Optimization in Machine Learning

Lecture 9: Subgradient calculus concluded, Necessary and sufficient conditions for optimization with and without Convexity, Lipschitz Continuity

Ganesh Ramakrishnan

Department of Computer Science
Dept of CSE, IIT Bombay
https://www.cse.iitb.ac.in/~ganesh

February, 2025

Outline

- Understanding the Convexity of Machine Learning Loss Functions [Done]
- First Order Conditions for Convexity [Done]
  - Direction Vector, Directional derivative
  - Quasi convexity & Sub-level sets of convex functions
  - Convex Functions & their Epigraphs
  - First-Order Convexity Conditions [Done
- Second Order Conditions for Convexity [Done]
- Basic Subgradient Calculus: Subgradients for non-differentiable convex functions [Almost Done]
- Convex Optimization Problems and Basic Optimality Conditions
- Lipschitz Properties of functions

[RECAP]

General pointwise maximum: if $f(\mathbf{x}) = \max_{s \in S} f_s(\mathbf{x})$, then

under some regularity conditions (on S, f_s), $\partial f(\mathbf{x}) = cl\left\{conv\left(\bigcup_{s: f_s(\mathbf{x}) = f(\mathbf{x})} \partial f_s(\mathbf{x})\right)\right\}$

What does CLOSURE cl{...} mean above?

See https://www.cse.iitb.ac.in/~ganesh/cs769/ and specifically, refer to pages 3-11 of www.cse.iitb.ac.in/~ganesh/cs769/notes/enotes/6-3-08-2018-separating-supporting-hyperplane-ellipsoidalgo-matrixnorms-annotated.pdf

Definition

[Closure of a Set]: Let $\mathcal{S} \subseteq \Re^n$. The closure of $\mathcal{S}$, denoted by *closure*$(\mathcal{S})$ is given by

$$closure(\mathcal{S}) = \{\mathbf{y} \in \Re^n | \forall\ \epsilon > 0, \mathcal{B}(\mathbf{y}, \epsilon) \cap \mathcal{S} \neq \emptyset\}$$

RECALL CAUCHY SHWARZ

$$x^T z \le |x^T z| \le \|x\|_2 \|z\|_2$$ with equality iff x = z

is $$\|x\|_2 = \max_{\|z\|_2 \le 1} x^T z$$

Generalized to

$$\|x\|_p = \max_{\|z\|_q \le 1} x^T z$$

HOLDER'S INEQUALITY

$$|z^T x| \le \|z\|_q \|x\|_p \quad \forall \; \frac{1}{p} + \frac{1}{q} = 1$$

HOLDER'S INEQUALITY (and our first exposure to duality)

$$|z^T x| \le \|z\|_q \|x\|_p$$

$$\forall \; \frac{1}{p} + \frac{1}{q} = 1$$

Two ways of making a scultpure (or in this case, of defining a norm)

1) PRIMAL : Casting - fill up a mould $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$

2) DUAL: Chiselling - carving out unwanted material from the base object (by discarding)

$$\|x\|_p = \max_{\|z\|_q \le 1} z^T x$$

Recap: Subgradient of $\|\mathbf{x}\|_1$

Assume $\mathbf{x} \in \Re^n$. Then

- $\|\mathbf{x}\|_1 = \max\limits_{\mathbf{s} \in \{-1,+1\}^n} \mathbf{x}^T\mathbf{s}$ which is a pointwise maximum of 2^n functions
- Let $\mathcal{S}^* \subseteq \{-1,+1\}^n$ be the set of $\mathbf{s}$ such that for each $\mathbf{s} \in \mathcal{S}^*$, the value of $\mathbf{x}^T\mathbf{s}$ is the same max value.
- Thus, $\partial\|\mathbf{x}\|_1 = conv\left(\bigcup\limits_{\mathbf{s}\in\mathcal{S}^*} \mathbf{s}\right)$.

