Optimization in Machine Learning

Lecture 8: Subgradient and its calculus, 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

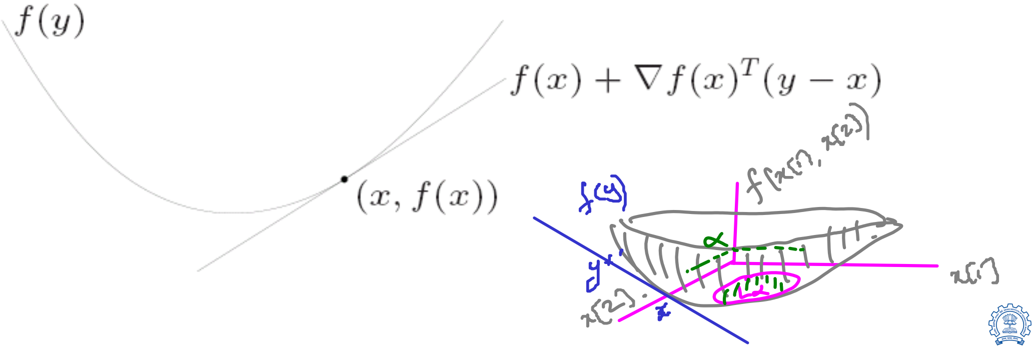
January, 2025

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
- Second Order Conditions for Convexity [Almost Done]
- Basic Subgradient Calculus: Subgradients for non-differentiable convex functions
- Convex Optimization Problems and Basic Optimality Conditions
- Lipschit Properties of functions

First-Order Convexity Conditions: The complete statement

The geometrical interpretation of this theorem is that at any point, the linear approximation based on a local derivative gives a lower estimate of the function, *i.e.* the convex function always lies above the supporting hyperplane at that point. This is pictorially depicted below:



Second Order Conditions of Convexity

Can we use the Hessian to prove that the logSumExp function is Convex?
Answer is YES
Boyd's book uses the fact that Hessian being positive semi-definite is necessary and sufficient condition for convexity

- Recall the Hessian of a continuous function:

$$\frac{\partial^2 \exp(x)}{\partial x^2} = \exp(x)$$

$$\text{sumexp} = \sum \exp(x_i)$$

$$\nabla^2 \text{sumexp} = \begin{bmatrix} \exp(x_i) & & 0 \\ & \ddots & \\ 0 & & \ddots \end{bmatrix}$$

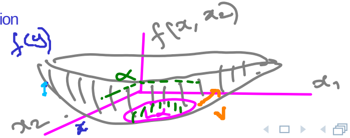
$$\nabla^2 f(w) = \begin{pmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_n} \\ \frac{\partial^2 f}{\partial w_2 \partial w_1} & \frac{\partial^2 f}{\partial w_2^2} & \cdots & \frac{\partial^2 f}{\partial w_2 \partial w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial w_n \partial w_1} & \frac{\partial^2 f}{\partial w_n \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_n^2} \end{pmatrix}$$

INTUITION:
1) First order condition: The directional derivative is non-decreasing in every direction
2) Second order condition: The curvature is positive in every direction!

- f is convex if and only if, a) $dom(f)$ is convex, and for all $x \in dom(f)$, $\nabla^2 f(x) \geq 0$ (i.e. $\nabla^2 f(x)$ is positive semi-definite).

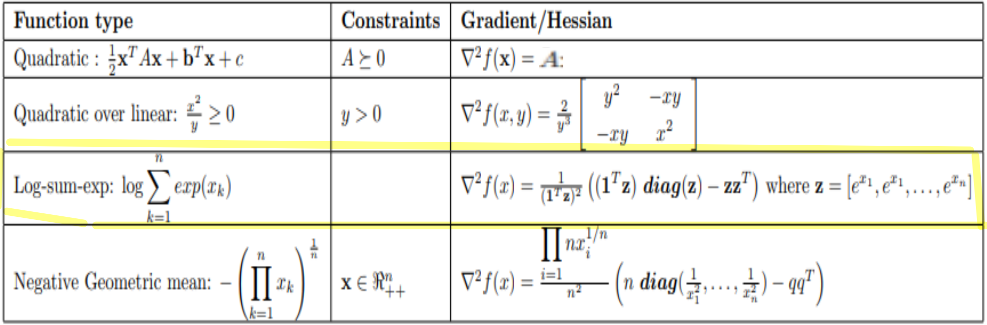
To show that LogSumExp is convex, can we prove that the quadrative expression is always non-negative

$$v^T \nabla^2 \text{logSumExp}\, V =$$ EXPAND AS HOMEWORK!



