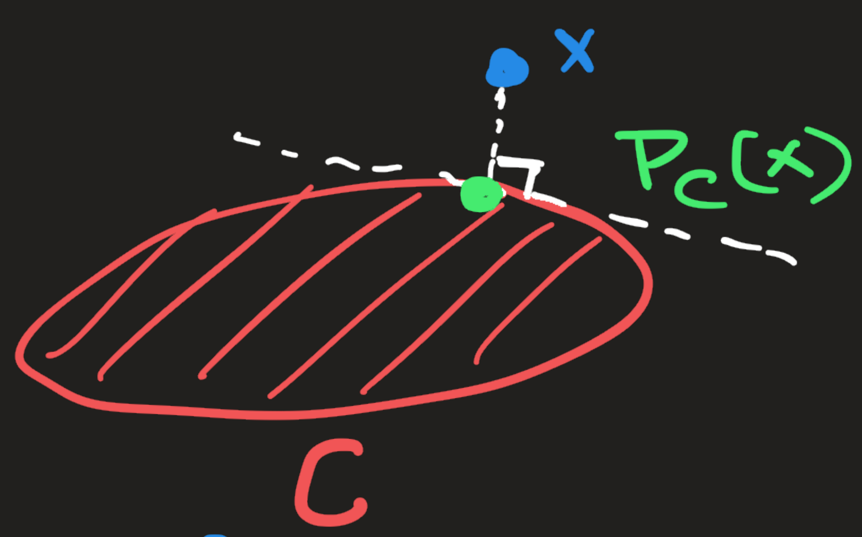
ML & Opt lecture 8

- Cauchy-Schwarz Inequality
- More examples of convex optim probs
- Optimality Conditions for Smooth Cvx Optim Probs
    - Constrained
    - Unconstrained

Ex

Projection onto convex set

- given a point C, find the closest pt in C to x



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

$P_C(x)$ is called the projection of x onto C, and is a well-defined function because the objective $z \mapsto \|z-x\|_2^2$ is strictly convex in z

Fact: if $\nabla^2 f(x) \succ 0$ then f is strictly convex

Consider $f(z) = \|z - x\|_2^2 = \langle z-x, z-x \rangle$

$$= z^T z - 2x^T z + x^T x$$

$$\nabla f(z) = 2z - 2x = 2(z-x)$$

$$\nabla^2 f(z) = 2 \cdot I$$

since $x \neq 0 \Rightarrow x^T \nabla^2 f(z) x = \|x\|_2^2 > 0$

we see that $\nabla^2 f(z) \succ 0$, so $f(z)$ is strictly convex.

Note also that $f(z) = \|z\|_2^2 - 2\langle x, z \rangle + \|x\|_2^2$

strictly cvx

cvx

ignorable b/c we are minimizing w.r.t. z

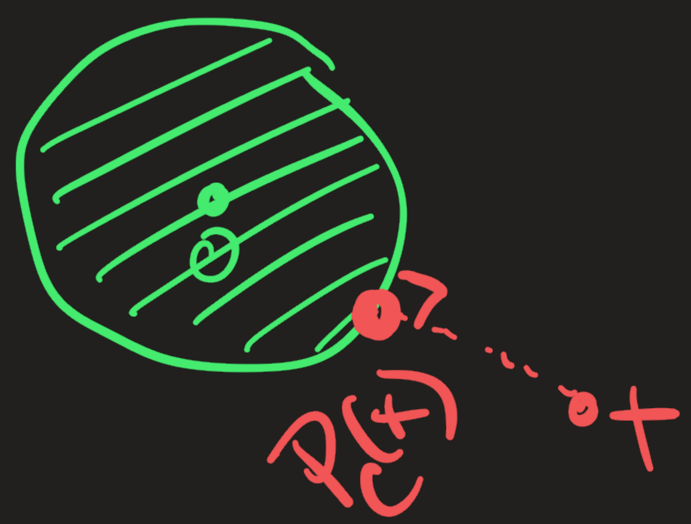
hence, strictly cvx w.r.t. z

We will show later in the lecture that if

$$C = B_{\|\cdot\|_2}(0,1) = \{ z \mid \|z\|_2 \le 1 \}$$

then

$$P_C(x) = \begin{cases} x & \text{if } x \in C \\ \dfrac{x}{\|x\|_2} & \text{if } x \notin C \end{cases}$$



