ML & Opt Lecture 9

- Ex of the smooth optimality condition: $P_C(x)$ where $C = B_{\|\cdot\|_2}(0,1)$
- Subgradients, subdifferentials, and Fermat's condition for nonsmooth convex opt problem
- Subgradient Descent Algorithm for nonsmooth convex opt.
- Projected Subgradient Descent Alg for constrained nonsmooth cvx opt
- Rules for manipulating subgradients

Ex

Let $C = B_{\|\cdot\|_2}(0,1) = \{x \mid \|x\|_2^2 \leq 1\}$

$$P_C(x) = \underset{z \in C}{\operatorname{argmin}} \frac{1}{2}\|x - z\|_2^2$$

Claim:

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

when $C = B_{\|\cdot\|_2}(0,1)$

Prf

Fermat's optimality condition for the optimizer $z^* = P_C(x)$ is

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

$$\forall z \in C: \langle z^* - x, z - z^* \rangle \geq 0$$

If $x \in C$, then $z^* = x$ gives

$$\langle 0, z - x \rangle \geq 0$$

so $P_C(x) = x$ when $x \in C$

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

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

This implies

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

Since $z^* \in C$, $\|z^*\|_2^2 = 1$ and $\langle x, \frac{x}{\|x\|_2} \rangle = \frac{\|x\|_2^2}{\|x\|_2} = \|x\|_2$

so

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

or equivalently

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

or equivalently

$$\langle x, z^*\rangle \geq \|x\|_2$$

and by CS-Ineq,

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

so

$$\langle x, z^*\rangle = \|x\|_2 = \|x\|_2 \|z^*\|_2$$

so by CS-Ineq $z^* = \alpha x$ for some $\alpha \in \mathbb{R}$

Further,

$$\|z^*\| = 1 \quad \text{and} \quad \|z^*\| = |\alpha| \, \|x\|_2$$

so $|\alpha| = \dfrac{1}{\|x\|_2}$

conclude that $\alpha = \dfrac{1}{\|x\|_2}$ and $z^* = \dfrac{x}{\|x\|_2}$

$$\boxed{\begin{aligned} &\nabla_z \langle z, b\rangle = b \quad (*) \\ &\qquad \Vert \\ &\frac{\partial}{\partial z_j} \sum z_i b_i = b_j \end{aligned}}$$

$$\boxed{\nabla_z \left[z^T A^T A z\right] = 2A^T A z \quad (**)}$$

$$2A^T(Az-q)$$

$$\begin{aligned}
\nabla_z \|Az-q\|_2^2 &= \nabla_z \langle Az-q, Az-q\rangle \\
&= \nabla_z \left[z^T A^T A z - \underbrace{2q^T A z}_{(*)} + q^T q\right] \\
&= \nabla_z \left[z^T A^T A z - 2\langle z, A^T q\rangle\right] \\
&= \nabla_z \left[\underbrace{z^T A^T A z}_{(**)}\right] - 2A^T q = 2A^T A z - A^T q
\end{aligned}$$

$$x^T M y = $$

$$= \sum_{i=1}^{d} x_i (My)_i$$

$$= \sum_{i=1}^{d} x_i \left[\sum_{j=1}^{r} M_{ij} y_j \right]$$

$$= \sum_{i,j=1}^{d,r} M_{ij} x_i y_j$$

$$\nabla_z \{ z^T A^T A z \} = \nabla_z \left[\sum_{ij} (A^T A)_{ij} z_i z_j \right]$$

$$\left(\nabla_z \{ z^T A^T A z \} \right)_k = \frac{\partial}{\partial z_k} \left[\sum_{ij} (A^T A)_{ij} z_i z_j \right]$$

$$= \frac{\partial}{\partial z_k} \left[\sum_{\substack{j=1 \\ j \neq k}}^{d} (A^T A)_{kj} z_k z_j + \sum_{\substack{i=1 \\ i \neq k}}^{d} (A^T A)_{ik} z_i z_k \right.$$

$$\left. + (A^T A)_{kk} z_k^2 \right]$$

$$= \sum_{\substack{j=1 \\ j \neq k}}^{d} (A^T A)_{kj} z_j + \sum_{\substack{i=1 \\ i \neq k}}^{d} (A^T A)_{ik} z_i + 2 (A^T A)_{kk} z_k$$

