For HW2:

You compute a confusion matrix $C \in \mathbb{R}^{k \times k}$

C_{ij} = # of objects <u>actually</u> in class i that are classified by the model as class j

Ideal confusion matrix

Bad confusion matrix:

- many objects in class i were misclassified
- many objects not of class j were classified as class j

Convex Optimization

We've seen how to convert ML problems to optim probs using RERM framework.

Question: how do we solve these?

Optimization of general fcns over general sets is intractable.

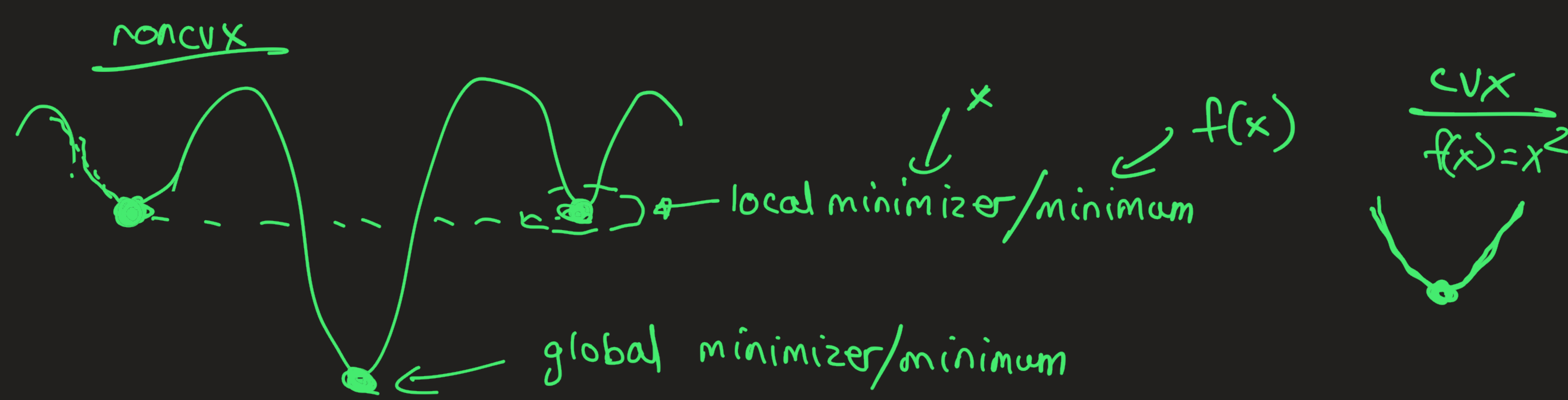
convex

We restrict ourselves first to the class of cvx optim problems:

$$x^* = \underset{x \in C}{\operatorname{argmin}}\ f(x)$$

f – cvx function

C – cvx set



Why do we restrict to cvx optim probs?

- Very expressive class
- They are <u>generally tractable</u>
- Many classical ML problems are cvx:

  in particular if ℓ is cvx in its first argument Ω is cvx, and $\mathcal{F}$ is cvx, then

$$\hat{f}_{\mathcal{F}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \Omega(f)$$

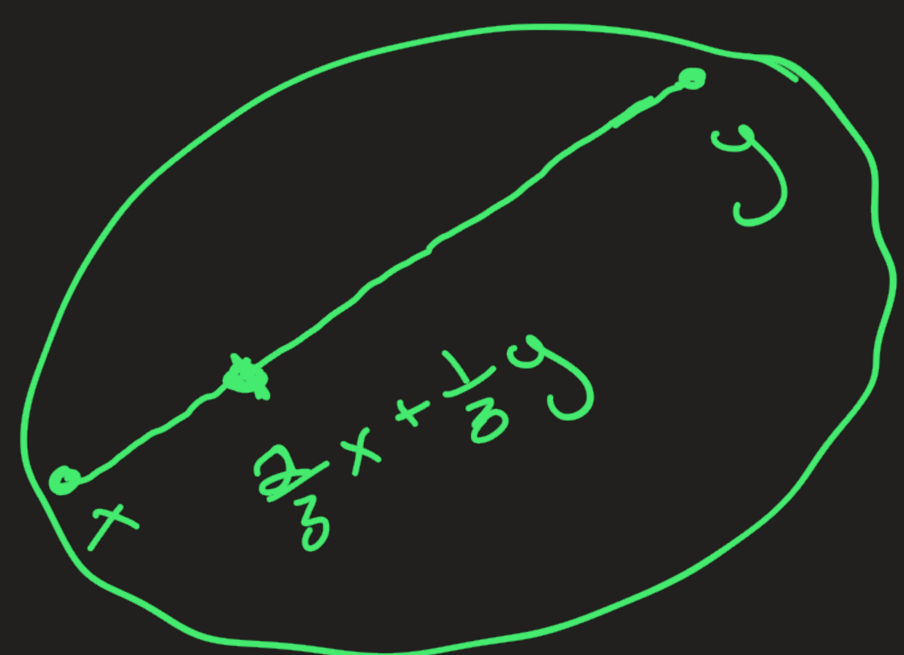
  is a cvx optim problem.

