# ML and Optimization Lecture 10

- More examples of convex optimization problems
- Cauchy-Schwarz Inequality
- Optimality Conditions for Smooth Convex Optimization
    - Constrained
    - Unconstrained

<u>Ex</u> Projection onto a convex set

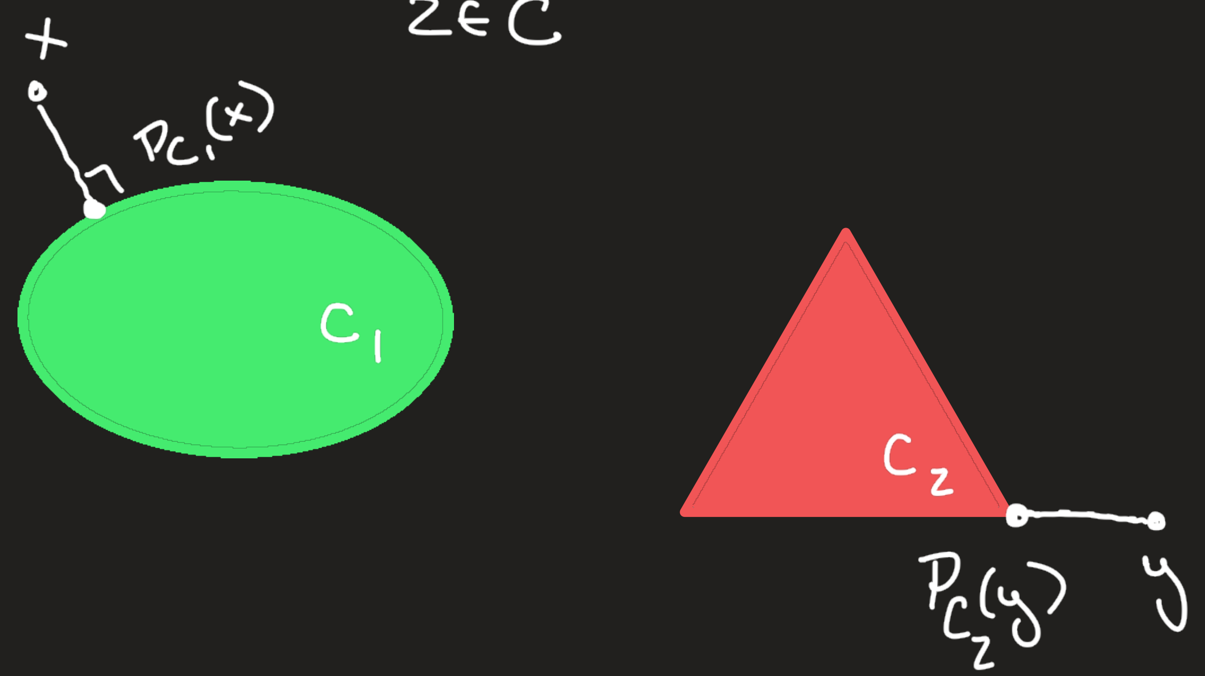
Let $C$ be a non-empty convex set. Define the function $P_C$ that maps an arbitrary point $x$ to the closest point in $C$ as follows:

$$P_C(x) = \underset{z \in C}{\operatorname{argmin}} \ \frac{1}{2}\|x - z\|_2^2$$

$\leftarrow$ convex b/c it is the composition of a squared norm and an affine function of $z$

$$= \underset{z \in C}{\operatorname{argmin}} \ \frac{1}{2}\left(\|x\|_2^2 - 2\langle x, z\rangle + \|z\|_2^2\right)$$

$$= \underset{z \in C}{\operatorname{argmin}} \ \|z\|_2^2 - 2\langle x, z\rangle$$