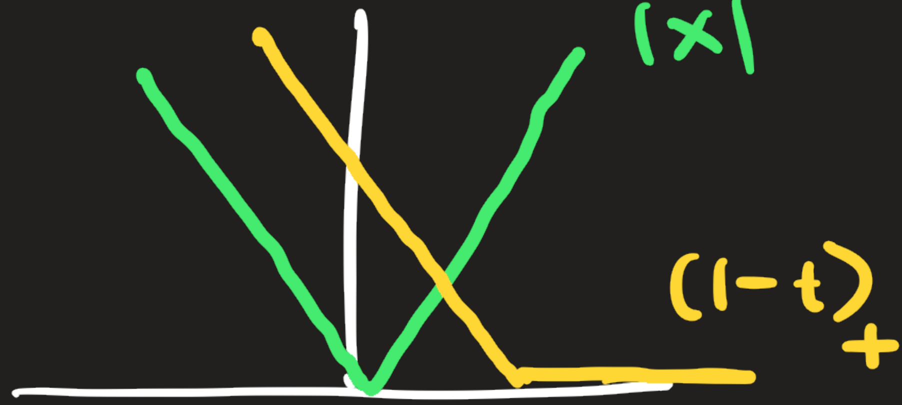
$$= \sum_{j=1}^{d} (A^T A)_{kj} z_j + \sum_{i=1}^{d} (A^T A)_{ik} z_k$$

$$= 2 \sum_{j=1}^{d} (A^T A)_{kj} z_j = 2 (A^T A z)_k$$

$$\nabla_z (z^T A^T A z) = 2 A^T A z$$

Non-smooth convex optimization

f is cvx but not differentiable



Ex:

$$f(\omega) = \frac{1}{n}\sum_{i=1}^{n}(1 - y_i\omega^T x_i)_+ + \lambda\|\omega\|_2$$

Introduce a subdifferential

$$\partial f(\omega) := \{ v \mid f(y) \geq f(\omega) + \langle v, y-\omega\rangle \}$$

Facts

- If f is differentiable at ω, $\partial f(\omega) = \{\nabla f(\omega)\}$
- If f is convex on $\mathrm{dom}(f)$ and $\omega \in \mathrm{dom}(f)$ is "nice" then $\partial f(\omega)$ is non-empty

Fermat's Condition for Optimality of Nonsmooth Convex Optim Probs

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

Prf

$\Leftarrow$

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

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

so x^* is a minimizer

$\Rightarrow$

Since x^* is a minimizer, for all $y \in \mathrm{dom}(f)$,

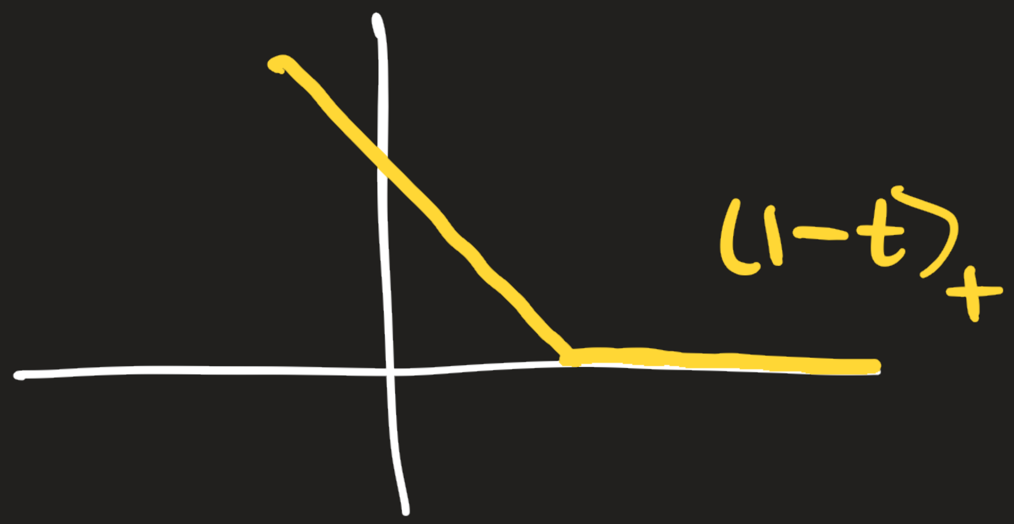
$$f(y) \geq f(x^*) = f(x^*) + \langle 0, y - x^* \rangle$$ so $0 \in \partial f(x^*)$

Ex

$$\partial |x| = \begin{cases} \{1\} & \text{if } x > 0 \\ \{-1\} & \text{if } x < 0 \\ [-1,1] & \text{if } x = 0 \end{cases}$$

Check that $[-1,1] \subseteq \partial|0|$: pick $v \in [-1,1]$ then for any $y \in \mathbb{R}$ we have

$$f(y) = |y| \geq v \cdot y = 0 + v(y-0)$$
$$= f(0) + \langle v, y-0 \rangle$$



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

Ex

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

by Cauchy-Schwarz Ineq

Extended value cvx fcns

We allow cvx fcns to take value ∞

Let C be a convex set, define the convex indicator

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

Ex: check that i_C is a convex func w/ dom(f)