- efficient algorithms exist, tailored for different settings

- local minima <u>are</u> global minima

Convexity

We say a set C is convex if whenever $x, y \in C$, the line segment $[x, y] = \{\underbrace{\alpha x + (1-\alpha) y}_{\text{convex combination of } x \,\&\, y} : \alpha \in [0,1]\}$ is in C.

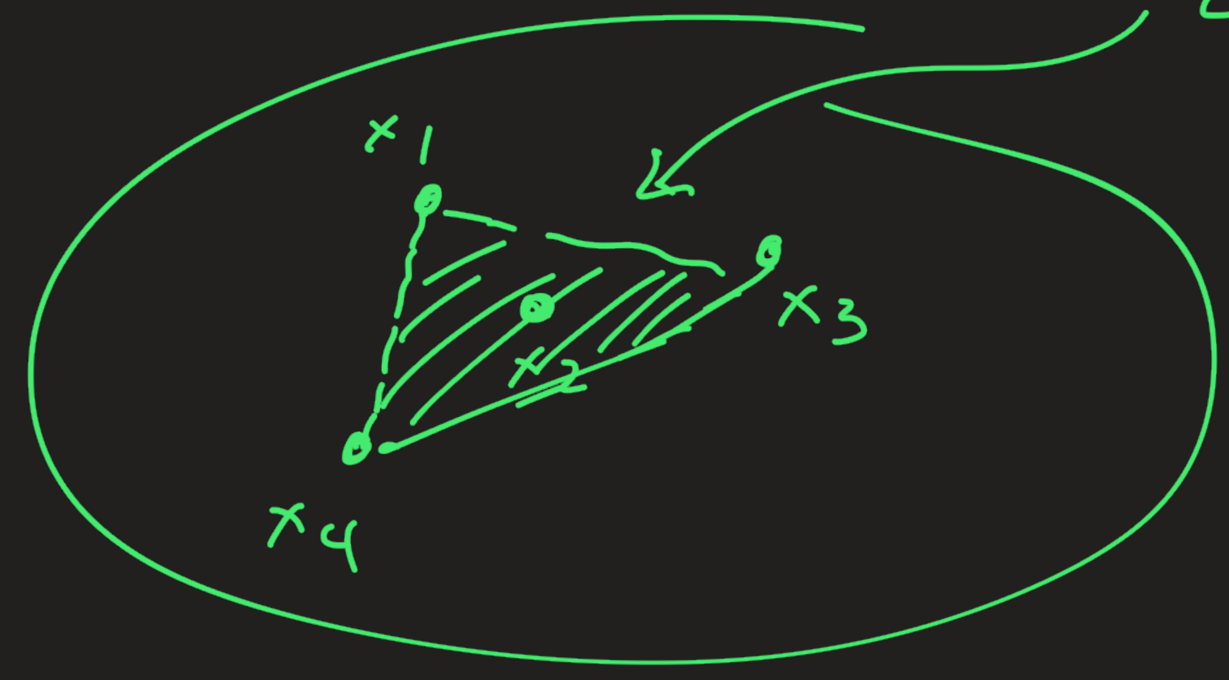


cvx set

noncvx

We can show that C is convex iff for all $x_1, \dots, x_k \in C$ and any numbers $\theta_1, \dots, \theta_k$ s.t. $\theta_i \geq 0$ and $\sum \theta_i = 1$ the convex combination

$$\theta_1 x_1 + \cdots + \theta_k x_k \in C$$



convex combinations of x_1, x_2, x_3, x_4

Interpretation/Fact

If C is a convex set and X is a r.v. taking values in C, then $\mathbb{E}X \in C$

