Today:

- Projects
- Optimality Conds for smooth cvx optim
- Gradient Descent
- Optimality Conds for non-smooth cvx optim
- Subdifferentials
- Subgradient Descent

Smooth cvx optimization

f, our cvx is differentiable: $\nabla f(x)$ exists at every x in our domain of optimization

Unconstrained Optim

$$\min_{x \in \mathbb{R}^d} f(x) \quad \text{or just} \quad \min_x f(x)$$

Constrained Optim

$$\min_{x \in C} f(x)$$

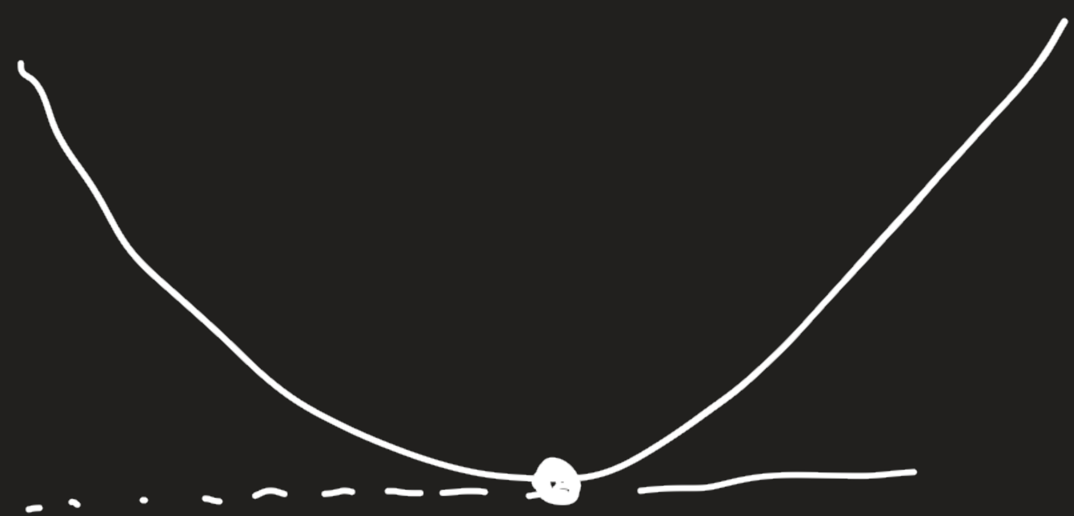
where C is convex

Optimality conditions

Unconstrained: $\nabla f(x_*) = 0$

sufficient & necessary

<u>Sufficiency</u> ($\nabla f(x) = 0 \Rightarrow x$ is a minimizer)



