ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 2-2: Convexity

Jia (Kevin) Liu

Assistant Professor
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, USA

Spring 2022

Outline

Today:

- Convex sets
- Convex functions
- Key properties
- Operations preserving convexity

Recap the Very First Lecture

Mathematical optimization problem:

$$\begin{array}{ll} \text{Minimize} & f_0(\mathbf{x}) \\ \text{subject to} & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \end{array}$$

- $\mathbf{x} = [x_1, \ldots, x_N]^\top \in \mathbb{R}^N$: decision variables
- $f_0 : \mathbb{R}^N \to \mathbb{R}$: objective function
- $f_i : \mathbb{R}^N \to \mathbb{R}, i = 1, \ldots, m$: constraint fucntions

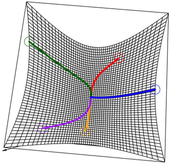
Solution or **optimal point** $\mathbf{x}^*$ has the smallest value of f_0 among all vectors that satisfy the constraints

Watershed between Problem Hardness: Convexity

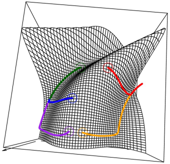
Why Do We Care About Convexity?

For convex optimization problem, local minima are global minima

Formally: Let $\mathcal{D}$ be the feasible domain defined by the constraints. If $\mathbf{x} \in \mathcal{D}$ satisfies the following local condition: $\exists\ d > 0$ such that for all $\mathbf{y} \in \mathcal{D}$ satisfying $\|\mathbf{x} - \mathbf{y}\|_2 \leq d$, we have $f_0(\mathbf{x}) \leq f_0(\mathbf{y})$. $\Rightarrow$ $f_0(\mathbf{x}) \leq f_0(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{D}$.



Convex



Nonconvex

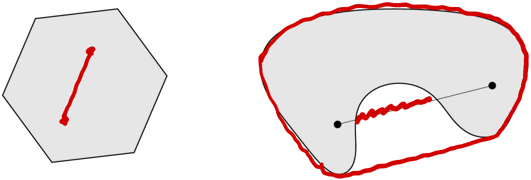
A crucial fact that would significantly reduce the complexity in optimization!

Convex Sets

Convex set: A set $\mathcal{D} \in \mathbb{R}^n$ such that

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{D} \quad \Rightarrow \quad \mu\mathbf{x} + (1-\mu)\mathbf{y} \in \mathcal{D}, \quad \forall 0 \leq \mu \leq 1$$

Geometrically, line segment joining any two points in $\mathcal{D}$ lies in entirely in $\mathcal{D}$



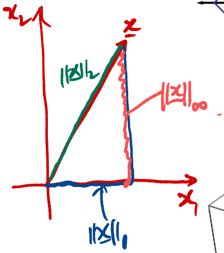
Convex combination: A linear combination $\mu_1\mathbf{x}_1 + \cdots + \mu_k\mathbf{x}_k$ for $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$, with $\mu_i \geq 0$, $i = 1, \ldots, k$ and $\sum_{i=1}^k \mu_i = 1$.

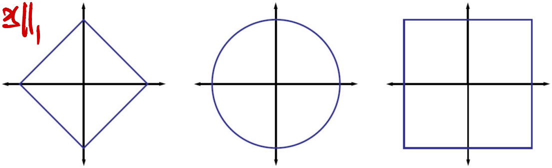
Convex hull: A set defined by all convex combinations of elements in a set $\mathcal{D}$.

Examples of Convex Sets

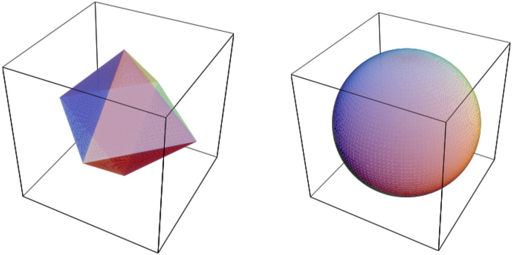
1) Norm balls: Radius r ball in l_p norm $\mathcal{B}_p = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p \leq r\}$

$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$





$p = 1$ $p = 2$ $p = \infty$

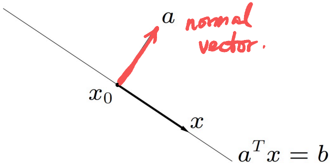


$p = 1$ $p = 2$

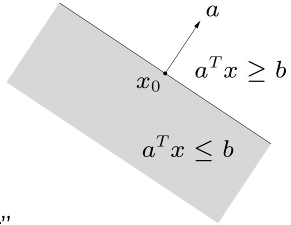
Examples of Convex Sets

2) Hyperplane and haflspaces

- Hyperplane: Set of the form $\{\mathbf{x}|\mathbf{a}^\top\mathbf{x} = b\}$ with $\mathbf{a} \neq \mathbf{0}$