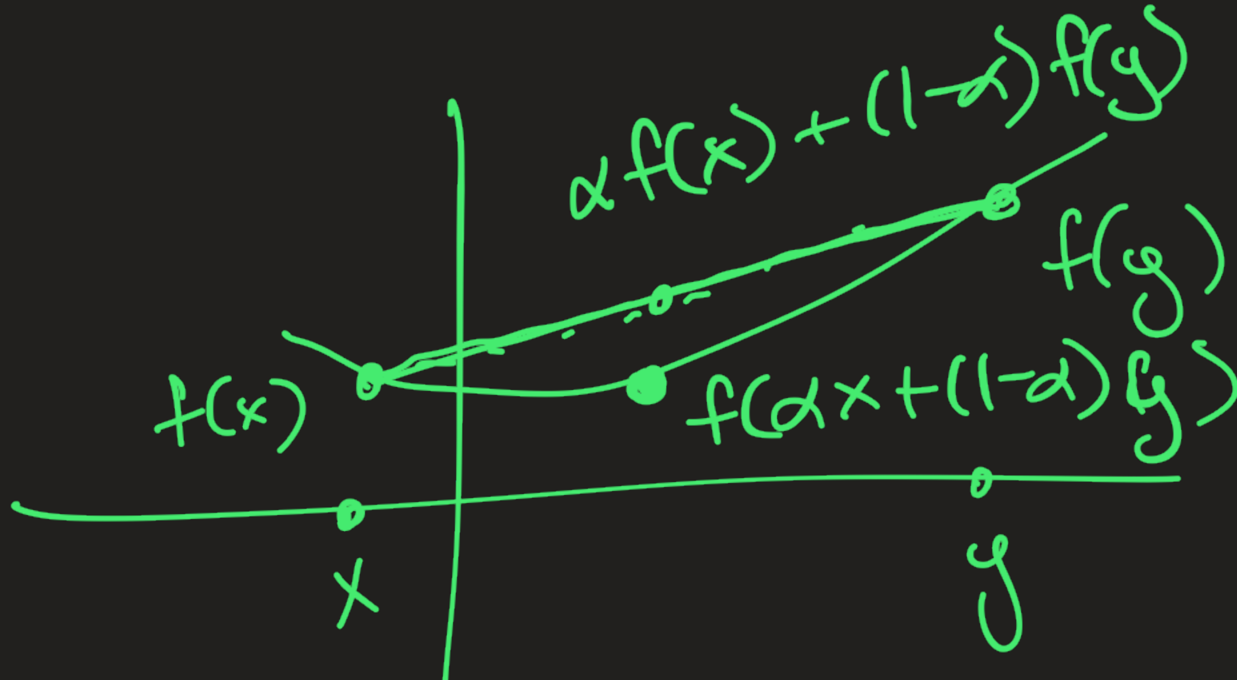
Convex functions

A function f defined on a convex set C

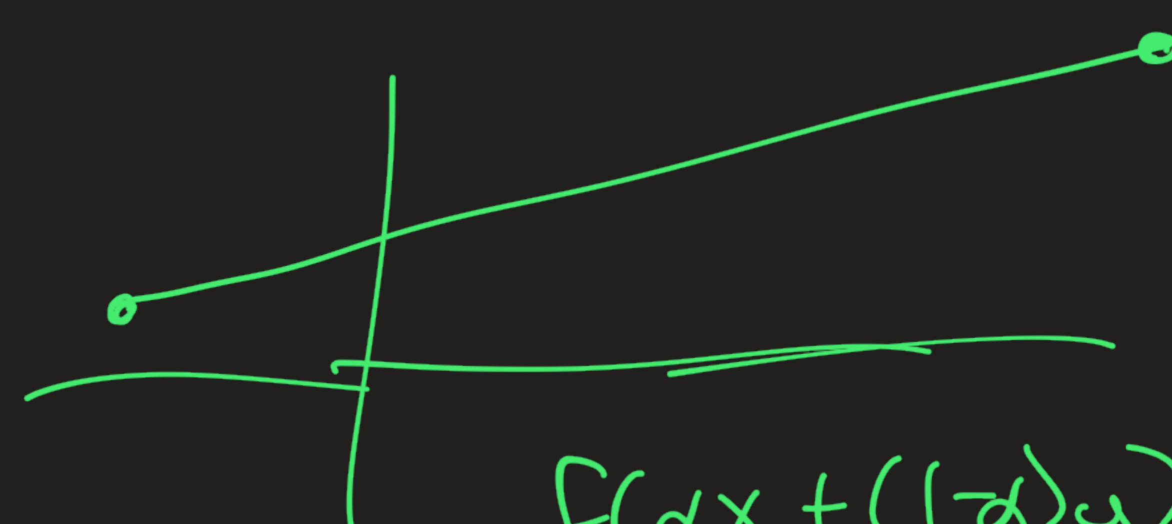
is called convex iff it satisfies

$$f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$$

for all $\alpha \in [0,1]$ and $x, y \in C$



The convex function f lies under all of its chords



linear functions are cvx, b/c

$$f(\alpha x + (1-\alpha) y) = \alpha f(x) + (1-\alpha) f(y)$$

Strict convexity: $f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y)$

strict inequality

Very important consequences of convexity

1) Local minima of convex fcns. are global minima: implies we can design our algs to just locally minimize, and they will reach global minima.

Prf (By contradiction) Let f be convex and assume x is a local minimizer (there is a neighborhood of x s.t. $f(x) < f(y)$ for any y in that neighborhood) but it is not a global minimizer, that is, there exists a z such that $f(z) < f(x)$

By convexity, for any $\varepsilon > 0$,

$$y = (1-\varepsilon)x + \varepsilon z$$

that is close to x, we have

$$f(y) \leq (1-\varepsilon) f(x) + \varepsilon f(z)$$

and since $f(z) < f(x)$

$$\Rightarrow \quad f(y) < (1-\varepsilon) f(x) + \varepsilon f(x) = f(x)$$

Since ε can be arbitrarily small, and

$$\|y - x\|_2 = \varepsilon \|z - x\|_2$$

we see we can find a pt y arbitrarily close to x s.t. $f(y) < f(x)$. This contradicts the fact that x is a local minimizer. We conclude that x must be a global minimizer.

- Another fact: if f is <u>strictly</u> cvx then the minimizer (if it exists) is unique.

- Another <u>Jensen's Inequality</u>: how convex functions and random variables interact.

