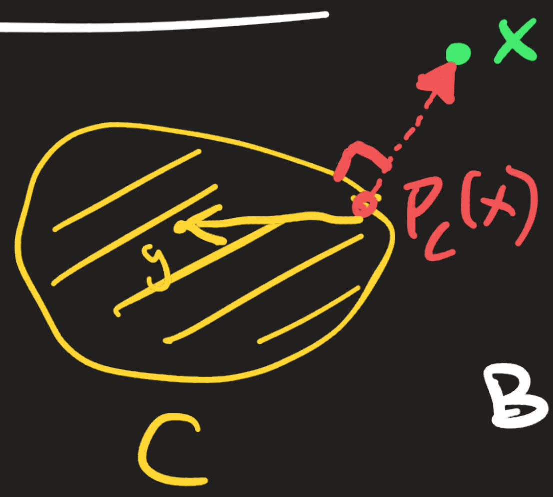
Lecture 9 Machine Learning and Optimization

- Properties of projection operator
- Show $P_C(x) = \begin{cases} x/\|x\|_2 & x \notin C \\ x & x \in C \end{cases}$ for C the Euclidean unit norm ball
- Subgradients and subdifferentials
- Fermat's condition for nonsmooth convex optimization

Last Time



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

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

By Fermat's condition for smooth constrained optimization:

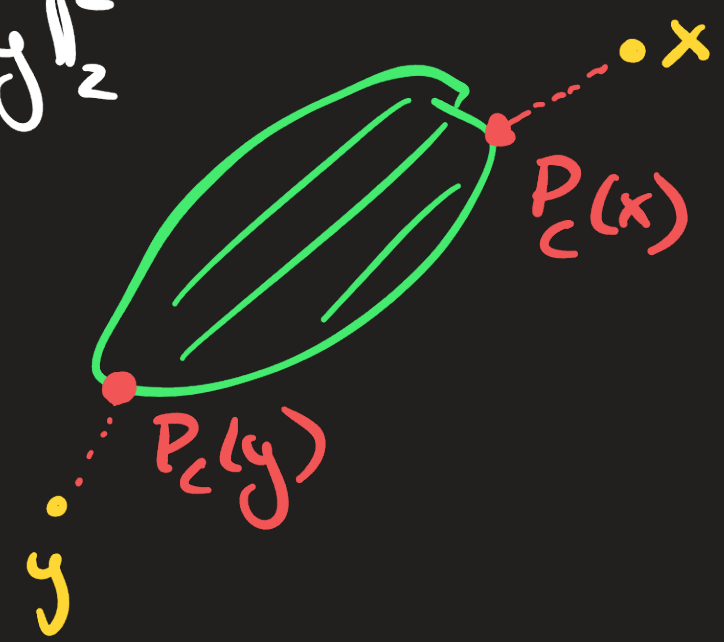
$$\forall y \in C: \ \langle y - P_C(x), \nabla f(P_C(x)) \rangle \geq 0$$

$$\iff \langle y - P_C(x), 2(P_C(x) - x) \rangle \geq 0$$

$$\iff \boxed{\langle y - P_C(x), P_C(x) - x \rangle \geq 0}$$

Projections are contractive

$$\| P_C(x) - P_C(y) \|_2^2 \leq \|x - y\|_2^2$$



Pf

WLOG assume $x \neq y$.

By optimality conditions

1) $\langle P_C(x) - P_C(y), P_C(y) - y \rangle \geq 0$

2) $\langle P_C(y) - P_C(x), P_C(x) - x \rangle \geq 0$

or equivalently

2') $\langle P_c(x) - P_C(y), x - P_C(x) \rangle \geq 0$

Adding 1) and 2') gives

$$\langle P_C(x) - P_C(y), x - y + P_C(y) - P_C(x) \rangle \geq 0$$

or equivalently

$$\langle P_C(x) - P_C(y), x - y \rangle - \| P_C(x) - P_C(y) \|_2^2 \geq 0$$

which implies

$$\langle P_C(x) - P_C(y), x - y \rangle \geq \| P_C(x) - P_C(y) \|_2^2$$

By C-S inequality,

$$\| P_C(x) - P_C(y) \|_2 \| x - y \|_2^2 \geq \langle P_C(x) - P_C(y), x - y \rangle$$
$$\geq \| P_C(x) - P_C(y) \|_2^2$$