- Halfspace: Set of the form $\{\mathbf{x}|\mathbf{a}^\top\mathbf{x} \leq b\}$ with $\mathbf{a} \neq \mathbf{0}$

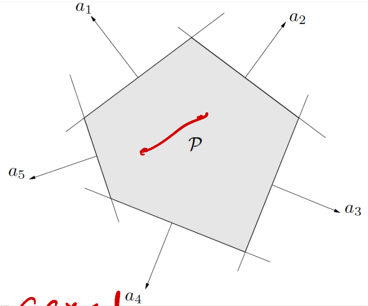


- $\mathbf{a}$ is called "normal vector"

Examples of Convex Sets

$\underline{x} \leq \underline{y} \Leftrightarrow x_i \leq y_i, \forall i$

3) Polyhedron: $\{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$, whre $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\leq$ is component-wise inequality



$$\begin{cases} \underline{C}\underline{x} \leq \underline{d} \\ \underline{C}\underline{x} \geq \underline{d} \rightarrow -\underline{C}\underline{x} \leq -\underline{d} \end{cases}$$

Note:

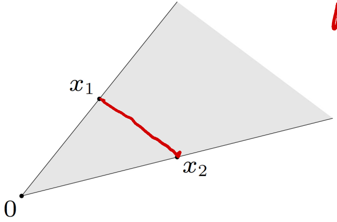
- $\{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{C}\mathbf{x} = \mathbf{d}\}$ is also a polyhedron (Why?)
- Polyhedron is an intersection of finite number of halfspaces and hyperplanes

Examples of Convex Sets

Cones: $\mathcal{K} \subseteq \mathbb{R}^n$ such that $\mathbf{x} \in \mathcal{K} \Rightarrow t\mathbf{x} \in \mathcal{K}, \quad \forall t \geq 0$

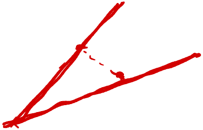
Convex Cones: A cone that is convex, i.e.,

$$\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{K} \quad \Rightarrow \quad \mu_1\mathbf{x}_1 + \mu_2\mathbf{x}_2 \in \mathcal{K}, \quad \forall \mu_1, \mu_2 \geq 0$$



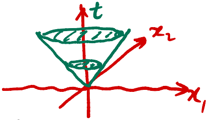
$\mu_1 + \mu_2 = 1$

$\mu_1, \mu_2 \in \{0,1\}$.



Conic Combination: For $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$, a linear combination $\mu_1\mathbf{x}_1 + \cdots + \mu_k\mathbf{x}_k$ with $\mu_i \geq 0$, $i = 1, \ldots, k$. Conic hull collects all conic combinations

Examples of Convex Sets



ice cream cone.

- Norm Cones: $\{(\mathbf{x}, t) \in \mathbb{R}^{d+1} : \|\mathbf{x}\| \le t\}$ for some norm $\|\cdot\|$ (the norm cone for l_2 norm is referred to as second-order cone)

- Normal Cone: Given any set $\mathcal{C}$ and at a boundary point $\mathbf{x} \in \mathcal{C}$, we define

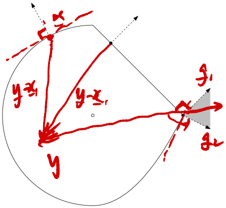
$$\mathcal{N}_{\mathcal{C}}(\mathbf{x}) = \{\mathbf{g} : \mathbf{g}^\top (\mathbf{y} - \mathbf{x}) \le 0, \ \forall \mathbf{y} \in \mathcal{C}\}$$

pick $\underline{x}_1, \underline{x}_2$

$\angle(\underline{x}, \underline{y})$

$= \cos^{-1}\left(\frac{\underline{x}^T \underline{y}}{\|\underline{x}\|_2 \|\underline{y}\|_2}\right) \le 0$

$\Rightarrow \in [\frac{\pi}{2}, \pi]$



This is always a convex cone, regardless of $\mathcal{C}$

symmetric n×n matrices.