Recap: More of Basic Subgradient Calculus

- Scaling: $\partial(af) = a \cdot \partial f$ provided $a > 0$. The condition $a > 0$ makes function f remain convex.
- Addition: $\partial(f_1 + f_2) = \partial(f_1) + \partial(f_2)$
- Affine composition: if $g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$, then $\partial g(\mathbf{x}) = A^T \partial f(A\mathbf{x} + b)$
- Norms: important special case, $f(\mathbf{x}) = ||\mathbf{x}||_p = \max\limits_{||\mathbf{z}||_q \leq 1} \mathbf{z}^T\mathbf{x}$ where q is such that $1/p + 1/q = 1$. Then $\partial f(\mathbf{x}) = \left\{ \mathbf{y} : ||\mathbf{y}||_q \leq 1 \text{ and } \mathbf{y}^T x = \max\limits_{||\mathbf{z}||_q \leq 1} \mathbf{z}^T\mathbf{x} \right\}$
- Can we derive the sub-differential of $||x||_1$?

Applying q = ∞ to

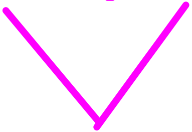
$$\left\{ \mathbf{y} : ||\mathbf{y}||_q \le 1 \text{ and } \mathbf{y}^T x = \max_{||\mathbf{z}||_q \le 1} \mathbf{z}^T \mathbf{x} \right\}$$

we get s

st $s_i = \text{sign}(x_i)$ if $x_i \neq 0$

& $s_i \in [-1, +1]$ if $x_i = 0$

$$||x||_1 = \max_{s \in \{-1,+1\}^n} s^T x$$



issue is at 0

$\partial ||x||_1 = \text{conv}\left(\bigcup_i s^i\right)$ st s^i is max at x

Recap: More of Basic Subgradient Calculus

- Scaling: $\partial(af) = a \cdot \partial f$ provided $a > 0$. The condition $a > 0$ makes function f remain convex.
- Addition: $\partial(f_1 + f_2) = \partial(f_1) + \partial(f_2)$
- Affine composition: if $g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$, then $\partial g(\mathbf{x}) = A^T \partial f(A\mathbf{x} + b)$
- Norms: important special case, $f(\mathbf{x}) = ||\mathbf{x}||_p = \max_{||\mathbf{z}||_q \leq 1} \mathbf{z}^T\mathbf{x}$ where q is such that $1/p + 1/q = 1$. Then $\partial f(\mathbf{x}) = \left\{ \mathbf{y} : ||\mathbf{y}||_q \leq 1 \text{ and } \mathbf{y}^T x = \max_{||\mathbf{z}||_q \leq 1} \mathbf{z}^T\mathbf{x} \right\}$
- Can we derive the sub-differential of $||x||_1$?
  - Let $\mathcal{S}^* \subseteq \{-1, +1\}^n$ be the set of $\mathbf{s}$ such that for each $\mathbf{s} \in \mathcal{S}^*$, the value of $\mathbf{x}^T\mathbf{s}$ is the same max value.
  - Thus, $\partial\|\mathbf{x}\|_1 = conv\left(\bigcup_{\mathbf{s} \in \mathcal{S}^*} \mathbf{s} \right)$.

Subgradients for the 'Lasso' Problem in Machine Learning

We use Lasso ($\min_{\mathbf{x}} \ f(\mathbf{x})$) as an example to illustrate subgradients of affine composition:

$$f(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathbf{x}||^2 + \lambda||\mathbf{x}||_1$$

The subgradients of $f(\mathbf{x})$ are

Subgradients for the ‘Lasso’ Problem in Machine Learning

We use Lasso ($\min_{\mathbf{x}} f(\mathbf{x})$) as an example to illustrate subgradients of affine composition:

Simplified Ax to x (since A^T premultiplication can be invoked as per calculus of subgradients)

$$f(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathbf{x}||^2 + \underbrace{\lambda||\mathbf{x}||_1}_{s}$$

