262B-Lecture 15

Date created: 2021.03.11
N. of Pages: 16

$$\min f(x) \quad \text{s.t.} \quad x \in X \implies M_{\alpha,f}(z) = \inf_{x \in X} \underbrace{\left(f(x) + \frac{1}{2\alpha} \|x - z\|^2 \right)}_{g(x,z)}$$

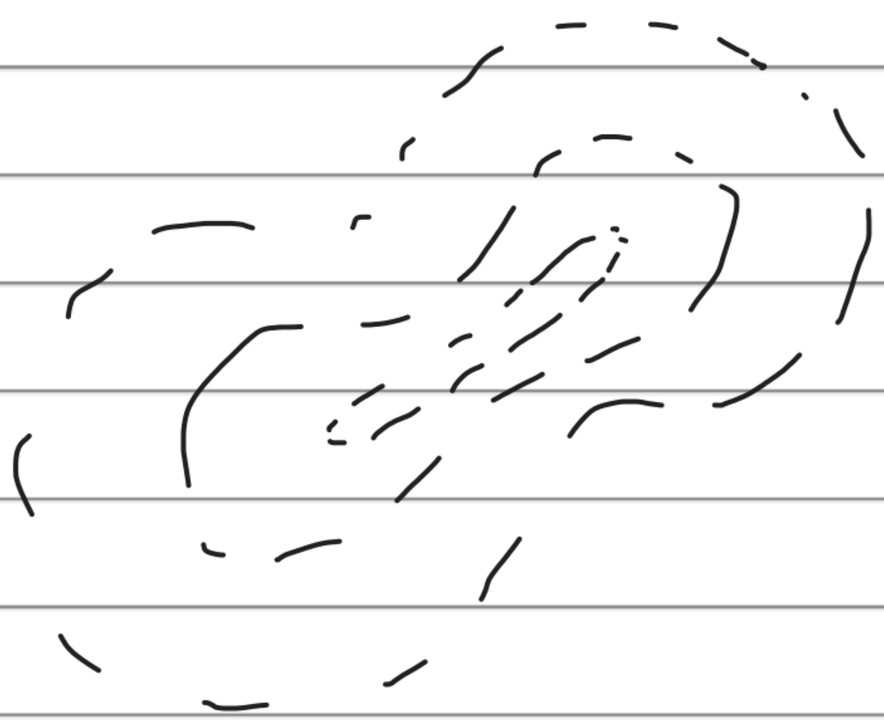
2- Prox operator:

$$\text{prox}_{\alpha,f}(z) = \underset{x \in X}{\arg\min}\ g(x,z)$$

$\text{prox}_{\alpha,f}(z)$ exists & is unique at every point z.

idea: $\{ x \mid f(x) + \frac{1}{2} \|x - z\|^2 \leq \text{any constant} \}$

is compact due to $\|x - z\|^2$.



as the constant gets smaller, the set shrinks and becomes of measure zero $\rightarrow$ existence of solution

$\downarrow$

a single point due to strict Convexity of $\|x - z\|$

Short hand notation: $\text{prox}_{\alpha,f}(z) = x(z)$

3 - $\nabla_z M_{\alpha,f}(z) = \nabla_z g(x,z)\Big|_{x=x(z)}$

$= \dfrac{z - x(z)}{\alpha} = \dfrac{z - \text{prox}_{\alpha,f}(z)}{\alpha}$ $\quad \forall z \in \mathbb{R}^n$

(not just $z \in X$)

Special case: assume $f(\cdot)$ is differentiable and X = entire space

$\min_x g(x,z) \Rightarrow \nabla_x g(x,z) = 0$

$\nabla f(x(z)) + \dfrac{x(z) - z}{\alpha} = 0 \Rightarrow$

$$\nabla_z M_{\alpha,f}(z) = \nabla f(\text{prox}_{\alpha,f}(z)) \qquad \forall z \in \mathbb{R}^n$$

4 - $\inf_{x \in X} f(x) \leq M_{\alpha,f}(z) \leq f(z) \qquad \forall z \in X$

may only be defined over X

(a)

(b)

(a): $M_{\alpha,f}(z) = \inf_{x \in X} \left(f(x) + \dfrac{1}{2\alpha} \|x - z\|^2 \right)$ $\quad z \in X$

$\leq f(z) + \dfrac{1}{2} \|z - z\|^2 = f(z)$

(b): $M_{\alpha,f}(z) = \inf_{x \in X} \left(f(x) + \frac{1}{2} \|x - z\|^2 \right)$

$$\geq \inf_{x \in X} \left(f(x) + 0 \right) = \inf_{x \in X} f(x)$$

5_ set of minima of $\min f(x)$ s.t. $x \in X$

$$= \text{set of minima of } \min M_{\alpha,f}(z) \text{ s.t. } \boxed{z \in \mathbb{R}^n}$$

proof: $\min_{z \in \mathbb{R}^n} M_{\alpha,f}(z) = \min_{z \in \mathbb{R}^n} \min_{x \in X} \left(f(x) + \frac{1}{2\alpha} \|x - z\|^2 \right)$

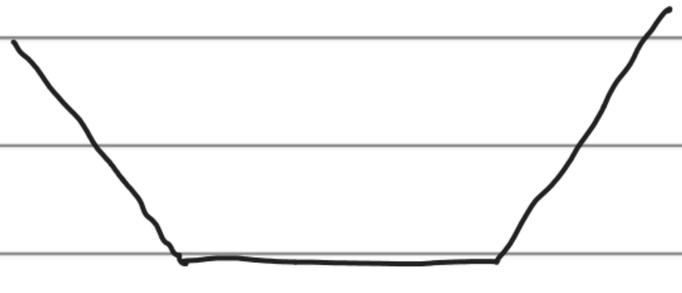
$$= \min_{\substack{x \in X \\ z \in \mathbb{R}^n}} f(x) + \frac{1}{2\alpha} \|x - z\|^2$$

$$= \min_{x \in X} \min_{z \in \mathbb{R}^n} \left(f(x) + \frac{1}{2\alpha} \|x - z\|^2 \right) = \min_{x \in X} f(x)$$