Cauchy-Schwarz inequality

For any two vectors $x, y \in \mathbb{R}^d$,

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

or equivalently

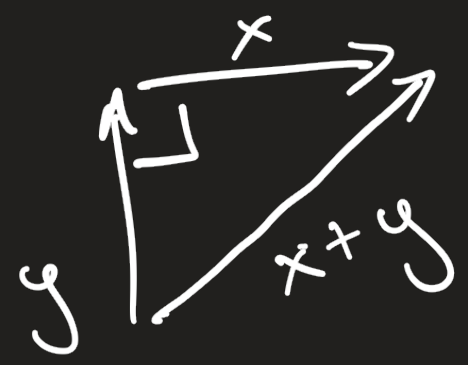
$$\left|\left\langle \frac{x}{\|x\|_2}, \frac{y}{\|y\|_2} \right\rangle\right| \le 1$$

and $|\langle x, y\rangle| = \|x\|_2 \|y\|_2$ iff $x = \alpha y$ for some $\alpha \in \mathbb{R}$

The CS-ineq allows us to define angles b/w vectors in high-dims, as

$$\cos(\theta \angle x, y) = \left\langle \frac{x}{\|x\|_2}, \frac{y}{\|y\|_2} \right\rangle$$

Intuition

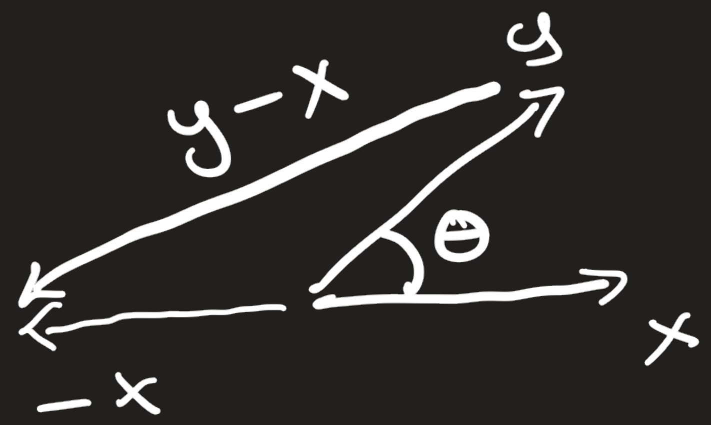


Pythagorean Theorem

$\|x+y\|_2^2 = \|x\|_2^2 + \|y\|_2^2$ follows from CS:

$$
\begin{aligned}
\|x+y\|_2^2 &= \langle x+y, x+y\rangle \\
&= x^Tx + \underbrace{2x^Ty}_{=0 \text{ b/c } \cos\angle\theta(x,y)=0} + y^Ty \\
&= \|x\|_2^2 + \|y\|_2^2
\end{aligned}
$$

$\Rightarrow \langle x, y\rangle = 0$



$$
\begin{aligned}
\|y-x\|_2^2 &= x^Tx + y^Ty - 2x^Ty \\
&= \|x\|_2^2 + \|y\|_2^2 - 2\|x\|_2\|y\|_2\cos\theta
\end{aligned}
$$

Law of Cosines

Types of Smooth Convex Optimization Problems

Let f be a differentiable convex func

Unconstrained

$$(U) \quad \min_{x \in \mathbb{R}^d} f(x) \quad \text{requires } \operatorname{dom}(f) = \mathbb{R}^d$$

Constrained

$$(C) \quad \min_{x \in C} f(x) \quad \text{requires } C \subseteq \operatorname{dom}(f)$$