$\nabla f(x) = 0$

Recall f is convex & differentiable $\Rightarrow$

<u>for all y</u>

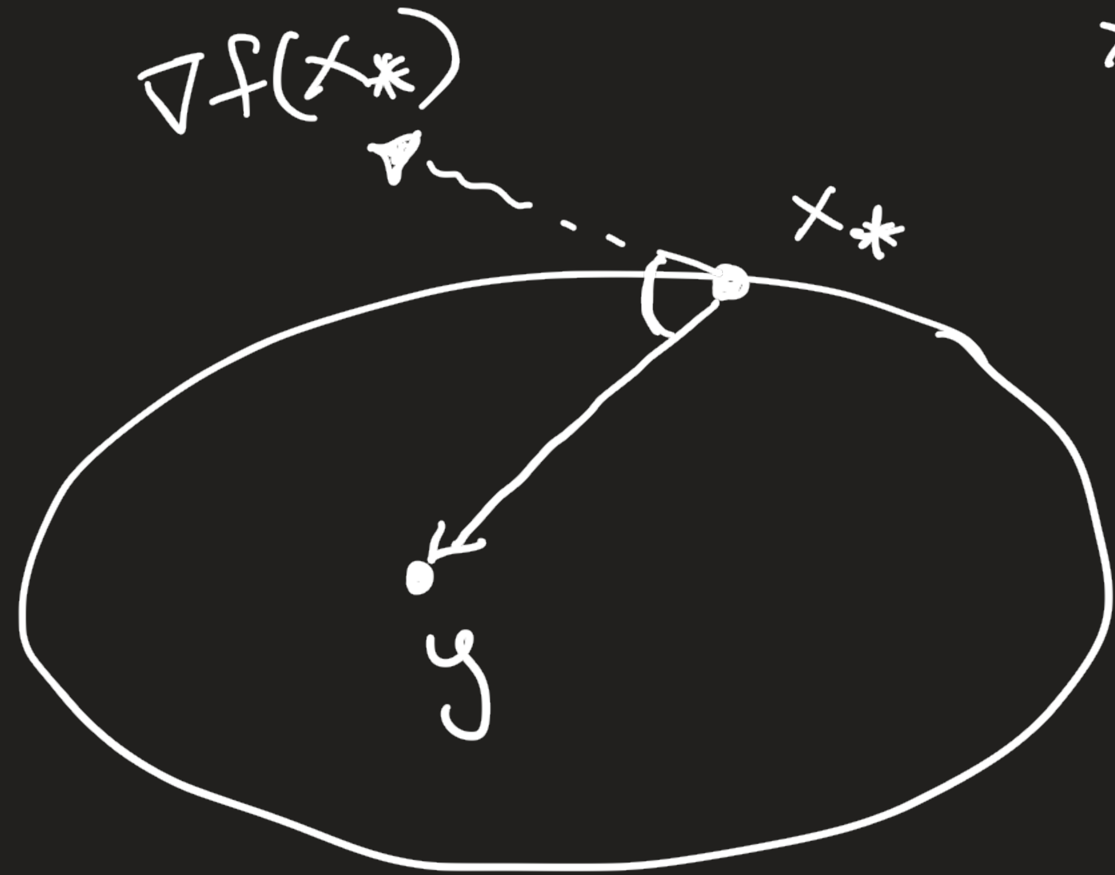
$$f(y) \geq f(x) + \underline{\nabla f(x)^T (y - x)}$$

so $\nabla f(x) = 0 \Rightarrow$

$$f(y) \geq f(x)$$

so x is a minimizer.

Constrained Optimization



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

$$\Longleftrightarrow$$

$$\langle \nabla f(x_*), y - x_* \rangle \geq 0$$

for all $y \in C$

Sufficiency:

Assume $\langle \nabla f(x), y - x \rangle \geq 0$ for all $y \in C$

then

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \geq f(x)$$

so x is a minimizer

Aside inner-products & Cauchy-Schwarz ineq.

$$|\langle x, y\rangle| \leq \|x\|_2 \|y\|_2$$

Cauchy-Schwarz ineq.

(C-S ineq.)

$$x^T x = \|x\|_2^2$$

higher-dim analog of the fact that

$$|ab| \leq |a||b|$$

$$\Rightarrow \left|\left\langle \frac{x}{\|x\|_2}, \frac{y}{\|y\|_2}\right\rangle\right| \leq 1$$

$\Rightarrow$ can interpret $\left\langle \frac{x}{\|x\|_2}, \frac{y}{\|y\|_2}\right\rangle$ as $\cos\angle x, y$

