ML and Optimization Lecture 10

- Projection onto $B_{\|\cdot\|_2}(0,1)$ 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

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

Claim

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

$$P_C(x) = \operatorname*{argmin}_{z \in C} \|x - z\|_2^2 = \begin{cases} x & \text{if } x \in C \\ x/\|x\|_2 & \text{if } x \notin C \end{cases}$$

Pf

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

$$\forall z \in C: \langle \nabla_z \tfrac{1}{2}\|x - z^*\|_2^2, z - z^* \rangle \geq 0$$

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 $\|z^*\|_2 = 1$. Why? Assume $\|z^*\|_2 < 1$, then it is in the interior of C,

This means that if we take $\Delta = \varepsilon(x - z^*)$ for $\varepsilon > 0$ small enough, $z^* + \Delta$ and $z^* - \Delta$ are both in C.

By Fermat's condition, this means

$$\langle z^* - x, z^* + \Delta - z^* \rangle = \langle z^* - x, \Delta \rangle =$$

$$-\epsilon \langle z^* - x, z^* - x \rangle = -\epsilon \|z^* - x\|_2^2 \geq 0$$

this implies $z^* - x = 0$ or $z^* = x$

This contradicts the facts that $x \notin C$ while $z^* \in C$.

There 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) $\langle z^*, u\rangle = \frac{1}{\|x\|_2}\langle z^*, x\rangle$

2) $\langle x, u\rangle = \frac{1}{\|x\|_2}\langle x, x\rangle = \|x\|_2$

3) $\langle z^*, z^*\rangle = \|z^*\|_2^2 = 1$

so in fact

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

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 CS-ineq, this means $z^* = \alpha x$ for some α.

Since $\|z^*\|_2 = 1 = \alpha \|x\|_2$, we conclude that $\alpha = \frac{1}{\|x\|_2}$.

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

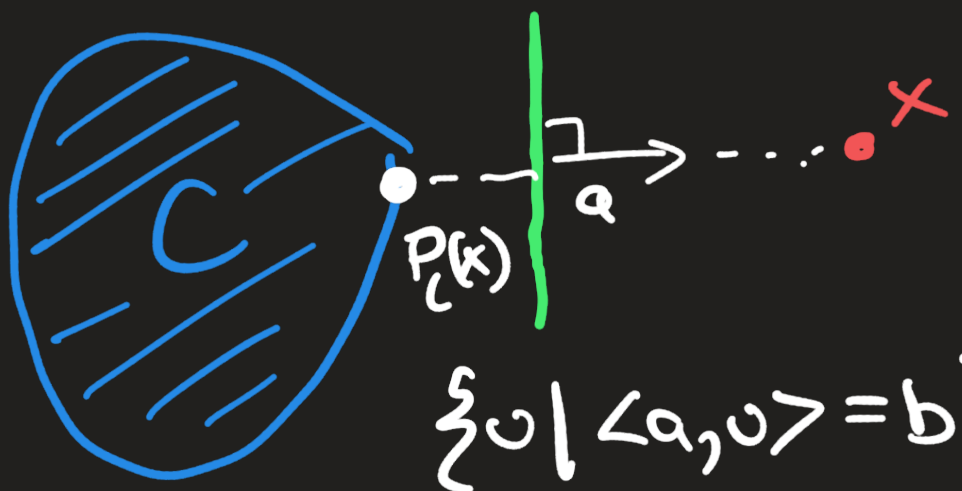
$$P_C(x) = \begin{cases} x & \text{if } x \in C \\ \frac{x}{\|x\|_2} & \text{if } x \notin C \end{cases}$$

Separating Hyperplanes

Fact (Convex Separation Thm)

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

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



$\{u \mid \langle a, u\rangle = b\}$

separating hyperplane

This is a consequence of Fermat's condition for $P_C(x)$

Prf

By Fermat's condition

$$\forall z \in C: \langle P_C(x) - x, z - P_C(x) \rangle \geq 0$$

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

$$\forall z \in C: \quad \langle -a, z - P_C(x) \rangle \geq 0$$

$$\Rightarrow \langle a, z - P_C(x) \rangle \leq 0$$

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

Meanwhile,

$$\langle a, x \rangle + b = \langle a, x - P_C(x) \rangle = \|x - P_C(x)\|_2^2 > 0$$