given x, min over z : $z = x$

$\Rightarrow$ x_* : min of $f(x)$ over X, $z_* = x_*$

$f(x)$ $\longrightarrow$ $M_{\alpha,f}(x)$: Moreau envelope



smooth

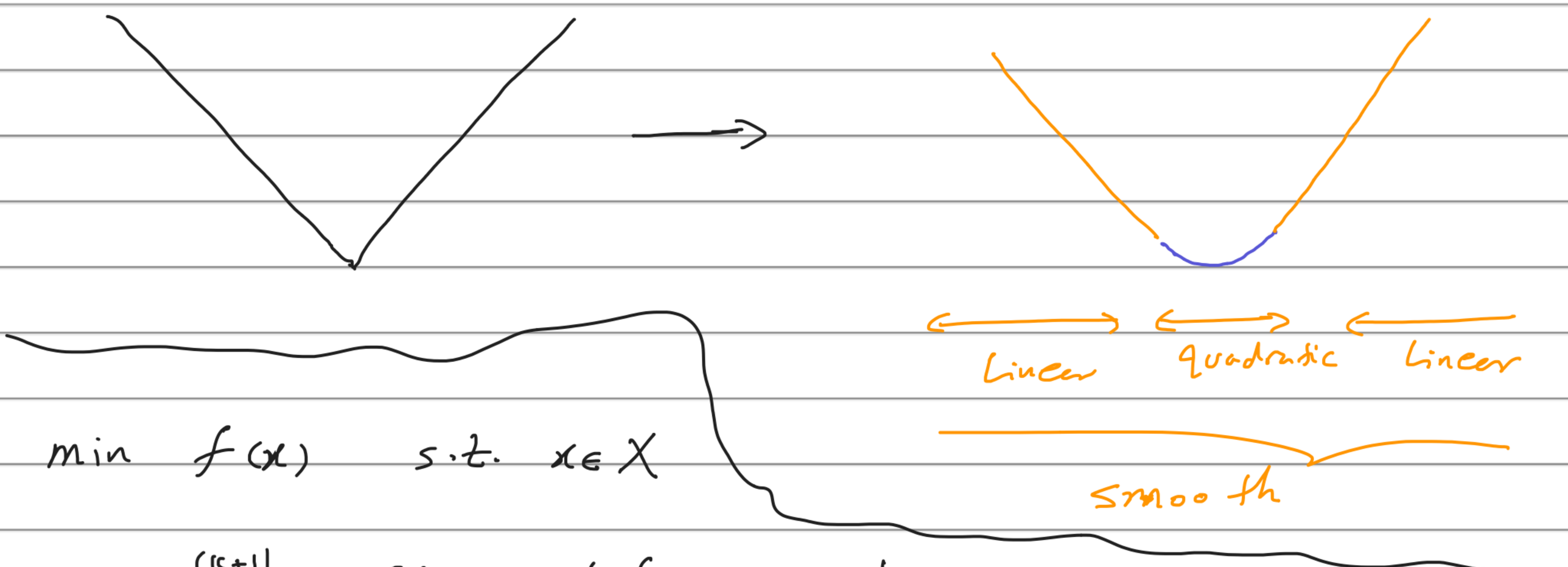
set same of minima

Ex: $f(x) = |x| \qquad x \in \mathbb{R}$

$$M_{\alpha, f}(x) = \inf_y \left(|y| + \frac{1}{2\alpha} |y - x|^2 \right)$$

$$= \begin{cases} \frac{1}{2\alpha} x^2 & |x| \le \alpha \longrightarrow \text{quadratic} \\ |x| - \frac{\alpha}{2} & |x| > \alpha \longrightarrow \text{linear} \end{cases}$$

$\rightarrow$ Huber function



$$\min \quad f(x) \quad \text{s.t.} \quad x \in X$$

$$\Rightarrow x^{(k+1)} = \arg\min_{x \in X} \left(f(x) + \frac{1}{2\alpha^{(k)}} \| x - x^{(k)} \|^2 \right)$$

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

Equivalently:

$$\nabla M_{\alpha^{(k)},f}(x^{(k)}) = \frac{x^{(k)} - Prox_{\alpha^{(k)},f}(x^{(k)})}{\alpha^{(k)}}$$

$$= \frac{x^{(k)} - x^{(k+1)}}{\alpha^{(k)}}$$

Assume: $\alpha^{(k)} = \alpha$

$$\Rightarrow \boxed{x^{(k+1)} = x^{(k)} - \alpha \nabla M_{\alpha,f}(x^{(k)})}$$

$\Rightarrow$ Iterates of proximal algorithm on possibly non-smooth $f(x)$ over X are the same as iterates of gradient algorithm on smooth $M_{\alpha,f}(x)$ over $\mathbb{R}^n$.

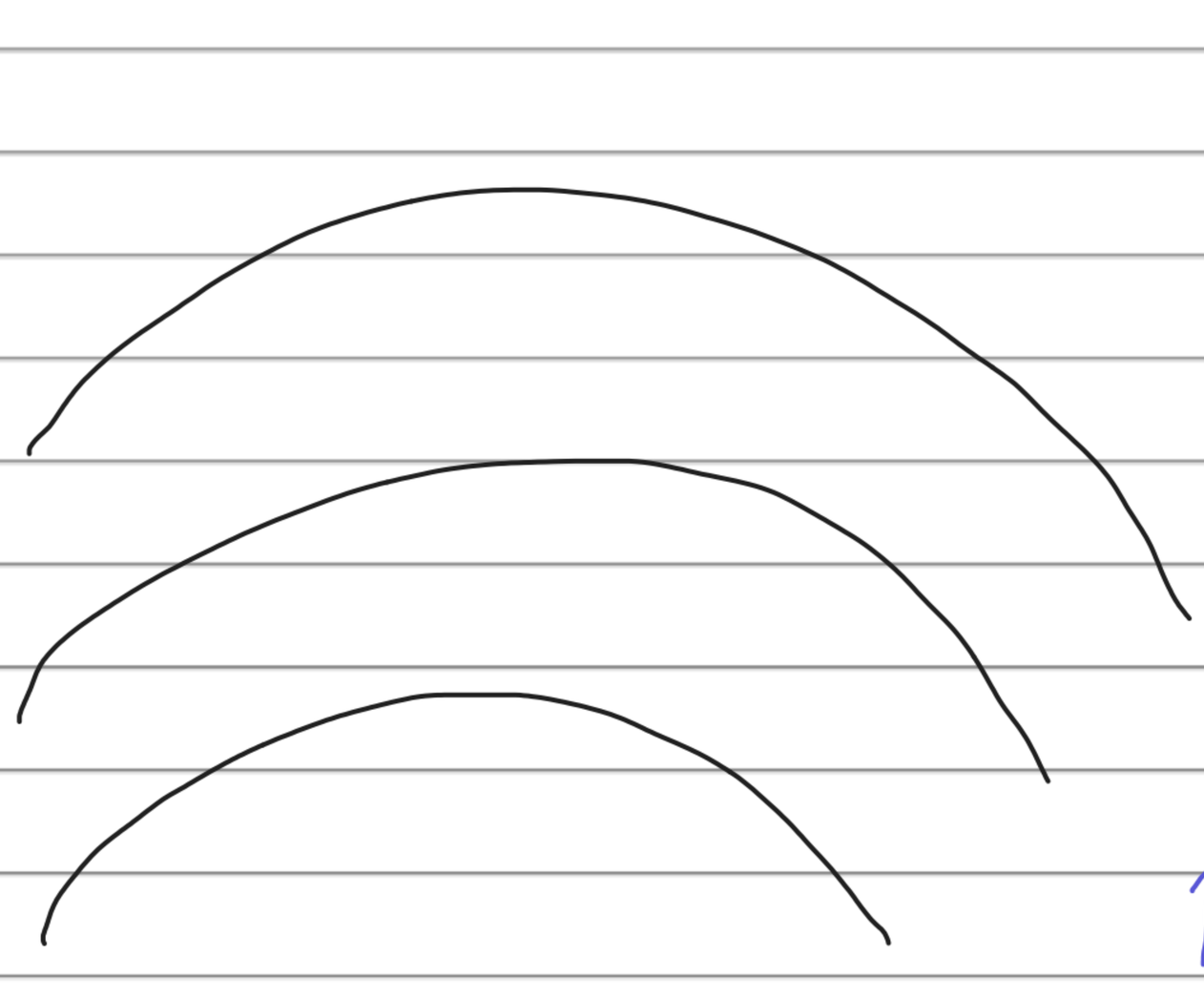
$$x^{(k+1)} = \underset{x \in X}{\operatorname{argmin}} \left(f(x) + \frac{1}{2\alpha^{(k)}} \|x - x^{(k)}\|^2 \right)$$