and we have $\langle x, y\rangle = (\cos\angle x, y)\|x\|_2 \|y\|_2$

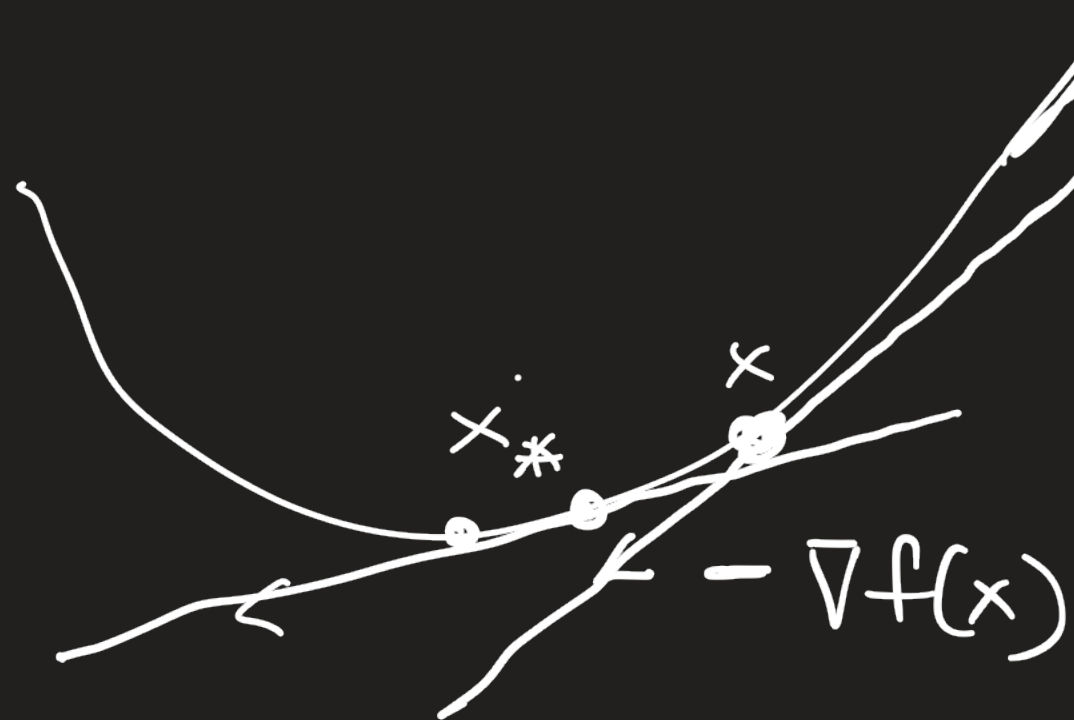
Note the optimality conditions give us measures of optimality

$(*)$

<u>Unconst</u> $\|\nabla f(x)\|_2 \to 0$ as $x \to x_*$

<u>Const</u> $\operatorname{dist}(\nabla f(x), S) \to 0$ (hand-waving)

Optimization Alg for smooth cvx optim: Gradient Descent



The basic idea is to note that

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

so if we want to decrease f by moving to a point $y = x + \Delta$, we have

$$f(y) \approx f(x) + \nabla f(x)^T \Delta$$

taking $\Delta = -\alpha \nabla f(x)$, we have

$$f(y) \approx f(x) - \alpha \|\nabla f(x)\|_2^2 \leq f(x)$$

Caveat: 1) f is nice
2) α is small enough that our linear approx is accurate

Gradient Descent alg

Input: f — convex function on $\mathbb{R}^d$

∇f — gradient

T — number of iterations

x_0 — starting guess

$\alpha_1, \dots, \alpha_T$ — step sizes for each iteration

Output: $x_1, \dots, x_T \in \mathbb{R}^d$ estimates of x_*

for $t = 1$ to T

$$x_{t+1} = x_t - \alpha_{t+1} \nabla f(x_t)$$

For now we want to show GD converges under some reasonable assumptions on f; that tell us

- how to choose our stepsizes
- how many steps we need to take to converge in the sense that

$$f(x_T) \leq f(x_*) + \varepsilon$$

A reasonable assumption is that our function's gradient is at least Lipschitz-continuous:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2$$

We call the gradient β-Lipschitz and we call the function β smooth.

<u>Facts</u>

1) $\|\nabla f(x)\|_2 \leq \beta \|x - x_*\|_2$ (take $y = x_*$ and note that $\nabla f(y) = \nabla f(x_*) = 0$)