$\forall v.$ $v^T \nabla^2 f(x) v \geq 0$

On page 67 of the Convex Optimization Notes: https://moodle.iitb.ac.in/mod/resource/view.php?id=18791
Table of Hessians for some Convex Optimization Problems



Function type	Constraints	Gradient/Hessian
Quadratic : $\frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$	$A \succeq 0$	$\nabla^2 f(\mathbf{x}) = A$
Quadratic over linear: $\frac{x^2}{y} \geq 0$	$y > 0$	$\nabla^2 f(x,y) = \frac{2}{y^3}\begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix}$
Log-sum-exp: $\log \sum_{k=1}^{n} exp(x_k)$		$\nabla^2 f(x) = \frac{1}{(\mathbf{1}^T\mathbf{z})^2}\left((\mathbf{1}^T\mathbf{z})\ \boldsymbol{diag}(\mathbf{z}) - \mathbf{z}\mathbf{z}^T\right)$ where $\mathbf{z} = [e^{x_1}, e^{x_1}, \ldots, e^{x_n}]$
Negative Geometric mean: $-\left(\prod_{k=1}^{n} x_k\right)^{\frac{1}{n}}$	$\mathbf{x} \in \Re^n_{++}$	$\nabla^2 f(x) = \frac{\prod_{i=1} nx_i^{1/n}}{n^2}\left(n\ \boldsymbol{diag}(\frac{1}{x_1^2}, \ldots, \frac{1}{x_n^2}) - qq^T\right)$

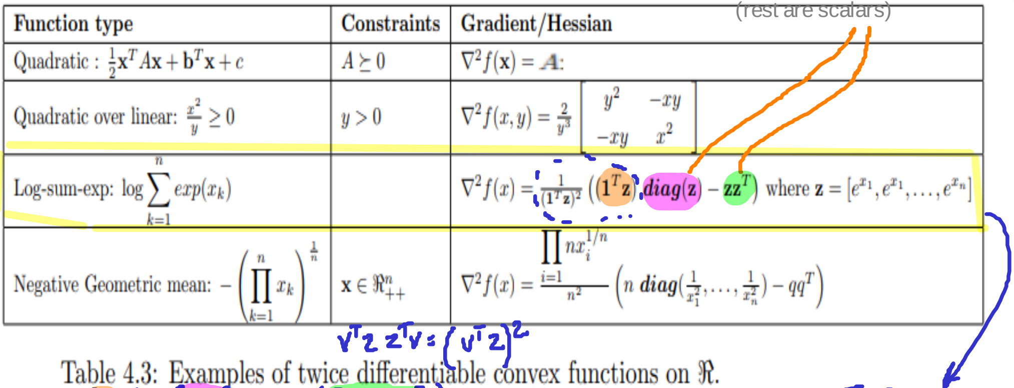
Table 4.3: Examples of twice differentiable convex functions on $\Re$.

On page 67 of the Convex Optimization Notes: https://moodle.iitb.ac.in/mod/resource/view.php?id=67925

Table of Hessians for some Convex Optimization Problems

The v can be pushed into multiplication with the highlighted matrices in pink & green (rest are scalars)

Note: x_k could itself be linearly transformed



Function type	Constraints	Gradient/Hessian
Quadratic : $\frac{1}{2}\mathbf{x}^TA\mathbf{x}+\mathbf{b}^T\mathbf{x}+c$	$A \succeq 0$	$\nabla^2 f(\mathbf{x}) = A$
Quadratic over linear: $\frac{x^2}{y} \geq 0$	$y > 0$	$\nabla^2 f(x,y) = \frac{2}{y^3}\begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix}$
Log-sum-exp: $\log \sum_{k=1}^{n} exp(x_k)$		$\nabla^2 f(x) = \frac{1}{(\mathbf{1}^T\mathbf{z})^2}\left((\mathbf{1}^T\mathbf{z})\, \boldsymbol{diag}(\mathbf{z}) - \mathbf{z}\mathbf{z}^T\right)$ where $\mathbf{z} = [e^{x_1}, e^{x_1}, \ldots, e^{x_n}]$
Negative Geometric mean: $-\left(\prod_{k=1}^{n} x_k\right)^{\frac{1}{n}}$	$\mathbf{x} \in \Re^n_{++}$	$\nabla^2 f(x) = \frac{\prod_{i=1} n x_i^{1/n}}{n^2}\left(n\, \boldsymbol{diag}(\frac{1}{x_1^2}, \ldots, \frac{1}{x_n^2}) - qq^T\right)$

$v^Tz\, z^Tv = (v^Tz)^2$

Table 4.3: Examples of twice differentiable convex functions on $\Re$.