Useful fact: $f(x) = \|x\|_2^2$ is strictly convex

Prf: (algebra, feel free to skip)

$$
\begin{aligned}
\|\alpha u + (1-\alpha) v\|_2^2 &= \alpha^2 \|u\|_2^2 + (1-\alpha)^2 \|v\|_2^2 + 2\alpha(1-\alpha)\langle u, v\rangle \\
&= \alpha \|u\|_2^2 + (\alpha^2 - \alpha)\|u\|_2^2 + (1-\alpha)\|v\|_2^2 \\
&\quad + \left[(1-\alpha)^2 - (1-\alpha)\right]\|v\|_2^2 + 2\alpha(1-\alpha)\langle u, v\rangle \\
&= \alpha\|u\|_2^2 + (1-\alpha)\|v\|_2^2 + \alpha(\alpha-1)\|u\|_2^2 + \\
&\quad \alpha(\alpha-1)\|v\|_2^2 + 2\alpha(1-\alpha)\langle u, v\rangle \\
&= \alpha\|u\|_2^2 + (1-\alpha)\|v\|_2^2 + \alpha(1-\alpha)\Big[2\langle u, v\rangle \\
&\qquad - \|u\|_2^2 - \|v\|_2^2\Big] \\
&= \alpha\|u\|_2^2 + (1-\alpha)\|v\|_2^2 - \alpha(1-\alpha)\|u - v\|_2^2 \\
&< \alpha\|u\|_2^2 + (1-\alpha)\|v\|_2^2
\end{aligned}
$$

$\blacksquare$

It follows that $f(z) = \frac{1}{2}\|x - z\|_2^2$ is strictly convex

(Exercise: verify this!)

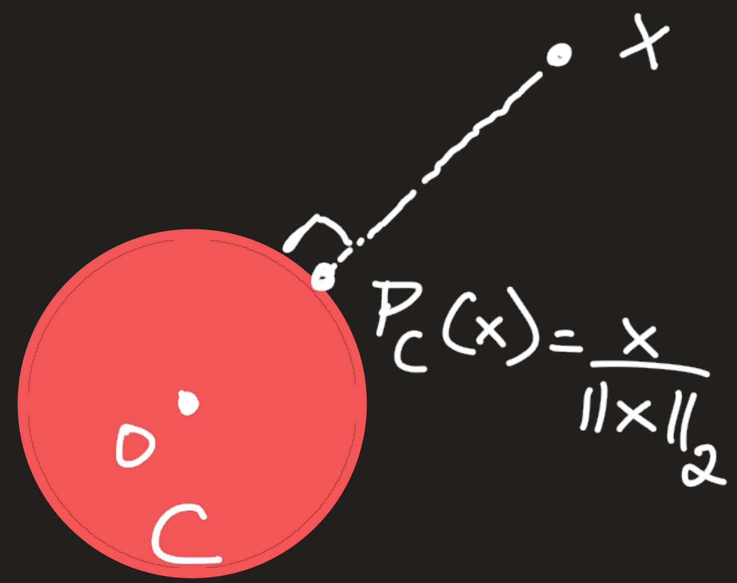
Therefore $P_C(x) = \underset{x \in C}{\operatorname{argmin}} f(z)$ is a well-defined function:

$P_C$ maps $x$ to the single unique point in $C$ closest to $x$.

<u>Ex</u> Let $C = B_{\|\cdot\|_2}(0,1) = \{ z \mid \|z\|_2 \leq 1 \}$

We will show at end of lecture that

$$P_C(x) = \begin{cases} x & \text{if } \|x\|_2 \leq 1 \\ \dfrac{x}{\|x\|_2} & \text{if } \|x\|_2 > 1 \end{cases}$$



## Ex   Min-max regression

Often we want to fit a regression model $y \approx \beta^T x$ by minimizing the maximum pointwise absolute error on our dataset (nb: this is not robust, as outliers will have a huge impact)

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \boxed{\max_{i=1,\dots,n} |y_i - \beta^T x_i|}$$

to see this is convex write it as

$$f(\beta) = \max_{i=1,\dots,n} f_i(\beta)$$

where

$$f_i(\beta) \equiv |y_i - \beta^T x_i|$$

are convex b/c composition of abs value with an affine function of $\beta$. Then we know pointwise maximum of convex functions is a convex function

$$= \underset{\beta}{\operatorname{argmin}} \underbrace{\| X\beta - y \|_\infty}$$

can see is convex b/c is composition of norm w/ affine function of $\beta$

where $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$ and $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ and $\|v\|_\infty = \max_i |v_i|$

<u>Ex</u> RERM

Consider RERM of the form

$$w^* = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} l_i(w) + \lambda \Omega(w)$$

where we write $l_i(w)$ to indicate the loss of the model with parameters $w$ on the $i^{th}$ training point.