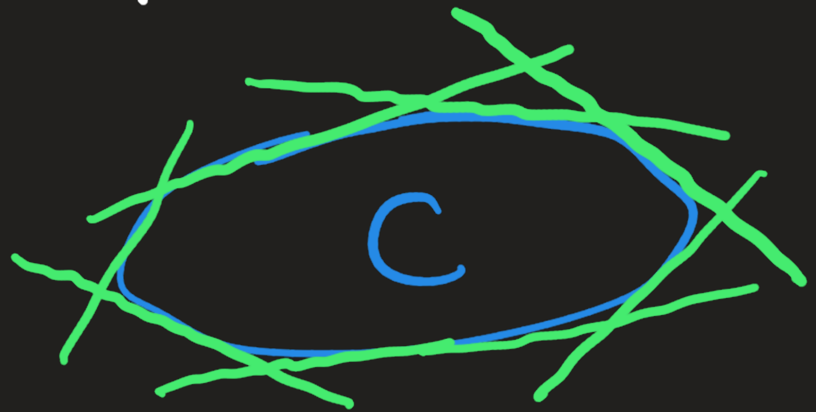
because $x \notin C$

so $\langle a, z \rangle + b < \langle a, x \rangle + b$ for every $z \in C$

$\blacksquare$

Cor

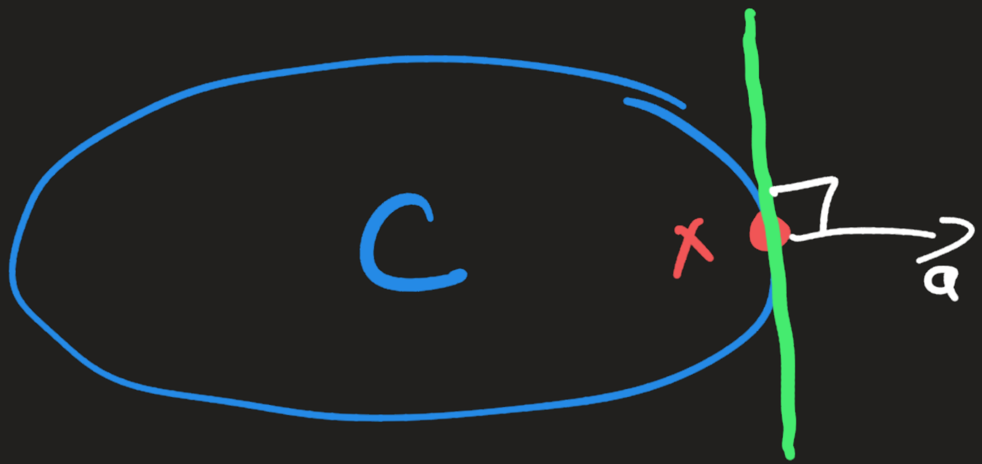
Every closed convex set C is the intersection of closed half-spaces of the form $a^T z + b \leq 0$