$v^T\nabla^2 f v \geq 0$

We will show that the quadratic expression on the RHS is >=0 for all v

$$= \frac{1}{(1^Tz)^2}\left(\left(\sum_{i=1}^{n} z_i\right)\left(\sum_{i=1}^{n} v_i^2 z_i\right) - \left(\sum_{i=1}^{n} v_i z_i\right)^2\right)$$

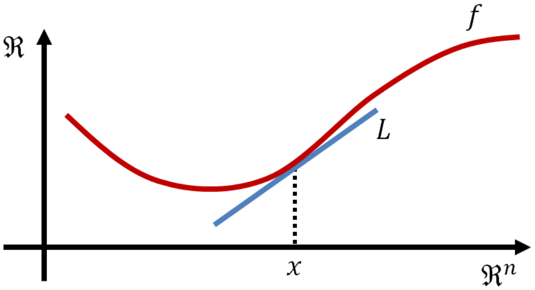
$x_i := v_i\sqrt{z_i}$

$y_i := \sqrt{z_i}$

$$(y^Ty)(x^Tx) - (x^Ty)^2 \geq 0$$

By Cauchy Schwarz inequality.. Thus we have proved that logSumExp is strictly(?) convex using a different metholodology

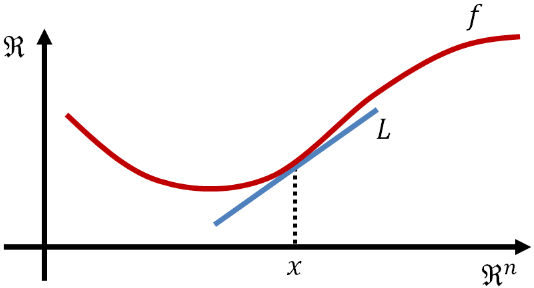
(Sub)Gradients and Convexity (contd)

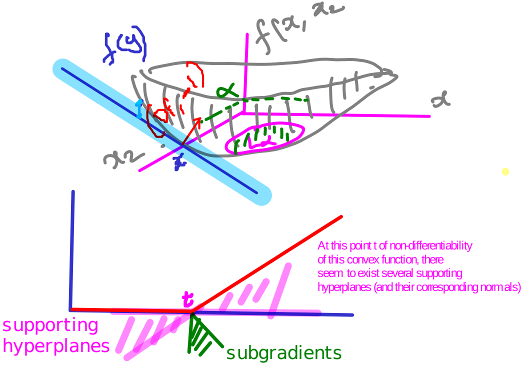


To say that a function $f : \Re^n \mapsto \Re$ is differentiable at $\mathbf{x}$ is to say that there is a single unique linear tangent that under estimates the function:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \triangledown f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}), \ \ \forall \mathbf{x}, \mathbf{y}$$

(Sub)Gradients and Convexity (contd)



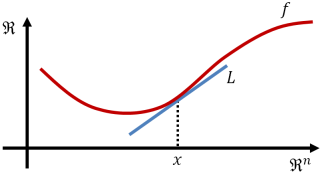


At this point t of non-differentiability of this convex function, there seem to exist several supporting hyperplanes (and their corresponding normals)

To say that a function $f : \Re^n \mapsto \Re$ is differentiable at $\mathbf{x}$ is to say that there is a single unique linear tangent that under estimates the function:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \bigtriangledown f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}), \;\; \forall \mathbf{x}, \mathbf{y}$$

([Homework discussion] Sub)Gradients and Convexity (contd)



To say that a function $f : \Re^n \mapsto \Re$ is differentiable at $\mathbf{x}$ is to say that there is a single unique linear tangent that under estimates the function:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \bigtriangledown f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}), \ \forall \mathbf{x}, \mathbf{y}$$

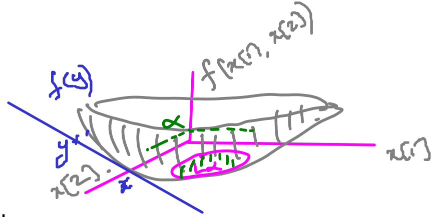
Homework 1: Is the subgradient guaranteed to exist at each point of the domain for a convex function even if the function is non-differentiable?
Optional Homework 2: How do we show that for a differentiable convex function, the only subgradient will be its gradient?