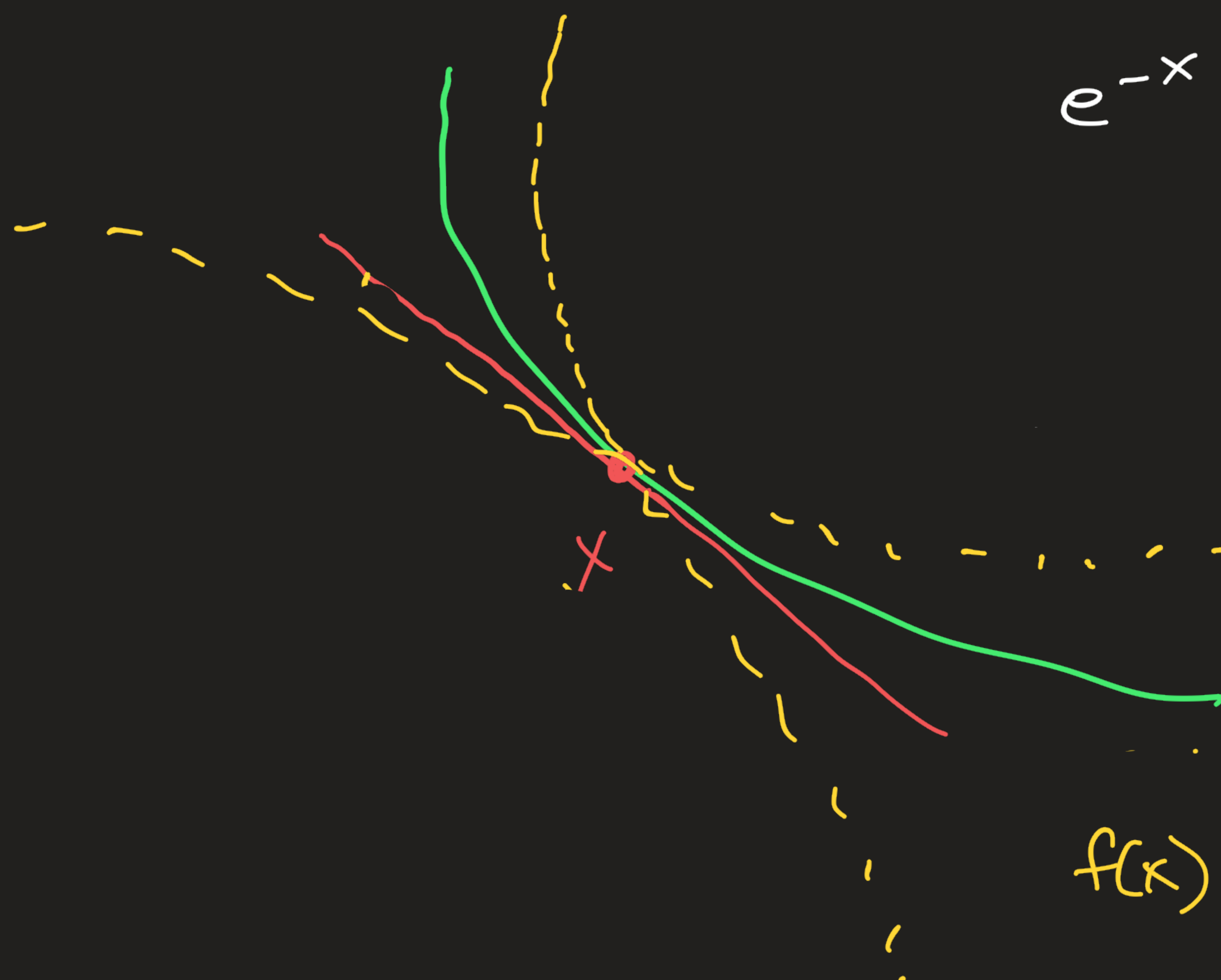
2) This implies a concrete bound on how good the approximation

$$f(y) \approx f(x) + \nabla f(x)^T (y - x) \quad \text{is}$$

$$\left| f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right| \leq \frac{\beta}{2} \|x - y\|_2^2$$

