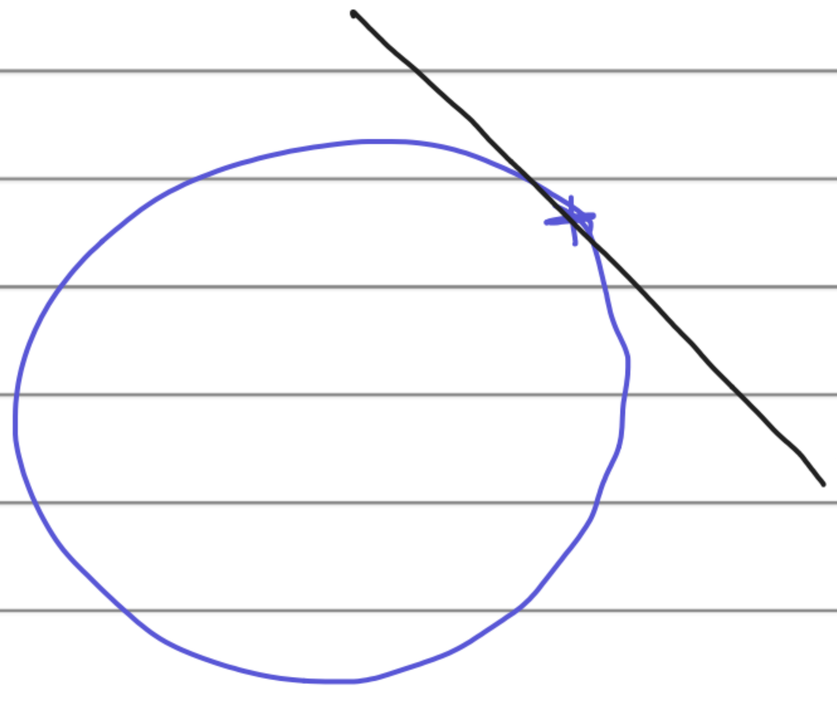
Ex:





not feasible, but almost feasible

$\Rightarrow$



$\Rightarrow V(y)$

$V(y)$ is the same as Tangent plane at y if y is a regular point.

Def: y is a regular point if $\nabla h_1(y), \dots, \nabla h_m(y)$ are linearly independent.

FOC (necessary): If x_* is a regular point and a local min for $\min f(x)$ s.t. $h(x) = 0$,

then $\exists\ \lambda_1^*, \dots, \lambda_m^* \in \mathbb{R}$ s.t.

$$\nabla f(x_*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x_*) = 0 \qquad (*)$$