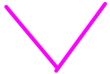
$\nabla \frac{1}{2}||x - y||^2$

$= (x - y)$

s st $s_i = \text{sign}(x_i)$ if $x_i \neq 0$

& $s_i \in [-1, +1]$ if $x_i = 0$

$||x||_1 = \max_{s \in \{-1,+1\}^n} s^T x$



issue is at 0

$\partial ||x||_1 = \text{conv}\left(\bigcup s^i\right)$ st s^i is max at x

The subgradients of $f(\mathbf{x})$ are

$(x - y) + \lambda s$

Subgradients for the 'Lasso' Problem in Machine Learning

We use Lasso ($\min_{\mathbf{x}} f(\mathbf{x})$) as an example to illustrate subgradients of affine composition:

$$f(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathbf{x}||^2 + \lambda||\mathbf{x}||_1$$

The subgradients of $f(\mathbf{x})$ are

$$\mathbf{h} = \mathbf{x} - \mathbf{y} + \lambda\mathbf{s},$$

where $s_i = sign(x_i)$ if $x_i \neq 0$ and $s_i \in [-1, 1]$ if $x_i = 0$.

More Subgradient Calculus: Composition

Following functions, though convex, may not be differentiable everywhere. How does one compute their subgradients? (what holds for subgradient also holds for gradient)

- **Composition with functions:** Let $p : \Re^k \rightarrow \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{d}om\ h$ q and $q : \Re^n \rightarrow \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument

Following functions, though convex, may not be differentiable everywhere. How does one compute their subgradients? (what holds for subgradient also holds for gradient)

- **Composition with functions:** Let $p : \Re^k \rightarrow \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{dom}\ h$ and $q : \Re^n \rightarrow \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument

  Some examples illustrating this property are:
  - $exp\ q(\mathbf{x})$ is convex if q is convex
  - $\sum_{i=1}^{m} \log q_i(\mathbf{x})$ is concave if q_i are concave and positive
  - $\log \sum_{i=1}^{m} \exp q_i(\mathbf{x})$ is convex if q_i are convex
  - $1/q(\mathbf{x})$ is convex if q is concave and positive

More Subgradient Calculus: Composition (contd)

- **Composition with functions:** Let $p : \Re^k \rightarrow \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{dom}\ h$ and $q : \Re^n \rightarrow \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument
- Subgradients for the first case (second one is homework):

More Subgradient Calculus: Composition (contd)

- **Composition with functions:** Let $p : \Re^k \rightarrow \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{dom}\ h$ and $q : \Re^n \rightarrow \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument
- Subgradients for the first case (second one is homework):
  - $f(\mathbf{y}) = p\left(q_1(\mathbf{y}), \ldots, q_k(\mathbf{y})\right) \geq p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right)$
    Where $\mathbf{h}_{q_i} \in \partial q_i(\mathbf{x})$ for $i = 1..k$ and since $p(.)$ is non-decreasing in each argument.

- **Composition with functions:** Let $p : \Re^k \rightarrow \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin$ **dom** h and $q : \Re^n \rightarrow \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - 1 q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument
- Subgradients for the first case (second one is homework):
  - $f(\mathbf{y}) = p\left(q_1(\mathbf{y}), \ldots, q_k(\mathbf{y})\right) \geq p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right)$
    Where $\mathbf{h}_{q_i} \in \partial q_i(\mathbf{x})$ for $i = 1..k$ and since $p(.)$ is non-decreasing in each argument.

$f(y) \geq p(q(x)) + h_f^T(x)(y-x)$

$p(q_1(y), \ldots q_k(y))$

(1) $q_i(y) \geq \underbrace{q_i(x) + h_{q_i}^T(y-x)}_{\text{a scalar}}$ since each q_i is convex

$$p\begin{bmatrix} q_1(x) \\ \vdots \\ q_k(x) \end{bmatrix} \geq p\left(\begin{bmatrix} q_1(x) \\ \vdots \\ q_k(x) \end{bmatrix} + \begin{bmatrix} h_{q_1}^T(y-x) \\ \vdots \\ h_{q_k}^T(y-x) \end{bmatrix}\right)$$

r $\quad$ r $\quad$ ω

(2) p is non-decreasing in each argument

(3) p is convex and it also has its subgradients $p(r+\omega) \geq p(r) + h_p^T(r)(\omega)$