Define: $$\gamma^{(k)} = \min_{x \in X} \left(f(x) + \frac{1}{2\alpha^{(k)}} \|x - x^{(k)}\|^2 \right)$$

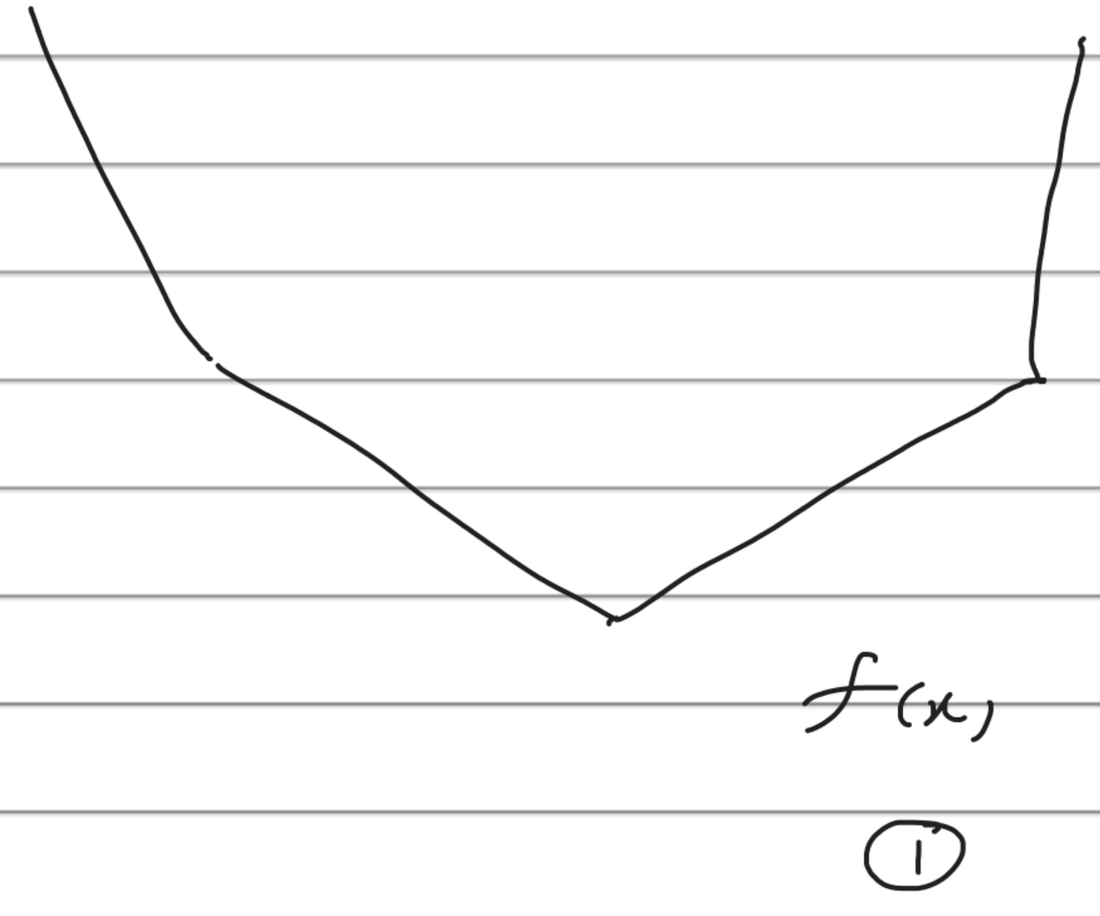
Consider 1- $f(x)$ over X

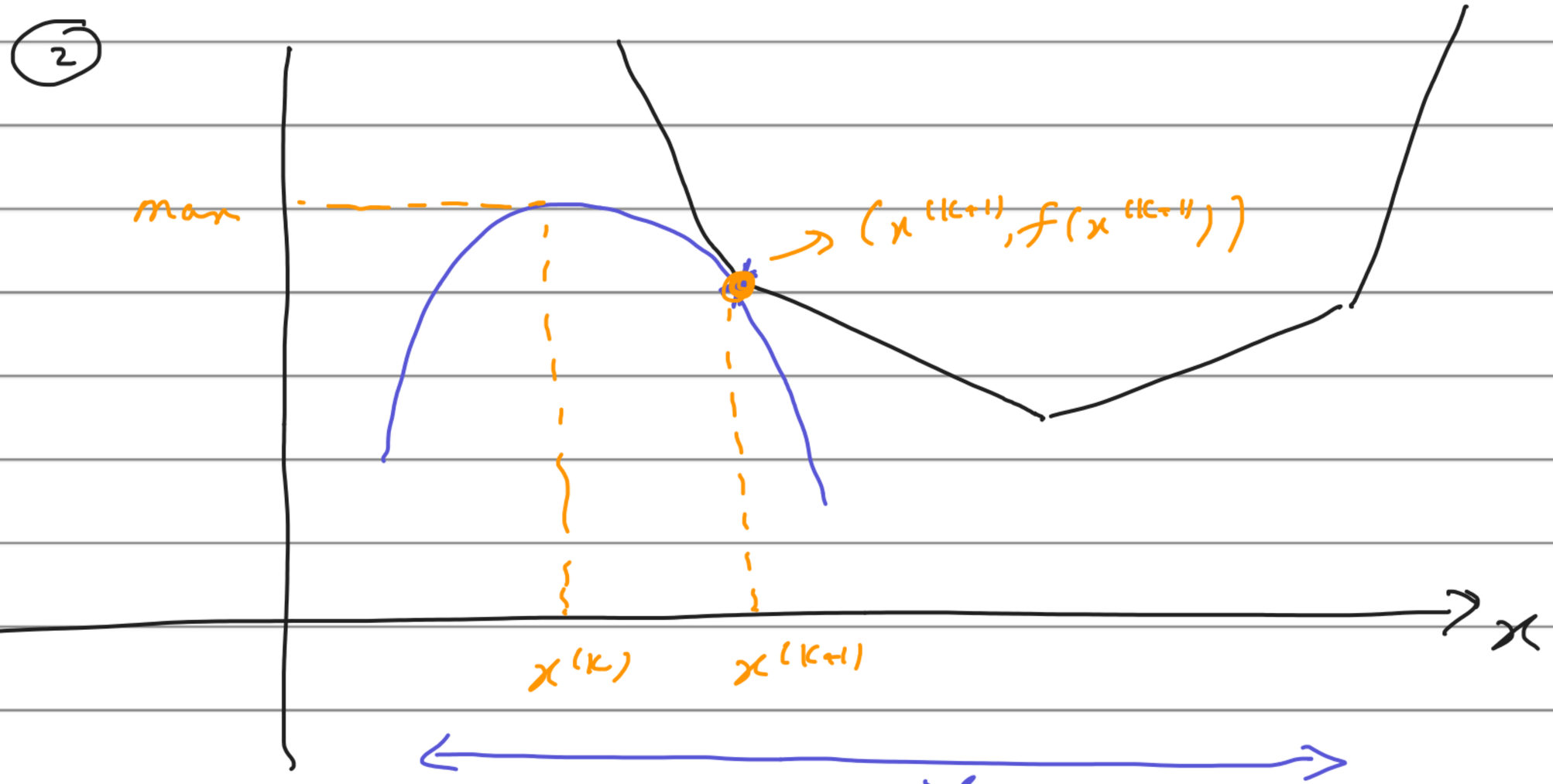
2- $\gamma - \frac{1}{2\alpha^{(k)}} \|x - x^{(k)}\|^2$ over X

