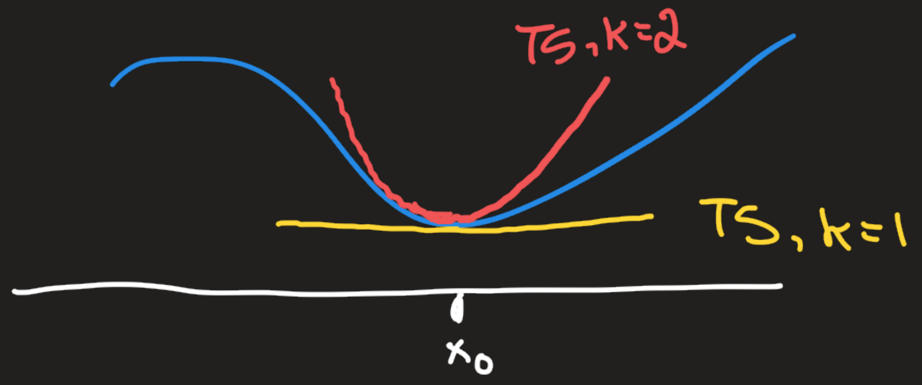
ML and Optimization Lecture 9

- Project details (groups due/assigned today)
- 2nd order characterization of convexity
- Jensen's ineq
- Operations that preserve convexity
- Examples of convex optim probs

Taylor Series expansions



$f: \mathbb{R} \to \mathbb{R}$ smooth

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2} f''(x_0)(x-x_0)^2$$
$$+ \cdots + \frac{1}{k!} f^{(k)}(x_0)(x-x_0)^k + \frac{1}{(k+1)!} f^{(k+1)}(z)(x-x_0)^{k+1}$$

where $z \in [x, x_0]$

if $f^{(k+1)}(z)$ is bounded (on whatever interval we care about)

i.e. $\exists L$ s.t. $|f^{(k+1)}(z)| \leq L$, then

$$f(x) = f(x_0) + \sum_{i=1}^{k} \frac{f^{(i)}(x_0)}{i!}(x-x_0)^i + o(|x-x_0|^k)$$

T.S. expansion of order k

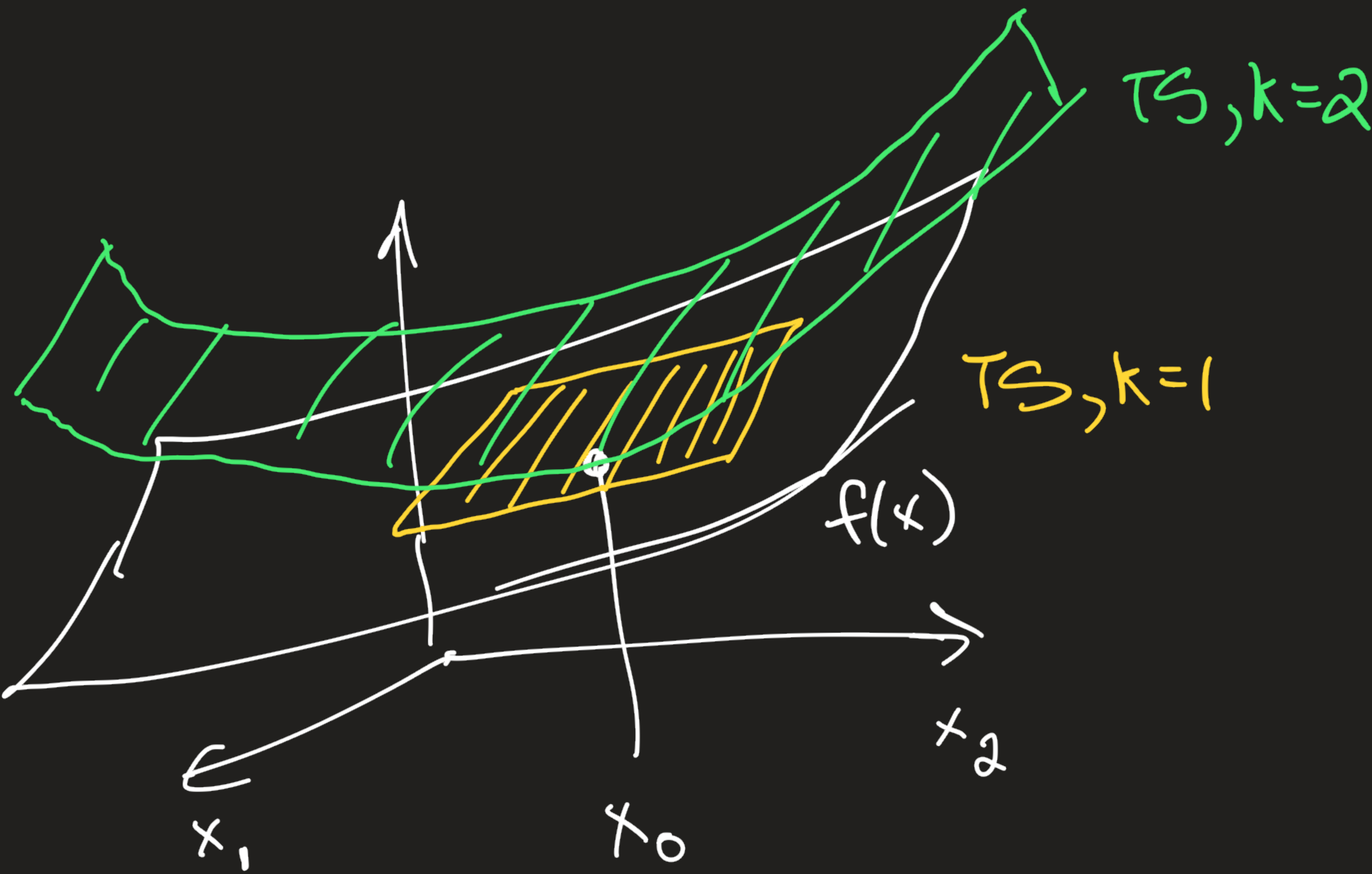
$f: \mathbb{R}^d \to \mathbb{R}$ then $\overset{\text{"gradient"}}{\nabla f(x)} \in \mathbb{R}^d$ and $[\nabla f(x)]_i = \frac{\partial f(x)}{\partial x_i}$