More Subgradient Calculus: Composition (contd)

- **Composition with functions:** Let $p : \Re^k \rightarrow \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{dom}\ h$ and $q : \Re^n \rightarrow \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument
- Subgradients for the first case (second one is homework):
  - $f(\mathbf{y}) = p\left(q_1(\mathbf{y}), \ldots, q_k(\mathbf{y})\right) \geq p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right)$
    Where $\mathbf{h}_{q_i} \in \partial q_i(\mathbf{x})$ for $i = 1..k$ and since $p(.)$ is non-decreasing in each argument.
  - $p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right) \geq$
    $p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right) + \mathbf{h}_p^T\left(\mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right)$
    Where $\mathbf{h}_p \in \partial p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right)$

More Subgradient Calculus: Composition (contd)

- **Composition with functions:** Let $p : \Re^k \rightarrow \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{dom}\ h$ and $q : \Re^n \rightarrow \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument
- Subgradients for the first case (second one is homework):
  - $f(\mathbf{y}) = p\left(q_1(\mathbf{y}), \ldots, q_k(\mathbf{y})\right) \geq p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right)$
    Where $\mathbf{h}_{q_i} \in \partial q_i(\mathbf{x})$ for $i = 1..k$ and since $p(.)$ is non-decreasing in each argument.
  - $p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right) \geq$
    $p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right) + \mathbf{h}_p^T\left(\mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right)$
    Where $\mathbf{h}_p \in \partial p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right)$
  - $p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right) + h_p^T\left(h_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, h_{q_k}^T(\mathbf{y}-\mathbf{x})\right) = f(\mathbf{x}) + \sum_{i=1}^{k}(h_p)_i h_{q_i}(\mathbf{x})$

That is, $\sum_{i=1}^{k}(h_p)_i h_{q_i}(\mathbf{x})$ is a subgradient of the composite function at $\mathbf{x}$.

Subgradient Calculus: Second Composition [Homework - Understand]

- **Composition with functions:** Let $p : \Re^k \to \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \mathbf{dom}\ q$ and $q : \Re^n \to \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument [Homework]
    Recall: Concavity of each q_i (or convexity of $-q_i$) means $q_i(\mathbf{y}) \leq q_i(\mathbf{x}) + \mathbf{h}_{q_i}^T(\mathbf{y} - \mathbf{x})$

Subgradient Calculus: Second Composition [Homework - Understand]

- **Composition with functions:** Let $p : \Re^k \to \Re$ with $q(x) = \infty, \forall\ \mathbf{x} \notin \textbf{dom}\ q$ and $q : \Re^n \to \Re^k$. Define $f(\mathbf{x}) = p(q(\mathbf{x}))$. f is convex if
  - q_i is convex, p is convex and nondecreasing in each argument
  - or q_i is concave, p is convex and nonincreasing in each argument [Homework]
    Recall: Concavity of each q_i (or convexity of $-q_i$) means $q_i(\mathbf{y}) \le q_i(\mathbf{x}) + \mathbf{h}_{q_i}^T(\mathbf{y}-\mathbf{x})$
- Subgradients for the second case (first case already solved):
  - $f(\mathbf{y}) = p\left(q_1(\mathbf{y}), \ldots, q_k(\mathbf{y})\right) \ge p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right)$
    Where $\mathbf{h}_{q_i} \in \partial\left[-q_i(\mathbf{x})\right]$ for $i = 1..k$ and since $p(.)$ is non-increasing in each argument.
  - $p\left(q_1(\mathbf{x}) + \mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, q_k(\mathbf{x}) + \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right) \ge$
    $p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right) + \mathbf{h}_p^T\left(\mathbf{h}_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, \mathbf{h}_{q_k}^T(\mathbf{y}-\mathbf{x})\right)$
    Where $\mathbf{h}_p \in \partial p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right)$
  - $p\left(q_1(\mathbf{x}), \ldots, q_k(\mathbf{x})\right) + h_p^T\left(h_{q_1}^T(\mathbf{y}-\mathbf{x}), \ldots, h_{q_k}^T(\mathbf{y}-\mathbf{x})\right) = f(\mathbf{x}) + \sum_{i=1}^{k} (h_p)_i h_{q_i}(\mathbf{x})$

That is, $\sum_{i=1}^{k} (h_p)_i h_{q_i}(\mathbf{x})$ is still a subgradient of the composite function at $\mathbf{x}$.

