ML and Optimization Lecture 10

- More examples of convex optimization problems
- Cauchy-Schwarz Inequality
- Optimality Conditions for Smooth Convex Optimization
  - Constrained
  - Unconstrained
\textbf{Ex} \quad \text{Projection onto a convex set}

\textit{Let } C \text{ be a convex set. Define the function } P_C \text{ that maps an arbitrary point } x \text{ to the closest point in } C \text{ as follows:}

\[ P_C(x) = \arg \min_{z \in C} \frac{1}{2} \| x - z \|^2 \]

\[ = \arg \min_{z \in C} \frac{1}{2} \| x \|^2 - 2 \langle x, z \rangle + \| z \|^2 \]

\[ = \arg \min_{z \in C} \| z \|^2 - 2 \langle x, z \rangle \]
Useful fact: \( f(x) = \|x\|_2^2 \) is strictly convex

Proof: (algebra, feel free to skip)

\[
\|\alpha u + (1-\alpha) v\|_2^2 = \alpha^2 \|u\|_2^2 + (1-\alpha)^2 \|v\|_2^2 + 2\alpha (1-\alpha) \langle u, v \rangle
\]

\[
= \alpha \|u\|_2^2 + (\alpha^2 - \alpha) \|u\|_2^2 + (1-\alpha) \|u\|_2^2
\]

\[
+ \left[ (1-\alpha)^2 - (1-\alpha) \right] \|v\|_2^2 + 2\alpha (1-\alpha) \langle u, v \rangle
\]

\[
= \alpha \|u\|_2^2 + (1-\alpha) \|v\|_2^2 + \alpha (\alpha-1) \|u\|_2^2 +
\]

\[
\alpha (\alpha-1) \|v\|_2^2 + 2\alpha (1-\alpha) \langle u, v \rangle
\]

\[
= \alpha \|u\|_2^2 + (1-\alpha) \|v\|_2^2 + \alpha (1-\alpha) \left[ 2 \langle u, v \rangle
\]

\[
- \|u\|_2^2 - \|v\|_2^2 \right]
\]

\[
= \alpha \|u\|_2^2 + (1-\alpha) \|v\|_2^2 - \alpha (1-\alpha) \|u-v\|_2^2
\]

\[
\langle \alpha \|u\|_2^2 + (1-\alpha) \|v\|_2^2
\]
It follows that \( f(z) = \frac{1}{2} \| x - z \|_2^2 \) is strictly convex
(Exercise: verify this!)

Therefore \( P_C(x) = \arg\min_{x \in C} f(z) \) is a well-defined function:

\( P_C \) maps \( x \) to the single unique point in \( C \) closest to \( x \).

**Ex** Let \( C = B_{\|\cdot\|_a}(0,1) = \{ z | \| z \|_a \leq 1 \} \)

We will show at end of lecture that

\[
P_C(x) = \begin{cases} 
  x & \text{if } \| x \|_2 \leq 1 \\
  \frac{x}{\| x \|_a} & \text{if } \| x \|_2 > 1
\end{cases}
\]
**Ex: Min-max regression**

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

\[
\beta^* = \arg\min_{\beta} \max_{i=1,\ldots,n} |y_i - \beta^T x_i|
\]

\[
= \arg\min_{\beta} \|X\beta - y\|_\infty
\]

Can see is convex b/c is composition of norm w/ affine function of \( \beta \).

To see this is convex write it as:

\[
f(\beta) = \max_{i=1,\ldots,n} f_i(\beta)
\]

where

\[
f_i(\beta) = |y_i - \beta^T x_i|
\]

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.

Where

\[
X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad \|v\|_\infty = \max_i |v_i|
Consider RERM of the form

$$\omega^* = \arg\min_{\omega \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} l_i(x) + \lambda J(\omega)$$

where we write $l_i(x)$ to indicate the loss of the model with parameters $\omega$ on the $i$th training point.

If $l_i$ is convex, $J$ is convex, and $\lambda \geq 0$, this is a convex optimization problem because positively weighted sum of convex functions is convex.
**Ex** logistic regression

Recall in this case the pointwise loss is given by

\[ l_i(\omega) = \log(1 + \exp(-y_i < x_i, \omega>)) \]

To see this is convex, write it as

\[ l_i(\omega) = \log(e^0 + e^{-y_i < x_i, \omega>}) \]

\[ = \log \text{sumexp} \left( \begin{bmatrix} 0^T \\ -y_i x_i^T \end{bmatrix} \omega \right) \]

We wrote \( l_i \) 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 \( x \) and \( y \) are parallel or anti-parallel (\( \exists \alpha \in \mathbb{R} : x = \alpha y \) )

\[\text{Prf (you can skip this)}\]

If \( x \) or \( y \) is 0, this is vacuously true.