$$f(x) = f(x_0) + \underset{\sum_{i=1}^{d} \nabla f(x_0)_i (x-x_0)_i}{\underset{\shortparallel}{\langle \nabla f(x_0), x - x_0 \rangle}} + \frac{1}{2}(x-x_0)^T \nabla^2 f(x_0)(x-x_0) + o(\|x-x_0\|_2^2)$$

$\underset{\text{"Hessian"}}{\nabla^2 f(x_0)} \in \mathbb{R}^{d \times d}$ and $[\nabla^2 f(x_0)]_{ij} = \frac{\partial^2 f(x_0)}{\partial x_i \partial x_j}$

so fact: $x^T A x = \sum_{i=1}^{d} x_i x_j A_{ij}$

$$(x-x_0)^T \nabla^2 f(x_0)(x-x_0) = \sum_{i=i}^{d} (x-x_0)_i (x-x_0)_j \frac{\partial^2 f(x_0)}{\partial x_i \partial x_j}$$


TS, k=2
TS, k=1
f(x)
x₂
x₁
x₀

2^{nd} order characterization of convexity

A twice differentiable function f defined on a convex set $C \subseteq \mathbb{R}^d$ is convex iff at each point in C,

$$\nabla^2 f(x) \succcurlyeq 0$$

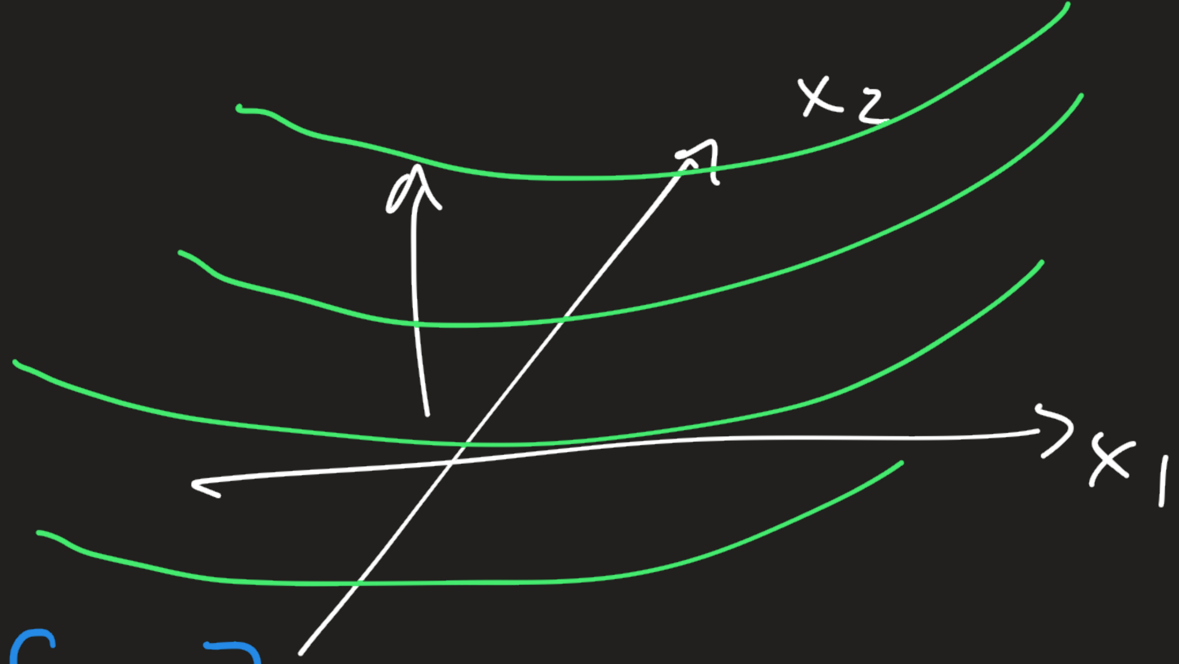
symmetric

Recall: $M \in \mathbb{R}^{d \times d}$ is positive semidefinite/PSD $(M \succcurlyeq 0)$

1) for every vector $x \in \mathbb{R}^d$, $x^T M x \geq 0$
2) $M = CC^T$ for some matrix C
3) M has eigenvalues all nonnegative

Ex:

$$f(x) = f(x_1, x_2) = x_1^2$$



$$\nabla f(x) = \begin{bmatrix} \dfrac{\partial f(x_1,x_2)}{\partial x_1} \\ \dfrac{\partial f(x_1,x_2)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 0 \end{bmatrix}$$

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x_1,x_2)}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\left[\frac{\partial f(x_1,x_2)}{\partial x_j}\right] = \frac{\partial}{\partial x_i}\left[\nabla f(x)_j\right]$$

$$\Rightarrow \nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

is PSD because: (i) eigenvalues are all nonnegative, (ii) its square root is $\begin{bmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix}$, (iii) for any z, $z^T \nabla^2 f(x) z = 2z_1^2 \geq 0$

Ex of a matrix whose entries are all nonnegative, but it is not PSD

$$x^T \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} x = x_1^2 + x_2^2 + 4x_1x_2$$

take $x_1 = 1$, $x_2 = -1$ then $= 2 - 4 = -2$

so $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ is not psd

Jensen's Inequality

Recall