More Subgradient Calculus: Proximal Operator

Following functions are again convex, but again, may not be differentiable everywhere. How does one compute their subgradients at points of non-differentiability?

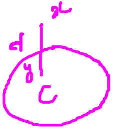
- **Infimum:** If $c(x,y)$ is convex in (x,y) and $\mathcal{C}$ is a convex set, then $d(x) = \inf_{y \in \mathcal{C}} c(x,y)$ is convex. For example:
  - Let $d(\mathbf{x}, \mathcal{C})$ that returns the distance of a point $\mathbf{x}$ to a convex set $\mathcal{C}$. That is $d(\mathbf{x}, \mathcal{C}) = \inf_{y \in \mathcal{C}} ||\mathbf{x} - \mathbf{y}||^2$. Then $d(\mathbf{x}, \mathcal{C})$ is a convex function.
  - $\operatorname{argmin}_{y \in \mathcal{C}} d(\mathbf{x}, \mathcal{C})$ is a special case of the proximity operator: $prox_f(\mathbf{x}) = \operatorname{argmin}_{y} PROX_f(\mathbf{x})$ of a convex function $f(\mathbf{x})$. Here, $PROX_f(\mathbf{x}) = f(\mathbf{y}) + \frac{1}{2}||\mathbf{x} - \mathbf{y}||^2$ The special case is when

More Subgradient Calculus: Proximal Operator

Following functions are again convex, but again, may not be differentiable everywhere. How does one compute their subgradients at points of non-differentiability?

- **Infimum:** If $c(x,y)$ is convex in (x,y) and $\mathcal{C}$ is a convex set, then $d(x) = \inf\limits_{y\in\mathcal{C}} c(x,y)$ is convex. For example:

If C is an open convex set, we need infinimum
For example: Minimum is not defined
when the point x is not in the same hyperplane as C



  - Let $d(\mathbf{x},\mathcal{C})$ that returns the distance of a point $\mathbf{x}$ to a convex set $\mathcal{C}$. That is $d(\mathbf{x},\mathcal{C}) = \inf\limits_{y\in\mathcal{C}} ||\mathbf{x}-\mathbf{y}||^2$. Then $d(\mathbf{x},\mathcal{C})$ is a convex function.
  - $\underset{y\in\mathcal{C}}{\operatorname{argmin}}\, d(\mathbf{x},\mathcal{C})$ is a special case of the proximity operator: $prox_f(\mathbf{x}) = \underset{y}{\operatorname{argmin}}\, PROX_f(\mathbf{x})$ of a convex function $f(\mathbf{x})$. Here, $PROX_f(\mathbf{x}) = f(\mathbf{y}) + \frac{1}{2}||\mathbf{x}-\mathbf{y}||^2$ The special case is when

$$\min a(x) \quad \text{s.t. } x \in C, \; C \text{ is convex set} \;=\; \min_x \; a(x) + I_C(x)$$

$$I_C(x) = 0 \text{ if } x \in C$$
$$= \infty \text{ o/w}$$

large finite value

More Subgradient Calculus: Proximal Operator

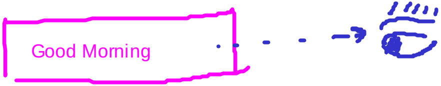
Following functions are again convex, but again, may not be differentiable everywhere. How does one compute their subgradients at points of non-differentiability?