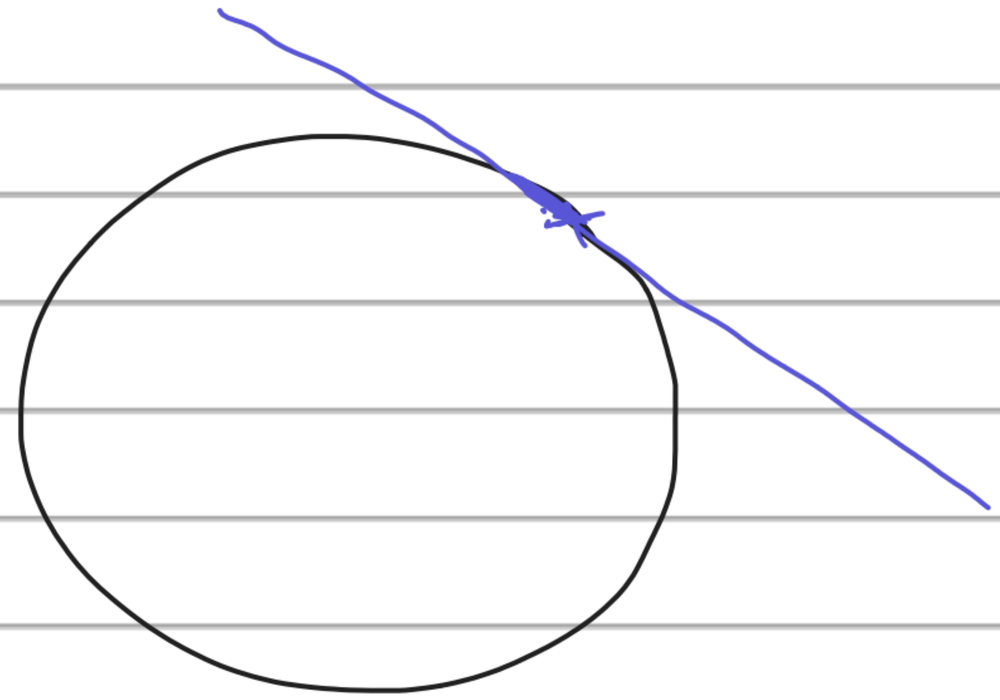
λ_i^*: Lagrange multipliers

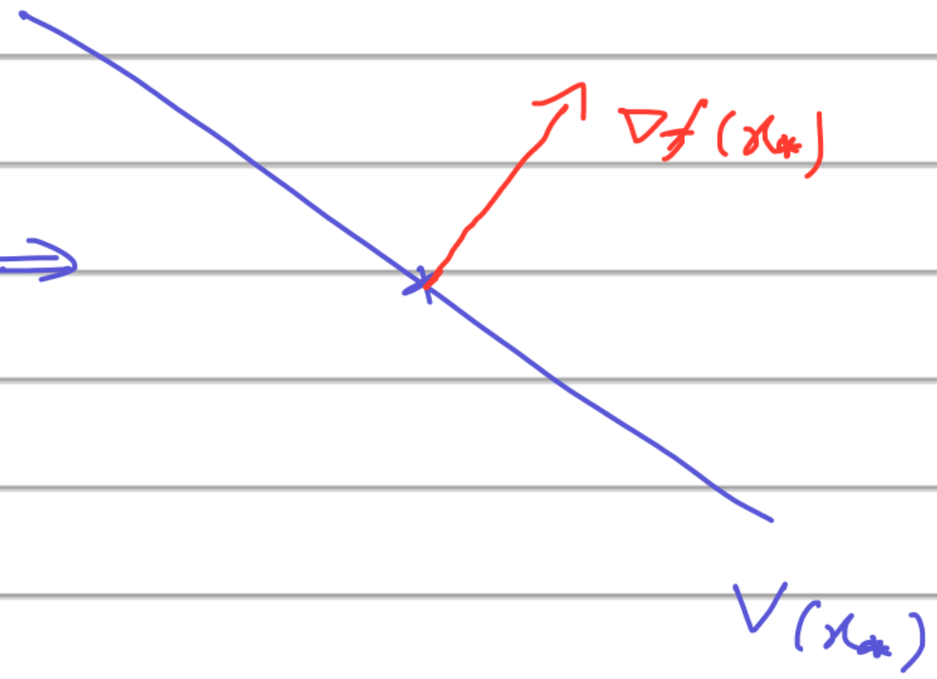
Interpretations:

1- $\nabla f(x_*) \in$ space spanned by gradients of constraints

2- $\nabla f(x_*)$ is orthogonal to $V(x_*)$.

Ex:

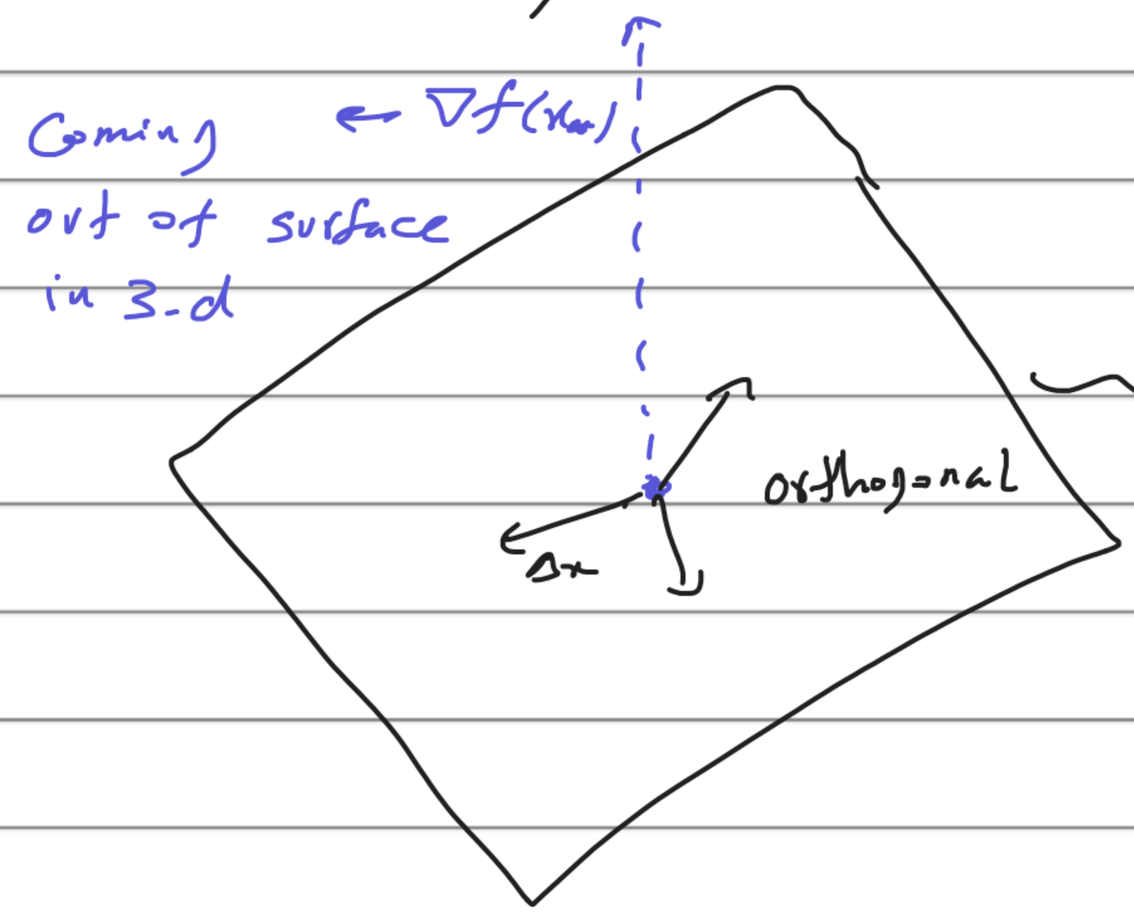




Proof: (*) $\Rightarrow$ If Δx satisfies

$\nabla h_i(x_*)^T \Delta x = 0 \quad i = 1, \dots, m$, then $\nabla f(x_*)^T \Delta x = 0$

$\Rightarrow$ every $\Delta x \in V(x_*)$ is orthogonal to $\nabla f(x_*)$



$\rightsquigarrow$ surface defined by gradients of constraints

SOC (necessary): If x_* is a regular point and a local min, then $\lambda_1^*, \dots, \lambda_m^*$ satisfying FOC also satisfy:

$$\Delta x^T \left(\nabla^2 f(x_*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x_*) \right) \Delta x \geq 0$$

