MC and Optimization Lecture 10

- Projection onto $B_{11 \cdot 11}$ $(0,1)$ using Fermat's condition
- Separating and supporting hyperplane
- Subdifferentials and subgradients
- Extended value convex functions and convex indicator function
- Examples of computing subdifferentials
- Fermat's condition for optimality of non-smooth convex optimization
- Rules for manipulating subdifferentials
- Examples of computing subdifferentials with these rules
Ex. of applying Fermat's condition for constrained, smooth convex optimization

\[ x^* \in \arg\min_{x \in C} f(x) \iff \langle \nabla f(x^*), y - x \rangle > 0 \quad \text{for all } y \in C \]

Claim

If \( C = B_{\| \cdot \|_2}(0, 1) = \{ x \mid \| x \|_2 \leq 1 \} \), then

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

Proof:

Fermat's condition says \( z^* = P_C(x) \) satisfies

\[ \forall z \in C : \langle \frac{1}{2} \| x - z \|_2^2, z - z^* \rangle \geq 0 \]

i.e.

\[ \forall z \in C : \langle z^* - x, z - z^* \rangle \geq 0 \]
Consider first the case that \( x \in C \), then we see that taking \( z^* = x \) satisfies Fermat's condition:

\[
\langle z^* - x, z - z^* \rangle = \langle 0, z - z^* \rangle = 0
\]

so indeed if \( x \in C \) then \( P_C(x) = x \) as claimed.

Now consider the case that \( x \notin C \). Note that this means that \( \|z^*\|_{\alpha} = 1 \). Why? Assume \( \|z^*\|_{\alpha} < 1 \), then it is in the interior of \( C \).

This means that if we take \( \Delta = \epsilon(x - z^*) \) for \( \epsilon > 0 \) small enough, \( z^* + \Delta \) and \( z^* - \Delta \) are both in \( C \).
By Fermat's condition, this means

\[ \langle z^* - x, z^* + \Delta - z^* \rangle = \langle z^* - x, \Delta \rangle = -\epsilon \langle z^* - x, z^* - x \rangle = -\epsilon \| z^* - x \|^2 \geq 0 \]

This implies \( z^* - x = 0 \) or \( z^* = x \).

This contradicts the facts that \( x \notin C \) while \( z^* \in C \). Therefore we conclude that if \( x \notin C \), then \( \| z^* \|_2 = 1 \).

Now that we know \( \| z^* \|_2 = 1 \), we can guess and check that \( u = \frac{x}{\|x\|_2} \) is indeed equal to \( z^* \).

Recall that since \( u \in C \), Fermat's condition implies

\[ \langle z^* - x, u - z^* \rangle \geq 0 \]

or equivalently,

\[ \langle z^*, u \rangle = \langle x, u \rangle - \langle z^*, z^* \rangle + \langle x, z^* \rangle \geq 0 \]
Notice that:

1) \( \langle z^*, v \rangle = \frac{1}{\| x \|_1} \langle z^*, x \rangle \)

2) \( \langle x, v \rangle = \frac{1}{\| x \|_2} \langle x, x \rangle = \| x \|_2 \)

3) \( \| z^* \|_2^2 = 1 \)

so in fact

\( \frac{1}{\| x \|_2} \langle z^*, x \rangle - \| x \|_1 - 1 + \langle x, z^* \rangle \geq 0 \)

Re-arranging, \( \left( \frac{1}{\| x \|_2} + 1 \right) \langle x, z^* \rangle \geq 1 + \| x \|_2 \)

or equivalently \( \langle x, z^* \rangle \geq \| x \|_2 \)

But the C-S inequality gives

\( \langle x, z^* \rangle \leq \| x \|_2 \| z^* \|_2 = \| x \|_2 \)

so

\( \langle x, z^* \rangle = \| x \|_2 \| z^* \|_2 \)
Again by the (S− ineq.), this means $z^* = \alpha x$ for some $\alpha$. Since $\|z^*\|_2 = 1 = \alpha \|x\|_2$, we conclude that $\alpha = \frac{1}{\|x\|_2}$.

Thus, as claimed, $z^* = \frac{x}{\|x\|_2}$.