Dividing both sides by $\| P_C(x) - P_C(y) \|_2$, we conclude that

$$\| P_C(x) - P_C(y) \|_2^2 \leq \| x - y \|_2$$

$$\nabla_z \|x-z\|_2^2 = \begin{bmatrix} \frac{\partial}{\partial z_1}\|x-z\|_2^2 \\ \vdots \\ \frac{\partial}{\partial z_d}\|x-z\|_2^2 \end{bmatrix}$$

$$\|x-z\|_2^2 = \sum_{i=1}^{d} (x_i - z_i)^2$$

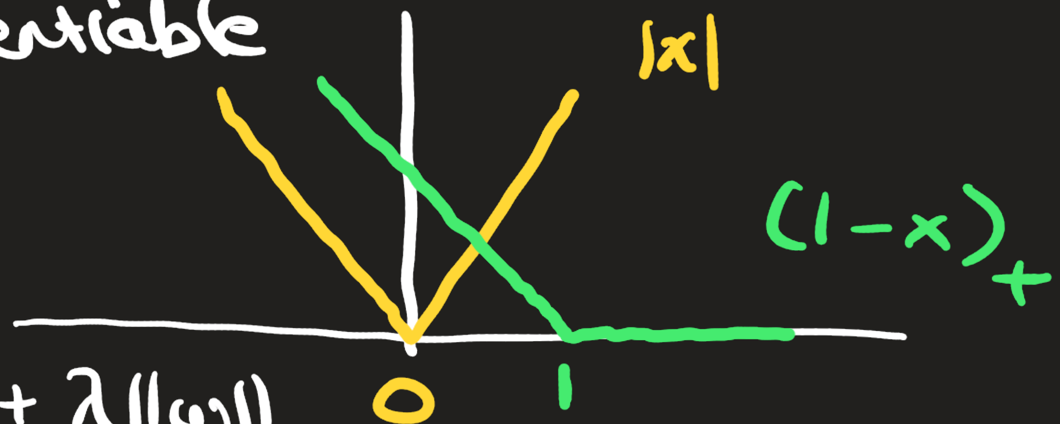
$$\nabla_x \|x\|_2^2 = 2x$$

$$= \begin{bmatrix} \sum_{i=1}^{d} \frac{\partial}{\partial z_1}(x_i-z_i)^2 \\ \vdots \\ \sum_{i=1}^{d} \frac{\partial}{\partial z_d}(x_i-z_i)^2 \end{bmatrix} = \begin{bmatrix} 2(z_1-x_1) \\ \vdots \\ 2(z_d-x_d) \end{bmatrix}$$

$$=2(z-x)$$

Non-smooth convex functions & subgradients

f is convex but not differentiable



Ex:

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

Recall the first-order characterization of smooth convex functions:

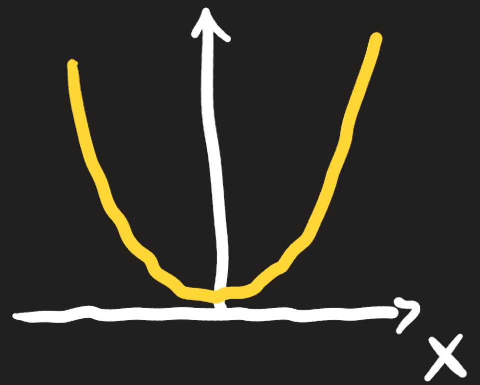
$$\forall y \in \mathrm{dom}(f): \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

Generalize to arbitrary convex functions by introducing the subdifferential

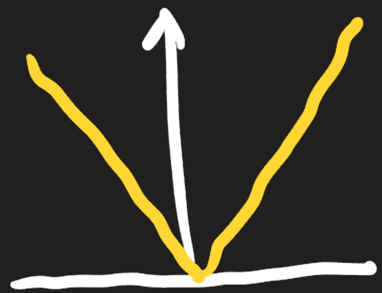
$$\partial f(x) := \{ v \mid f(y) \geq f(x) + \langle v, y-x \rangle \ \forall y \in \mathrm{dom}(f) \}$$