Note

$$\operatorname*{argmin}_{x \in C} f(x) \iff \operatorname*{argmin}_{x} f(x) + i_C(x)$$

Fermat's Condition

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

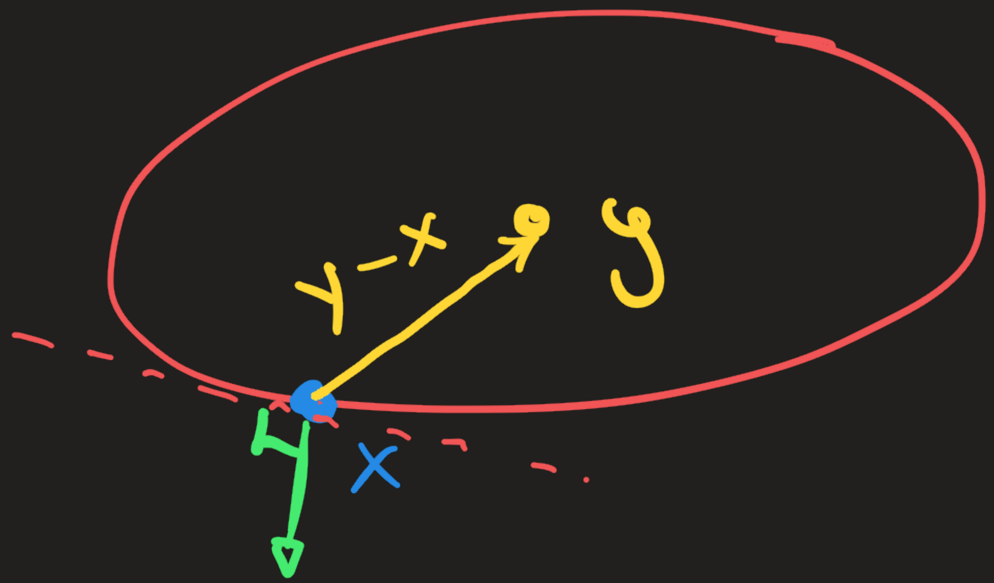
Two things we need for this to be useful:

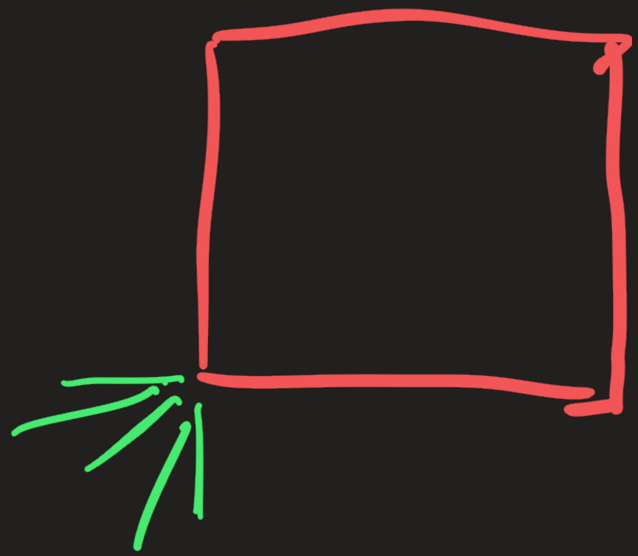
1) What is ∂i_C?

2) What are some rules for manipulating subdifferentials? E.g. $\partial(f+g)$

Subdifferential of convex indicators

Claim: $\partial i_C(x) = \begin{cases} \emptyset & \text{if } x \notin C \\ \{v : \langle v, y-x \rangle \le 0 \text{ for any } y \in C\} \end{cases}$





Weak & Strong Subgradient Calculus Rules

- Multiplication by a positive scalar

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

- Differentiability

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

- Sum rule of sub-differentiable calc.