As a caveat, (U) and (C) in general have multiple optimizers, but the minimal value is unique

Optimality Conditions

We want to know when x^* is a solution of (U) or (C), because:

1) If it is a simple enough condition, we can directly use them to find x^*

2) we can use these conditions to
   - i) say a priori qualities x^* has, like sparsity
   - ii) verify the quality of approximate solns
   - iii) design algorithms for solving (U) & (C)

Optimality Condition for (U)

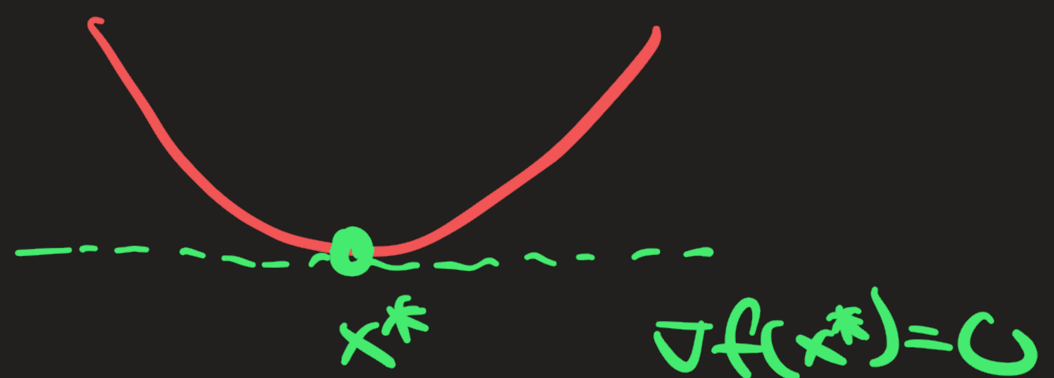
Fermat's Condition

$$x^* \in \underset{x \in \mathbb{R}^d}{\operatorname{argmin}}\ f(x) \quad \text{iff} \quad \nabla f(x^*) = 0$$

Prf

$\Leftarrow$ Assume $\nabla f(x^*) = 0$. Because f is convex

$$\forall y: \quad f(y) \geq f(x^*) + \nabla f(x^*)^T (y - x)$$
$$= f(x^*)$$



$\Rightarrow$ Let x^* be a minimizer.

Now assume that $\nabla f(x^*) \neq 0$. Then take $y = x^* - \varepsilon \dfrac{\nabla f(x^*)}{\|\nabla f(x^*)\|_2}$ for an $\varepsilon \in (0,1)$.

Then we know that

$$f(y) = f(x^*) + \langle \nabla f(x^*), y - x^* \rangle + O(\|y - x^*\|_2^2)$$

$$= f(x^*) + \left\langle \nabla f(x^*), -\frac{\varepsilon \nabla f(x^*)}{\|\nabla f(x^*)\|_2} \right\rangle + O\left(\varepsilon^2 \left\| \frac{\nabla f(x^*)}{\|\nabla f(x^*)\|_2} \right\|_2^2 \right)$$

$$= f(x^*) - \varepsilon \|\nabla f(x^*)\|_2 + O(\varepsilon^2)$$

So if $\varepsilon \ll \|\nabla f(x^*)\|_2$ then $\varepsilon^2 \ll \varepsilon \|\nabla f(x^*)\|_2$