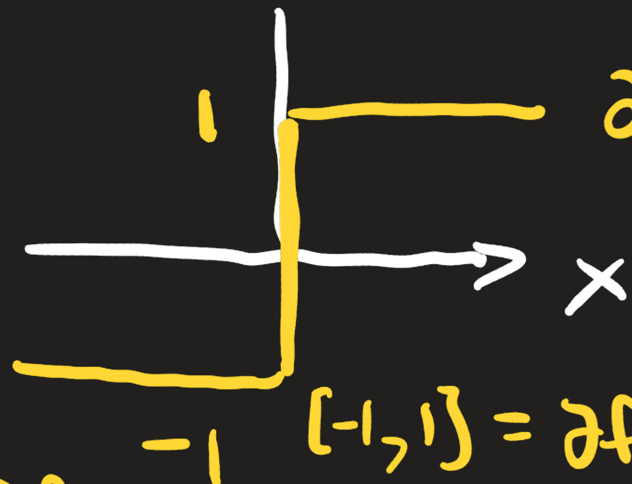
Ex:

$f(x)=x^2$



$f'(x)=2x$





$\partial f(x)=\{1\}$ when $x>0$

$[-1,1]=\partial f(0)$

$\partial f(x)=\{-1\}$ when $x<0$

<u>Facts</u> (let f be a convex function and $x \in dom(f)$)

- $\partial f(x)$ is a singleton iff f is differentiable at x, and

$$\partial f(x) = \{ \nabla f(x) \}$$

- If f is convex on $dom(f)$ and $x \in dom(f)$ is "nice" then $\partial f(x)$ is nonempty

Fermat's Condition for Optimality of Nonsmooth Convex Optimization Problems

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

Prf

$\Leftarrow$

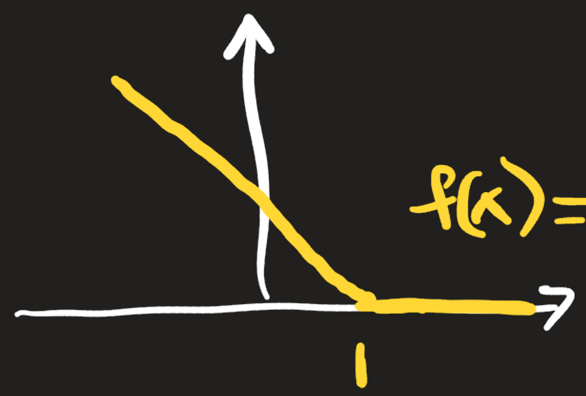
Since $0 \in \partial f(x^*)$ then for all $y \in \operatorname{dom}(f)$,

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

so x^* is a minimizer

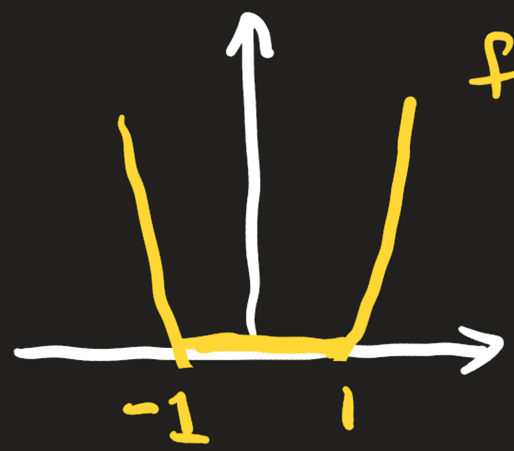
$\Rightarrow$

Since x^* is a minimizer $f(y) \geq f(x^*) + \langle 0, y - x^* \rangle$ for all y, so $0 \in \partial f(x^*)$



