Machine Learning & Optimization Lecture 8

- Optimization as minimization
- Convex Sets, Functions ; Convex Optim Probs
- Jensen's Inequality
- First & Second Order Characterizations of Cvxty
- Examples
- Operations that preserve convexity
- Examples of Cvx Optim Problems

Convex Optimization

- We've seen how to convert ML problems to optim probs using RERM and MLE frameworks

- We also know we only need to ensure

$$R_n(f_T) \le R_n(\hat{f}_{\mathcal{F}}) + O(\varepsilon)$$

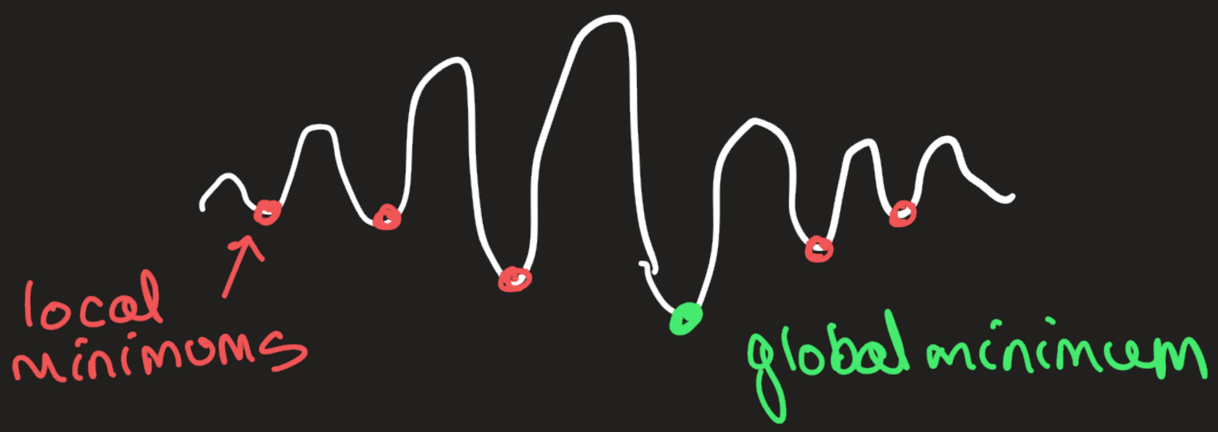
determined by approximate error and generalization error

- In general optimization is NP-Hard so we restrict ourselves to <u>convex optimization problems</u> (for now)

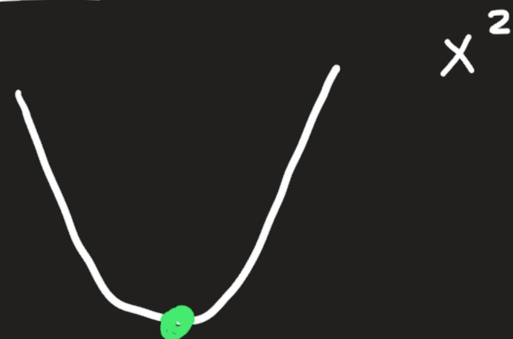
$$x^* = \underset{x \in C}{\operatorname{argmin}} f(x)$$

f – cvx function
C – cvx set

<u>noncvx</u>



<u>cvx</u>



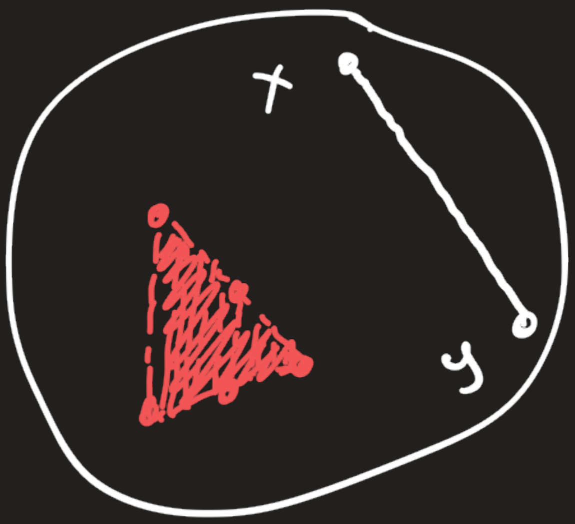
Why restrict to cvx optim probs

- Very expressive model class
- They are generally tractable
- local minima are global minima