Outline of next few topics

- Subgradients, Subgradient Calculus and Convexity
- Local and Global Minimum
- Sufficient Subgradient condition for Global Minimum
- Convexity and Local & Global Minimum
- Rates of Convergence, Lipschitz Continuity and Smoothness
- Algorithms for Optimization: First Order and thereafter

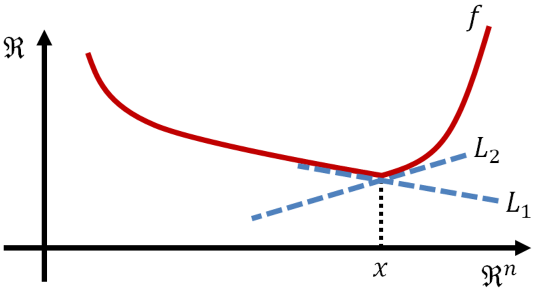
Outline of next few topics

- Subgradients, Subgradient Calculus and Convexity
- Local and Global Minimum
- Sufficient Subgradient condition for Global Minimum
- Convexity and Local & Global Minimum
- Rates of Convergence, Lipschitz Continuity and Smoothness
- Algorithms for Optimization: First Order and thereafter



What if the function is not differentiable everywhere?
Yet is convex?
Supporting Hyperplane theorem (there is a supporting hyperplane
to the epi(f) at every point) holds even if the convex f is not differentiable
==> There is a generalization of the gradient called the subgradient

(Sub)Gradients and Convexity (contd)

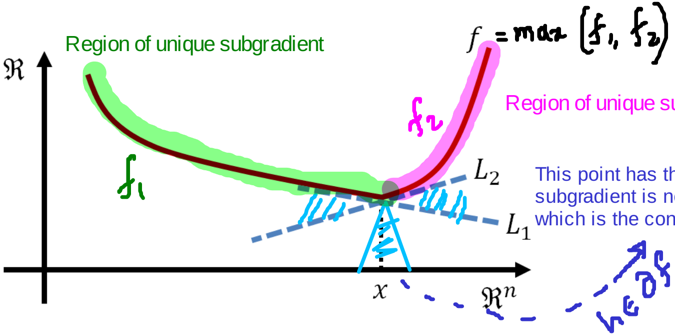


In this figure we see the function f at $\mathbf{x}$ has many possible linear tangents that may fit appropriately. Then a **subgradient** is any $\mathbf{h} \in \Re^n$ (same dimension as x) such that:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{h}^T(\mathbf{y} - \mathbf{x}), \ \forall \mathbf{y}$$

Thus, intuitively, if a convex function is differentiable at a point $\mathbf{x}$ then

(Sub)Gradients and Convexity (contd)



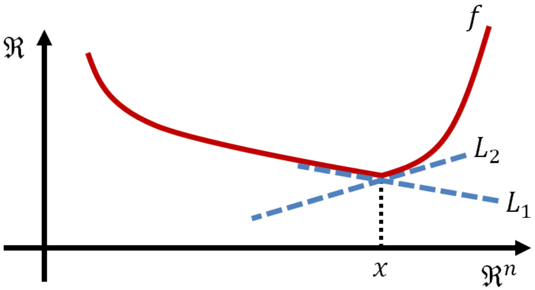
This point has the two function values coinciding.. Hence subgradient is not unique. In fact the subgradients form a set which is the convex hull of the gradients to f1 and f2 at that point

In this figure we see the function f at $\mathbf{x}$ has many possible linear tangents that may fit appropriately. Then a **subgradient** is any $\mathbf{h} \in \Re^n$ (same dimension as x) such that:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{h}^T(\mathbf{y} - \mathbf{x}), \ \forall \mathbf{y}$$

Thus, intuitively, if a convex function is differentiable at a point $\mathbf{x}$ then

(Sub)Gradients and Convexity (contd)