Let f be a cvx fcn on a set C, and let X take values in C, then

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

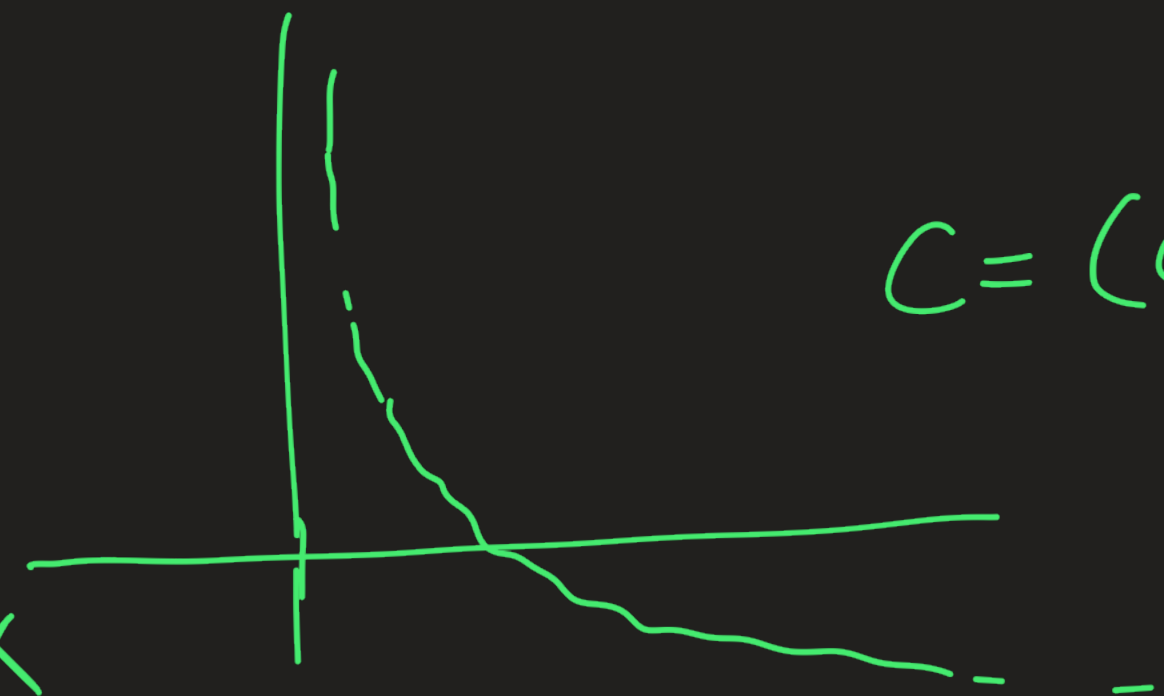
Ex: $f = x^2 \underset{\text{Jensen's}}{\Rightarrow} \mathbb{E}(X^2) \geq (\mathbb{E}X)^2$

(also a conseq of
$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2 \geq 0$)

Ex: $f = -\log x$

Jensen's $\Rightarrow$

$$\mathbb{E} -\log X \geq -\log \mathbb{E} X$$



$C = (0, \infty)$

$f = -\log x$

(Ex: cvx fcn w/ no minimizer)

Ex application of Jensen's ineq

If $\theta \in \mathbb{R}^d$ is a probability vector $\left(\theta_i \geq 0 \;\&\; \sum_{i=1}^{d} \theta_i = 1\right)$ and $v \in \mathbb{R}^d$ is a vector, then

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

Jensen w/ this r.v. and $f = x^2$

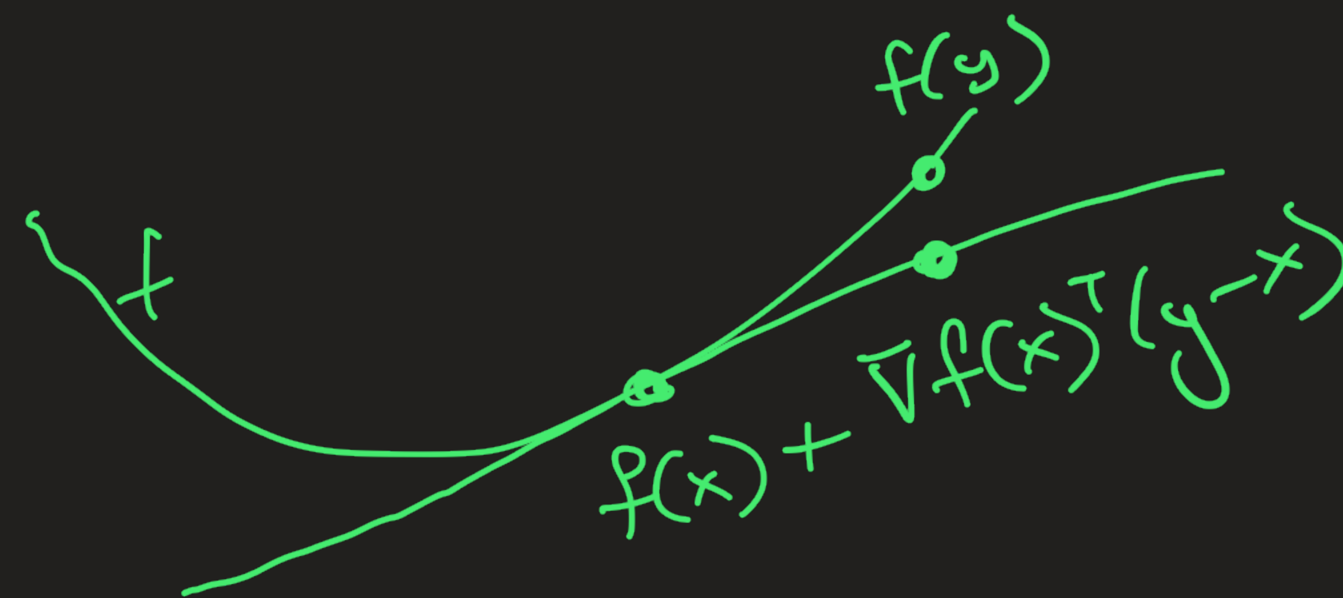
Define r.v. $X = v_i$ w.p. θ_i $\Rightarrow$ $\mathbb{E}X = \sum v_i \theta_i$