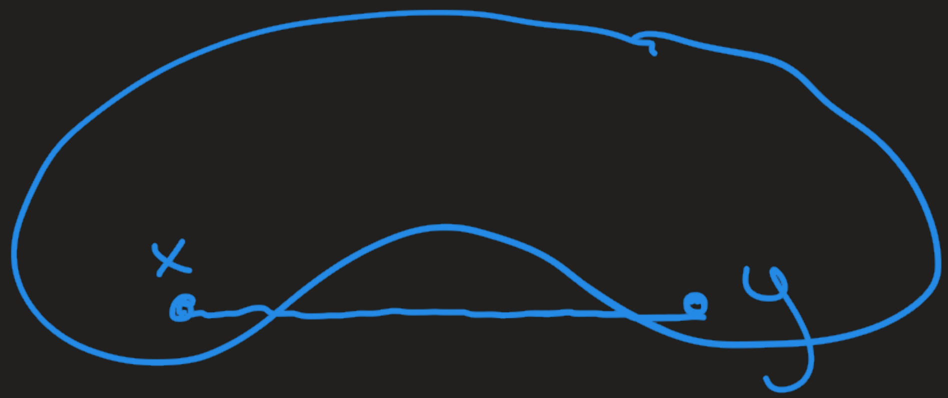
- Many ML problems are cvx:
  - GLMs fit using MLE gives cvx optim probs
  - many efficient algs exist, tailored for different settings

Convexity

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



C – convex



C – nonconvex

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. $\theta_i \geq 0$ for all i and $\sum \theta_i = 1$, the convex combination

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

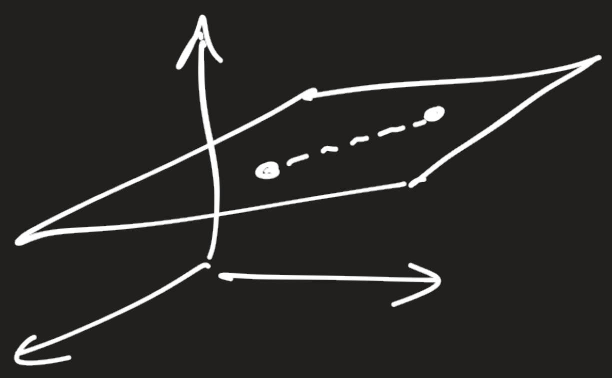
Given $x_1, \dots, x_k$ the set of all convex combinations of these points is called their convex hull

Observe:

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

Examples of convex sets

- empty set
- $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}^d$, rays, lines, line segments
- hyperplanes: $\{x \mid a^T x = b\} = H$