- **Infimum:** If $c(x, y)$ is convex in (x, y) and $\mathcal{C}$ is a convex set, then $d(x) = \inf_{y \in \mathcal{C}} c(x, y)$ is convex. For example:
  - Let $d(\mathbf{x}, \mathcal{C})$ that returns the distance of a point $\mathbf{x}$ to a convex set $\mathcal{C}$. That is $d(\mathbf{x}, \mathcal{C}) = \inf_{y \in \mathcal{C}} ||\mathbf{x} - \mathbf{y}||^2$. Then $d(\mathbf{x}, \mathcal{C})$ is a convex function.
  - $\operatorname{argmin}_{y \in \mathcal{C}} d(\mathbf{x}, \mathcal{C})$ is a special case of the proximity operator: $prox_f(\mathbf{x}) = \operatorname{argmin}_{y} PROX_f(\mathbf{x})$ of a convex function $f(\mathbf{x})$. Here, $PROX_f(\mathbf{x}) = f(\mathbf{y}) + \frac{1}{2}||\mathbf{x} - \mathbf{y}||^2$ The special case is when $f(\mathbf{y})$ is the indicator function $I_C(\mathbf{y})$, introduced to eliminate the constraints of an optimization problem.
    - ★ Note that $\partial I_C(\mathbf{y}) = N_C(\mathbf{y}) = \{\mathbf{h} \in \Re^n : \mathbf{h}^T\mathbf{y} \geq \mathbf{h}^T\mathbf{z} \text{ for any } \mathbf{z} \in C\}$
    - ★ The subdifferential $\partial PROX_f(\mathbf{x}) = \partial f(\mathbf{y}) + \mathbf{y} - \mathbf{x}$ which can now be obtained for the special case $f(\mathbf{y}) = I_C(\mathbf{y})$.
    - ★ We will invoke this when we discuss the **proximal gradient descent** algorithm

More Subgradient Calculus: Perspective (Advanced & Optional)

Following functions are again convex, but again, may not be differentiable everywhere. How does one compute their subgradients at points of non-differentiability?

- **Perspective Function:** The perspective of a function $f : \Re^n \to \Re$ is the function $g : R^n \times \Re \to \Re$, $g(x,t) = tf(x/t)$. Function g is convex if f is convex on $\mathbf{d}omg = \{(x,t) | x/t \in \mathbf{d}omf, t > 0\}$. For example,
    - The perspective of $f(x) = x^T x$ is (quadratic-over-linear) function $g(x,t) = \frac{x^T x}{t}$ and is convex.
    - The perspective of negative logarithm $f(x) = -\log x$ is the relative entropy function $g(x,t) = t \log t - t \log x$ and is convex.



Example of perspective transformation

Following functions are again convex, but again, may not be differentiable everywhere. How does one compute their subgradients at points of non-differentiability?

- **Perspective Function:** The perspective of a function $f : \Re^n \rightarrow \Re$ is the function $g : R^n \times \Re \rightarrow \Re$, $g(x,t) = tf(x/t)$. Function g is convex if f is convex on $\mathbf{d}om g = \{(x,t) | x/t \in \mathbf{d}om f, t > 0\}$. For example,
  - The perspective of $f(x) = x^T x$ is (quadratic-over-linear) function $g(x,t) = \frac{x^T x}{t}$ and is convex.
  - The perspective of negative logarithm $f(x) = -\log x$ is the relative entropy function $g(x,t) = \underbrace{t \log t - t \log}_{\text{Cross entropy}} x$ and is convex.

More on SubGradient kind of functions: Monotonicity

A differentiable function $f : \Re \rightarrow \Re$ is (strictly) convex, iff and only if $f'(x)$ is (strictly) increasing. Is there a closer analog for $f : \Re^n \rightarrow \Re$?