"f is β-smooth"

e^{-x}



$$f(x) + \nabla f(x)^T (y - x) + \beta \|x - y\|_2^2$$

$$f(x) + \nabla f(x)^T (y - x) - \beta \|x - y\|_2^2$$

Now assume f is β-smooth. Take $\alpha_t = \frac{1}{\beta}$

Then we see

$$f\left(x_t - \frac{1}{\beta}\nabla f(x_t)\right) - \left[f(x_t) + \left\langle \nabla f(x_t), -\frac{1}{\beta}\nabla f(x_t)\right\rangle\right]$$

$$\leq \frac{\beta}{2}\left\| -\frac{1}{\beta}\nabla f(x_t)\right\|_2^2 = \frac{\beta}{2\beta^2}\|\nabla f(x_t)\|_2^2$$

$$= \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

$$f\left(x_t - \frac{1}{\beta}\nabla f(x_t)\right) - \left[f(x_t) + \left\langle \nabla f(x_t), -\frac{1}{\beta}\nabla f(x_t)\right\rangle\right]$$

$$\leq \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

note $\left\langle \nabla f(x_t), -\frac{1}{\beta}\nabla f(x_t)\right\rangle = -\frac{1}{\beta}\|\nabla f(x_t)\|_2^2$

$\Rightarrow$

$$\boxed{f\left(x_t - \frac{1}{\beta}\nabla f(x_t)\right) - f(x_t) \leq -\frac{1}{2\beta}\|\nabla f(x_t)\|_2^2}$$

so taking a step of size $\frac{1}{\beta}$ in the negative gradient direction decreases f!

Note: we didn't make use of the fact that the function is convex

Ref: Bubeck. Convex Optimization: Algorithms & Complexity sec 3.2 for a full proof.

<u>Fact</u> If f is convex & β-smooth then gradient descent with $\alpha_t = \frac{1}{\beta}$ guarantees that

$$f(x_t) - f(x^*) \leq \frac{2\beta \|x_0 - x^*\|_2^2}{t}$$

rate of convergence is $O\left(\frac{1}{t}\right)$

In general we don't know $\|x_0 - x_*\|_2^2$ so we usually express this result as

$$f(x_T) - f(x^*) \leq \varepsilon$$

"ε-suboptimality"

is achieved when $T = O\left(\frac{\beta}{\varepsilon}\right)$ iterations

What about constrained optimization?

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

Projected Gradient Descent

Inputs: f, ∇f, T, $\alpha_1, \ldots, \alpha_T$, x_0

P_C — projection operator onto C

Output: $x_1, \ldots, x_T$

$$P_C(x) = \underset{z \in C}{\operatorname{argmin}} \; \|x - z\|_2^2$$

Alg:

for t in 1 to T:

$$x_{t+1} = P_C\left(x_t - \alpha_t \nabla f(x_t)\right)$$

Non-smooth convex optimization

f is cvx but not differentiable

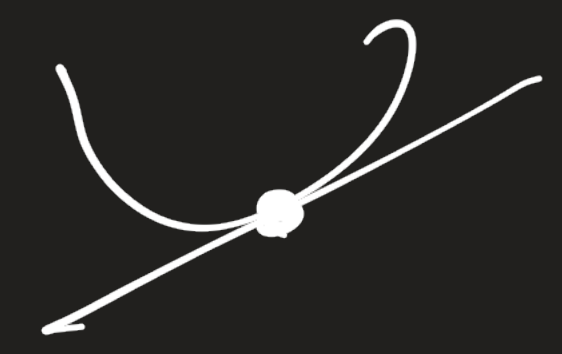
Ex:

$$F(\omega) = \frac{1}{\lambda} \sum_{i=1}^{n} (1 - y_i \omega^T x_i)_+ + \lambda \|\omega\|_1$$

In this case, we have no gradient, so can't move in the negative gradient direction to minimize F.