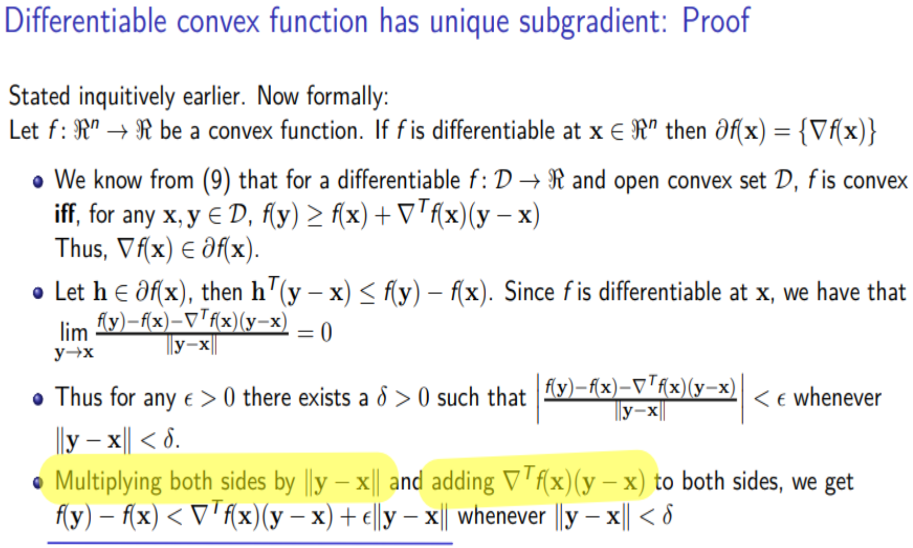
In this figure we see the function f at $\mathbf{x}$ has many possible linear tangents that may fit appropriately. Then a **subgradient** is any $\mathbf{h} \in \Re^n$ (same dimension as x) such that:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{h}^T(\mathbf{y} - \mathbf{x}), \;\; \forall \mathbf{y}$$

Thus, intuitively, if a convex function is differentiable at a point $\mathbf{x}$ then it has a unique subgradient at that point ($\bigtriangledown f(\mathbf{x})$). Formal Proof?

ADVANCED AND OPTIONAL:

Proof is using the limit definition of derivatives on pages 6-9 (slides 63-64) of
https://www.cse.iitb.ac.in/~ganesh/cs769/notes/enotes/12-28-08-2018-firstordergradient-why-what-how-subgradient-annotated.pdf



Differentiable convex function has unique subgradient: Proof

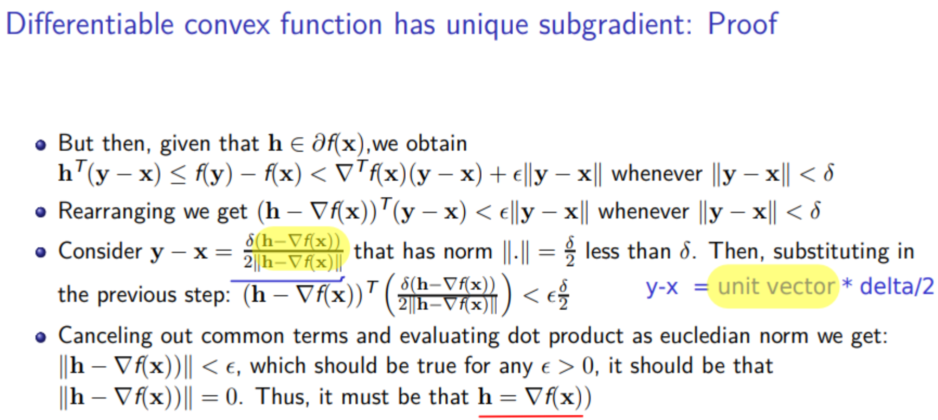
Stated inquitively earlier. Now formally:

Let $f: \Re^n \rightarrow \Re$ be a convex function. If f is differentiable at $\mathbf{x} \in \Re^n$ then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$

- We know from (9) that for a differentiable $f: \mathcal{D} \rightarrow \Re$ and open convex set $\mathcal{D}$, f is convex **iff**, for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x})$
  Thus, $\nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$.
- Let $\mathbf{h} \in \partial f(\mathbf{x})$, then $\mathbf{h}^T(\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x})$. Since f is differentiable at $\mathbf{x}$, we have that $\lim_{\mathbf{y} \rightarrow \mathbf{x}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x})}{\|\mathbf{y}-\mathbf{x}\|} = 0$
- Thus for any $\epsilon > 0$ there exists a $\delta > 0$ such that $\left|\frac{f(\mathbf{y}) - f(\mathbf{x}) - \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x})}{\|\mathbf{y}-\mathbf{x}\|}\right| < \epsilon$ whenever $\|\mathbf{y} - \mathbf{x}\| < \delta$.
- Multiplying both sides by $\|\mathbf{y} - \mathbf{x}\|$ and adding $\nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x})$ to both sides, we get $f(\mathbf{y}) - f(\mathbf{x}) < \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \epsilon\|\mathbf{y} - \mathbf{x}\|$ whenever $\|\mathbf{y} - \mathbf{x}\| < \delta$