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

Fact (Convex Separation Thm)

Given a closed convex set $C$ and a point $x \notin C$, there exists a **separating hyperplane**, i.e., $a$ and $b$ s.t.

\[
\langle a, z \rangle + b < \langle a, x \rangle + b \quad \text{for all } z \in C
\]

This is a consequence of **Farkas’s condition** for $P_C(x)$. 
Proof

By Fermat's condition

∀z ∈ C : \langle P_c(x) - x, z - P_c(x) \rangle \geq 0

let a = x - P_c(x)

b = -\langle x - P_c(x), P_c(x) \rangle = -\langle a, P_c(x) \rangle

then Fermat's condition is equivalent to

∀z ∈ C : \langle -a, z - P_c(x) \rangle \geq 0

⇒ \langle a, z - P_c(x) \rangle \leq 0

⇒ \langle a, z \rangle + b \leq 0 \quad \text{for all } z \in C

Meanwhile:

\langle a, x \rangle + b = \langle a, x - P_c(x) \rangle = \|x - P_c(x)\|_a^2 > 0

because x ∈ C

so \langle a, z \rangle + b < \langle a, x \rangle + b \text{ for every } z \in C
Cor
Every closed convex set $C$ is the intersection of closed half-spaces of the form $a^Tz + b \leq 0$

Ex of convex duality

Cor (Supporting hyperplanes)
Given a point $x$ on the boundary of a convex set $C$, there exists a supporting hyperplane of the form $\langle a, z \rangle + b$ such that for all $z \in C$, $\langle a, z \rangle + b \leq \langle a, x \rangle + b$
Now we can generalize gradients of smooth convex functions to subgradients of non-smooth convex functions.

Claim: If $f$ is convex and differentiable at $x$, then

$a = (\nabla f(x), -1) \in \mathbb{R}^{d+1}$, \hspace{1cm} b = 0$

is a supporting hyperplane to $\text{epi}(f)$ at $(x, f(x))$. 
\textbf{Prf}

Consider any \((y, t) \in \text{epi}(f)\) then
\[ t \geq f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle \]
so
\[ \langle \nabla f(x), x \rangle - f(x) \geq \langle \nabla f(x), y \rangle - t \]
or equivalently,
\[ \langle (\nabla f(x), -1), (x, f(x)) \rangle \geq \langle (\nabla f(x), 1), (y, t) \rangle \]

This means the hyperplane
\[ a = (\nabla f(x), -1) \]
\[ b = 0 \]
supports \(\text{epi}(f)\) at \((x, f(x))\)
Even if \((\text{cuv})\) \(f\) is not differentiable at \(x\), because \(\text{epi}(f)\) is a convex set, it has supporting hyperplanes at \((x, f(x))\), of the form \((u > -1) \in \mathbb{R}^{d+1}\).

These give us the subgradients of \(f\). A vector \(u \in \mathbb{R}^d\) is a subgradient of \(f\) at \(x\) if

\[\forall y \in \text{dom}(f) : f(y) \geq f(x) + \langle u, y - x \rangle\]

Equivalently, \((u > -1)\) is a supporting hyperplane to \(\text{epi}(f)\) at \((x, f(x))\).
Notice there may be multiple subgradients. We define the subdifferential $\partial f(x)$ to be the set of all subgradients at $x$:

$$
\partial f(x) = \left\{ u \mid \forall g \in \text{dom}(f): f(g) \geq f(x) + \langle u, g - x \rangle \right\}
$$

Facts:

- $\partial f(x)$ is a convex set
- If $f$ is differentiable at $x$, $\partial f(x) = \left\{ \nabla f(x) \right\}$
- If $f$ is convex, $\partial f(x)$ is nonempty for every $x$ in the interior of dom $(f)$
Ex

\[ f(x) = |x| \]

\[ \text{sgn}(x) = \begin{cases} 
  1, & x > 0 \\
  -1, & x < 0 \\
  0, & x = 0 
\end{cases} \]

We abuse notation for convenience.

Ex

\[ f(x) = (1 - x)_+ \]

\[ \text{sgn}(x) = \begin{cases} 
  0, & x > 1 \\
  1, & 0 < x \leq 1 \\
  -1, & x \leq 0 
\end{cases} \]
Ex

\[ f(x) = \|x\|_2 \]  

Notice \( f(x) = \sqrt{\|x\|_2^2} \) so is differentiable (by the chain rule) everywhere except \( x = 0 \)

When \( x \neq 0 \), \( \nabla f(x) = \frac{1}{2\sqrt{\|x\|_2^2}} \cdot dx = \frac{x}{\|x\|_2} \)

When \( x = 0 \), we will use the definition of \( \partial f(0) \) to find the subdifferential

\[ u \in \partial f(0) \iff f(y) \geq f(0) + \langle u, y - 0 \rangle \]

\[ \iff \|y\|_2 \geq \langle u, y \rangle \]

By the CS-ineq, \( \langle u, y \rangle \leq \|y\|_2 \|u\|_2 \) so if \( \|u\|_2 < 1 \), \( u \in \partial f(0) \)

That is,

\[ B_{\|\cdot\|_2} (0, 1) \subseteq \partial f(0) \]
In fact, \( \partial f(0) = B_{\|\cdot\|_a}(0,1) \). To show this, it suffices to show that \( \|u\|_a > 1 \Rightarrow u \notin \partial f(0) \).