$$\forall \ \Delta x \in V(x_*)$$

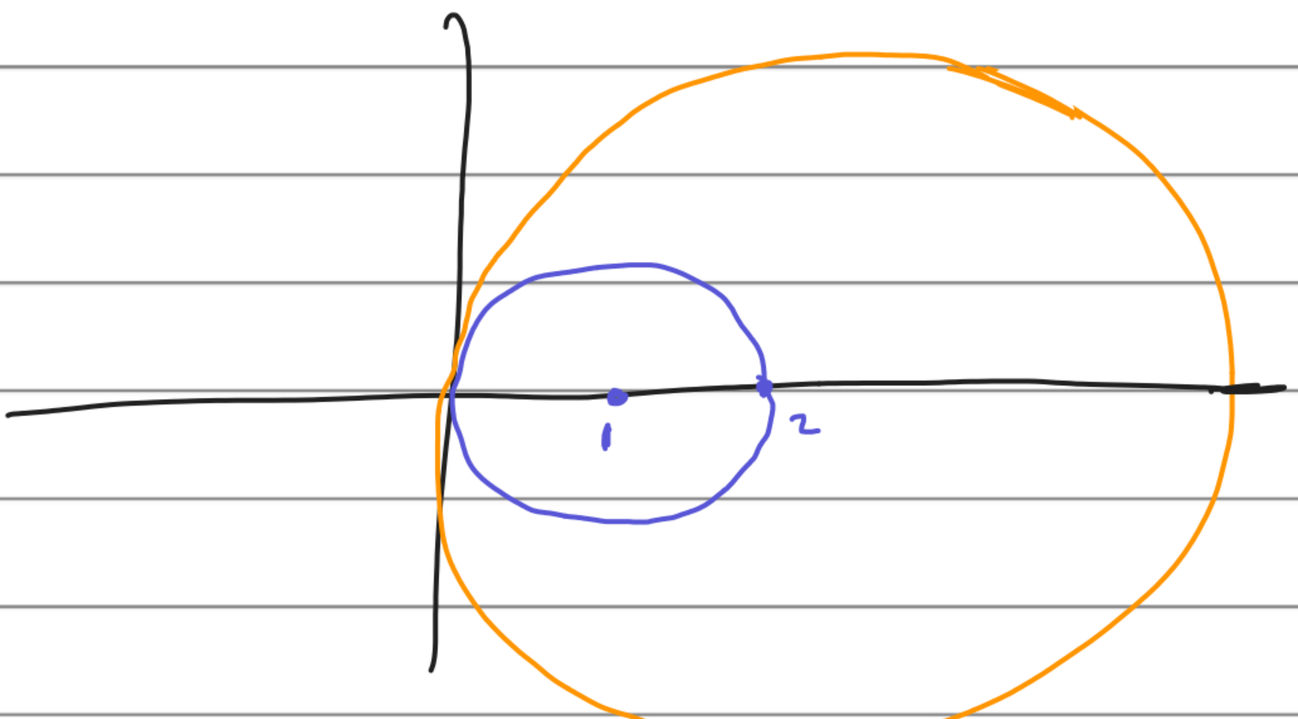
Why is regularity important:

$$\min \ x_1 + x_2$$

$$\text{s.t.} \ (x_1 - 1)^2 + x_2^2 - 1 = 0$$

$$(x_1 - 2)^2 + x_2^2 - 4 = 0$$

$\Rightarrow$ feasible set



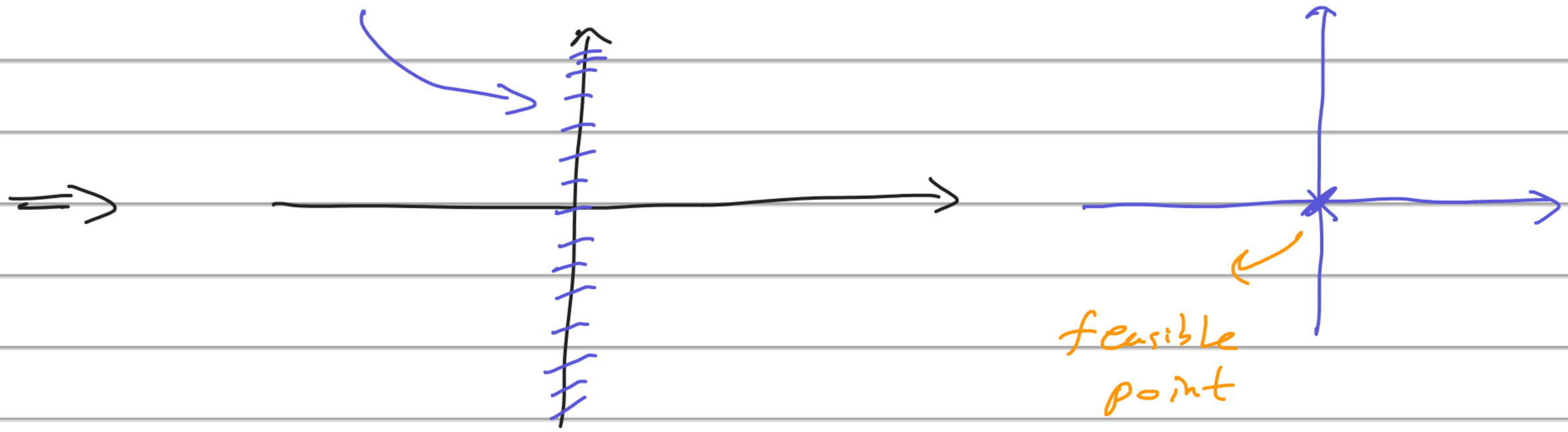
$\Rightarrow$

X

origin = feasible point

$\Rightarrow \quad x_* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Also, $V(x_*) = \{ \Delta x \mid \begin{bmatrix} -2 & 0 \end{bmatrix} \Delta x \geq 0, \begin{bmatrix} -4 & 0 \end{bmatrix} \Delta x \geq 0 \}$



Line is not tangent plane for a single point

x_* : Not a regular point ($\begin{bmatrix} -2 \\ 0 \end{bmatrix}, \begin{bmatrix} -4 \\ 0 \end{bmatrix}$: linearly dependent)

$\nabla f(x_*) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

FoC : $\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \lambda_1^* \begin{bmatrix} -2 \\ 0 \end{bmatrix} + \lambda_2^* \begin{bmatrix} -4 \\ 0 \end{bmatrix} = 0$

This has no solution.

Prove FOC, SOC :

{ 1 - penalty method

2 - Elimination method

---

1 - penalty method:

Define: $\underbrace{F^k(x)}_{\text{surrogate for original optimization}} = f(x) + \underbrace{\frac{k}{2}\|h(x)\|^2}_{\text{Constraint violation}} + \underbrace{\frac{\alpha}{2}\|x - x_*\|^2}_{\text{proximal term}}$

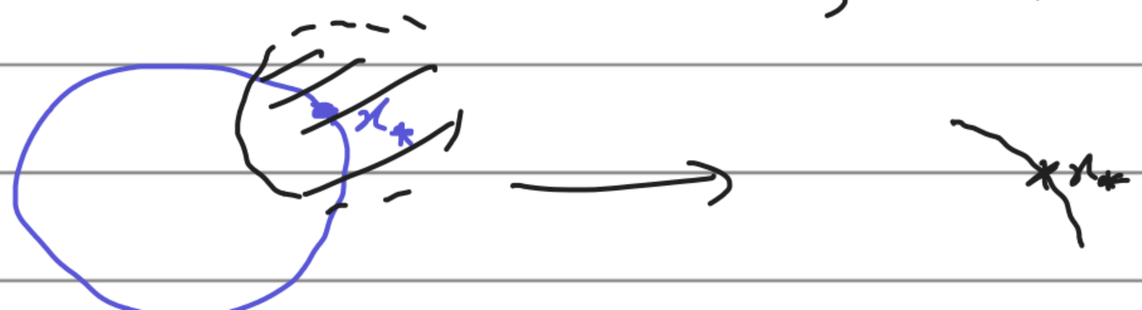
k : positive integer , $\alpha > 0$

$F^1(x) \longrightarrow F^2(x) \longrightarrow \cdots \longrightarrow F^{k(x)} \longrightarrow \cdots$

Since x_* is a local min, $\exists\ \varepsilon > 0$ s.t

$f(x_*) \leq f(x) \quad \forall x \in S$ s.t. $h(x) = 0$

where $S = \{ x \mid \|x - x_*\| \leq \varepsilon \}$



Define: $x^{(k)}$ = global min for $\min F^k(x)$ s.t. $x \in S$

no $h(x)=0$ in this optimization

Note: $F^k(x^{(k)}) \leq F^k(x_*) = f(x_*)$

arbitrary point in S

$$+ \frac{k}{2} \| h(x_*) \|^2 + \frac{\alpha}{2} \| x_* - x_* \|^2$$

$$\Rightarrow F^k(x^{(k)}) \leq f(x_*)$$

$$\Rightarrow \left(F^k(x^{(k)}) = f(x^{(k)}) + \frac{k}{2} \| h(x^{(k)}) \|^2 + \frac{\alpha}{2} \| x^{(k)} - x_* \|^2 \right) \leq f(x_*) \quad (*)$$