If $l_i$ is convex, $\Omega$ is convex, and $\lambda \geq 0$, this is a convex optim prob b/c positively weighted sum of cvx fcns is cvx

Ex logistic regression

Recall in this case the pointwise loss is given by

$$\ell_i(\omega) = \log\left(1 + \exp\left(-y_i \langle x_i, \omega \rangle\right)\right)$$

To see this is convex, write it as

$$\ell_i(\omega) = \log\left(e^0 + e^{-y_i \langle x_i, \omega \rangle}\right)$$

$$= \operatorname{logsumexp}\left(\begin{bmatrix} 0^T \\ -y_i x_i^T \end{bmatrix} \omega\right)$$

We wrote $\ell_i$ as the composition of a convex function with an affine function of $\omega$, so it is convex.

# Cauchy-Schwarz Inequality (CS-Ineq)

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

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

and

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

iff $x$ & $y$ are parallel or anti-parallel $(\exists \alpha \in \mathbb{R} : x = \alpha y)$

Why we care: the CS-ineq lets us interpret the inner product as measuring angles

Prf (you can skip this)

If $x$ or $y$ is $0$, this is vacuously true.

It now suffices to show that if $\|u\|_2 = \|v\|_2 = 1$, then $|\langle u, v \rangle| \leq 1$, because for arbitrary nonzero $x$ & $y$ we have that $\left\| \frac{x}{\|x\|_2} \right\|_2 = 1$ and $\left\| \frac{y}{\|y\|_2} \right\|_2 = 1$, so we have that

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

So now assume that $\|u\|_2 = \|v\|_2 = 1$, then because

$$\|u+v\|_2^2 = \|u\|_2^2 + 2\langle u,v\rangle + \|v\|_2^2$$

$$= 2 + 2\langle u,v\rangle \geq 0$$

we see that

$$\langle u,v\rangle \geq -1$$

and similarly, because

$$\|u-v\|_2^2 = \|u\|_2^2 - 2\langle u,v\rangle + \|v\|_2^2$$

$$= 2 - 2\langle u,v\rangle \geq 0$$

we see that

$$\langle u,v\rangle \leq 1$$

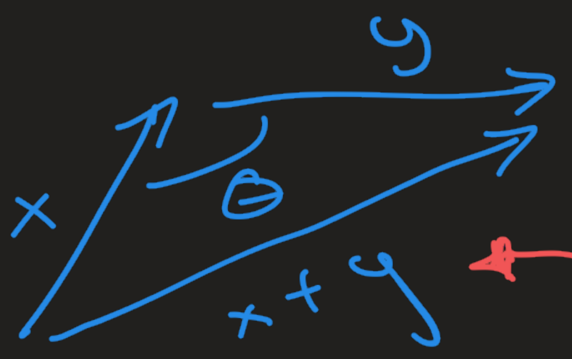
so

$$|\langle u,v\rangle| \leq 1$$

$\blacksquare$

Optional
(Exercise: prove the claim about when equality occurs)

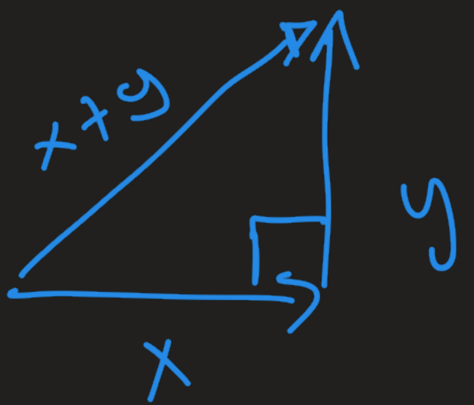
# Geometric Interpretation




Let $\theta = \angle(x, y)$, then

$$\|x+y\|_2^2 = \|x\|_2^2 + \underline{2\langle x, y\rangle} + \|y\|_2^2$$

$$= \|x\|_2^2 + \|y\|_2^2 + \underline{2\|x\|_2\|y\|_2 \cos\theta}$$




E.g. if $x \perp y$ so $\theta = \frac{\pi}{2}$, then $\cos\theta = 0$ and we recover