$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 \geq 0$$

replace $f(x) = x^2$ with any convex function

$$\boxed{\mathbb{E}f(X) \geq f(\mathbb{E}X)}$$

Imagine that X took on only two values

$$X = \begin{cases} a & p \\ b & (1-p) \end{cases}$$

b/c f is cvx

$$f(pa + (1-p)b) \leq pf(a) + (1-p)f(b)$$

so

$$f(\mathbb{E}X) \leq \mathbb{E}f(X)$$

Prf (assume f is differentiable)

Because f is differentiable,

$$f(X) \geq f(\mathbb{E}X) + \langle \nabla f(\mathbb{E}X), X - \mathbb{E}X \rangle$$

by first-order condition.

It follows by monotonicity of expectation that

$$\mathbb{E}f(X) \geq f(\mathbb{E}X) + \mathbb{E}\langle \nabla f(\mathbb{E}X), X - \mathbb{E}X \rangle$$

and by linearity of expectation

$$= f(\mathbb{E}X) + \langle \nabla f(\mathbb{E}X), \mathbb{E}(X - \mathbb{E}X) \rangle$$

$$= f(\mathbb{E}X)$$

$\blacksquare$

Ex using 2^{nd}-order characterization of convexity

$$f(x) = \text{logsumexp}(x) = \log\left(\sum_{i=1}^{d} e^{x_i}\right)$$

Two steps:

1) compute $\nabla^2 f(x)$
2) argue $\nabla^2 f(x) \succcurlyeq 0$ at each x

Step 1

$$[\nabla f(x)]_i = \frac{\partial f}{\partial x_i}(x) = \frac{1}{\sum_{i=1}^{d} e^{x_i}} e^{x_i} = \frac{e^{x_i}}{1^T e^x}$$

$$\nabla f(x) = \text{softmax}(x)$$

$$\left[\nabla^2 f(x)\right]_{ij} = \frac{\partial f(x)}{\partial x_j \partial x_i} = \frac{\partial}{\partial x_j}\left[\nabla f(x)_i\right]$$

$$= \frac{\partial}{\partial x_j}\left[\frac{\exp(x_i)}{1^T \exp(x)}\right]$$

$$= \begin{cases} \dfrac{\partial}{\partial x_j}\left[\dfrac{\exp(x_j)}{1^T \exp(x)}\right] & j = i \\[2ex] \dfrac{\partial}{\partial x_j}\left[\dfrac{\exp(x_i)}{1^T \exp(x)}\right] & j \neq i \end{cases}$$

$$= \begin{cases} \dfrac{\exp(x_j)\left(1^T \exp(x)\right) - \exp(x_j)^2}{\left(1^T \exp(x)\right)^2} & j = i \\[2ex] -\dfrac{\exp(x_j)\exp(x_i)}{\left(1^T \exp(x)\right)^2} & j \neq i \end{cases}$$

$$\nabla^2 f(x) = \mathrm{Diag}\left(\frac{\exp(x)}{1^T\exp(x)}\right) - \left[\frac{\exp(x)}{1^T\exp(x)}\right]\left[\frac{\exp(x)}{1^T\exp(x)}\right]^T$$

$$= \mathrm{Diag}(\mathrm{softmax}(x)) - \mathrm{softmax}(x)\cdot\mathrm{softmax}(x)^T$$