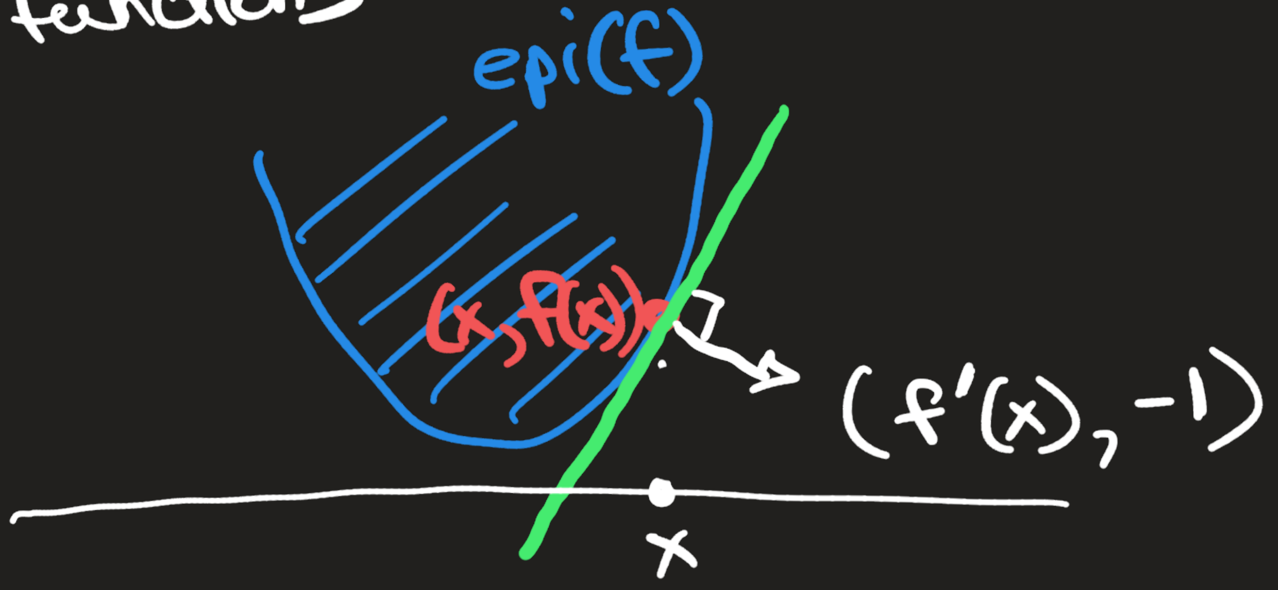
Ex of convex duality

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 <u>subgradients</u> of non-smooth convex functions



<u>Claim</u> If f is convex and differentiable at x, then

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

is a supporting hyperplane to epi(f) at $(x, f(x))$

Prf

Consider any $(y,t) \in \mathrm{epi}(f)$ 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

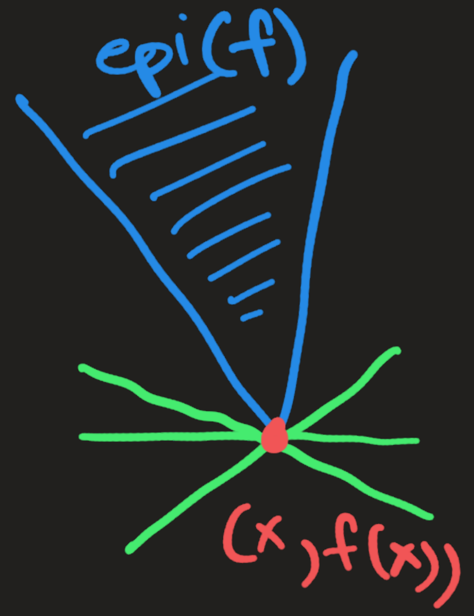
$a = (\nabla f(x), -1)$

$b = 0$

supports $\mathrm{epi}(f)$ at $(x, f(x))$

$\blacksquare$

Even if (cvx) f is not differentiable at x, because $\mathrm{epi}(f)$ 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 <u>subgradients</u> of f. A vector $u \in \mathbb{R}^d$ is a subgradient of f at x if

$$\forall y \in \mathrm{dom}(f): \quad f(y) \geq f(x) + \langle u, y - x \rangle$$

Equivalently, $(u, -1)$ is a supporting hyperplane to $\mathrm{epi}(f)$ at $(x, f(x))$

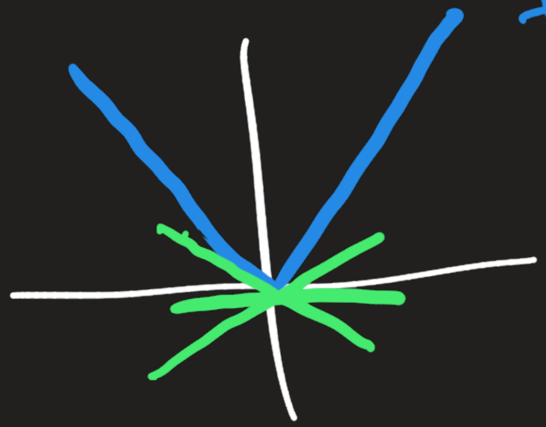
Notice there may be multiple subgradients.
We define the <u>subdifferential</u> $\partial f(x)$ to be the set of <u>all</u> subgradients at x:

$$\partial f(x) = \{ v \mid \forall y \in \mathrm{dom}(f) : f(y) \geq f(x) + \langle v, y - x \rangle \}$$

Facts

- $\partial f(x)$ is a convex set
- If f is differentiable at x, $\partial f(x) = \{\nabla f(x)\}$
- If f is convex, $\partial f(x)$ is nonempty for every x in the <u>interior</u> of $\mathrm{dom}(f)$