$\Rightarrow \gamma^{(k)}$ is smallest γ s.t. the two functions (1) and (2) intersects.



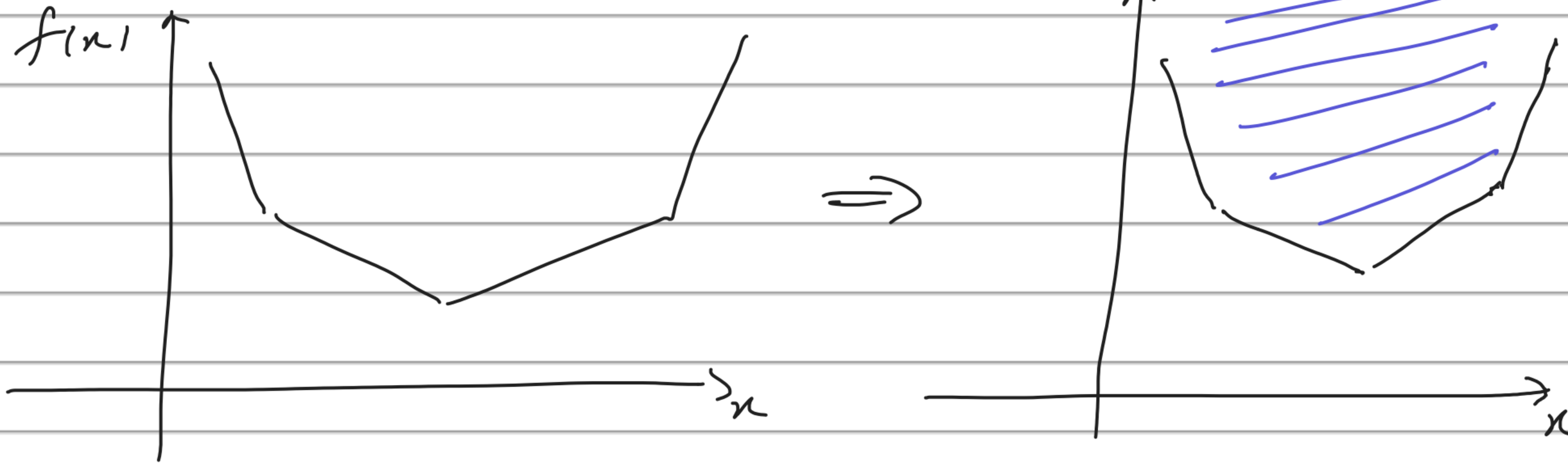
$\uparrow \gamma$ increases





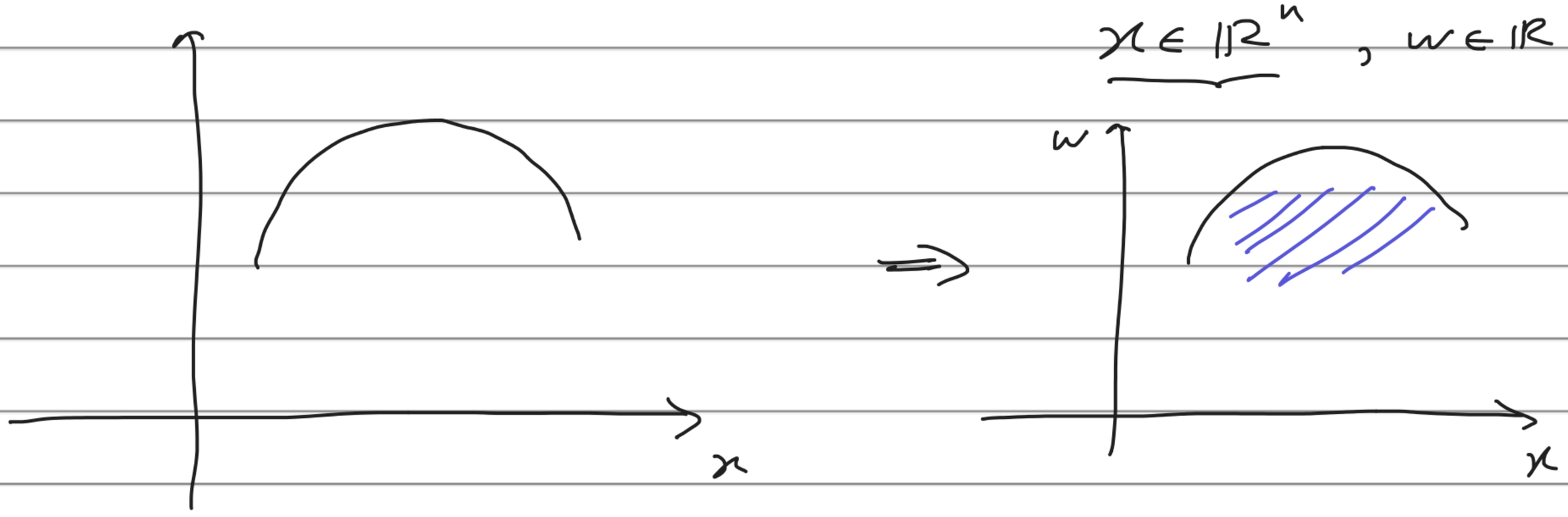
Define :