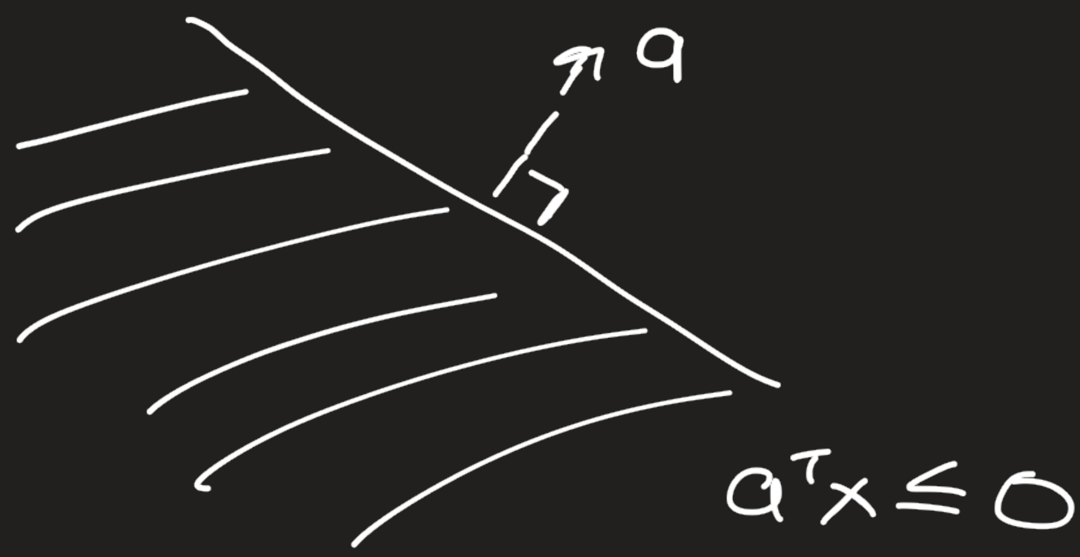
if x & y in this hyperplane, then

$$a^T[\alpha x + (1-\alpha) y] = \alpha a^T x + (1-\alpha) a^T y = \alpha b + (1-\alpha) b = b$$

so $[x, y] \subseteq H$

- half-spaces are convex $\{x \mid a^Tx \le b\} = H_{\le}$

fix $x, y \in H_{\le}$ and $\alpha \in [0,1]$, then



$$a^T[\alpha x + (1-\alpha)y]$$
$$= \alpha a^Tx + (1-\alpha)a^Ty$$
$$\le \alpha b + (1-\alpha)b = b$$

so $[x,y] \in H_{\le}$

- convex hull of an arbitrary set S

$$\mathrm{conv}(S) = \{\theta_1 x_1 + \cdots + \theta_k x_k : x_i \in S, \theta_i \ge 0 \text{ and } \textstyle\sum \theta_i = 1\}$$

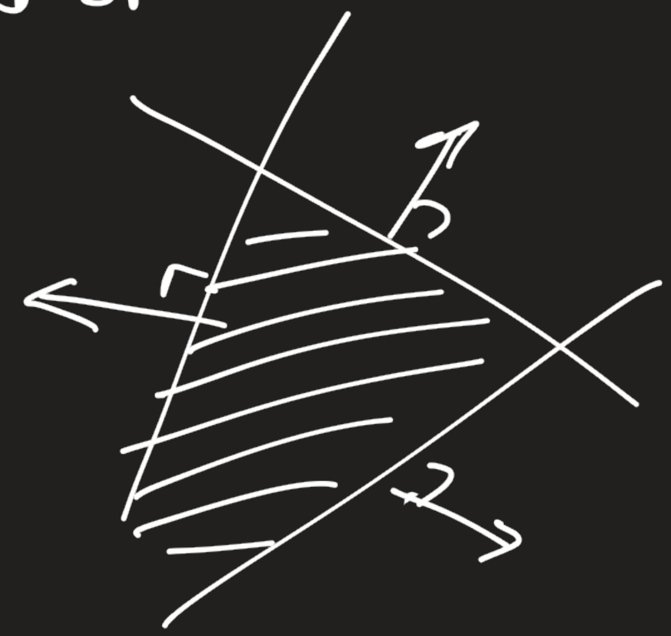
is a convex set. This is the <u>smallest</u> convex set containing S.