It now suffices to show that if \( \|x\|_2 = \|y\|_2 = 1 \) then

\[
|\langle x, y \rangle| \leq 1
\]

because for arbitrary nonzero \( x \) and \( y \), we have that \( \|x\|_2 = 1 \) and \( \|y\|_2 = 1 \), so we have that

\[
|\langle \frac{x}{\|x\|_2}, \frac{y}{\|y\|_2} \rangle| \leq 1 \implies |\langle x, y \rangle| \leq \|x\|_2 \|y\|_2.
\]
So now assume that \( |u|_2 = |v|_2 = 1 \), then because

\[
|u + v|_2^2 = |u|_2^2 + 2<u, v> + |v|_2^2 \\
= 2 + 2<u, v> \geq 0
\]

we see that

\[
<u, v> \geq -1
\]

and similarly, because

\[
|u - v|_2^2 = |u|_2^2 - 2<u, v> + |v|_2^2 \\
= 2 - 2<u, v> \geq 0
\]

we see that

\[
<u, v> \leq 1
\]

so

\[
|<u, v>| \leq 1
\]

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

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

$$
\|x + y\|^2 = \|x\|^2 + 2 \langle x, y \rangle + \|y\|^2
= \|x\|^2 + \|y\|^2 + 2\|x\|\|y\|\cos\theta
$$

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

and we recover

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

the usual Pythagorean identity

This is a high-dimensional analog of the Law of Cosines. We define in high dimensions

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

By Cauchy-Schwarz, $\cos \theta \in [-1, 1]$ 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 that inner-products are preserved by rotations.

\[
\begin{align*}
\langle Rx, Ry \rangle &= (Rx)^T Ry = x^T R^T Ry = x^T y = \langle x, y \rangle
\end{align*}
\]

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

\[
\langle Rx, Ry \rangle = \langle x, y \rangle
\]

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 \( x \) 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 \( RR^T = R^TA = I \) and

\[
R \frac{x}{\|x\|_2} = \begin{bmatrix} \cos \gamma \cos \beta + \sin \gamma \sin \beta \\ \cos \gamma \sin \beta - \sin \gamma \cos \beta \end{bmatrix} = \begin{bmatrix} \cos(\beta - \delta) \\ \sin(\beta - \delta) \end{bmatrix}
\]
This implies that
\[
\langle x, y \rangle = \langle \frac{x}{\|x\|_2}, \frac{y}{\|y\|_2} \rangle
\]
\[
= \|x\|_2 \|y\|_2 \left\langle \frac{x}{\|x\|_2}, \frac{y}{\|y\|_2} \right\rangle
\]
\[
= \|x\|_2 \|y\|_2 \left\langle \frac{R}{\|x\|_2} \cdot \frac{y}{\|y\|_2}, \frac{R y}{\|y\|_2} \right\rangle
\]
\[
= \|x\|_2 \|y\|_2 \left\langle \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} \cos(\beta - \varphi) \\ \sin(\beta - \varphi) \end{bmatrix} \right\rangle
\]
\[
= \|x\|_2 \|y\|_2 \cos(\beta - \varphi)
\]
\[
= \|x\|_2 \|y\|_2 \cos \varphi
\]

so indeed we get that for vectors in \( \mathbb{R}^2 \)
\[
\frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2} = \cos \varphi
\]

so the CS-norm definition of cost between vectors coincides with the trig defn
Takeaway:

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

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

so there is an acute angle between $x$ and $y$.