Soln: move in the negative dir of a <u>subgradient</u>

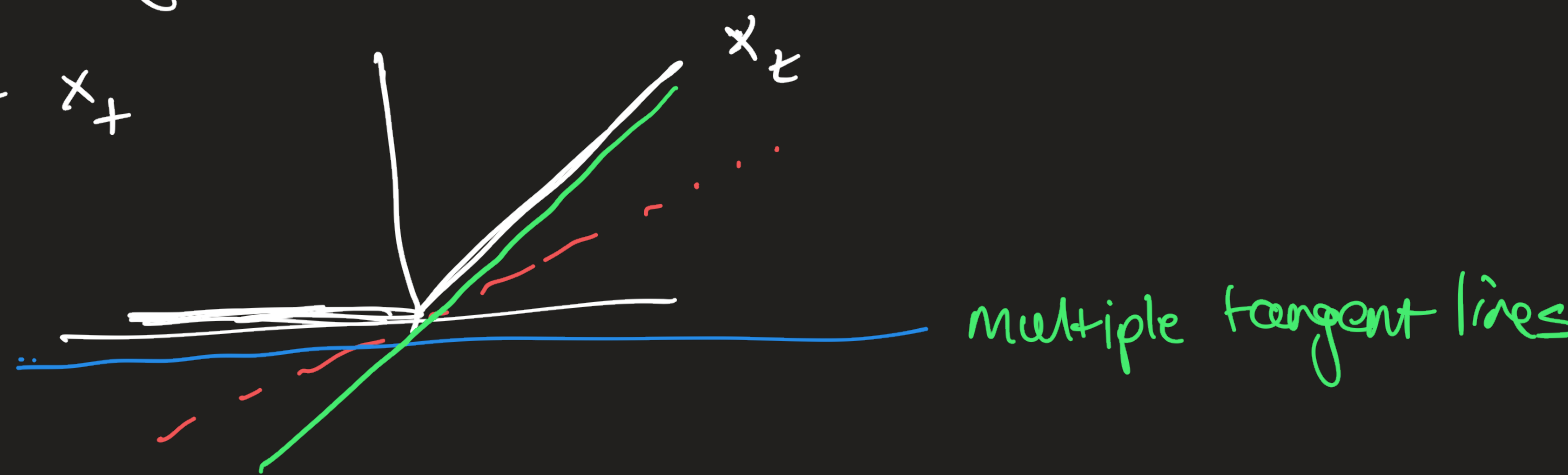
Recall for a smooth cvx fcn



$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

If f is cvx but not smooth, it may have multiple tangent lines:

consider x_+



In general we define the subdifferential of a function f at x:

$$\partial f(x) = \{ u \mid \text{for all } y, \; f(y) \geq f(x) + \langle u, y - x \rangle \}$$

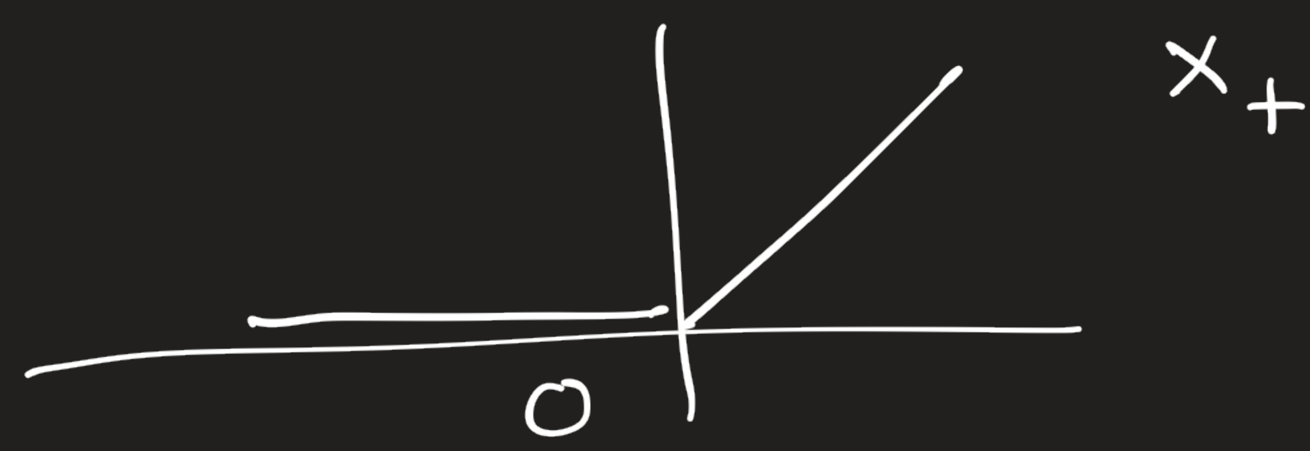
This is a <u>set</u>

<u>Fact</u> For convex functions, the subdifferential is nonempty, and if f is differentiable,