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

- Affine transformation rule

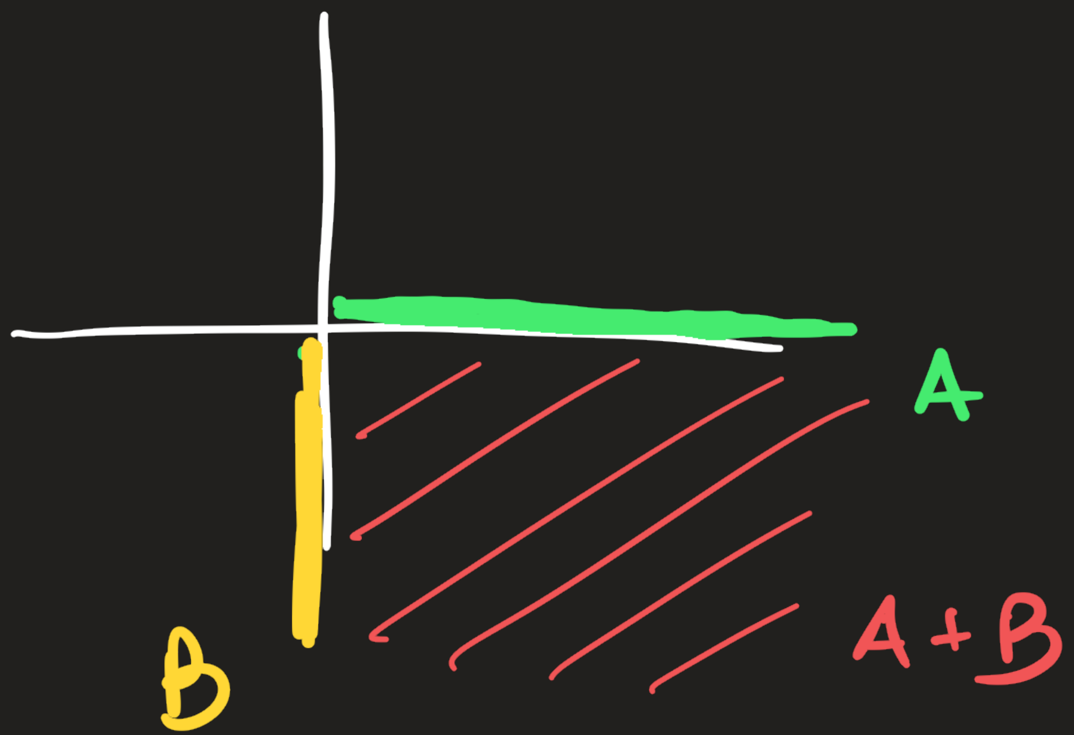
  $h(x) = f(Ax+b)$ where f is convex

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

if A & B are sets, then

$$A+B = \{ v : v = a+b \text{ for some } a \in A, b \in B \}$$

"Minkowski sum"



$$A = \{ v : v = \lambda e_1, \lambda \geq 0 \}$$

$$B = \{ v : v = \lambda e_2, \lambda \leq 0 \}$$

- Chain rule of subdiff calculus

  f – convex

  g – nondecreasing, differentiable, convex

$h = g \circ f \quad (h(x) = g(f(x)))$

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

- Max rule

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

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

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

Ex

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

To find $\partial f(w)$, apply multiplication & sum rules:

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

Notice

$\|w\|_1 = \sum_i |w_i|$ so by the sum rule

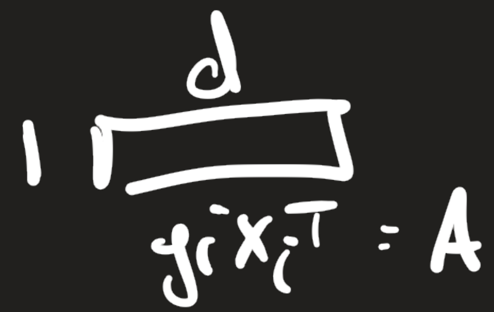
$$\partial\|w\|_1 = \sum_i \partial |w_i|$$

$$= \sum_{i=1}^{d} \left\{ \nu e_i : \nu \in \begin{cases} -1 & \text{if } w_i < 0 \\ 1 & \text{if } w_i > 0 \\ [-1,1] & \text{if } w_i = 0 \end{cases} \right\}$$

we consider $|w_i|$ as functions $w \mapsto |w_i|$

so $\partial\|\omega\|_1 = \{ v : v_i \in \operatorname{sign}(\omega_i) \}$

where $\operatorname{sign}(x) = \begin{cases} -1 & \text{if } x<0 \\ 1 & \text{if } x>0 \\ [-1,1] & \text{if } x=0 \end{cases}$



Consider one of the data terms $h_i(\omega) = (1 - y_i x_i^T \omega)_+$

Use the affine transformation rule

$$\partial h_i(\omega) = -y_i x_i \cdot \partial p(z)\Big|_{z = 1 - y_i x_i^T \omega} \quad \text{where } p(z) = z_+$$

$$= -y_i x_i \begin{cases} 0 & 1 - y_i x_i^T \omega < 0 \\ 1 & 1 - y_i x_i^T \omega > 0 \\ [0,1] & 1 = y_i x_i^T \omega \end{cases}$$