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

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

Constrained

(\text{C}) \quad \min_{x \in C} f(x) \quad \text{and we require } f \text{ 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 solutions
Optimality Conditions for \( (U) \) (unconstrained)

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

In the case of convex functions 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 for $x^*$ is a minimizer iff all directions are acute to $-\nabla f(x^*)$. The only vector that is acute to all directions is zero.

\[ \nabla f(x^*) = 0, \text{ then for all } y \in \mathbb{R}^d, \]
\[ f(y) \geq f(x^*) + \langle \nabla f(x^*), y-x^* \rangle = f(x^*) \]

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

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

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

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)

$$\beta^* = \arg\min_\beta \|X\beta - y\|_2^2 \iff \nabla_\beta \|X\beta^* - y\|_2^2 = 0$$

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

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

$$\beta^* = (X^TX)^{-1}X^Ty$$
\( \text{Ex} \ (\text{Requires more effort}) \)

\[ x^* = \min_{x \in \mathbb{R}^d} \|x\|_2^2 + \log \sum \exp(x) \]

\[ \iff 2x + \text{softmax}(x) = 0 \]

\[ \iff \text{softmax}(x) = -2x \]

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.,

$$\arg\min_{x \in [-1, 1]} x^2 = 1$$, but $$\nabla_x x^2 \bigg|_{x=1} = 2 \neq 0$$

However

$$\arg\min_{x \in [-1, 1]} x^2 = 0$$, and $$\nabla_x x^2 \bigg|_{x=0} = 0$$

so we see the optimality conditions depend on the interaction of $f$ with $C$. 
Geometric Intuition

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

$x^* = \min_{x \in C} \|x\|_2$
Note that $\mathcal{L}_{\|x^*\|^2_0}$ is the level set of lowest value that intersects $C$: if a level set with lower value intersected $C$, say $\mathcal{L}_\gamma$, then that would mean there is a $y \in C$ for which $\|y\|^2_0 = \gamma < \|x^*\|^2_0$, 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

\[(C) \min_{x \in C} f(x)\]

\(x^*\) is a minimizer of \((C)\) iff every "feasible direction"

has an acute angle with \(\nabla f(x)\).

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 usually:

we see if

\[x^* = \arg\min_{x \in C} \|x\|_2\]

then

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

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 \( z \in [x, y] \), \( z = (1-\epsilon)x^* + \epsilon y = x^* + \epsilon(y-x^*) \)
by choosing \( \epsilon \) small enough, for which
\[ f(z) = f(x^*) + \langle \nabla f(x^*), z - x^* \rangle + o(\|z - x^*\|_2) \]

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

\[ \leq f(x^*) + \frac{\varepsilon_0}{\alpha} \langle \nabla f(x^*), y - x^* \rangle \]

\[ < f(x^*) \quad \text{(because} \quad \langle \nabla f(x^*), y - x^* \rangle < 0 \text{)} \]

which contradicts the minimality of \( x^* \).

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

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

Consider projection onto a convex set

\[ P_C(x) = \arg\min_{z \in C} \| x - z \|_2^2 = f(z) \]

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

\[ \nabla f(z^*) = z^* - x \]

so the optimality condition is that

\[ \langle z^* - x, z - z^* \rangle \geq 0 \quad \text{for all } z \in C \]
Consider $C = B_{1,1}((0,1))$

Then the optimality condition is (from the last example)

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

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

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

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

This implies

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

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

This is our guess for $P_c(x)$, and we are checking it.
Now \( \|P_c(x)\|_a = 1 \) and \( \langle x, \frac{x}{\|x\|_a} \rangle = \|x\|_a \), so

\[
\frac{1}{\|x\|_a} \langle x, P_c(x) \rangle - \|x\|_a - 1 + \langle x, P_c(x) \rangle \geq 0,
\]
or equivalently

\[
(1 + \frac{1}{\|x\|_a}) \langle x, P_c(x) \rangle \geq 1 + \|x\|_a,
\]
or equivalently

\[
\langle x, P_c(x) \rangle \geq \|x\|_a.
\]

From the CS-ineq, we know

\[
\langle x, P_c(x) \rangle \leq \|x\|_a \|P_c(x)\|_a = \|x\|_a,
\]
so we conclude that

\[
\langle x, P_c(x) \rangle = \|x\|_a.
\]

Again by CS-ineq we can conclude that \( P_c(x) = \alpha x \) for some \( \alpha \in \mathbb{R} \).
Since \( \langle x, P_c(x) \rangle = \langle x, \alpha x \rangle = \alpha \|x\|_a^2 = \|x\|_2^2 \)
we have that \( \alpha = 1 \) and \( P_c(x) = \frac{x}{\|x\|_2} \).

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

\[ P_c(x) = \arg\min_{\|z\|_a^2 \leq 1} \frac{1}{2} \|x - z\|_2^2 = \begin{cases} 
2, & \text{if } z \in C \\
\frac{z}{\|z\|_a}, & \text{if } z \notin C
\end{cases} \]

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