Assume \( \|u\|_a > 1 \), then take \( y = 0 \).

Observe that
\[
f(y) = \|y\|_a = \|u\|_a < \|u\|^2_a = \langle u, u \rangle = \langle u, y \rangle \]

so
\[
f(y) < f(0) + \langle u, y - 0 \rangle
\]
and we see that \( u \notin \partial f(0) \).

Therefore \( \partial f(0) = B_{\|\cdot\|_a}(0,1) \)

\[
\partial \|\cdot\|_a = \begin{cases} \frac{x}{\|x\|_a}, & x \neq 0 \\ B_{\|\cdot\|_a}(0,1), & x = 0 \end{cases}
\]
Extended value convex function

For convenience, given a convex function $f$ with $\text{dom}(f)$, we define

$$f(x) = \begin{cases} f(x) & x \in \text{dom}(f) \\ \infty & x \notin \text{dom}(f) \end{cases}$$

Let $C$ be a convex, define the convex indicator function

$$i_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$

Then

$$\arg\min_{x \in C} f(x) = \arg\min_{x \in C} f(x) + i_C(x)$$

So we convert constrained optimization into unconstrained optimization.
Ex. subdifferential of convex indicator functions

Let $x \in C$, then

$\forall u \in \partial i_C(x) \iff \forall y \in \text{dom}(i_C) : i_C(y) \geq i_C(x) + \langle u, y-x \rangle$

$\iff \forall y \in C : i_C(y) \geq i_C(x) + \langle u, y-x \rangle$

$\iff \forall y \in C : 0 \geq \langle u, y-x \rangle$

So we see

$\partial i_C(x) = \{ u \mid \forall y \in C, \langle u, y-x \rangle \leq 0 \}$
Fermat's Condition for Optimality

If $f$ is convex, then

$$x^* \in \arg\min_x f(x) \iff 0 \in \partial f(x^*)$$

**Proof**

$$\iff (0 \in \partial f(x^*) \implies x^* \in \arg\min_x f(x))$$

If $0 \in \partial f(x^*)$ then

$$\forall g \in \text{dom}(f): \quad f(g) \geq f(x^*) + \langle 0, g - x^* \rangle = f(x^*)$$

So $f(g) \geq f(x^*)$ for all feasible $g$. 
\[
\Rightarrow \ (x^* \in \text{argmin}_x f(x) \Rightarrow 0 \in \partial f(x^*)) \\

x^* \in \text{argmin}_x f(x) \Rightarrow \forall y \in \text{dom}(f) : f(y) \geq f(x^*) \\
\Rightarrow \forall y \in \text{dom}(f) : f(y) \geq f(x^*) + \langle o, y - x^* \rangle \\
\Rightarrow 0 \in \partial f(x^*)
\]
What about **constrained** non-smooth convex optimization?

**Fact.** (generalizes fact - \( \nabla (f + g)(x) = \nabla f(x) + \nabla g(x) \))

If \( f \) and \( g \) are extended value convex functions, then for all \( x \in \text{dom}(f) \cap \text{dom}(g) \)

\[
\nabla (f + g)(x) = \nabla f(x) + \nabla g(x)
\]

where the right-hand side is the Minkowski sum of two sets

\[
A + B = \{ z = u + v \mid u \in A, v \in B \}
\]

1 - satisfying the qualification condition \( \text{ri}(\text{dom}(f)) \cap \text{ri}(\text{dom}(g)) \neq \emptyset \)
Now we can see that

\[ \arg\min_{x \in C} f(x) = \arg\min_{x \in C} (f + \psi_C)(x) \]

so

\[ x^* \in \arg\min_{x \in C} f(x) \iff 0 \in \partial (f + \psi_C)(x^*) \]

Recall \( \partial \psi_C(x) = \{ u | \langle u, y - x \rangle \leq 0 \text{ for all } y \in C \} \)

so we have

\[ x^* \in \arg\min_{x \in C} f(x) \iff \text{There is a } u \in \partial f(x^*) \text{ such that} \]

\[ \langle u, y - x^* \rangle \geq 0 \text{ for all } y \in C \]
Fermat's Conditions

**Smooth, Unconstrained**

\[ x^* \in \arg\min_{x} f(x) \]
\[ \iff \nabla f(x^*) = 0 \]

**Non-smooth, Unconstrained**

\[ x^* \in \arg\min_{x} f(x) \]
\[ \iff 0 \in \partial f(x^*) \]