ADVANCED AND OPTIONAL:

Proof is using the limit definition of derivatives on pages 6-9 (slides 63-64) of https://www.cse.iitb.ac.in/~ganesh/cs769/notes/enotes/12-28-08-2018-firstordergradient-why-what-how-subgradient-annotated.pdf



Differentiable convex function has unique subgradient: Proof

- But then, given that $\mathbf{h} \in \partial f(\mathbf{x})$,we obtain $\mathbf{h}^T(\mathbf{y}-\mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x}) < \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \epsilon\|\mathbf{y}-\mathbf{x}\|$ whenever $\|\mathbf{y}-\mathbf{x}\| < \delta$
- Rearranging we get $(\mathbf{h} - \nabla f(\mathbf{x}))^T(\mathbf{y}-\mathbf{x}) < \epsilon\|\mathbf{y}-\mathbf{x}\|$ whenever $\|\mathbf{y}-\mathbf{x}\| < \delta$
- Consider $\mathbf{y}-\mathbf{x} = \frac{\delta(\mathbf{h}-\nabla f(\mathbf{x}))}{2\|\mathbf{h}-\nabla f(\mathbf{x})\|}$ that has norm $\|.\| = \frac{\delta}{2}$ less than δ. Then, substituting in the previous step: $(\mathbf{h} - \nabla f(\mathbf{x}))^T \left(\frac{\delta(\mathbf{h}-\nabla f(\mathbf{x}))}{2\|\mathbf{h}-\nabla f(\mathbf{x})\|}\right) < \epsilon\frac{\delta}{2}$

y-x = unit vector * delta/2

- Canceling out common terms and evaluating dot product as eucledian norm we get: $\|\mathbf{h} - \nabla f(\mathbf{x}))\| < \epsilon$, which should be true for any $\epsilon > 0$, it should be that $\|\mathbf{h} - \nabla f(\mathbf{x}))\| = 0$. Thus, it must be that $\underline{\mathbf{h} = \nabla f(\mathbf{x}))}$

Detour: Convexity and Continuity

- Let f be a convex function and suppose $dom(f)$ is open. Then f is continuous.
- How *wild* can non-differentiable convex functions be?
- While there are continuous functions which are nowhere differentiable, (see `https://en.wikipedia.org/wiki/Weierstrass_function`), convex functions cannot be pathological!
- Infact, a convex function is differentiable *almost* everywhere. In other words, the set of points where f is non-differentiable is of measure 0.
- However we cannot ignore the non-differentiability, since a) the global minima could easily be a point of non differentiability and b) with any optimization algorithms, you can stumble upon these "kinks".

(Sub)Gradients and Convexity (contd)

- A **subdifferential** is the closed convex set of all subgradients of the convex function f:

$$\partial f(\mathbf{x}) = \{\mathbf{h} \in \Re^n : \mathbf{h} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}$$

Note that this set is guaranteed to be nonempty unless f is not convex.

(Sub)Gradients and Convexity (contd)

- A **subdifferential** is the closed convex set of all subgradients of the convex function f:

$$\partial f(\mathbf{x}) = \{\mathbf{h} \in \Re^n : \mathbf{h} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}$$

  Note that this set is guaranteed to be nonempty unless f is not convex.

- **Pointwise Maximum:**. if $f(\mathbf{x}) = \max\limits_{i=1\ldots m} f_i(\mathbf{x})$, then $\partial f(\mathbf{x}) = conv\left(\bigcup\limits_{i:f_i(\mathbf{x})=f(\mathbf{x})} \partial f_i(\mathbf{x})\right)$, which is the convex hull of union of subdifferentials of all active functions at x.