Ex



$f(x) = |x|$

$$\partial f(x) = \begin{cases} \{1\}, & x > 0 \\ \{-1\}, & x < 0 \\ [-1,1], & x = 0 \end{cases}$$

we abuse notation for convenience and say

$$\partial f(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \\ [-1,1] & x = 0 \end{cases} := \operatorname{sgn}(x)$$

Ex

$f(x) = (1-t)_+$

$$\partial f(x) = \begin{cases} 0, & x > 1 \\ [-1,0], & x = 1 \\ -1, & x < 1 \end{cases}$$

Ex

$f(x) = \|x\|_2$

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

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

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

$$u \in \partial f(0) \iff \forall y : f(y) \geq f(0) + \langle u, y - 0 \rangle$$
$$\iff \|y\|_2 \geq \langle u, y \rangle$$

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

That is,

$$B_{\|\cdot\|_2}(0,1) \subseteq \partial f(0)$$

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

Assume $\|u\|_2 > 1$, then take $y = u$.
Observe that

$$f(y) = \|y\|_2 = \|u\|_2 < \|u\|_2^2 = \langle u, u \rangle = \langle u, y \rangle$$

so

$$f(y) < f(0) + \langle u, y - 0 \rangle$$

and we see that $u \notin \partial f(0)$.

Therefore $\partial f(0) = B_{\|\cdot\|_2}(0,1)$

$$\partial \|x\|_2 = \begin{cases} \dfrac{x}{\|x\|_2}, & x \neq 0 \\ B_{\|\cdot\|_2}(0,1) & x = 0 \end{cases}$$

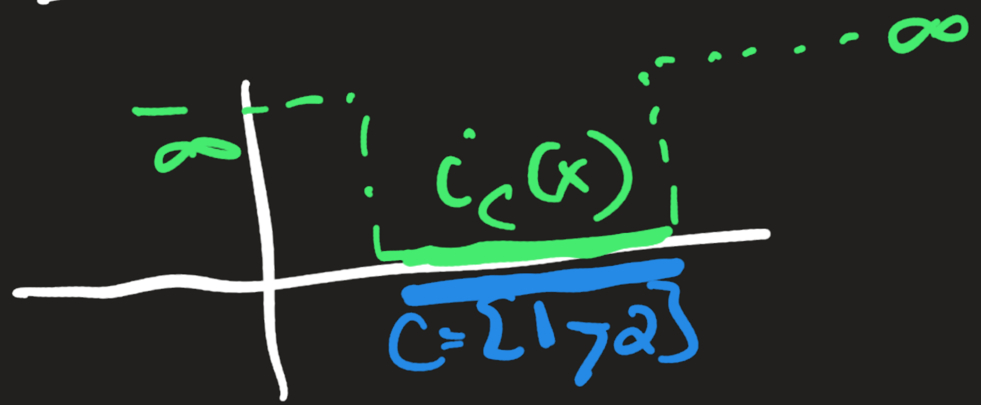
Extended value cvx function

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

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

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

$$\underset{x \in C}{\operatorname{argmin}}\, f(x) = \underset{x}{\operatorname{argmin}}\, f(x) + i_C(x)$$

So we convert constrained optimization into unconstrained optimization.

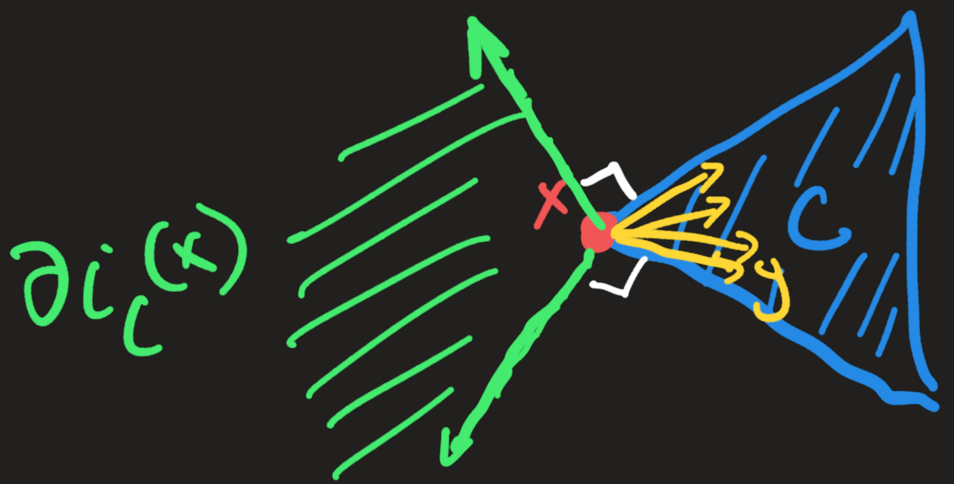
Ex subdifferential of convex indicator functions