$$\|x+y\|_2^2 = \|x\|_2^2 + \|y\|_2^2$$

the usual Pythagorean identity

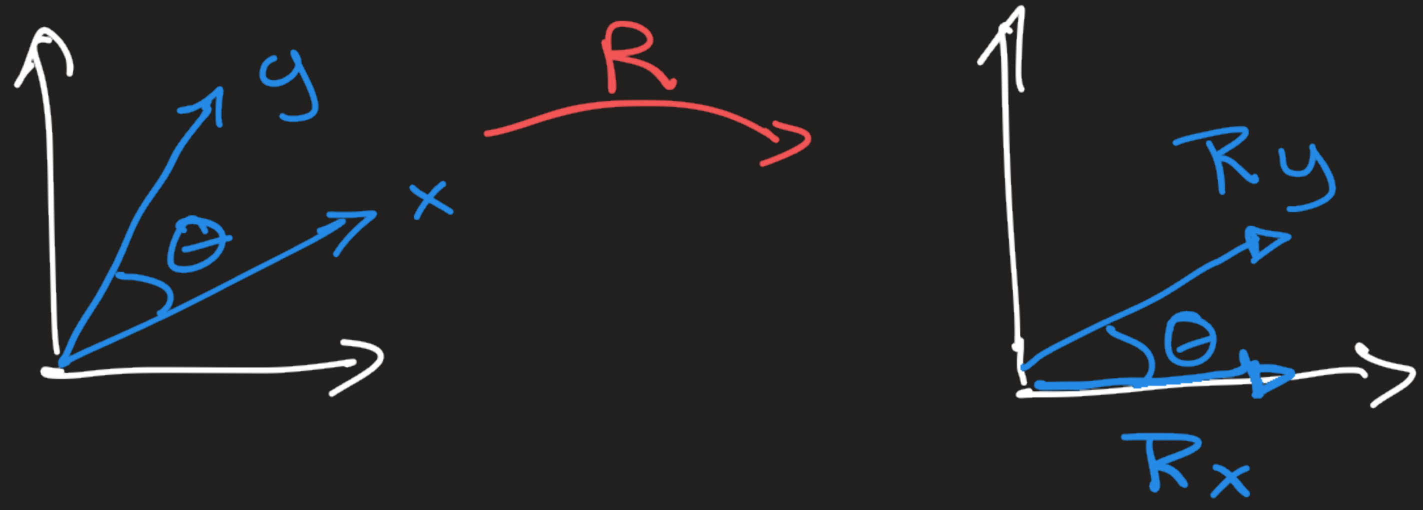
This is a high-dimensional analog of the Law of Cosines

We define in high dimensions

$$\cos\theta = \frac{\langle x, y\rangle}{\|x\|_2\|y\|_2}$$

By CS-Ineq, $\cos\theta \in [-1, 1]$ as it should be

Notice that in 2D, this expression for $\cos\theta$ indeed matches our usual definition. To see this, we need the fact inner-products are preserved by rotations



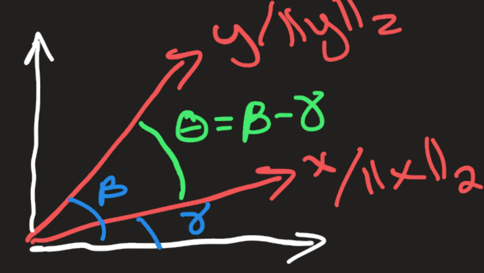

Remember rotations are matrices that satisfy $RR^T = R^TR = I$ and $\det R = 1$. Thus we have for any $x$ and $y$ and rotation $R$ that

$$\langle Rx, Ry\rangle = (Rx)^T Ry = x^T R^T R y = x^T y = \langle x, y\rangle$$

so inner products are preserved by rotations

Now observe that given $x$ and $y$ in $\mathbb{R}^2$, there is a Rotation matrix $R$ that rotates $\frac{x}{\|x\|_2}$ to the x-axis.

If $\frac{x}{\|x\|_2} = \begin{bmatrix} \cos\gamma \\ \sin\gamma \end{bmatrix}$ and $\frac{y}{\|y\|_2} = \begin{bmatrix} \cos\beta \\ \sin\beta \end{bmatrix}$, take




$$R = \begin{bmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{bmatrix}$$

Then you can verify that $RR^T = R^TR = I$ and $R\frac{x}{\|x\|_2} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, and

$$R\frac{y}{\|y\|_2} = \begin{bmatrix} \cos\gamma\cos\beta + \sin\gamma\sin\beta \\ \cos\gamma\sin\beta - \sin\gamma\cos\beta \end{bmatrix} = \begin{bmatrix} \cos(\beta-\gamma) \\ \sin(\beta-\gamma) \end{bmatrix}$$