$\mathbb{E}X^2 = \sum v_i^2 \theta_i$

First & Second-Order Characterizations of Convexity

1) First-Order: if f is differentiable, then f is convex iff $\forall y, x \in C$

$$f(y) \geq \underbrace{f(x) + \nabla f(x)^T (y-x)}$$



← first-order TS-expansion at x

"f is above its tangent lines"

Ex $f = x^2$ note $\frac{df}{dx} = 2x$

$$f(y) \geq f(x) + \frac{df}{dx}(y-x) \ ?$$

if true then $f(x) = x^2$ is convex on $\mathbb{R}$

$$\iff y^2 \geq x^2 + 2x(y-x)$$

$$\iff y^2 \geq 2xy - x^2$$

$$\iff x^2 + y^2 - 2xy \geq 0$$

$$\iff (x-y)^2 \geq 0 \quad \checkmark$$

conclude $f(x) = x^2$ is convex on $\mathbb{R}$ b/c we showed the first-order condition is satisfied.

2) Second-Order condition: if f is twice-differentiable then f is convex iff

$$\nabla^2 f(x) \succcurlyeq 0$$ (Hessian is positive semi-definite (PSD))

Recall: a matrix $M \in \mathbb{R}^{d \times d}$ is PSD (i.e. $M \succcurlyeq 0$) iff for all vectors $x \in \mathbb{R}^d$,

$$x^T M x \geq 0 \quad \text{and} \quad M = M^T$$

Characterizations of PSD matrices:

1) M is symmetric and all its eigenvalues are nonnegative.

2) $M = CC^T$ for some matrix C

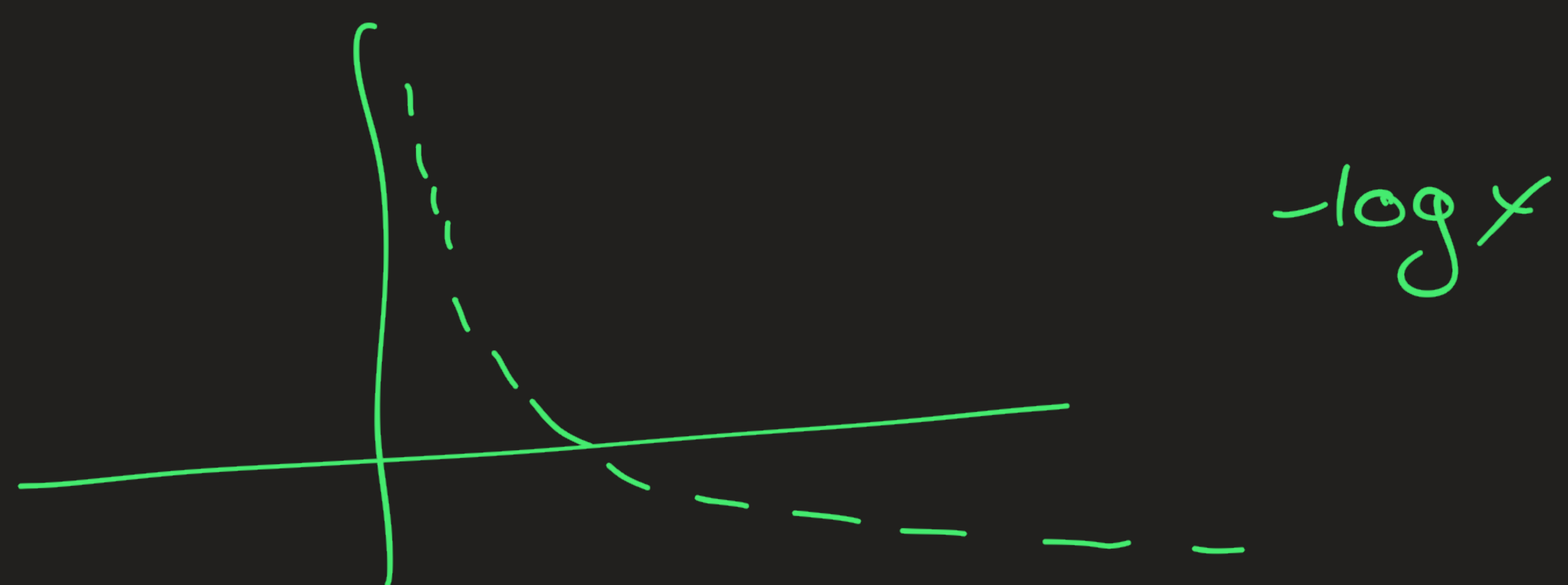
Note: if $\lambda_i \geq 0$ and $M_i \succcurlyeq 0$ then $\sum \lambda_i M_i \succcurlyeq 0$