More on SubGradient kind of functions: Monotonicity

A differentiable function $f : \Re \to \Re$ is (strictly) convex, iff and only if $f'(x)$ is (strictly) increasing. Is there a closer analog for $f : \Re^n \to \Re$? View subgradient as an instance of a general function $\mathbf{h} : \mathcal{D} \to \Re^n$ and $\mathcal{D} \subseteq \Re^n$. Then

More on SubGradient kind of functions: Monotonicity

A differentiable function $f : \Re \to \Re$ is (strictly) convex, iff and only if $f'(x)$ is (strictly) increasing. Is there a closer analog for $f : \Re^n \to \Re$? View subgradient as an instance of a general function $\mathbf{h} : \mathcal{D} \to \Re^n$ and $\mathcal{D} \subseteq \Re^n$. Then

Definition

1. $\mathbf{h}$ is *monotone* on $\mathcal{D}$ if for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$,

$$\left(\mathbf{h}(\mathbf{x}_1) - \mathbf{h}(\mathbf{x}_2)\right)^T (\mathbf{x}_1 - \mathbf{x}_2) \geq 0 \tag{1}$$

More on SubGradient kind of functions: Monotonicity (contd)

Definition

2. $\mathbf{h}$ is *strictly monotone* on $\mathcal{D}$ if for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$ with $\mathbf{x}_1 \neq \mathbf{x}_2$,

$$\left(\mathbf{h}(\mathbf{x}_1) - \mathbf{h}(\mathbf{x}_2)\right)^T (\mathbf{x}_1 - \mathbf{x}_2) > 0 \tag{2}$$

3. $\mathbf{h}$ is *uniformly* or *strongly monotone* on $\mathcal{D}$ if for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{D}$, there is a constant $c > 0$ such that

$$\left(\mathbf{h}(\mathbf{x}_1) - \mathbf{h}(\mathbf{x}_2)\right)^T (\mathbf{x}_1 - \mathbf{x}_2) \geq c||\mathbf{x}_1 - \mathbf{x}_2||^2 \tag{3}$$

(Sub)Gradients and Convexity

Relationship between convexity of a function and monotonicity of its (sub)gradient:

Theorem

Let $f : \mathcal{D} \rightarrow \Re$ with $\mathcal{D} \subseteq \Re^n$ be differentiable on the convex set $\mathcal{D}$. Then,

1. *f is convex on $\mathcal{D}$* **iff** *its gradient ∇f is monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$:* $\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) \geq 0$
2. *f is strictly convex on $\mathcal{D}$* **iff** *its gradient ∇f is strictly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$ with $\mathbf{x} \neq \mathbf{y}$:* $\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) > 0$
3. *f is uniformly or strongly convex on $\mathcal{D}$* **iff** *its gradient ∇f is uniformly monotone. That is, for all $\mathbf{x}, \mathbf{y} \in \Re$, $\left(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right)^T (\mathbf{x} - \mathbf{y}) \geq c||\mathbf{x} - \mathbf{y}||^2$ for some constant $c > 0$.*

While these results also hold for (more advanced proximal) subgradients $\mathbf{h}_p$ (see `https://moodle.iitb.ac.in/mod/resource/view.php?id=32806`), we will quickly show them only for gradients ∇f

Advanced: $\mathbf{h}_p$ is a proximal gradient of f at $\mathbf{x}$ iff, $\forall \mathbf{y} \in dmn(f)$, $f(y) \geq f(x) + \mathbf{h}_p(\mathbf{y} - \mathbf{x}) - \frac{\lambda}{2}||\mathbf{y} - \mathbf{x}||^2$

(Sub)Gradients and Convexity (contd)

Proof:
Necessity: Suppose f is strongly convex on $\mathcal{D}$. Then we know from an earlier result that for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{1}{2}c||\mathbf{y} - \mathbf{x}||^2$$
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla^T f(\mathbf{y})(\mathbf{x} - \mathbf{y}) + \frac{1}{2}c||\mathbf{x} - \mathbf{y}||^2$$

Adding the two inequalities,

(Sub)Gradients and Convexity (contd)

Proof:
Necessity: Suppose f is strongly convex on $\mathcal{D}$. Then we know from an earlier result that for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,

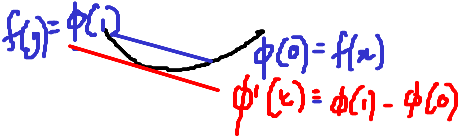
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{1}{2}c||\mathbf{y} - \mathbf{x}||^2$$
$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla^T f(\mathbf{y})(\mathbf{x} - \mathbf{y}) + \frac{1}{2}c||\mathbf{x} - \mathbf{y}||^2$$

Adding the two inequalities, we get uniform/strong monotonicity in definition (3). If f is convex, the inequalities hold with $c = 0$, yielding monotonicity in definition (1). If f is strictly convex, the inequalities will be strict, yielding strict monotonicity in definition (2).

