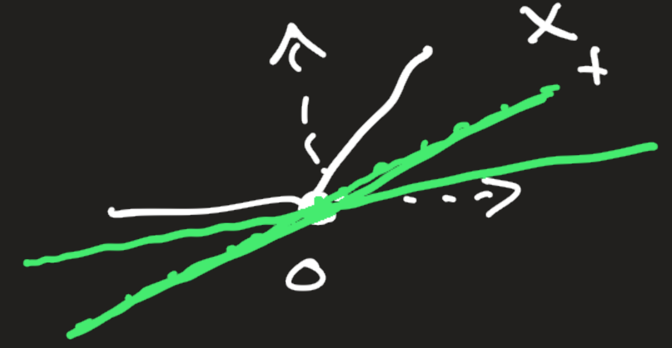
ML & Optim Lect 12

- Optimality Conditions for non-smooth cvx optim
- Subdifferentials & subgradients
  - Some rules for manipulating subdifferentials
- Projected Subgradient Descent

Non-smooth convex optimization



f is cvx but not differentiable

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

In this case we have no gradient, so can't move in $-\nabla f(x_t)$ to minimize f

Soln: move in the negative direction of a <u>subgradient</u>

g is a subgradient to f at point x iff

$$\forall y : \quad f(y) \geq f(x) + \langle g, y - x \rangle$$

The collection of subgradients at x is called the <u>subdifferential</u> $\partial f(x) = \{ g : \forall y \; f(y) \geq f(x) + \langle g, y - x \rangle \}$

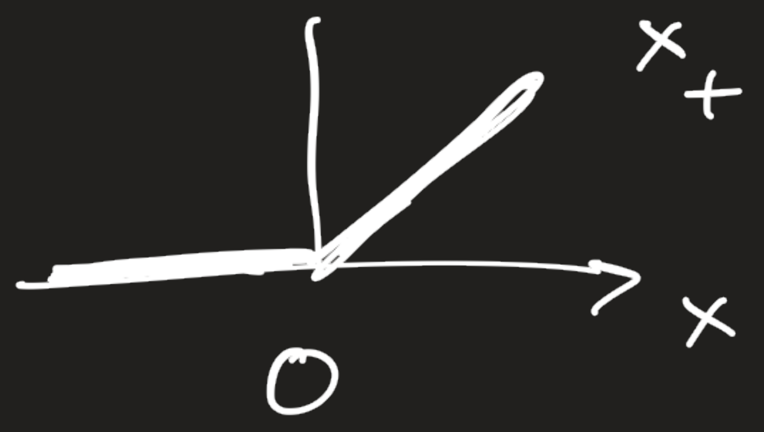
Fact

For convex functions, the subdifferential is non-empty (on the interior of the domain), and if f is differentiable at x, then

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

Ex

$f(x) = x_+$



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

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

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

Let $g \in [0,1]$ then to show $g \in \partial f(0)$ we must establish that $\forall y \in \mathbb{R}$

$$f(y) \geq f(0) + g(y-0)$$

$$\iff$$

$$y_+ \geq gy \qquad (*)$$

Two case:

(i) $y \leq 0$: then $y_+ = 0$ and $gy \leq 0$ so $(*)$ holds

(ii) $y > 0$: then $y_+ = y \geq gy$ b/c $g \in [0,1]$ so $(*)$ holds

– Show that $\partial f(0) \subseteq [0,1]$

i.e. if $g \notin [0,1]$ then $g \notin \partial f(0)$

i.e. if $g \notin [0,1]$ then there exists a $y \in \mathbb{R}$

s.t.
$$f(y) < f(0) + gy \quad (**)$$

Assume $g \notin [0,1]$. There are two cases:

(i) $g < 0$. Take $y = -1$ then

$$f(y) = (-1)_+ = 0 < -g = gy \quad \text{because } g < 0$$

(ii) $g > 1$. Take $y = 1$ then

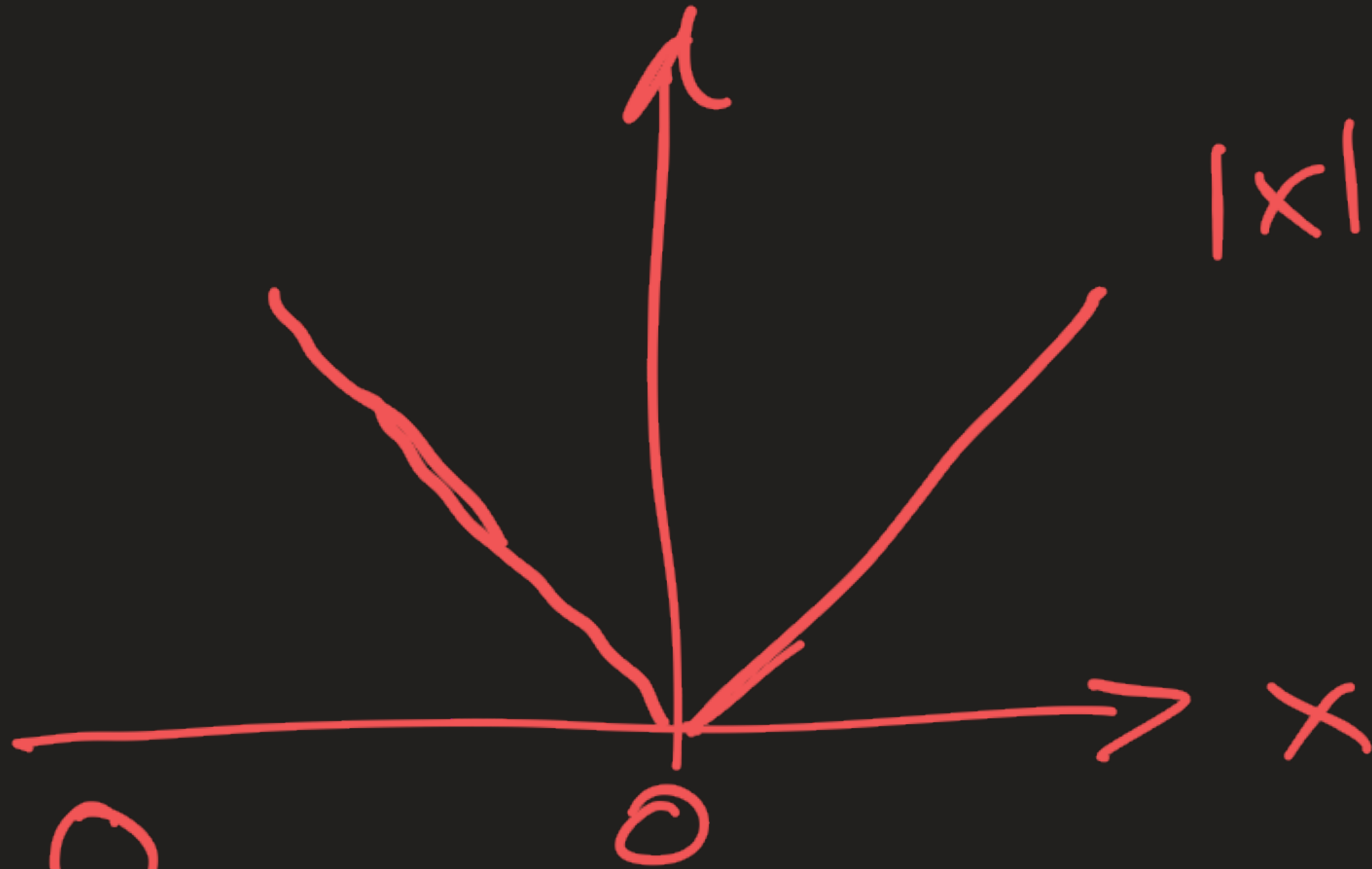
$$f(y) = (1)_+ = 1 < g = gy \quad \text{because } g > 1$$

so we have shown $\partial f(0) = [0,1]$

$\blacksquare$

Exercise

$$f(x) = |x|$$



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

Optimality Conditions

(U) $\quad x^* = \underset{x}{\operatorname{argmin}} \, f(x)$ $\qquad$ f-convex $\qquad$ Fermat's Optimality Condition

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

Prf

$\Leftarrow$ Assume $0 \in \partial f(x^*)$, this means that for all y

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

so x^* solves (U)