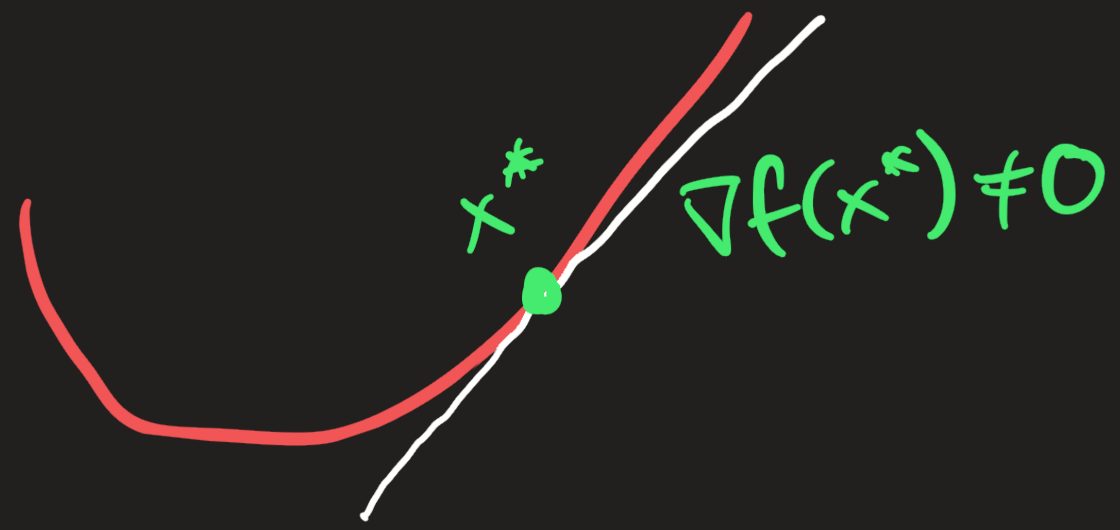
so $f(y) = f(x^*) - \varepsilon \|\nabla f(x^*)\|_2 + O(\varepsilon^2)$

$$< f(x^*).$$

This a contradiction, b/c we assumed x^* is a minimizer. Therefore it must be the case that

$\nabla f(x^*) = 0$

ω



Consequences

Solving the convex optimization problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

is as hard/easy as solving the set of equations

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

Ex Linear regression

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \|X\beta - y\|_2^2 \iff \nabla_\beta \|X\beta^* - y\|_2^2 = 0$$

$$\iff 2X^T(X\beta^* - y) = 0$$

$$\iff X^T X \beta^* = X^T y$$

$$\iff \beta^* = (X^T X)^{-1} X^T y$$

Ex

$$x^* = \arg\min_x \|x\|_2^2 + \text{logsumexp}(x)$$

$$\Updownarrow$$

$$\nabla_x \left(\|x^*\|_2^2 + \text{logsumexp}(x^*) \right) = 0$$

$$2x^* + \text{softmax}(x^*) = 0$$

Observations:

1) $-2x^*$ is a probability vector

2) objective is strictly cvx so x^* is unique, and permutation invariant so any permutation of entries of x^* is a minimizer $\Rightarrow$ x^* is a constant vector

$$x^* = c \cdot \mathbb{1}$$

Because $x^* = c \cdot \mathbb{1}$ and

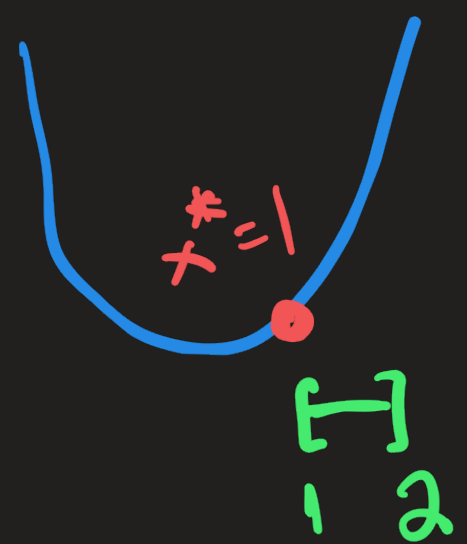
$$\text{softmax}(x^*) = -2x^*,$$

$$\frac{1}{d} = -2c \cdot \mathbb{1}$$

$$\Rightarrow \quad c = -\frac{1}{2d}$$

$$\Rightarrow \quad x^* = \begin{bmatrix} -1/2d \\ \vdots \\ -1/2d \end{bmatrix}$$

Optimality Conditions for (C)



First, note that $\nabla f(x^*)=0$ is not an optimality condition for (C) in general, e.g.