This implies that

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

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

$$= \|x\|_2 \|y\|_2 \left\langle R\frac{x}{\|x\|_2}, R\frac{y}{\|y\|_2} \right\rangle$$

$$= \|x\|_2 \|y\|_2 \left\langle \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} \cos(\beta-\gamma) \\ \sin(\beta-\gamma) \end{bmatrix} \right\rangle$$

$$= \|x\|_2 \|y\|_2 \cos(\beta-\gamma)$$

$$= \|x\|_2 \|y\|_2 \cos\theta$$

so indeed we get that for vectors in $\mathbb{R}^2$

$$\frac{\langle x, y\rangle}{\|x\|_2 \|y\|_2} = \cos\theta$$

so the CS-Ineq definition of $\cos\theta$ beteween vectors coincides with the trig defn

Takeaway:

if $\langle x, y \rangle \geq 0$, then

$$\cos \angle(x,y) = \frac{\langle x, y \rangle}{\|x\| \, \|y\|_2} \geq 0$$

so there is an acute angle between $x$ & $y$

This geometric interpretation of the CS-Ineq is important and widely used. We will use it to state optimality conditions for constrained optimization problems

## Types of Smooth Convex Optimization Problems

Let $f$ be a differentiable convex function

### Unconstrained

(U) $\displaystyle\min_{x \in \mathbb{R}^d} f(x)$ and we require $f$ is convex on $\mathbb{R}^d$

### Constrained

(C) $\displaystyle\min_{x \in C} f(x)$ and we require $f$ is convex on the convex set $C$

Again, in general (U) or (C) may have multiple solutions unless we know $f$ is strictly convex

# 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 solve it to determine $x^*$

2) we can use these conditions to design algorithms to solve (U) and (C), and to verify the quality of approximate solns

## Optimality Conditions for (U)

Recall from calculus that the (unconstrained) extrema (local minima/maxima and saddlepoints) can be found for smooth functions by setting $\nabla f(x^*) = 0$.

In the case of cvx fcns and unconstrained optimization, the only extrema are the global minima.

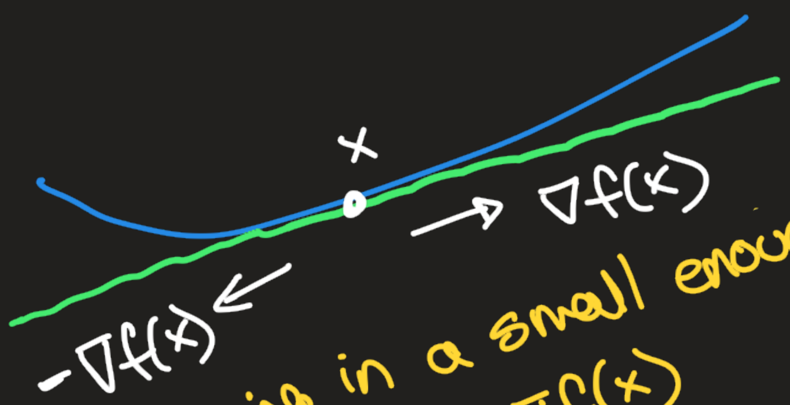
Consider

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

where $f$ is differentiable and convex. Then $x^*$ is a minimizer of (U) iff $\nabla f(x^*) = 0$

Intuition: moving away from $x$ in an acute direction to $-\nabla f(x)$ decreases the value of $f$, by a T.S. argument.

This means that




moving in a small enough step in the $-\nabla f(x)$ direction decreases the value of $f$

for $x^*$ is a minimizer iff <u>all</u> directions are acute to $-\nabla f(x^*)$. The only vector that is acute to all directions is zero.

<u>Prf</u>

Assume $\nabla f(x^*)=0$, then for all $y \in \mathbb{R}^d$,

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

so $x^*$ is a minimizer.

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

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

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

Since the remainder term goes to zero faster than $\varepsilon$, we know there is some sufficiently small $\varepsilon_0$ for which the corresponding

$$y = x^* - \varepsilon_0 \nabla f(x^*)$$

satisfies

$$f(y) \leq f(x^*) - \frac{\varepsilon_0}{2} \|\nabla f(x^*)\|_2 < f(x^*).$$

This shows that $x^*$ is not a minimizer if $\nabla f(x^*) \neq 0$.