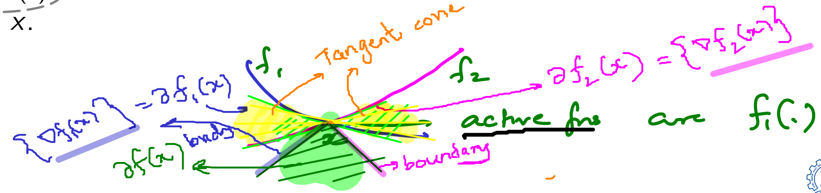
(Sub)Gradients and Convexity (contd)

- A **subdifferential** is the closed convex set of all subgradients of the convex function f:

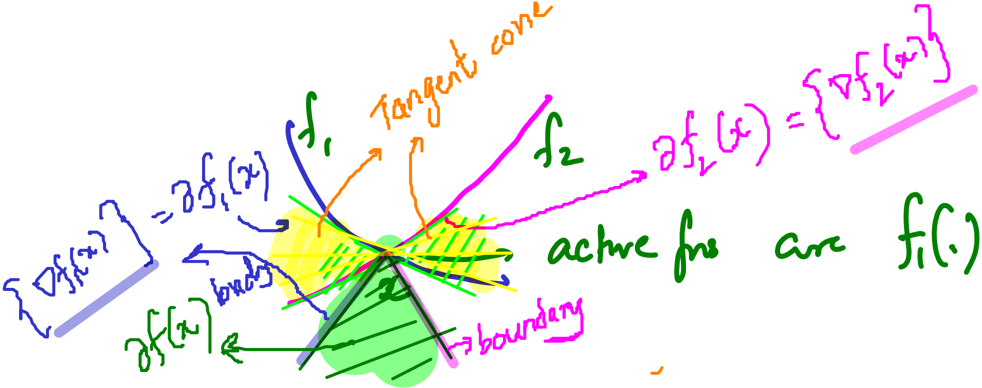
$$\partial f(\mathbf{x}) = \{\mathbf{h} \in \Re^n : \mathbf{h} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}$$

  Note that this set is guaranteed to be nonempty unless f is not convex.

- **Pointwise Maximum:**. if $f(\mathbf{x}) = \max\limits_{i=1\ldots m} f_i(\mathbf{x})$, then $\partial f(\mathbf{x}) = conv\left(\bigcup\limits_{i:f_i(\mathbf{x})=f(\mathbf{x})} \partial f_i(\mathbf{x})\right)$, which is the convex hull of union of subdifferentials of all active functions at x.



f_1
Tangent cone
f_2
$\{\nabla f_1(x)\} = \partial f_1(x)$
$\partial f_2(x) = \{\nabla f_2(x)\}$
bounds
x
$\partial f(x)$
boundary
active fns are $f_i(\cdot)$


f_1
tangent cone
f_2
$\partial f_2(x) = \{\nabla f_2(x)\}$
$\{\nabla f_1(x)\} = \partial f_1(x)$
$\partial f(x)$
x
boundary

active fns are $f_1(\cdot)$ & $f_2(\cdot)$

(Sub)Gradients and Convexity (contd)

- A **subdifferential** is the closed convex set of all subgradients of the convex function f:

$$\partial f(\mathbf{x}) = \{\mathbf{h} \in \Re^n : \mathbf{h} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}$$

  Note that this set is guaranteed to be nonempty unless f is not convex.

- **Pointwise Maximum:**. if $f(\mathbf{x}) = \max\limits_{i=1...m} f_i(\mathbf{x})$, then $\partial f(\mathbf{x}) = conv\left(\bigcup\limits_{i:f_i(\mathbf{x})=f(\mathbf{x})} \partial f_i(\mathbf{x})\right)$, which is the convex hull of union of subdifferentials of all active functions at x.

- **General pointwise maximum:** if $f(\mathbf{x}) = \max\limits_{s \in S} f_s(\mathbf{x})$, then under some regularity conditions (on S, f_s), $\partial f(\mathbf{x}) = cl\left\{conv\left(\bigcup\limits_{s:f_s(\mathbf{x})=f(\mathbf{x})} \partial f_s(\mathbf{x})\right)\right\}$

(Sub)Gradients and Convexity (contd)

- A **subdifferential** is the closed convex set of all subgradients of the convex function f:

$$\partial f(\mathbf{x}) = \{\mathbf{h} \in \Re^n : \mathbf{h} \text{ is a subgradient of } f \text{ at } \mathbf{x}\}$$

  Note that this set is guaranteed to be nonempty unless f is not convex.

- **Pointwise Maximum:**. if $f(\mathbf{x}) = \max\limits_{i=1...m} f_i(\mathbf{x})$, then $\partial f(\mathbf{x}) = conv\left(\bigcup\limits_{i:f_i(\mathbf{x})=f(\mathbf{x})} \partial f_i(\mathbf{x})\right)$, which is the convex hull of union of subdifferentials of all active functions at x.