- norm balls $B_{|||\cdot|||}(x, r) = \{ y \mid |||x - y||| \le r \}$

- polyhedrons: set of solutions to a finite number of linear equalities and inequalities

$$\left\{ x : \begin{array}{l} a_1^T x \le b_1 \\ a_2^T x \le b_2 \\ a_3^T x = b_3 \\ \quad \vdots \end{array} \right\}$$

$\to$ equiv to $a_3^T x \le b_3$ and $(-a_3)^T x \le -b_3$



$= \{ x : Ax \le b \}$ where $A = \begin{bmatrix} a_1^T \\ \vdots \end{bmatrix}$ and $b = \begin{bmatrix} b_1 \\ \vdots \end{bmatrix}$

- Intersection of arbitrary # of convex sets is convex

$$\bigcap_{i \in I} C_i \text{ is convex if } C_i \text{ is convex for all } i \in I$$

Convex function

A function f defined on a convex set C

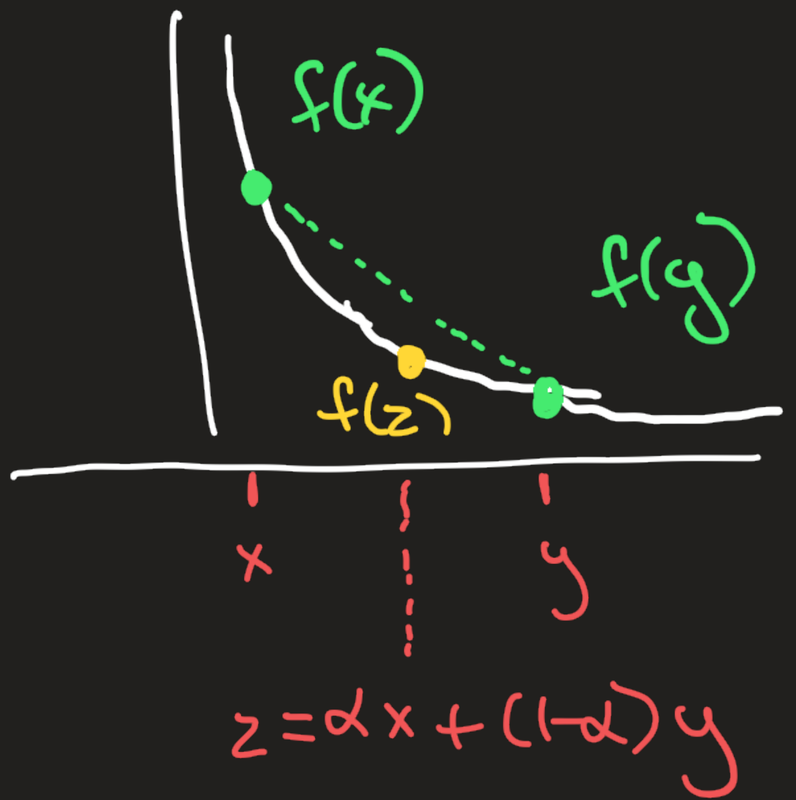
is called convex iff it satisfies

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

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

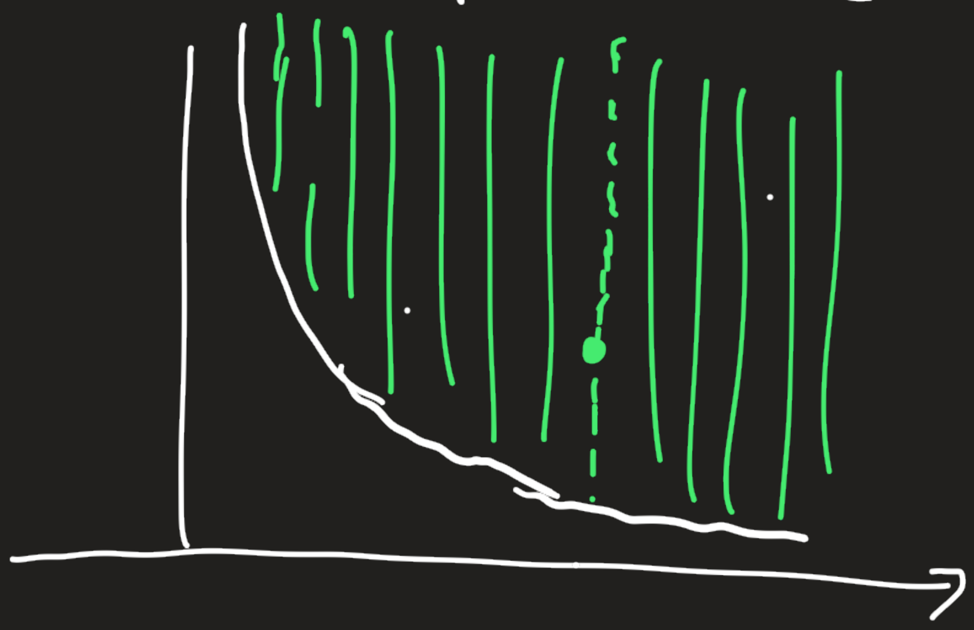
Graphically: a function is convex if it lies under its chords