$$\underset{x\in[1,2]}{\operatorname{argmin}}\ x^2 = 1, \text{ but } \nabla_x x^2\Big|_{x=1} = 2 \neq 0$$

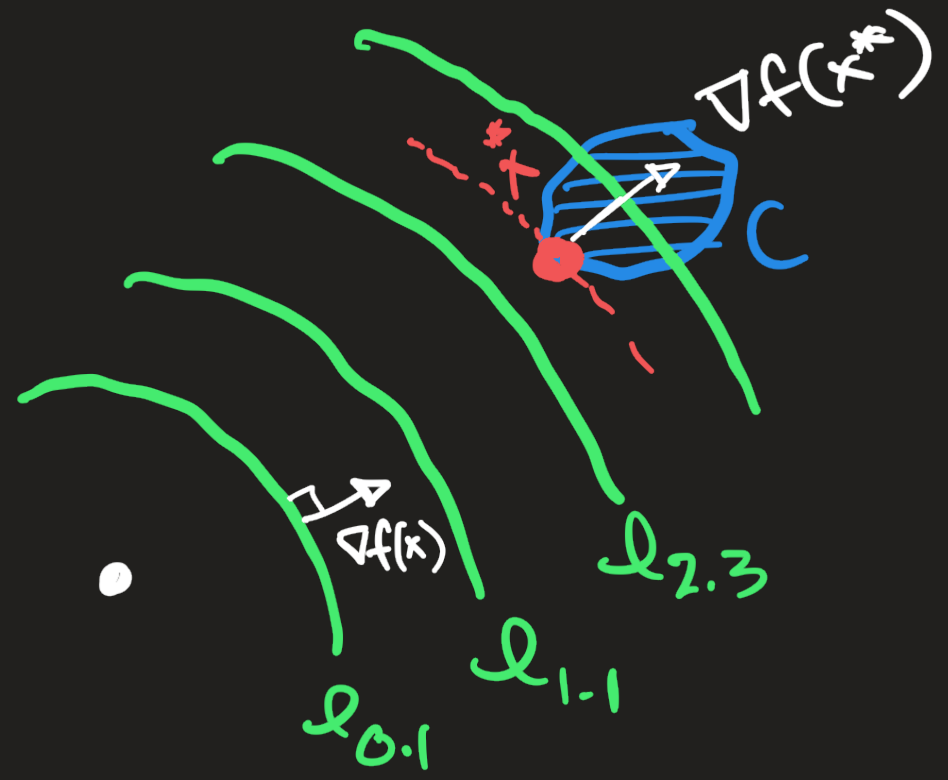
However

$$\underset{[-1,1]}{\operatorname{argmin}}\ x^2 = 0 \text{ and } \nabla_x x^2\Big|_{x=0} = 0$$