*easy to remember, all the others are special cases*
Subgradient Calculus Rules

- Multiplication by a positive scalar
  \[ \partial (af)(x) = a \partial f(x) \]

- Differentiability
  \[ f \text{ is differentiable at } x \text{ (and convex)} \Leftrightarrow \partial f(x) = \frac{1}{2} \nabla f(x) \]

- Sum rule
  \[ \partial \left( \sum_{i=1}^{m} f_i \right)(x) = \sum_{i=1}^{m} \partial f_i(x) \]

- Affine transformation rule
  \( h(x) = f(Ax + b) \) where \( f \) is convex \( \Rightarrow \)
  \[ \partial h(x) = A^T \partial f(Ax + b) \]
- Chain rule
  \( f \) - convex
  \( g \) - non-decreasing, differentiable, convex function at \( x \)
  \( \Rightarrow \ h = g \circ f \) (i.e. \( h(x) = g(f(x)) \))
  satisfies
  \( \nabla h(x) = g'(f(x)) \nabla f(x) \)

- Max rule
  \( h(x) = \max_{i=1, \ldots, m} \{ f_i(x), \ldots, f_m(x) \} \) where \( f_i \) are convex
  \( \Rightarrow \ nabla h(x) = \text{conv} \left( \bigcup_{i \in I(x)} \nabla f_i(x) \right) \)
  where \( I(x) = \{ i : f_i(x) = h(x) \} \)
\[ h(x) = (1-x)_+ = \max(1-x, 0) \]
- if \( x < 1 \) then \( h(x) = 1-x \), so by max-rule
  \[ \forall h(x) = 1 \]
- if \( x > 1 \) then \( h(x) = 0 \), so by max-rule
  \[ \forall h(x) = 0 \]
- if \( x = 1 \) then \( h(x) = 0 = \{ 1-x \} \), so since both functions achieve the max at \( x = 1 \), the max-rule gives
  \[ \forall h(1) = \text{conv}(\{ 1-x \}, 0) = \mathbb{R} \]
so
\[ \forall h(x) = \begin{cases} -1 & x < 1 \\ 0 & x > 1 \\ [-1, 0] & x = 1 \end{cases} \]
\[
\text{Ex}
\]
\[
f(\omega) = \frac{1}{n} \sum_{i=1}^{n} (1 - y_i x_i^T \omega) + \lambda \nabla \|\omega\|^2 \\
= \frac{1}{n} \sum_{i=1}^{n} f_i(\omega) + \lambda r(\omega)
\]

We use the sum rule and multiplication rule to see
\[
\nabla f(\omega) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\omega) + \lambda \nabla r(\omega)
\]

Consider a data-fitting term:
\[
f_i(\omega) = (1 - a_i^T \omega)_+ = h(a_i^T \omega) \text{ where } h(x) = (1 - x)_+ \text{ and } a_i = y_i x_i
\]

so by the affine transformation rule,
\[
\nabla f_i(\omega) = a_i \nabla h(a_i^T \omega) = y_i x_i \nabla h(a_i^T \omega)
\]
We just compute $\partial h(x)$, so
\[ \partial f_i(x) = g_i x_i \begin{cases} 0 & \text{if } g_i x_i^T \omega > 1 \\ -1 & \text{if } g_i x_i^T \omega < 1 \\ [1,0] & \text{if } g_i x_i^T \omega = 1 \end{cases} \]

Now compute the subdifferential of the regularizer:
\[ r(\omega) = \| \omega \|_1 = \sum_{i=1}^d |\omega_i| = \sum_{i=1}^d e_i^T \omega = \sum_{i=1}^d g(e_i^T \omega) \]
where $g(x) = |x|$.

By the sum-rule, affine transformation rule, and our knowledge of $\partial 1/1$:
\[ \partial r(\omega) = \sum_{i=1}^d e_i \partial g(e_i^T \omega) = \sum_{i=1}^d e_i \partial g(\omega_i) = \sum_{i=1}^d e_i \begin{cases} 1 & \omega_i > 0 \\ -1 & \omega_i < 0 \\ 0 & \omega_i = 0 \end{cases} \]
\[ = \sum_{i=1}^d e_i \text{ sign}(\omega_i) \]
Putting the pieces together, we see that

$$\nabla f(\omega) = \frac{1}{n} \sum_{i=1}^{n} y_i x_i \begin{cases} 0 & \text{if } y_i x_i^T \omega > 1 \\ -1 & \text{if } y_i x_i^T \omega < 1 \\ 0 & \text{if } y_i x_i^T \omega = 1 \end{cases}$$

$$+ \lambda \sum_{i=1}^{d} e_i \text{sgn}(\omega_i)$$