- **General pointwise maximum:** if $f(\mathbf{x}) = \max\limits_{s \in S} f_s(\mathbf{x})$, then under some regularity conditions (on S, f_s), $\partial f(\mathbf{x}) = cl\left\{conv\left(\bigcup\limits_{s:f_s(\mathbf{x})=f(\mathbf{x})} \partial f_s(\mathbf{x})\right)\right\}$

Eg: Vector induced Matrix norm

$$M_N(A) = \sup_{v \neq 0} \frac{N(Av)}{N(v)}$$

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

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

- $\|\mathbf{x}\|_1 =$

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

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

- $\|\mathbf{x}\|_1 = \max\limits_{v \in \{-1,+1\}^n} v^T x = \max \left\{ \begin{matrix} x_1 + x_2 \cdots \\ x_1 - x_2 \\ -x_1 + x_2 \cdots \\ -x_1 - x_2 \cdots \end{matrix} \right\} = |x_1| + |x_2| + \cdots$

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

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

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

- $\|\mathbf{x}\|_1 = \max_{\mathbf{s}\in\{-1,+1\}^n} \mathbf{x}^T\mathbf{s}$ which is a pointwise maximum of 2^n functions (or configurations of s)

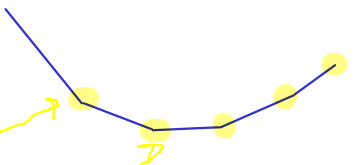
eg: $[-1, -1 \, . \, . \, . \, -1]$ $\overset{x}{[-5 \;\; -6 \;\; -7 \, . \, . \, . \, -10]}$

$[+1, +1 \; . \; +1]$ $[5 \;\; 6 \;\; 7 \;\; 10]$

$[-1, +1, -1, \, . \, . \, +1]$ $[-5 \;\; 6 \;\; -7 \;\; 10 \, . \, .]$

2^n fns.

2^n configs



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

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

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

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.

$x = [-1 \quad 0 \quad 5]$

$s \to [-1 \quad 1 \quad 1]$ $\quad x^T s = 6$

$s \to [-1 \quad -1 \quad 1]$ $\quad x^T s = 6$

=> At a point x, there
could be multiple
s indices which
yield that maximum value $\Big\}$ $S^*(x)$

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

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

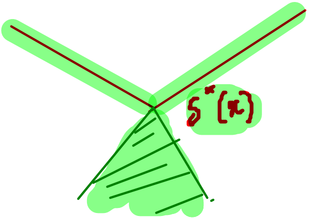
- $\|\mathbf{x}\|_1 = \max_{\mathbf{s} \in \{-1,+1\}^n} \mathbf{x}^T\mathbf{s}$ which is a pointwise maximum of 2^n functions
- Let $\mathcal{S}^* \subseteq \{-1,+1\}^n$ be the set of $\mathbf{s}$ such that for each $\mathbf{s} \in \mathcal{S}^*$, the value of $\mathbf{x}^T\mathbf{s}$ is the same max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

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

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

- $\|\mathbf{x}\|_1 = \max_{\mathbf{s} \in \{-1,+1\}^n} \mathbf{x}^T\mathbf{s}$ 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, $\partial\|\mathbf{x}\|_1 = conv\left(\bigcup_{\mathbf{s}\in\mathcal{S}^*} \mathbf{s}\right)$.

By invoking calculus of subgradients



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, $f(\mathbf{x}) = ||\mathbf{x}||_p$

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

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

HOMEWORK: Try and Derive the Subgradients.
For the Norm case, you can use the Holder's inequality stated on the next slide

RECALL CAUCHY SHWARZ

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

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

Generalized to

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

HOLDER'S INEQUALITY

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

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)

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

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, $f(\mathbf{x}) = ||\mathbf{x}||_p = \max\limits_{||\mathbf{z}||_q \le 1} \mathbf{z}^T\mathbf{x}$ where q is such that $1/p + 1/q = 1$. Then $\partial f(\mathbf{x}) = \left\{\mathbf{y} : ||\mathbf{y}||_q \le 1 \text{ and } \mathbf{y}^T x = \max\limits_{||\mathbf{z}||_q \le 1} \mathbf{z}^T\mathbf{x}\right\}$
- Can we derive the sub-differential of $||x||_1$?