We conclude $x^*$ is a minimizer iff $\nabla f(x^*) = 0$

## Consequence

Solving the optimization problem (finding a minimizer)

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

is as hard/easy as solving the set of equations

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

Ex Linear regression (Easy)

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

$$\Updownarrow$$

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

$$\Updownarrow$$

$$X^TX\beta^* = X^Ty$$

$$\Updownarrow$$

$$\beta^* = (X^TX)^{-1}X^Ty$$

Ex (Requires more effort)

$$x^* = \min_{x \in \mathbb{R}^d} \|x\|_2^2 + \text{logsumexp}(x)$$

$$\Leftrightarrow \quad 2x + \text{softmax}(x) = 0$$

$$\Leftrightarrow \quad \text{softmax}(x) = -2x$$

To solve this nonlinear equation we'll use our intuition:

1) $-2x$ is a probability vector
2) the objective function gives the same value for a given $x$ or any permutation of the entries of that $x$, so we expect each coordinate is equally important to minimize, so $x^*$ will be constant

This implies $x^* = \frac{-1}{2d}$ is a solution. (verify this!)

Note the objective is strictly convex, so this is the only solution.

## Optimality conditions for (C)

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

E.g.,

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

However

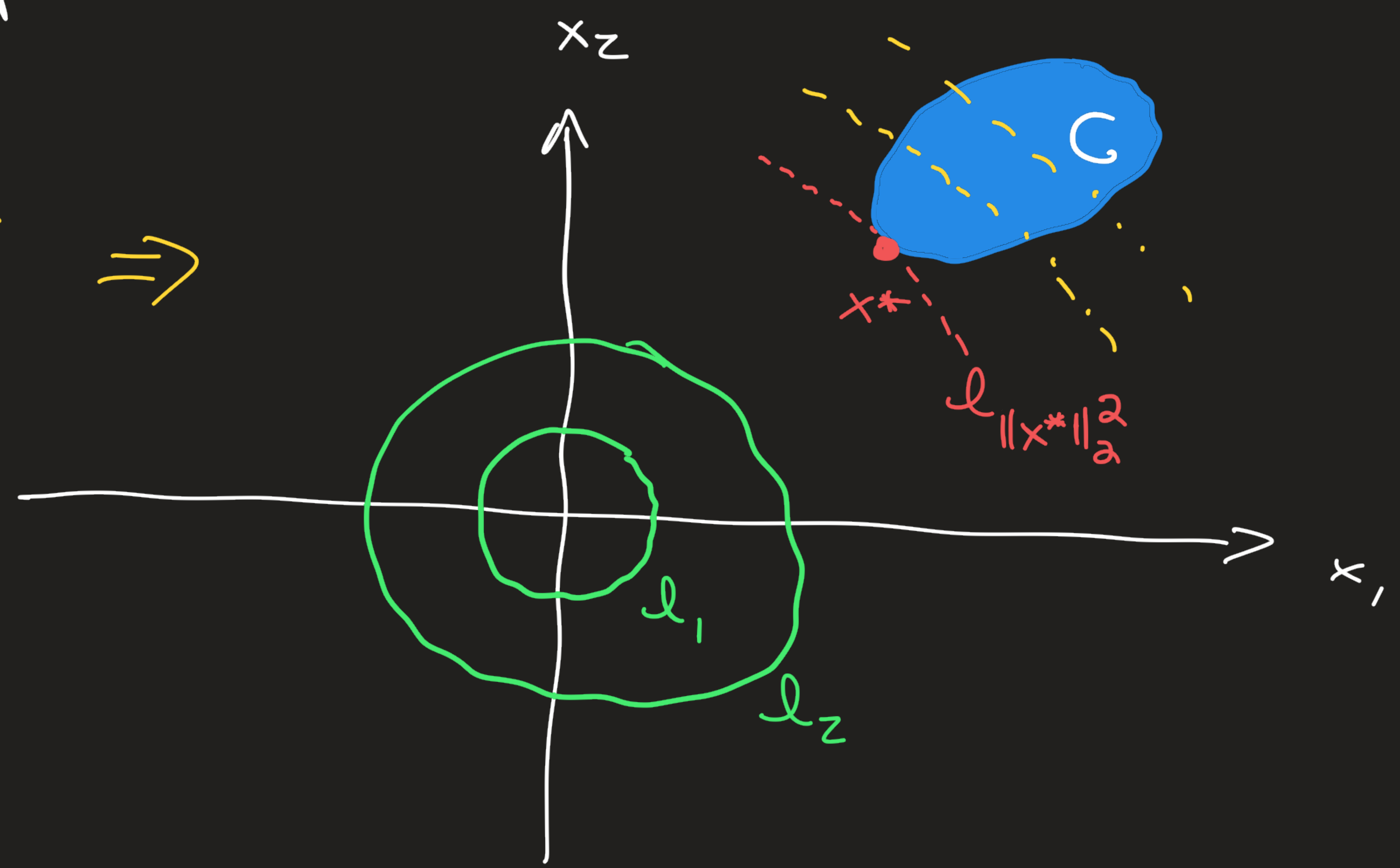
$$\underset{x\in[-1,1]}{\operatorname{argmin}}\ x^2 = 0, \quad \text{and} \quad \nabla_x x^2\Big|_{x=0} = 0$$