Let $x \in C$, then

$$v \in \partial i_C(x) \iff \forall y \in \mathrm{dom}(i_C): \; i_C(y) \geq i_C(x) + \langle v, y-x \rangle$$
$$\iff \forall y \in C: \; i_C(y) \geq i_C(x) + \langle v, y-x \rangle$$
$$\iff \forall y \in C: \; 0 \geq \langle v, y-x \rangle$$

So we see

$$\partial i_C(x) = \{ v \mid \forall y \in C, \; \langle v, y-x \rangle \leq 0 \}$$



Fermat's Condition for Optimality

If f is convex, then

$$x^* \in \underset{x}{\operatorname{argmin}}\ f(x) \iff 0 \in \partial f(x^*)$$

Pf

$\Leftarrow$ $\left(0 \in \partial f(x^*) \Rightarrow x^* \in \underset{x}{\operatorname{argmin}}\ f(x) \right)$

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

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

so $f(y) \geq f(x^*)$ for all feasible y.

$\underline{\Rightarrow}$ $\left(x^* \in \operatorname*{argmin}_x f(x) \Rightarrow 0 \in \partial f(x^*)\right)$

$x^* \in \operatorname*{argmin}_x f(x) \Rightarrow \forall y \in \operatorname{dom}(f): \; f(y) \geq f(x^*)$

$\Rightarrow \forall y \in \operatorname{dom}(f): \; f(y) \geq f(x^*) + \langle 0, y - x^* \rangle$

so $0 \in \partial f(x^*)$

$\blacksquare$

What about <u>constrained</u> non-smooth cvx optimization?

<u>Fact</u> (generalizes fact $\nabla(f+g)(x) = \nabla f(x) + \nabla g(x)$)

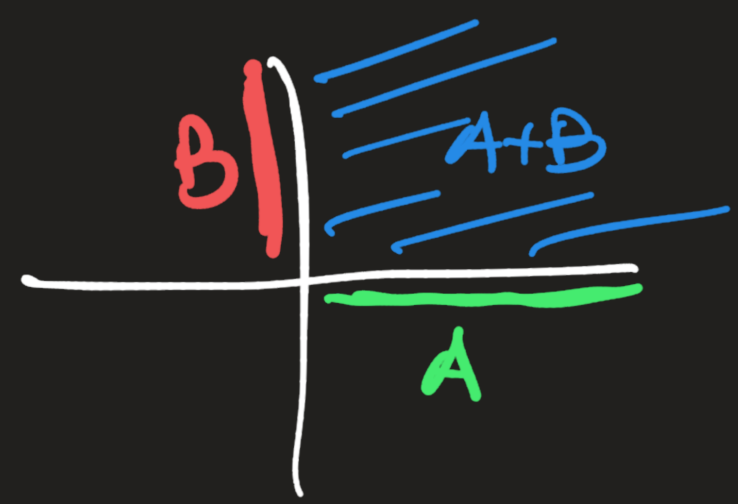
If f and g are extended value convex functions^{1}

then for all $x \in \mathrm{dom}(f) \cap \mathrm{dom}(g)$

$$\partial(f+g)(x) = \partial f(x) + \partial g(x)$$

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 $\mathrm{ri}(\mathrm{dom}(f)) \cap \mathrm{ri}(\mathrm{dom}(g)) \neq \emptyset$

Now we can see that

$$\underset{x \in C}{\operatorname{argmin}} f(x) = \underset{x}{\operatorname{argmin}} (f + i_C)(x)$$

so

$$x^* \in \underset{x \in C}{\operatorname{argmin}} f(x) \iff 0 \in \partial(f + i_C)(x^*)$$

$$\iff 0 \in \partial f(x^*) + \partial i_C(x^*)$$

Recall $\partial i_C(x) = \{ v \mid \langle v, y - x \rangle \leq 0 \text{ for all } y \in C \}$

so we have

$$x^* \in \underset{x \in C}{\operatorname{argmin}} f(x) \iff \text{There is a } v \in \partial f(x^*)$$