Step 2 (Showing $\nabla^2 f(x) \succeq 0$)

Fix a vector $z \in \mathbb{R}^d$

Consider $z^T \nabla^2 f(x) z = z^T \mathrm{Diag}(\theta) z - \underbrace{z^T\theta\theta^T z}_{(\theta^T z)^2}$

where for convenience $\theta \equiv \mathrm{softmax}(x)$

$$= \sum_{i=1}^{d} \theta_i z_i^2 - \left(\sum_{i=1}^{d} \theta_i z_i\right)^2$$

≥ 0 b/c θ is a probability vector

Ex of convex functions & operations preserving convexity

- $f(x) = Ax + b$ (affine functions) are convex

  Check

- $f(x) = \frac{1}{2} x^T A x + b^T x + c$ is convex

  when $A \succcurlyeq 0$ (b/c $\nabla^2 f(x) = A$)

- If $f_1, \dots, f_n$ are convex and $\alpha_1, \dots, \alpha_n \geq 0$ then

$$f = \sum \alpha_i f_i \quad \text{is convex}$$

Check (from def of convexity)

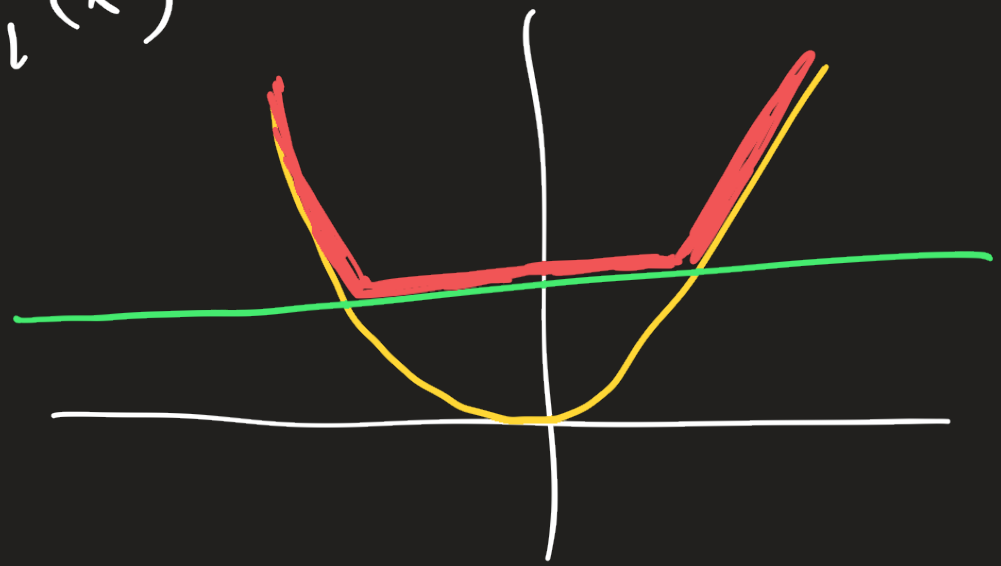
- If f is convex then $f(Ax+b)$ is convex

Check

- The pointwise maximum of convex functions is convex

$$f(x) = \max_i f_i(x)$$

Check



- If f is a convex & non-decreasing function ($x \geq y$ then $f(x) \geq f(y)$) and g is a convex function then $f \circ g$, defined pointwise as

$$(f \circ g)(x) = f(g(x)),$$

is convex

Ex: $f(x) = x^2$ is nondecreasing & convex on $\mathbb{R}_+$

so if g is convex into $\mathbb{R}_+$, then g^2 is convex on $\mathbb{R}_+$

<u>Cor</u>: square of any norm is convex

<u>Cor</u> square of any nonnegative convex function is convex

Examples of convex optim probs

1) Linear programming

$$\min \; c^T x$$ ← linear function

s.t.

$$Ax \leq b$$ ← intersection of half spaces

2) Least-squares

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2$$

3) Non-negative least squares

$$\min_{x \geq 0} \|Ax - b\|_2^2$$

4) Regularized least squares

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda \|x\|_1$$

5) Norm-constrained least squares

$$\min_{\|x\|_1 \leq \tau} \|Ax - b\|_2^2$$