$$C_1 = \{ (x, w) \mid f(x) \leq w \,,\; x \in X \,,\; w \in \mathbb{R} \}$$



$$C_2 = \{ (x, w) \mid w \leq \gamma^{(k)} - \frac{1}{2\alpha^{(k)}} \| x - x^{(k)} \|^2 \,,$$

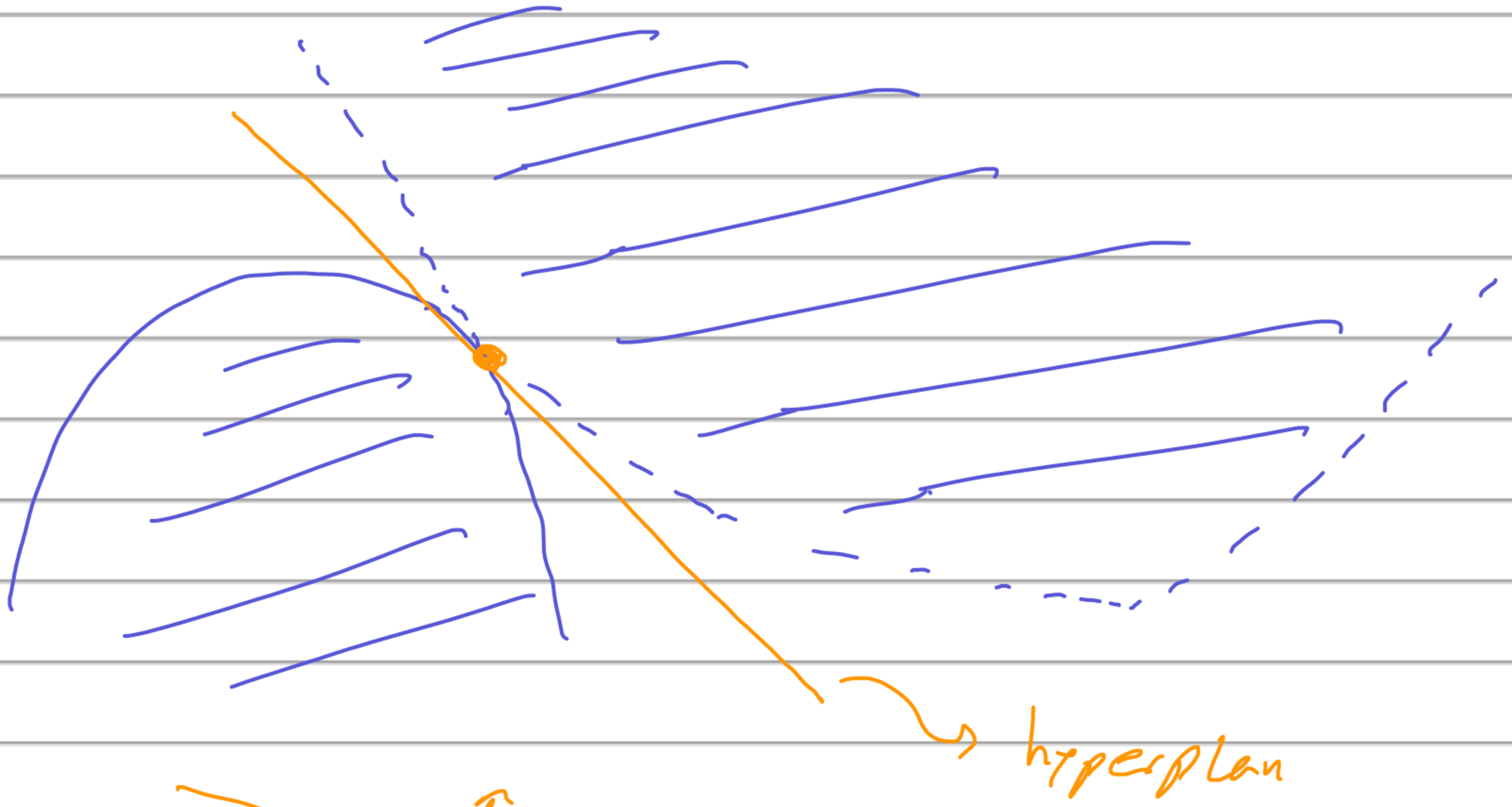
$$\underline{x \in \mathbb{R}^n} \,,\; w \in \mathbb{R} \}$$



$C_1 \cap C_2 = \emptyset$, $C_1, C_2 =$ Conven $\Longrightarrow$

(C_1 : strict inequality)

There is a separating hyperplane:

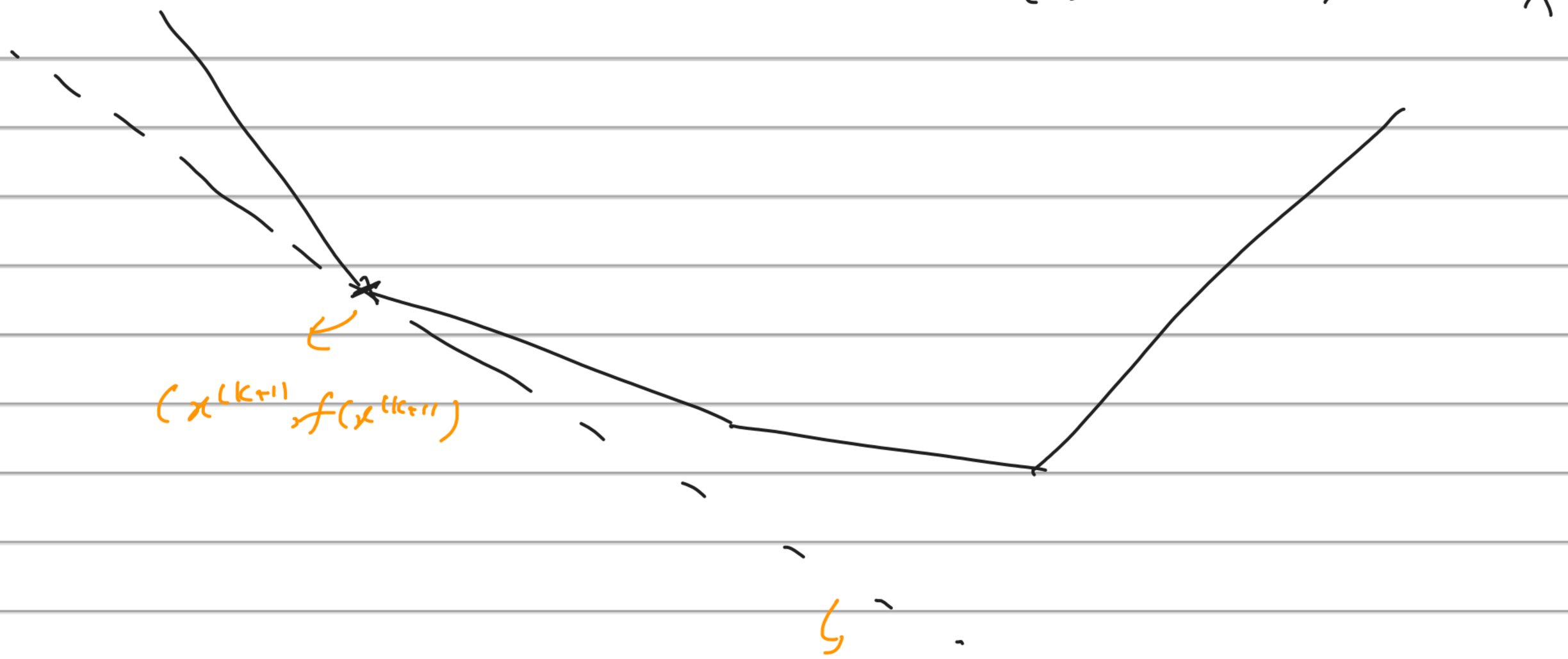


normal vector :

$$\nabla\left(\gamma^{(k)} - \frac{1}{2\alpha^{(k)}}\|x - x^{(k)}\|^2\right)\Big|_{x=x^{(k+1)}} = \frac{x^{(k)} - x^{(k+1)}}{\alpha^{(k)}}$$

since C_1 is above the hyperplane, so is its boundary, $f(x)$.

Thm: $f(x) \geq f(x^{(k+1)}) + \frac{1}{\alpha^{(k)}} (x^{(k)} - x^{(k+1)})^T (x - x^{(k+1)}) \quad \forall x \in X$



Define: $f_* = \inf_{x \in X} f(x)$ → note: f_* could be $-\infty$

$X_* =$ set of minima of $f(x)$ over X

→ note: X_* might be empty.

Thm: Consider proximal algorithm &

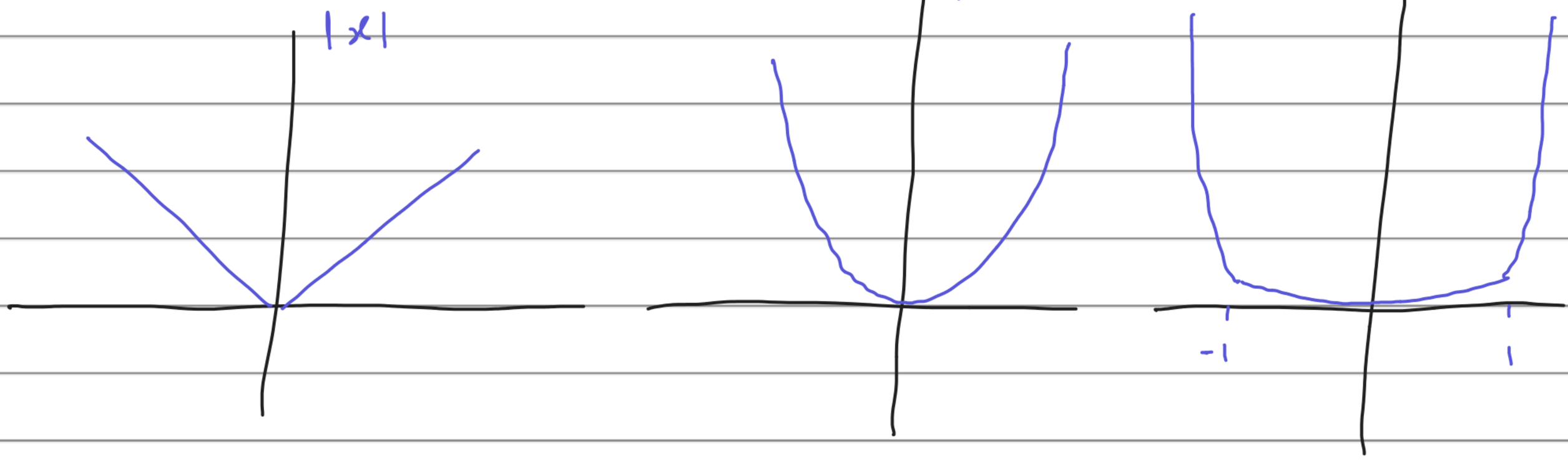
pick $\alpha^{(k)}$ s.t. $\sum_{k=0}^{\infty} \alpha^{(k)} = \infty$, then

- $f(x^{(k)}) \longrightarrow f_*$

- If $X_* \neq$ empty, then $\{x^{(k)}\}$ converges to a unique point in X_*

Rate of Convergence:

Example:



fast growth around origin

moderate growth around origin

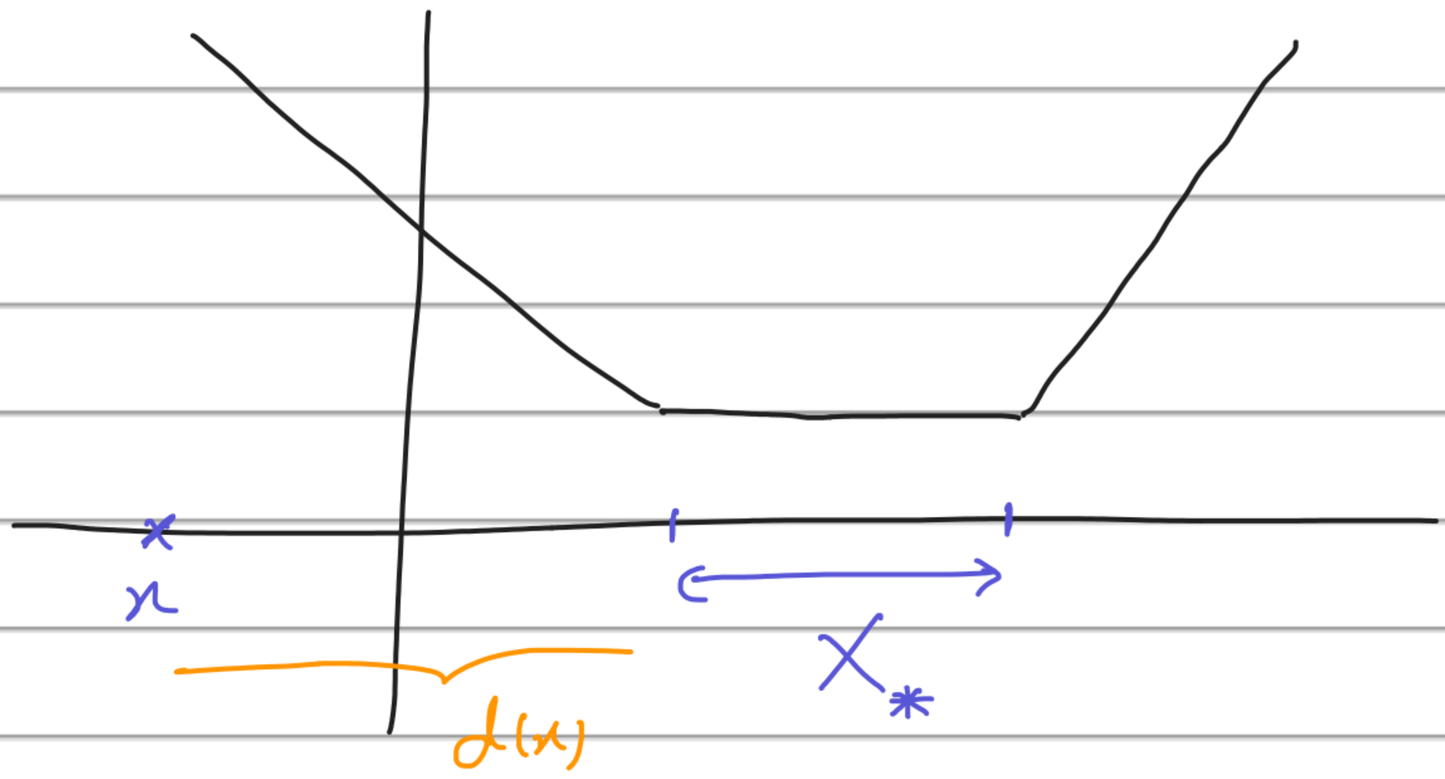
slow growth

($x = \varepsilon \rightarrow |x| = \varepsilon$ or ε^2 or ε^{100})

slow growth $\longrightarrow$ slow convergence

Distance function: $d(x) = \inf_{x_* \in X_*} \|x - x_*\|$

projection on X_* and finding the distance



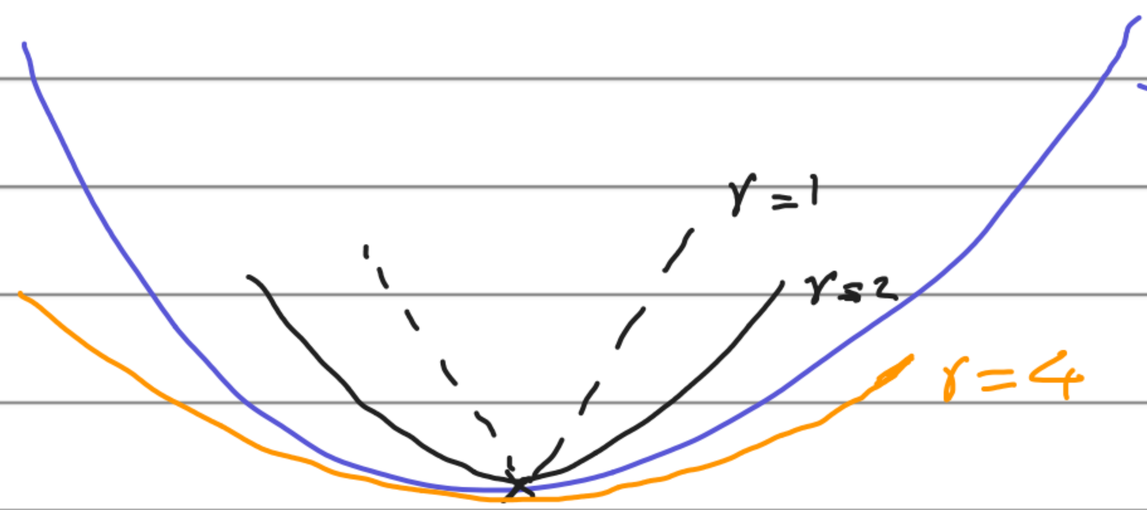
Assume X_* = non_empty $\in$ $\exists \beta > 0$, $\delta > 0$

, $\gamma \geq 1$, s.t.

$f_* + \beta (d(x))^\gamma \leq f(x) \quad \forall x \in X$ s.t. $d(x) \leq \delta$

Example: X_* = single point x_* in $\mathbb{R}$

$\Rightarrow \beta |x - x_*|^\gamma \leq f(x) - f_* \rightarrow$ Local region



$\Rightarrow \gamma$ = rate of growth

Thm: Assume $\sum_{k=0}^{\infty} \alpha^{(k)} = \infty$

$$\Rightarrow d(x^{(k+1)}) + \beta \alpha^{(k)} \left(d(x^{(k+1)})\right)^{\gamma - 1} \leq d(x^{(k)})$$

for all large values of k if $x^{(k+1)} \notin X_*$

---

left side : $d(x^{(k+1)})$

right side : $d(x^{(k)})$

Case $\underline{1}$: $1 < \gamma < 2$ and $x^{(k)} \notin X_*$ $\forall k$

$$d(x^{(k)}) \geq \beta \alpha^{(k)} d(x^{(k+1)})^{\gamma-1} + d(x^{(k+1)})$$

$$\geq \beta \alpha^{(k)} d(x^{(k+1)})^{\gamma-1}$$

$$\Rightarrow \text{Limsup} \left(\frac{d(x^{(k+1)})}{d(x^{(k)})^{\frac{1}{\gamma-1}}} \right) < \infty$$

if $\alpha^{(k)} \geq$ a positive Constant $\forall k$ $\Rightarrow$ Super Linear Convergence

Ex: $\gamma = 1.5 \Rightarrow \limsup_{k \to \infty} \frac{d(x^{(k+1)})}{d(x^{(k)})^2} < \infty$

