ML & Opt Lecture 8

- Cauchy-Schwarz Inequality
- More examples of convex optim prob
- Optimality Conditions for Smooth Conv Optim Probs
  - Constrained
  - Unconstrained
Ex

Projection onto convex set
- given a point $C$, find the closest pt in $C$ to $x$

$$P_C(x) = \arg\min_{z \in C} ||z - x||_2$$

$P_C(x)$ is called the projection of $x$ onto $C$, and is a well-defined function because the objective $z \mapsto ||z - x||_2$ is strictly convex in $z$
Fact: if $\nabla^2 f(x) > 0$ then $f$ is strictly convex

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

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

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

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

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

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

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

hence, strictly cvx w.r.t. $z$
We will show later in the lecture that if

\[ C = \mathcal{B}_{\| \cdot \|_a}(0, 1) = \{ z \mid \| z \|_a \leq 1 \} \]

then

\[ P_C(x) = \begin{cases} x & \text{if } x \in C \\ \frac{x}{\| x \|_a} & \text{if } x \notin C \end{cases} \]
**Cauchy-Schwarz Inequality**

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

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

or equivalently

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

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

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

$$\cos(\Theta \angle x, y) = \left\langle \frac{x}{\|x\|_2}, \frac{y}{\|y\|_2} \right\rangle.$$
Pythagorean Theorem

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

\[ \|x + y\|_2^2 = \langle x + y, x + y \rangle \]
\[ = x^T x + 2x^T y + y^T y \]
\[ = 0 \quad \text{b/c } \cos \theta(x, y) = 0 \]
\[ = \|x\|_2^2 + \|y\|_2^2 \]
\[ \Rightarrow \langle x, y \rangle = 0 \]

\[ \|y - x\|_2^2 = \quad x^T x + y^T y - 2x^T y \]
\[ = \|x\|_2^2 + \|y\|_2^2 - 2 \|x\|_2 \|y\|_2 \cos \theta \]

Law of Cosines
Types of Smooth Convex Optimization Problems

Let \( f \) be a differentiable convex function.

**Unconstrained**

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

**Constrained**

\[
(C) \quad \min_{x \in C} f(x) \quad \text{requires} \quad C \subseteq \text{dom}(f)
\]

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)

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

Proof:

$\subseteq$ Assume $\nabla f(x^*) = 0$. Because $f$ is convex, we have:

$f(y) \geq f(x^*) + \nabla f(x^*)^T(y - x^*) = f(x^*)$
\[ \Rightarrow \text{let } x^* \text{ be a minimizer.} \]

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

\[
f(y) = f(x^*) + \langle \nabla f(x^*), y - x^* \rangle \\
+ O(\|y - x^*\|_2^2) \\
= f(x^*) + \langle \nabla f(x^*), -\varepsilon \frac{\nabla f(x^*)}{\|\nabla f(x^*)\|_2} \rangle \\
+ O(\varepsilon^2 \left( \frac{\|\nabla f(x^*)\|_2^2}{\|\nabla f(x^*)\|_2^2} \right)) \\
= f(x^*) - \varepsilon \|\nabla f(x^*)\|_2 + O(\varepsilon^2)\]
So if \( \varepsilon < \| \nabla f(x^*) \|_2 \) then \( \varepsilon^2 < \varepsilon \| \nabla f(x^*) \|_2 \)

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

[Diagram showing a point \( x^* \) where \( \nabla f(x^*) \neq 0 \)]
**Consequences**

Solving the optimization problem

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

is as hard/easy as solving the set of equations

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

**Example Linear regression**

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

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

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

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

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

\[
\nabla_x (\| x^* \|_2^2 + \text{logsumexp}(x^*)) = 0
\]

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

Observations:

1) -2x* is a probability vector

2) Objective is strictly convex so x* is unique, and permutation invariant so any permutation of entries of x* is a minimizer \[ \Rightarrow x^* \text{ is a constant vector} \]

\[ x^* = c \cdot 1 \]
Because \( x^* = c \cdot 1 \) and

\[
\text{softmax}(x^*) = -2x^*;
\]

\[
\frac{d}{d} = -2c \cdot 1
\]

\[
\Rightarrow c = \frac{-1}{2d}
\]

\[
\Rightarrow x^* = \begin{bmatrix}
-\frac{1}{2d} \\
\vdots \\
-\frac{1}{2d}
\end{bmatrix}
\]
Optimality Conditions for (C)

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

$$\arg\min_{x \in [1, 2]} x^2 = 1,$$

but $\nabla x^2 \bigg|_{x=1} = 2 \neq 0$.

However,

$$\arg\min_{[-1, 1]} x^2 = 0,$$

and $\nabla x^2 \bigg|_{x=0} = 0$.

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

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

Consider the \( \alpha \) level sets of \( f \)

\[ L_{\alpha} = \{ x \mid f(x) = \alpha \} \]

The minimizer \( x^* \) must be on the level set of lowest value that intersects \( C \)

Notice that \( \nabla f(x) \perp L_{\alpha} \) at any point \( x \).
Imagine that there exists a $y \in C$ such that

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

then

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

$$+ o(\|y - x^*\|^2)$$

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

$$+ o(\text{(something negative)})^2)$$

$$\langle f(x^*) \rangle$$

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.,

$$H_{g-C} : \langle \nabla f(x^*), y - x^* \rangle \geq 0$$
Optimality Condition for (C)

\( x^* \) is a minimizer of

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

iff for all \( y \in C \),

\[
\langle \nabla f(x^*), y - x^* \rangle \geq 0
\]