Ex: $\frac{1}{x}$ is convex for $x \in \mathbb{R}_{++}$



Consider epigraph of f

$$\mathrm{epi}(f) = \{ (x, z) : f(x) \leq z \}$$



Fact: f is convex iff epi(f) is convex (on the set C)

<u>Prf</u> f cvx $\Rightarrow$ epi(f) cvx

Assume f is convex, and (x,z) and $(y,\omega) \in \mathrm{epi}(f)$ on C.

We need to show that if $\alpha \in [0,1]$ then $(\alpha x + (1-\alpha)y, \alpha z + (1-\alpha)\omega) \in \mathrm{epi}(f)$.

To see this note that

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

so $(\alpha x + (1-\alpha)y, \alpha z + (1-\alpha)\omega) \in \mathrm{epi}(f)$ as claimed

Strict convexity

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

$f = 0$

not strictly convex

$(1-x)_+$

not strictly convex

x^2

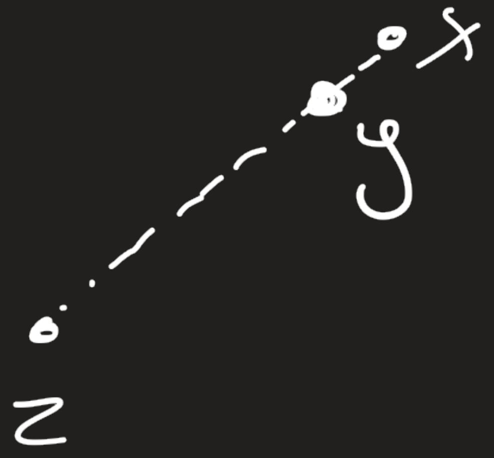
strictly convex

Most important consequence of convexity:

If a function f is convex on a convex set C, then any local minimum in C is a global minimum

Prf (contradiction)

Assume that x is a local minimizer (there is a neighborhood of x s.t. $f(x) \leq f(y)$ for any y in that neighborhood) but it is not a global minimizer. This means there exists a $z \in C$ such that $f(z) < f(x)$.



By convexity for an $\varepsilon \in (0,1]$,

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

then we have by convexity that

$$f(y) \leq (1-\varepsilon) f(x) + \varepsilon f(z) < f(x) \quad \text{because } f(z) < f(x)$$