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

## Geometric Intuition

Consider the $\alpha$-level sets of $f$, $l_\alpha = \{x \mid f(x) = \alpha\}$, plot these for various values of $\alpha$, and see where they intersect $C$. E.g., if $f(x) = \|x\|_2^2$, we get the following plot

$$x^* = \min_{x \in C} \|x\|_2^2 \Rightarrow$$



Note that $\ell_{\|x^*\|_2^2}$ is the level set of lowest value that intersects $C$: if a level set with lower value intersected $C$, say $\ell_\gamma$, then that would mean there is a $y \in C$ for which $\|y\|_2^2 = \gamma < \|x^*\|_2^2$, which contradicts the minimality of $x^*$.

Now note that if I can move from $x^*$ into $C$ in a direction acute to $-\nabla f(x^*)$ then I can decrease the value of $f$, so move to a lower-valued level set (by the same T.S. argument we used earlier)

So our optimality condition is that $\langle -\nabla f(x^*), y - x^* \rangle \leq 0$ for all $y \in C$. Equivalently, $\langle \nabla f(x^*), y - x^* \rangle \geq 0$

Consider

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

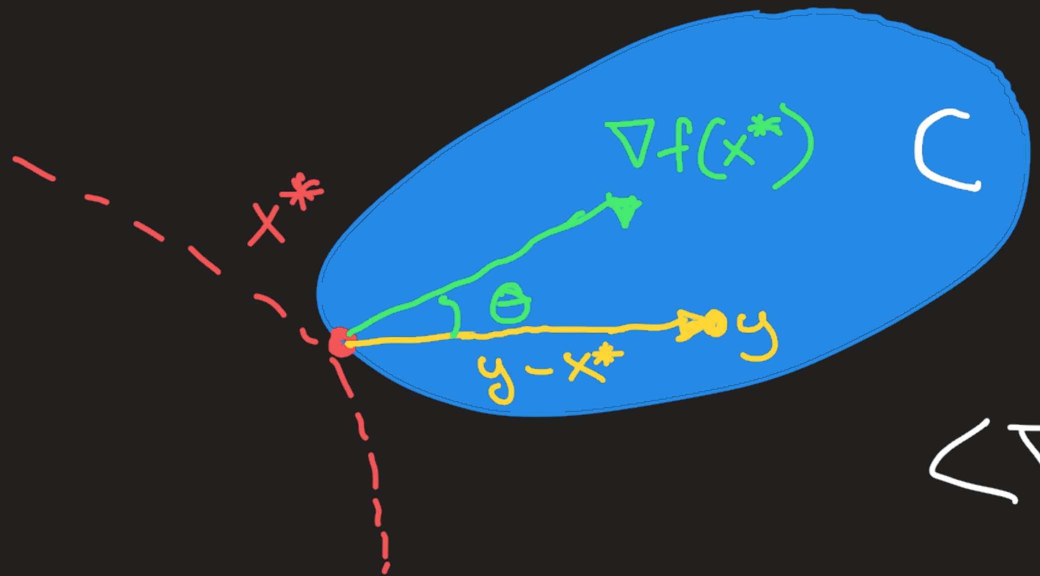
$x^*$ is a minimizer of $(C)$ iff every "feasible direction" has an acute angle with $\nabla f(x)$.

feasible directions are directions that move us from $x^*$ into $C$.