Ex

$$f(x) = -\log x \Rightarrow \frac{df}{dx} = \frac{-1}{x} \Rightarrow \frac{d^2f}{dx^2} = \frac{1}{x^2}$$

notice $\frac{d^2f}{dx^2} \geq 0$ when $x \in (0, \infty)$

so $-\log x$ is convex on $(0, \infty)$



Ex $f(x) = \text{logsumexp}(x)$ is cvx

$$= \log\left(\sum_{i=1}^{d} \exp(x_i)\right) = \log(\mathbb{1}^T \exp(x))$$

$$[\nabla f(x)]_i = \frac{\partial}{\partial x_i} f(x) = \frac{1}{\mathbb{1}^T \exp(x)} \frac{\partial}{\partial x_i}(\mathbb{1}^T \exp(x))$$

$$= \frac{\exp(x_i)}{\mathbb{1}^T \exp(x)}$$

$$\Rightarrow \nabla f(x) = \frac{\exp(x)}{\mathbb{1}^T \exp(x)} \quad \text{(prob vector!)}$$

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

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

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

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

need to show that $\nabla^2 f(x)$ is PSD on $\mathbb{R}^d$,

i.e. for $v \in \mathbb{R}^d$

$$v^T \nabla^2 f(x) v \geq 0$$

First note that take $\theta = \dfrac{\exp(x)}{1^T\exp(x)} = \nabla f(x)$

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

$$\nabla^2 f(x) = \mathrm{Diag}(\theta) + \theta\theta^T$$

$$\Rightarrow \quad v^T \nabla^2 f(x) v = v^T \mathrm{Diag}(\theta) v + (v^T\theta)(\theta^T v)$$

<u>Fact:</u> if $A \in \mathbb{R}^{d\times d}$ and $v \in \mathbb{R}^d$ then

$$v^T A v = \sum_{i=1}^{d} v_i (Av)_i = \sum_{i=1}^{d} v_i \left(\sum_{j=1}^{d} A_{ij} v_j \right)$$

$$\boxed{v^T A v = \sum_{i=1}^{d} \sum_{j=1}^{d} v_i v_j A_{ij}}$$

$$\Rightarrow \quad v^T \mathrm{Diag}(\theta) v = \sum_i v_i^2 \theta_i$$

and $v^T\theta = \sum v_i \theta_i$

$$\Rightarrow v^T \nabla^2 f(x) v = \sum v_i^2 \theta_i - \left(\sum v_i \theta_i\right)^2$$

recall we showed that

$$\left(\sum v_i^2 \theta_i\right) \geq \left(\sum v_i \theta_i\right)^2$$

by Jensen's in-eq.

$$\Rightarrow v^T \nabla^2 f(x) v \geq 0 \text{ for any } v \in \mathbb{R}^d$$

$$\Rightarrow \nabla^2 f(x) \succeq 0$$

$\Rightarrow f$ is convex on $\mathbb{R}^d$ by second-order convxty cond.