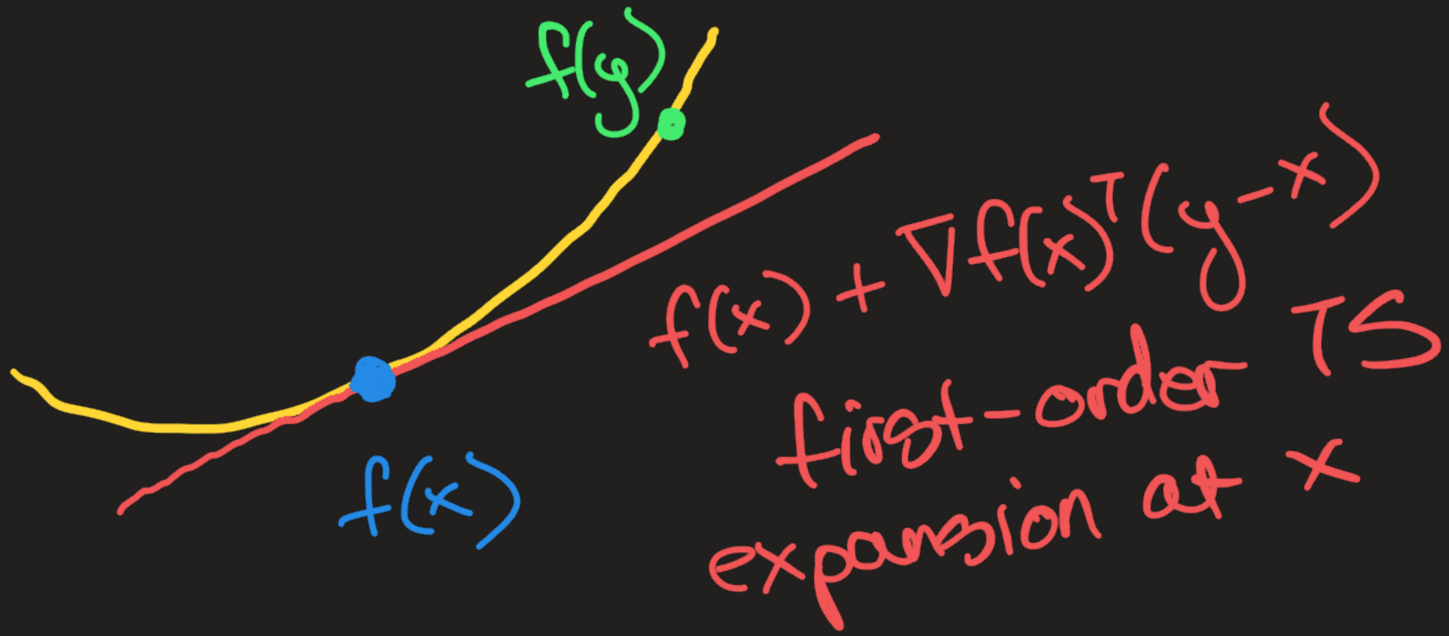
We've shown that we can a point arbitrarily close to x that satisfies $f(y) < f(x)$. This violates the assumption that x is a local minimizer. We conclude that its not possible to be a local minimizer without also being a global minimizer in C.

Fact: if f is strictly convex on C then the minimizer of f on C (if any exists) is unique.

First and Second-Order Characterizations of Convexity

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

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



That is, f is convex iff it lies above all of its tangent lines

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

note $\frac{df}{dx} = 2x$

Check $f(y) \geq f(x) + \frac{df}{dx}(x)(y-x)$ for all $x, y \in \mathbb{R}$:

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

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

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

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

conclude that $f(x) = x^2$ is convex on $\mathbb{R}$

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

$$\nabla^2 f(x) \succeq 0 \quad \left(\begin{array}{l}\text{Hessian is positive} \\ \text{semidefinite (PSD)}\end{array}\right)$$

Recall the Hessian of a function $f: \mathbb{R}^d \to \mathbb{R}$ is the matrix of second derivatives:

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

so $\nabla^2 f(x)$ is always symmetric

Characterizations of PSD matrices M $(M \succeq 0)$:

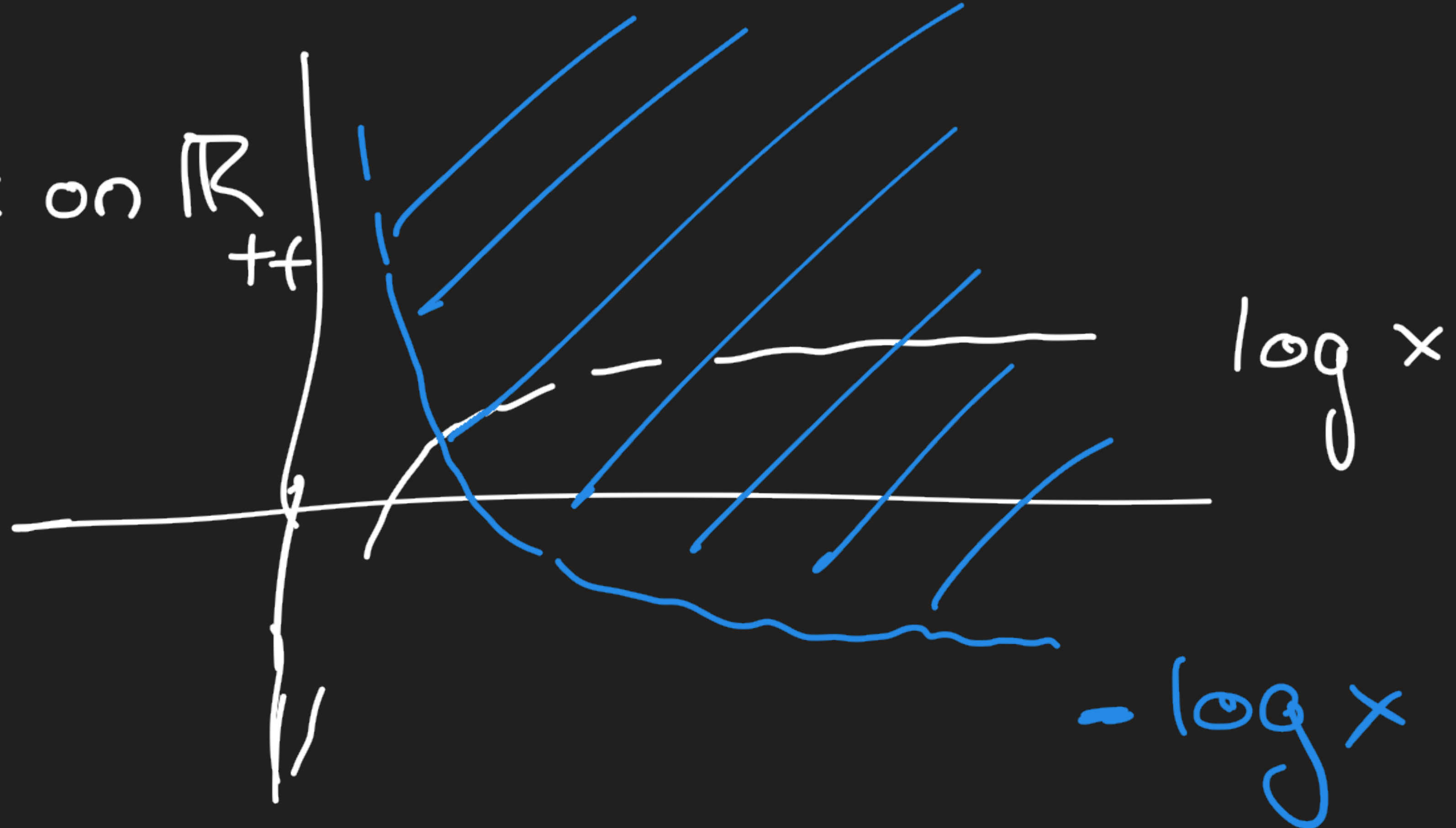
1) M is symmetric and all of its eigenvalues are nonnegative

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

If $f: \mathbb{R} \to \mathbb{R}$ then $\nabla^2 f = \{\nabla^2 f\}_{11} = \dfrac{d^2 f}{dx^2}$

<u>Ex</u>

$f(x) = -\log x$ is convex on $\mathbb{R}_{++}$



$$\frac{df}{dx} = -\frac{1}{x} = -x^{-1}$$

$$\frac{d^2 f}{dx^2} = \frac{1}{x^2} \geq 0 \text{ for } x \in \mathbb{R}_{++}$$

so by the 2nd-order condition, $-\log x$ is cvx on $\mathbb{R}_{++}$

Ex (next time) logsumexp(x) is convex on $\mathbb{R}^d$

- see Jensen's ineq first
- implies multiclass logistic regression is a convex optim problem