(Sub)Gradients and Convexity (contd)

Sufficiency: Suppose ∇f is monotone. For any fixed $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, consider the function $\phi(t) = f\left(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\right)$. By the mean value theorem applied to $\phi(t)$, we should have for some $t \in (0, 1)$,

Sufficiency: Suppose ∇f is monotone. For any fixed $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, consider the function $\phi(t) = f\left(\mathbf{x} + t(\mathbf{y} - \mathbf{x})\right)$. By the mean value theorem applied to $\phi(t)$, we should have for some $t \in (0, 1)$,



= Dot product

(Sub)Gradients and Convexity (contd)

Sufficiency: Suppose ∇f is monotone. For any fixed $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, consider the function $\phi(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$. By the mean value theorem applied to $\phi(t)$, we should have for some $t \in (0,1)$,

$$\phi(1) - \phi(0) = \phi'(t) \tag{4}$$

Letting $\mathbf{z} = \mathbf{x} + t(\mathbf{y} - \mathbf{x})$, (4) translates to

$$f(\mathbf{y}) - f(\mathbf{x}) = \nabla^T f(\mathbf{z})(\mathbf{y} - \mathbf{x}) \tag{5}$$

Also, by definition of monotonicity of ∇f,

$$(\nabla f(\mathbf{z}) - \nabla f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x}) = \frac{1}{t} (\nabla f(\mathbf{z}) - \nabla f(\mathbf{x}))^T (\mathbf{z} - \mathbf{x}) \geq 0 \tag{6}$$

(Sub)Gradients and Convexity (contd)

Combining (5) with (6), we get,

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{x}) &= (\nabla f(\mathbf{z}) - f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \\ &\geq \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) \end{aligned} \tag{7}$$

By a previous foundational result, this inequality proves that f is convex. Strict convexity can be similarly proved by using the strict inequality in (6) inherited from strict monotonicity, and letting the strict inequality follow through to (7).

(Sub)Gradients and Convexity (contd)

For the case of strong convexity, we have

$$\phi'(t) - \phi'(0) = \left(\nabla f(\mathbf{z}) - f(\mathbf{x})\right)^T (\mathbf{y} - \mathbf{x})$$
$$= \frac{1}{t}\left(\nabla f(\mathbf{z}) - f(\mathbf{x})\right)^T (\mathbf{z} - \mathbf{x}) \geq \frac{1}{t} c||\mathbf{z} - \mathbf{x}||^2 = ct||\mathbf{y} - \mathbf{x}||^2 \quad (8)$$

Therefore,

(Sub)Gradients and Convexity (contd)

For the case of strong convexity, we have

$$\phi'(t) - \phi'(0) = \left(\nabla f(\mathbf{z}) - f(\mathbf{x})\right)^T (\mathbf{y} - \mathbf{x})$$
$$= \frac{1}{t}\left(\nabla f(\mathbf{z}) - f(\mathbf{x})\right)^T (\mathbf{z} - \mathbf{x}) \geq \frac{1}{t}c||\mathbf{z} - \mathbf{x}||^2 = ct||\mathbf{y} - \mathbf{x}||^2 \tag{8}$$

Therefore,

$$\phi(1) - \phi(0) - \phi'(0) = \int_0^1 [\phi'(t) - \phi'(0)]dt \geq \frac{1}{2}c||\mathbf{y} - \mathbf{x}||^2 \tag{9}$$

which translates to

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{1}{2}c||\mathbf{y} - \mathbf{x}||^2$$

Thus, f must be strongly convex.