$\Rightarrow$ Assume x^* solves (U), then for all y

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

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

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

$$\iff \exists g \in \partial f(x^*) \text{ s.t. } \langle g, y - x^* \rangle \geq 0$$

for all $y \in C$

Fact : Fermat's Optimality Cond $\Rightarrow$ optimality conditions for (C)

Results :

1) If C is a convex set, define its convex indicator function

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

2)

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

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

3) If $f = h + g$, then $\partial f(x) = \partial h(x) + \partial g(x)$

$= \{ v : v = g_1 + g_2$ for

a $g_1 \in \partial h(x)$ and a $g_2 \in \partial g(x) \}$

"Minkowski sum of sets"

Apply Fermat's optimality condition to

$$x^* = \underset{x}{\operatorname{argmin}} \; f(x) + i_c(x)$$

$$\Longleftrightarrow \quad 0 \in \partial f(x^*) + \partial i_c(x^*)$$

Compute $\partial i_C(x)$:

Claim $\partial i_C(x) = \{ \nu : \langle \nu, y-x \rangle \leq 0$ for any $y \in C \}$

Assume ν satisfies this, then fix a $y \in C$, we have

$$\underbrace{i_C(y)}_{=0} \geq \underbrace{i_C(x)}_{=0} + \underbrace{\langle \nu, y-x \rangle}_{\leq 0}$$

Then

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

means there are $g \in \partial f(x^*)$ and $\nu \in \partial i_C(x^*)$ so that

$$0 = g + \nu \iff -g = \nu$$

that means that

$$\langle -g, y - x^* \rangle \leq 0 \quad \forall y \in C$$

$$\iff \langle g, y - x^* \rangle \geq 0 \quad \forall y \in C$$

so

$$x^* = \underset{x \in C}{\operatorname{argmin}} f(x) \iff \exists g \in \partial f(x^*) \text{ s.t. } \langle g, y - x^* \rangle \geq 0 \quad \forall y \in C$$

Subgradient Descent Algorithm

f – cvx function

$x_* = \underset{x}{\operatorname{argmin}}\ f(x)$

Inputs: x_0 – starting point

T – # of iterations

$\alpha_0, \dots, \alpha_{T-1}$ stepsizes

∂f – an oracle for subgradient

Outputs

$\hat{x} \approx x^*$

Alg

for $t = 0, \dots, T-1$:

$g_t \leftarrow$ a vector in $\partial f(x_t)$

$x_{t+1} \leftarrow x_t - \alpha_t g_t$

Guarantees

$f(\hat{x}) - f(x^*) \leq O\left(\frac{\log T}{T}\right)$

Return

$$\hat{x} = \frac{1}{T} \sum_{t=1}^{T} x_t \quad \text{(Polyak average)}$$

Weak & Strong Subgradient Calculus Results

- Multiplication by a <u>positive</u> scalar

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

- Differentiability

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

- Sum rule of sub-differentiable 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 of subdifferential 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

$$h(x) = \max_i (f_1(x), \dots, f_m(x))$$

then

$$\partial(\max(f_1, f_2, \dots, f_m))(x) = \operatorname{conv}\Big(\bigcup_{i \in I(x)} \partial f_i(x)\Big)$$

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

Ex of using subdifferential calculus

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

To find $\partial f(\omega)$, apply multiplication and sum rules:

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

Consider $\|\omega\|_1 = \sum_{i=1}^{d} |\omega_i|$, by the sum rule

$$\partial \|\omega\|_1 = \sum_{i=1}^{d} \partial |\omega_i|$$

where we consider $|\omega_i|$ as a function of ω

exercise

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

$$\Rightarrow \quad \partial \|\omega\|_1 = \operatorname{sgn}(\omega)$$

$$\text{where } \operatorname{sgn}(\omega)_i = \begin{cases} -1 & \text{if } \omega_i < 0 \\ 1 & \text{if } \omega_i > 0 \\ [-1, 1] & \text{if } \omega_i = 0 \end{cases}$$

Consider a term of the form $h_i(\omega) = (1 - y_i x_i^T \omega)_+$

Using the affine transformation rule,

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

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

Exercise

Implement a solver for the LASSO problem

$$\min_x \|Ax - b\|_2^2 + \lambda \|x\|_1$$

Key point is what is your subgradient oracle.

(will post as an alternative participation)