So we see that the optimality conditions depend on the interaction of f with C

Geometric Intuition

Consider $\min\limits_{x \in C} f(x)$



Consider the α level sets of f

$$\ell_\alpha = \{ x \mid f(x) = \alpha \}$$

The minimizer x_* must be on the level set of lowest value that intersects C

Notice that $\nabla f(x) \perp \ell_{f(x)}$ at any point x

Imagine that there exists a $y \in C$ such that

$\langle \nabla f(x^*), y - x^* \rangle < 0$, then

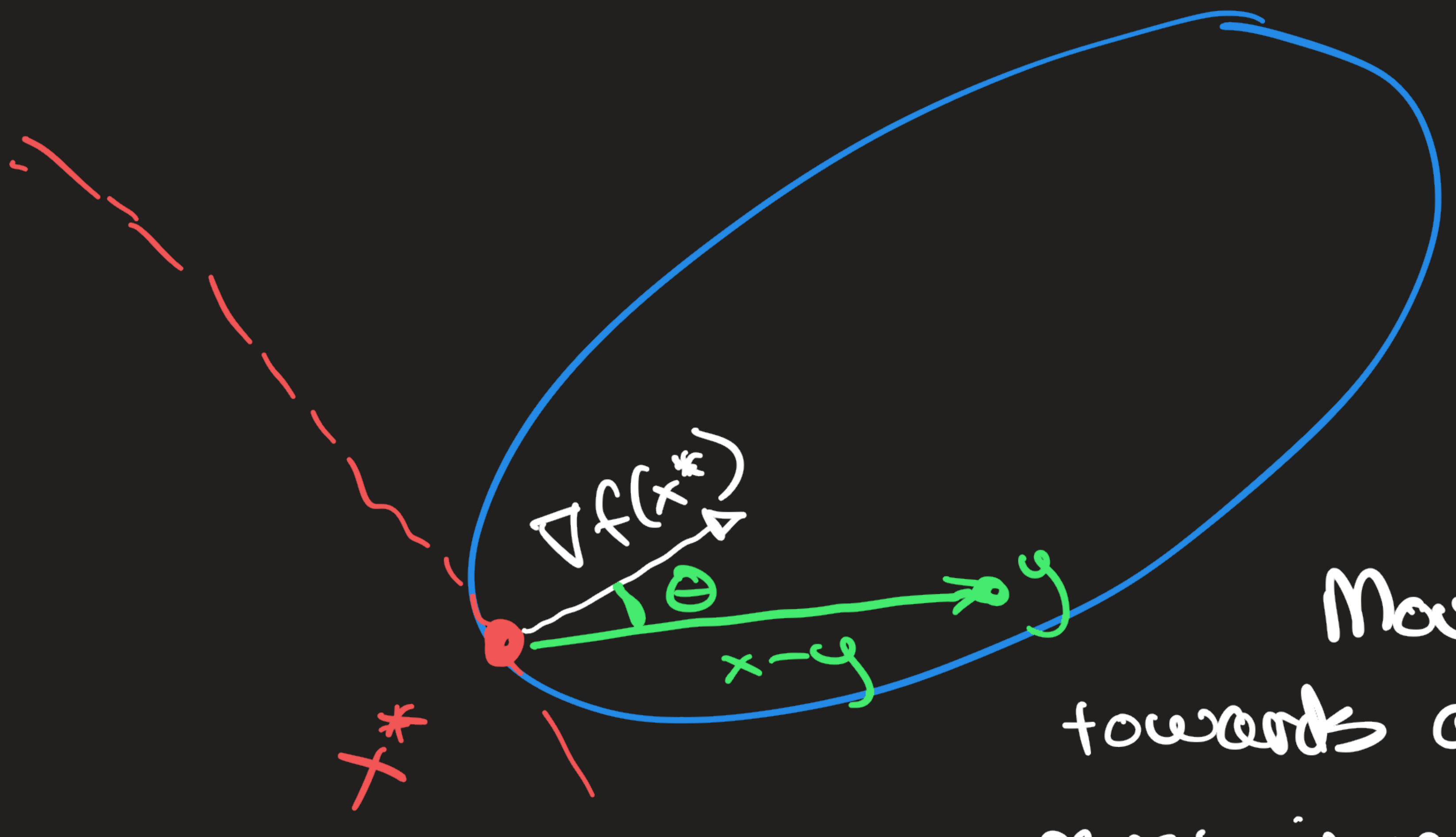
$$f(y) = f(x^*) + \langle \nabla f(x^*), y - x^* \rangle + o(\|y - x^*\|_2^2)$$

$$= f(x^*) + (\text{something negative}) + o((\text{something negative})^2)$$

$$< f(x^*)$$

so it must be the case that

$$\langle \nabla f(x^*), y - x^* \rangle \geq 0 \quad \forall y \in C$$



Moving from x^* towards any $y \in C$ must increase f, if x^* is a minimizer, i.e.

$$\forall y \in C: \langle \nabla f(x^*), y - x^* \rangle \geq 0$$

Optimality Condition for (C)

Fermat's Condition

x^* is a minimizer of

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

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