That is, $x^*$ is a minimizer of $(C)$ iff

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

Returning to our earlier example, notice this is the case visually:




we see if

$$x^* = \operatorname*{argmin}_{x \in C} \|x\|_2^2$$

then

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

Prf

Assume that it is the case that

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

Then for any $y \in C$,

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

so $x^*$ is a minimizer of $f$ on $C$.

Assume that there is a $y \in C$ for which

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

Then by a T.S. argument similar to the one from before, we can find a $z \in [x, y]$, $z = (1-\varepsilon)x^* + \varepsilon y = x^* + \varepsilon(y - x^*)$ by choosing $\varepsilon$ small enough, for which

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

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

$$\leq f(x^*) + \frac{\varepsilon_0}{2} \langle \nabla f(x^*), y - x^* \rangle$$

$$< f(x^*) \qquad (\text{because } \langle \nabla f(x^*), y - x \rangle < 0)$$

which contradicts the minimality of $x^*$.

Thus we conclude that $x^*$ is a minimizer of (C) iff

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

$\blacksquare$

Ex

Consider projection onto a convex set

$$P_C(x) = \underset{z \in C}{\operatorname{argmin}} \frac{1}{2}\|x - z\|_2^2 = f(z)$$

For convenience call $z^* = P_C(x)$. We have

$$\nabla f(z^*) = z^* - x$$

so the optimality condition is that

$$\langle z^* - x, z - z^* \rangle \geq 0 \quad \text{for all } z \in C$$

Ex

Consider $C = B_{\|\ \|_2}(0,1)$

Then the optimality condition is (from the last example)

$$\langle P_C(x) - x, z - P_C(x) \rangle \geq 0 \quad \text{for all } z \text{ s.t. } \|z\|_2 \leq 1$$

If $x \in C$, clearly $P_C(x) = x$ satisfies this condition.

If $x \notin C$, then $\|x\|_2 > 1$. Notice $\dfrac{x}{\|x\|_2} = 1$, so we have

here we use a guess & check method as in our unconstrained example

$$\left\langle P_C(x) - x, \frac{x}{\|x\|_2} - P_C(x) \right\rangle \geq 0$$

this is our guess for $P_C(x)$, and we are checking it

This implies

$$\left\langle P_C(x), \frac{x}{\|x\|_2} \right\rangle - \left\langle x, \frac{x}{\|x\|_2} \right\rangle - \langle P_C(x), P_C(x) \rangle + \langle x, P_C(x) \rangle \geq 0$$

Now $\|P_C(x)\|_2 = 1$ and $\langle x, \frac{x}{\|x\|_2} \rangle = \|x\|_2$, so

$$\frac{1}{\|x\|_2} \langle x, P_C(x) \rangle - \|x\|_2 - 1 + \langle x, P_C(x) \rangle \geq 0,$$

or equivalently

$$\left(1 + \frac{1}{\|x\|_2}\right) \langle x, P_C(x) \rangle \geq 1 + \|x\|_2,$$

or equivalently

$$\langle x, P_C(x) \rangle \geq \|x\|_2.$$

From the CS-Ineq, we know

$$\langle x, P_C(x) \rangle \leq \|x\|_2 \|P_C(x)\|_2 = \|x\|_2,$$

so we conclude that

$$\langle x, P_C(x) \rangle = \|x\|_2$$

Again by CS-Ineq we can conclude that $P_C(x) = \alpha x$ for some $\alpha \in \mathbb{R}$

Since

$$\langle x, P_C(x) \rangle = \langle x, \alpha x \rangle = \alpha \|x\|_2^2 = \|x\|_2$$

we have that $\alpha = \frac{1}{\|x\|_2}$ and $P_C(x) = \frac{x}{\|x\|_2}$.

To sum up, we have used the optimality condition of constrained smooth optimization to show that

$$P_C(x) = \underset{\|z\|_2^2 \leq 1}{\operatorname{argmin}} \ \frac{1}{2}\|x - z\|_2^2 = \begin{cases} z, & \text{if } z \in C \\ \frac{z}{\|z\|_2}, & \text{if } z \notin C \end{cases}$$

As you can see, even for simple optim problems, using the optimality conditions to find exact solutions is nontrivial.