$$\partial f(x) = \{ \nabla f(x) \}$$

We call the vectors in $\partial f(x)$ <u>subgradients</u>

Ex subdifferential of positive part $f(x) = x_+$



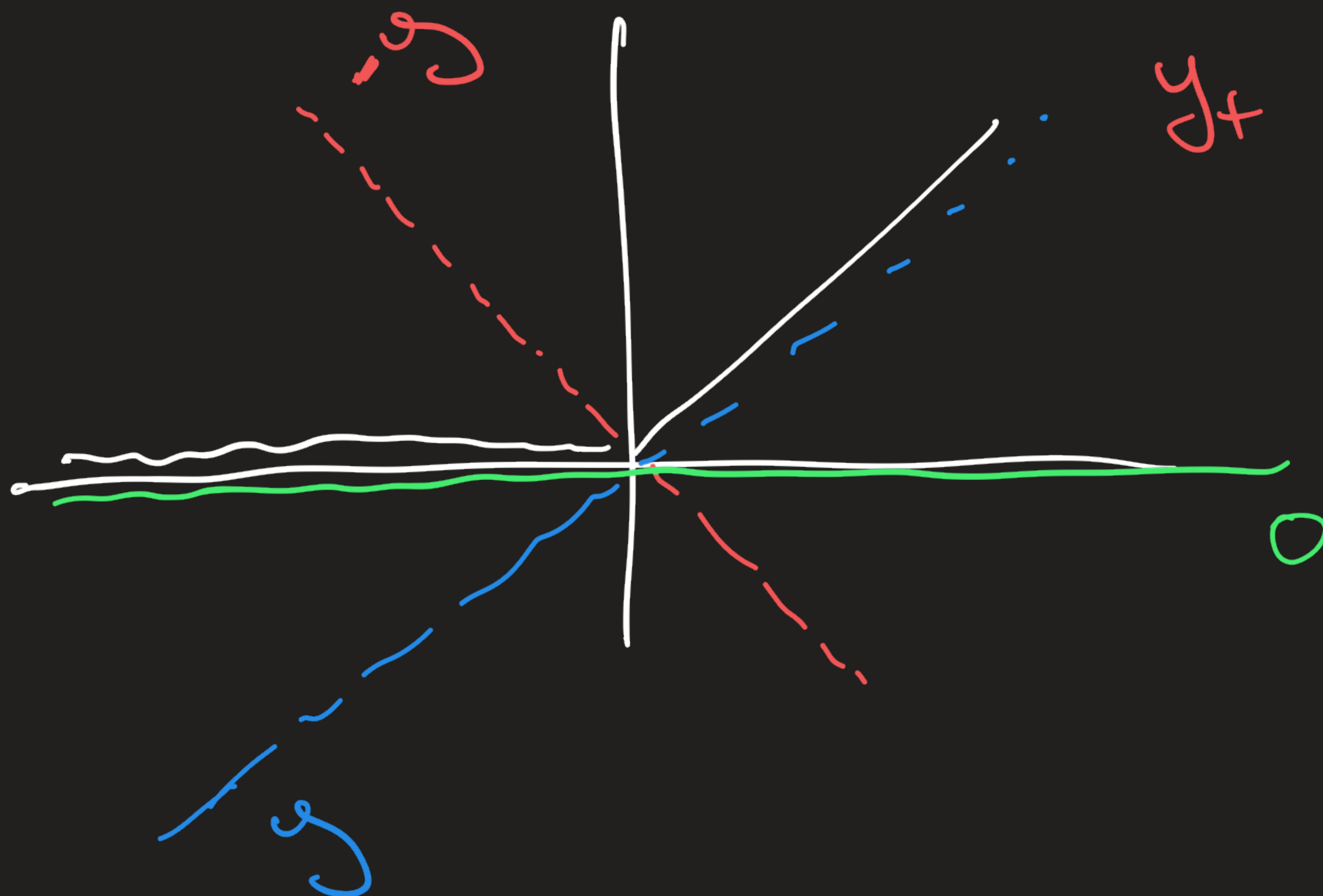
Clearly when $x < 0$, $\partial f(x) = \{0\}$