$\Rightarrow$ Quadratic Convergence

Case 2: $\gamma = 2$ and $x^{(k)} \notin X_*$ $\forall k$

$\Rightarrow d(x^{(k)}) \geq \beta \alpha^{(k)} d(x^{(k+1)})^{2-1} + d(x^{(k+1)})$

$\Rightarrow \frac{d(x^{(k+1)})}{d(x^{(k)})} \leq \frac{1}{\beta \alpha^{(k)} + 1}$

Scenario (a): $\alpha^{(k)} = \text{Constant} \Rightarrow$ Linear Convergence

Scenario (b): $\alpha^{(k)} \to \infty$

$\Rightarrow \limsup_{k \to \infty} \frac{d(x^{(k+1)})}{d(x^{(k)})} = 0 \rightarrow$ superLinear Convergence

Case 3: $\gamma > 2 \Rightarrow \limsup_{k \to \infty} \frac{d(x^{(k+1)})}{d(x^{(k)})^{2/\gamma}} < \infty$

(no assumption on $x^{(k)} \notin X_*$) $\Rightarrow$ sublinear Convergence

(easy to show: if $x^{(k)} \in X_* \Rightarrow x^{(k)} = x^{(k+1)} = x^{(k+2)} = \cdots$)

Case 4: $\gamma = 1$ and $\alpha^{(0)} = \text{Large}$

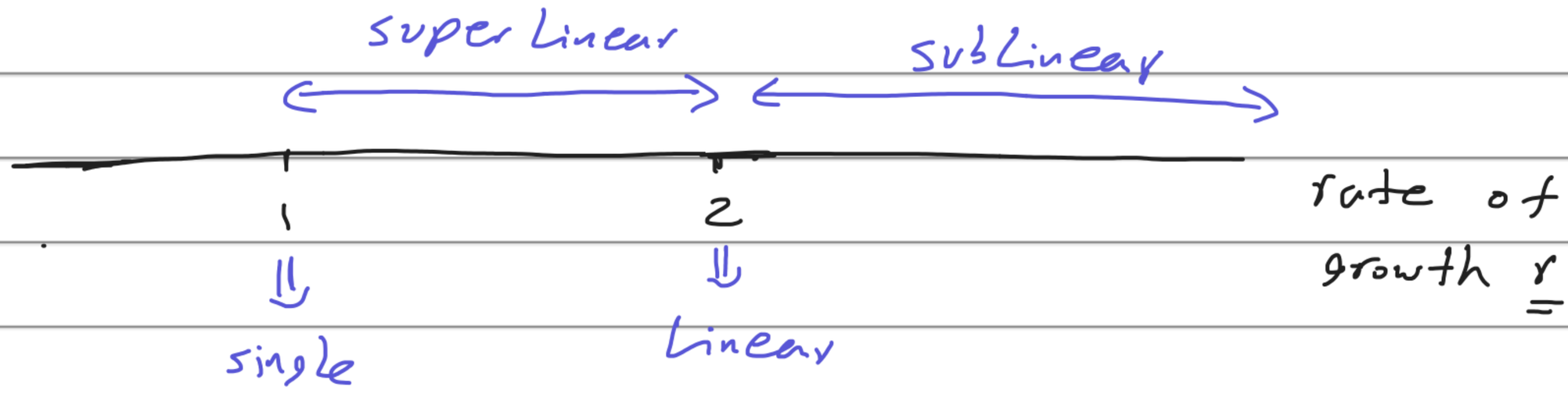
$$d(x^{(0)}) \geq \beta \alpha^{(0)} \underbrace{d(x^{(1)})^{1-1}}_{=1} + \underbrace{d(x^{(1)})}_{\geq 0}$$

$\Rightarrow d(x^{(0)}) \geq \beta \alpha^{(0)}$ but if $\alpha^{(0)}$ is Large, this can't be satisfied

$\Rightarrow x^{(1)} \in X_*$

$\Rightarrow$ Convergence in one iteration

$\Rightarrow$



Look at $f(x) + \frac{1}{2}\|x - z\|^2$

Proximal at rate 2

Dominates the shap

$\Rightarrow$ growth rate ≤ 2 is acceptable.

Proximal gradient algorithm:

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

$$\text{s.t.} \quad x \in X$$

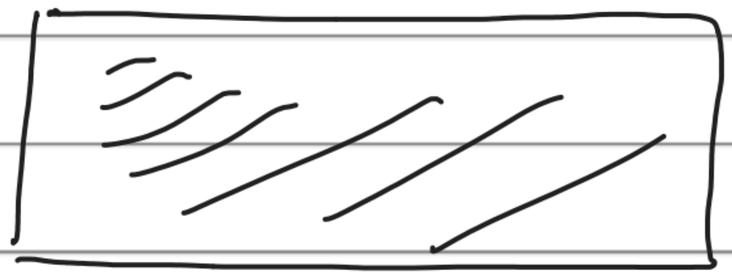
f: convex & differentiable

h: convex & possibly non-smooth

Ex: $\min \; f(x) + \lambda \|x\|_1$ → sparsity promoting

s.t. $x \in X$

Ex: Lasso $\min_x \; \|Ax - b\|^2 + \lambda \|x\|_1$

A: 

fat matrix

b: measurement

A: model of system

x: unknown state

$Ax = b \Rightarrow$ infinitely many solutions

$\Rightarrow$ find sparsest solution (Compressed sensing)

Recall projected gradient method for $\min_{x \in X} f(x)$:

$$x^{(k+1)} = \arg\min_{x \in X} \left\{ \nabla f(x^{(k)})^T (x - x^{(k)}) + \frac{1}{2\alpha^{(k)}} \|x - x^{(k)}\|^2 \right\}$$

what if we modify it for $\min_{x \in X} f(x) + h(x)$:

$\nabla h(x)$ doesn't exist $\Rightarrow$ use $h(x)$ directly:

$$x^{(k+1)} = \arg\min_{x \in X} \left\{ \underbrace{\nabla f(x^{(k)})^T (x - x^{(k)})}_{\text{Linear approximation of } f(x)} + \underbrace{\frac{1}{2\alpha^{(k)}} \|x - x^{(k)}\|^2}_{\text{Proximal}} + \underbrace{h(x)}_{h(\cdot) \text{ directly}} \right\}$$

$$= \arg\min_{x \in X} \left(h(x) + \frac{1}{2\alpha^{(k)}} \|x - \underbrace{(x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)}))}_{z^{(k)}}\|^2 \right)$$

$\Rightarrow$ proximal gradient algorithm:

$$\begin{cases} z^{(k)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)}) & \rightarrow \text{gradient on } f(\cdot) \\ x^{(k+1)} = \text{prox}_{\alpha^{(k)}, h}(z^{(k)}) & \rightarrow \text{proximal on } h(\cdot) \text{ and } X \end{cases}$$