such that

$$\langle v, y - x^* \rangle \geq 0$$

for all $y \in C$

Fermat's Conditions

Smooth, Unconstrained

$$x^* \in \operatorname*{argmin}_x f(x)$$

$$\iff$$

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

Smooth, Constrained

$$x^* \in \operatorname*{argmin}_{x \in C} f(x)$$

$$\iff$$

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

for all $y \in C$

Non-smooth, Unconstrained

$$x^* \in \operatorname*{argmin}_x f(x)$$

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

easy to remember, all the others are special cases

Non-smooth, Constrained

$$x^* \in \operatorname*{argmin}_{x \in C} f(x)$$

$$\iff$$

$\exists\, u \in \partial f(x^*)$ such that

$$\langle u, y - x^* \rangle \geq 0$$

for all $y \in C$

Subgradient Calculus Rules

- Multiplication by a positive scalar

$\partial(af)(x) = a\,\partial f(x)$

- Differentiability

f is differentiable at x (and convex) $\Leftrightarrow \partial f(x) = \{\nabla f(x)\}$

- Sum-rule

$\partial\left(\sum_{i=1}^{m} f_i\right)(x) = \sum_{i=1}^{m} \partial f_i(x)$

under some qualification conditions we will always assume hold in this course

- 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 - nondecreasing, differentiable, convex function at x

$\Rightarrow$ $h = g \circ f$ (i.e. $h(x) = g(f(x))$)

satisfies

$$\partial h(x) = g'(f(x))\, \partial f(x)$$

- Max-rule

$h(x) = \max_{i=1,\dots,m} (f_i(x), \dots, f_m(x))$ where f_i are convex

$$\Rightarrow \partial h(x) = \operatorname{conv}\left(\bigcup_{i \in I(x)} \partial f_i(x) \right)$$

where $I(x) = \{ i : f_i(x) = h(x) \}$

Ex

$h(x) = (1-x)_+ = \max(1-x, 0)$

- if $x<1$ then $h(x) = 1-x$, so by max-rule

  $\partial h(x) = 1$
- if $x>1$ then $h(x) = 0$, so by max-rule

  $\partial h(x) = 0$
- if $x=1$ then $h(x) = 0 = \begin{cases} 1-x \\ 0 \end{cases}$, so since both functions achieve the max at $x=1$, the max-rule gives

  $\partial h(1) = \operatorname{conv}(\{-1, 0\}) = [-1, 0]$

so

$$\partial h(x) = \begin{cases} -1 & x<1 \\ 0 & x>1 \\ [-1,0] & x=1 \end{cases}$$

Ex

$$f(\omega) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(1-y_i x_i^T\omega)_+}_{f_i(\omega)} + \lambda \underbrace{\|\omega\|_1}_{r(\omega)}$$

$$= \frac{1}{n}\sum_{i=1}^{n} f_i(\omega) + \lambda r(\omega)$$

We use the sum rule and multiplication rule to see

$$\partial f(\omega) = \frac{1}{n}\sum_{i=1}^{n} \partial f_i(\omega) + \lambda \partial r(\omega)$$

Consider a data-fitting term:

$$f_i(\omega) = (1-a_i^T\omega)_+ = h(a_i^T\omega) \quad \text{where } h(x) = (1-x)_+ \text{ and } a_i = y_i x_i$$

So by the affine transformation rule,

$$\partial f_i(\omega) = a_i \partial h(a_i^T\omega) = y_i x_i \partial h(a_i^T\omega)$$

We just compute $\partial h(x)$, so

$$\partial f_i(\omega) = y_i x_i \begin{cases} 0 & \text{if } y_i x_i^T \omega > 1 \\ -1 & \text{if } y_i x_i^T \omega < 1 \\ [-1, 0] & \text{if } y_i x_i^T \omega = 1 \end{cases}$$

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 $\partial |x|$:

$$\partial r(\omega) = \sum_{i=1}^{d} e_i \partial g(e_i^T \omega) = \sum_{i=1}^{d} e_i \partial g(\omega_i) = \sum_{i=1}^{d} e_i \begin{cases} 1 & \omega_i > 0 \\ -1 & \omega_i < 0 \\ [-1, 1] & \omega_i = 0 \end{cases}$$

$$= \sum_{i=1}^{d} e_i \operatorname{sgn}(\omega_i)$$

Putting the pieces together, we see that

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

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