$x > 0$, $\partial f(x) = \{1\}$

when $x = 0$,

$$\partial f(0) = \{ v : f(y) \geq f(0) + vy \ \text{for all } y \}$$

$$= \{ v : y_+ \geq vy \ \text{for all } y \}$$



try $v = -1$

$\Rightarrow v \notin \partial f(0)$

try $v = 1$

$\Rightarrow v \in \partial f(0)$

try $v = 0$

$\Rightarrow v \in \partial f(0)$

convince yourself that $\partial f(0) = [0, 1]$

$$\partial f(x) = \begin{cases} 1 & \text{if } x > 0 \\ [0,1] & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Optimality Conditions (Necessary & Sufficient)

Unconstrained

$$x_* = \underset{x}{\operatorname{argmin}}\ f(x)$$

$$\iff 0 \in \partial f(x_*)$$

Constrained

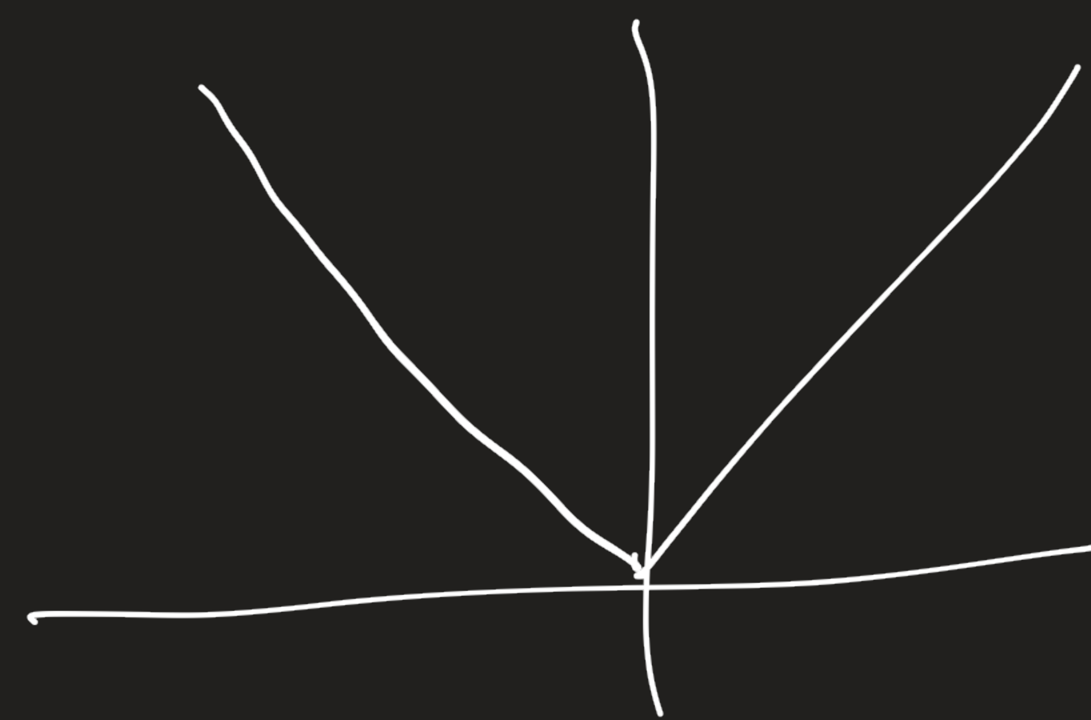
$$x_* = \underset{x \in C}{\operatorname{argmin}}\ f(x)$$

$\iff$ there is a subgradient g in the subdifferential at x_*, $g \in \partial f(x_*)$, for which

$$\langle g, y - x_* \rangle \geq 0$$

for all $y \in C$

Ex $f(x) = |x|$



$$\partial f(x) = \begin{cases} 1 & x > 0 \\ [-1, 1] & x = 0 \\ -1 & x < 0 \end{cases}$$