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

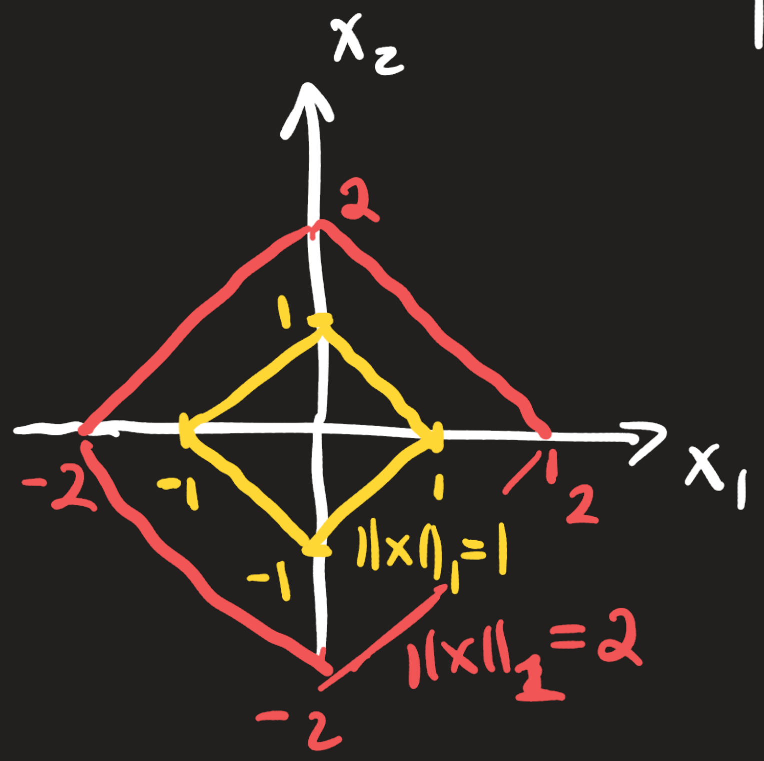
$$\partial f(x) = \begin{cases} -1 & x < 1 \\ 0 & x > 1 \\ [-1, 0] & x = 1 \end{cases}$$

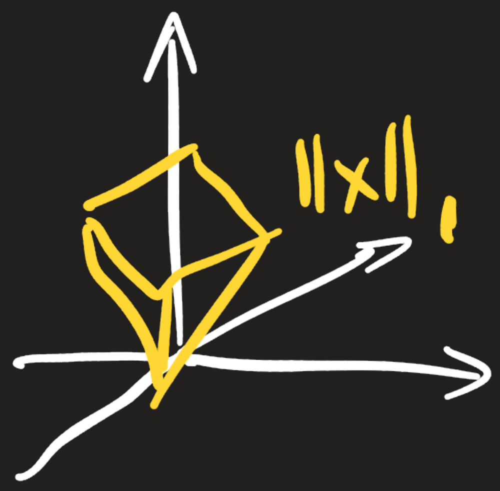


$f(x) = \max\{0, \frac{1}{2}(x^2 - 1)\}$

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

$$\|x\|_1 = |x_1| + |x_2|$$

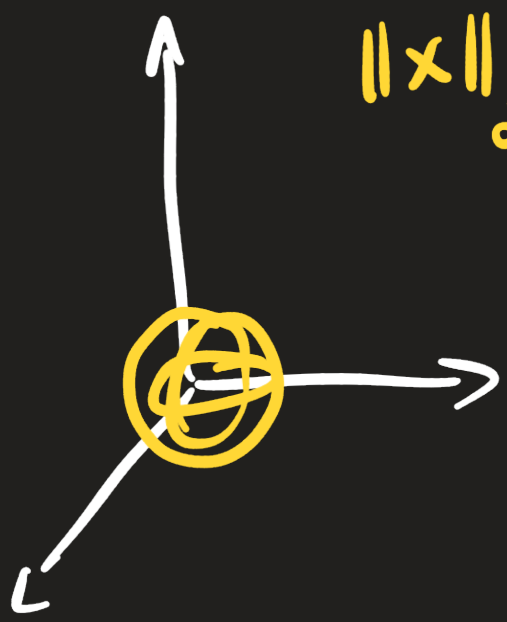




$$\partial \|x\|_1 = \left\{ v : \; v_i = \begin{cases} -1 & x_i < 0 \\ 1 & x_i > 0 \\ [-1,1] & x_i = 0 \end{cases} \right\}$$

$$= \left\{ v : \; v_i \in \operatorname{sgn}(x_i) \right\}$$

where $\operatorname{sgn}(z) = \begin{cases} 1 & z > 0 \\ -1 & z < 0 \\ [-1,1] & z = 0 \end{cases}$



$$\|x\|_2 = \sqrt{\|x\|_2^2}$$

$$\partial \|x\|_2 = \begin{cases} \frac{1}{2\|x\|_2}\cdot 2x = \frac{x}{\|x\|_2}, & x \neq 0 \\ B_{\|\cdot\|_2}(0,1), & x = 0 \end{cases}$$

By definition $\partial \|0\|_2 = \{ v : \|y\|_2 \geq \|0\|_2 + \langle v, y-0 \rangle \}$

$$= \{ v : \|y\|_2 \geq \langle v, y \rangle \}$$

By C-S in-eq.