$f(x^{(k)})$ is bounded on S, so the left side would blow up and violate $(*)$ when $k \to \infty$, unless $\| h(x^{(k)}) \| \to 0$

$$\Rightarrow \lim_{k\to\infty} \| h(x^{(k)}) \| = 0$$

let $\bar{x}$ denote an arbitrary Limit point of

$\{x^{(k)}\} \Rightarrow h(\bar{x}) = 0 \Rightarrow \bar{x}$: feasible point

$$(*) \Rightarrow f(x^{(k)}) + \frac{\alpha}{2} \| x^{(k)} - x_* \|^2 \leq f(x_*)$$

$$k \to \infty : \quad f(\bar{x}) + \frac{\alpha}{2} \| \bar{x} - x_* \|^2 \leq f(x_*)$$

If $\bar{x} \neq x_* \to \| \bar{x} - x_* \| > 0 \Rightarrow$

$f(\bar{x}) < f(x_*) \longrightarrow$ Contradiction

(x_* : best feasible point in the ball S)

$\Rightarrow \bar{x} = x_*$, $\longrightarrow$ every Limit point is x_*

$$\Rightarrow \begin{cases} \min \ F^{k}(x) \\ \text{s.t.} \quad x \in S \end{cases}$$

or

$$\begin{cases} \min f(x) + \frac{k}{2} \| h(x) \|^2 + \frac{\alpha}{2} \| x - x_k \|^2 \\ \text{s.t.} \quad x \in S \end{cases}$$

As $k \to \infty$, solution of new problem

Converges x_*.

Since $x_* \in$ interior of S

$\Rightarrow$ $x^{(k)} \in$ interior of S if k is large

$\Rightarrow$ $\min F^k(x)$ if k is large.

~~s.t. $x \in S$~~

$\Rightarrow$ Unconstrained optimization.

FOC, SOC for unconstrained optimization:

$$\nabla F^k(x^{(k)}) = 0 \quad , \quad \nabla^2 F^k(x^{(k)}) \succeq 0$$

Goal:

$= $ FOC for constrained opt as $k \to \infty$

$= $ SOC for constrained opt as $k \to \infty$

$$\nabla h(x) = \begin{bmatrix} \nabla h_1(x) & \nabla h_2(x) & \cdots & \nabla h_m(x) \end{bmatrix}$$

regularity: $\nabla h(x_*) =$ full column rank

$$\{x^{(k)}\} \longrightarrow x_*$$

$$\Rightarrow \nabla h(x^{(k)}) = \text{full column rank if } k = \text{Large}$$

$$\Rightarrow \nabla h(x^{(k)})^T \nabla h(x^{(k)}) = \text{invertible matrix}$$

$$0 \overset{\text{Foc}}{=} \nabla F^k(x^{(k)}) = \nabla f(x^{(k)}) + k \nabla h(x^{(k)}) h(x^{(k)}) + \alpha(x^{(k)} - x_*)$$

$$\Rightarrow k h(x^{(k)}) = -(\nabla h(x^{(k)})^T \nabla h(x^{(k)}))^{-1} \nabla h(x^{(k)})^T (\nabla f(x^{(k)}) + \alpha(x^{(k)} - x_*))$$

$$k \to \infty : \lim_{k \to \infty} k h(x^{(k)}) = \lambda_*$$

where: $\lambda_* = -(\nabla h(x_*)^T \nabla h(x_*))^{-1} \nabla h(x_*)^T \nabla f(x_*)$

explicitly found Lagrange multipliers

(**) $\longrightarrow$ Take Limit:

$$\nabla f(x_*) + \underbrace{\nabla h(x_*) \lambda_*}_{\sum_{i=1}^{m} \nabla h_i(x_*) \lambda_i^*} = 0$$