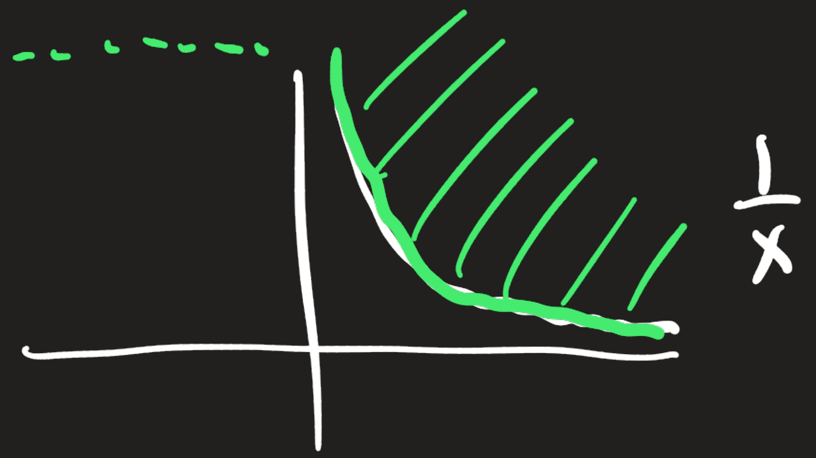
$$\langle v, y \rangle \leq \|y\|_2 \|v\|_2 \qquad \Rightarrow B_{\|\cdot\|_2}(0,1) \subseteq \partial \|0\|_2$$

so if $\|v\|_2 = 1$ then $\langle v, y \rangle \leq \|y\|_2$

Constrained Convex Optimization & Extended Value Convex Functions

Given a convex function $f: \mathbb{R}^d \to \mathbb{R}$ w/ dom(f) define its extended value version

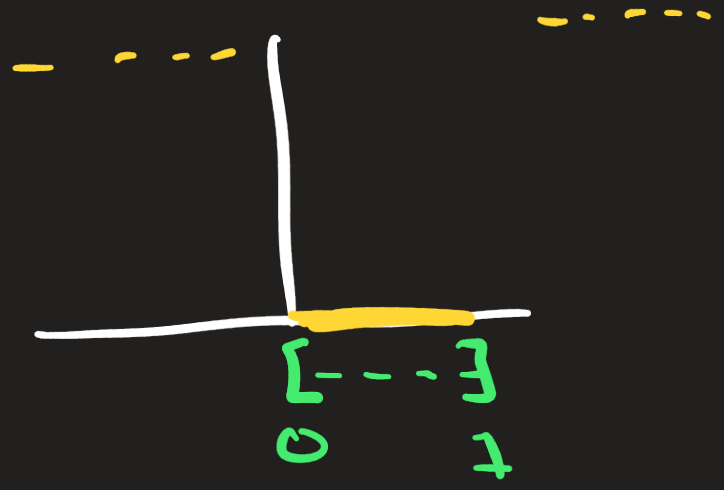
$$\bar{f}(x) = \begin{cases} f(x) & x \in \mathrm{dom}(f) \\ \infty & x \notin \mathrm{dom}(f) \end{cases} \qquad \bar{f}: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$$



We say f is proper if there is an x s.t. $f(x) \neq \infty$

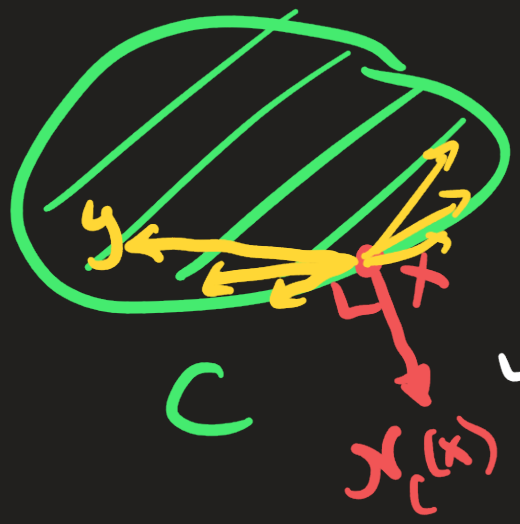
Let C be a closed convex set. We define its convex indicator function

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



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

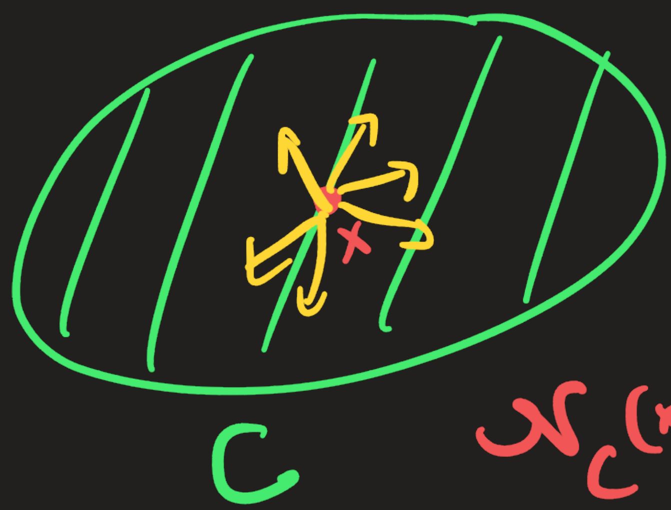
Subdifferential of convex indicator functions



$$\partial \iota_C(x) = \begin{cases} \emptyset, & x \notin C \\ N_C(x), & x \in C \end{cases}$$

where $N_C(x)$ is the collection of vectors that are normal to C at x.

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



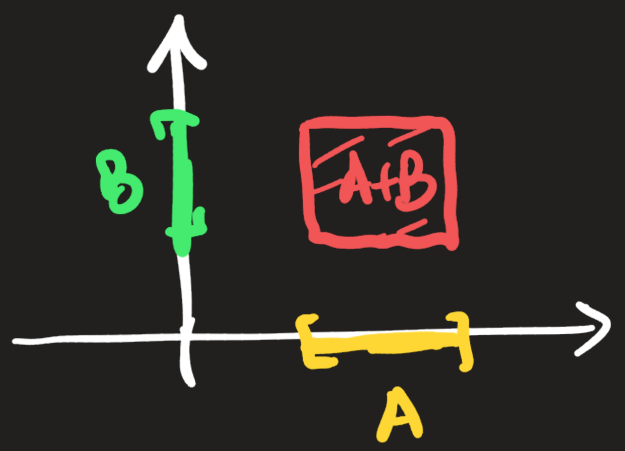
$N_C(x) = \{0\}$ if $x \in \text{int}(C)$

Fact

Sum-rule of subdifferential calculus

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

where we sum two sets $A + B = \{ v = a + b : a \in A, b \in B \}$



Applying Fermat's Condition

$$x^* \in \underset{x}{\operatorname{argmin}}\ f(x) \quad \text{where } f \text{ is differentiable}$$

$$\iff 0 \in \partial f(x^*) = \{\nabla f(x^*)\}$$

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

$$x^* \in \underset{x \in C}{\operatorname{argmin}}\ f(x) \quad \text{where } f \text{ is differentiable}$$

$$\iff \underset{x}{\operatorname{argmin}}\ f(x) + i_C(x)$$

$$\iff 0 \in \partial(f + i_C)(x^*) = \{\nabla f(x^*)\} + \mathcal{N}_C(x^*)$$

$$\iff \exists v \in \mathcal{N}_C(x^*) \text{ s.t. } 0 = \nabla f(x^*) + v$$

Recall v satisfies $\langle v, y-x^* \rangle \leq 0$ for all $y \in C$

so since $v = -\nabla f(x^*)$, we see

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

Weak & Strong Subgradient Calculus Rules

- Multiplication by a positive scalar

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

- Differentiability

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

- Sum rule of sub-differential calculus

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

- Affine transformation rule

$h(x) = f(Ax+b)$ where f is convex

$$\partial h(x) = A^T \partial f(Ax+b)$$

- Chain rule

  f - convex

  g - nondecreasing, differentiable, convex

  $$h = g(f(x))$$

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

- Max rule

  $$h(x) = \max_{i=1,\ldots,m} (f_1(x), \ldots, f_m(x))$$

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

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

Ex

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

work out $\partial f(\omega)$ as an exercise