$\underline{x}^T \underline{A}\, \underline{x} \ge 0.$

- Positive Semidefnite Cone: $\mathbb{S}^n_+ \triangleq \{\mathbf{X} \in \mathbb{S}^n : \mathbf{X} \succeq 0\}$, where $\mathbf{X} \succeq 0$ represents $\mathbf{X}$ is positive semidefinite and $\mathbb{S}^n$ is the set of $n \times n$ symmetric matrices. $\underline{x}^T \underline{X}_2 \underline{x} \ge 0$

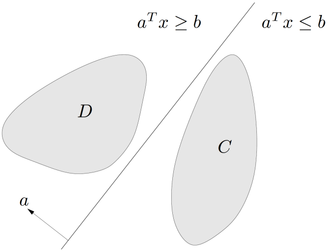
pick 2 elements $\underline{X}_1, \underline{X}_2 \in \mathbb{S}^n_+$

$\underline{x}^T(\mu \underline{X}_1 + (1-\mu)\underline{X}_2)\underline{x} = \mu \underbrace{\underline{x}^T \underline{X}_1 \underline{x}}_{\ge 0} + (1-\mu) \checkmark$

$\ge 0.$

Key Properties of Convex Sets

- Separating hyperplane theorem: Two disjoint convex sets have a separating hyperplane between them

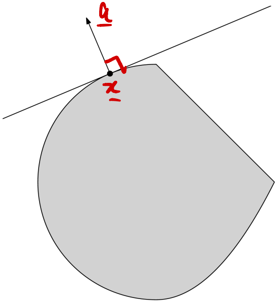


- More precisely, if $\mathcal{C}$ and $\mathcal{D}$ are non-empty convex sets with $\mathcal{C} \cap \mathcal{D} = \varnothing$, then there exists $\mathbf{a}$ and b such that:

$$\mathcal{C} \subseteq \{\mathbf{x} : \mathbf{a}^\top \mathbf{x} \leq b\}, \quad \mathcal{D} \subseteq \{\mathbf{x} : \mathbf{a}^\top \mathbf{x} \geq b\},$$

Key Properties of Convex Sets

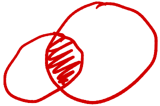
- Supporting hyperplane theorem: A boundary point of a convex set has a supporting hyperplane passing through it



- More precisely, if $\mathcal{C}$ is a non-empty convex set and $\mathbf{x}_0 \in \partial\mathcal{C}$, there exists a vector $\mathbf{a}$ such that:

$$\mathcal{C} = \{\mathbf{x} : \mathbf{a}^\top(\mathbf{x} - \mathbf{x}_0) \le 0\}$$

Operations That Preserve Convexity of Sets



- Intersection: The intersection of convex sets is convex
- Scaling and Translation: If $\mathcal{C}$ is convex, then $a\mathcal{C} + \mathbf{b} \triangleq \{a\mathbf{x} + \mathbf{b} : \mathbf{x} \in \mathcal{C}\}$ is also convex for any a and $\mathbf{b}$.

  scaling translation.

- Affine image and preimage: If $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ and $\mathcal{C}$ is convex, then

$$f(\mathcal{C}) \triangleq \{f(\mathbf{x}) : \mathbf{x} \in \mathcal{C}\}$$

  is also convex. If $\mathcal{D}$ is convex, then

$$f^{-1}(\mathcal{D}) \triangleq \{\mathbf{x} : f(\mathbf{x}) \in \mathcal{D}\}$$

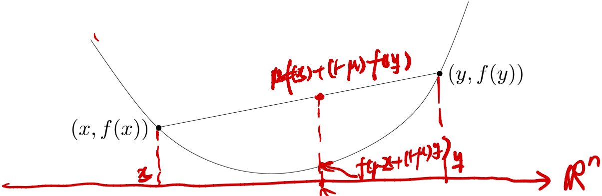
  is also convex

Convex Functions

- Convex function: $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ is convex if $\mathrm{dom}(f) \in \mathbb{R}^n$ is convex and

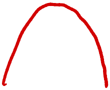
$$f\big(\mu\mathbf{x} + (1-\mu)\mathbf{y}\big) \le \mu f(\mathbf{x}) + (1-\mu) f(\mathbf{y})$$

for all $\mu \in [0,1]$ and for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$.



In words, f lies below the line segment that joins any $f(\mathbf{x})$ and $f(\mathbf{y})$.

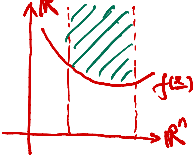
- Concave function: f concave $\Longleftrightarrow$ $-f$ convex



Key Properties of Convex Functions

- Epigraph characterization: A function f is convex if and only if its epigraph

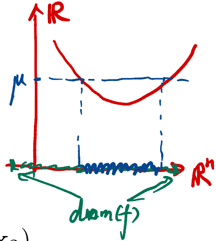
$$\mathrm{ep}(f) \triangleq \{(\mathbf{x}, \mu) \in \mathrm{dom}(f) \times \mathbb{R} : f(\mathbf{x}) \leq \mu\}$$



  is a convex set

- Convex sublevel set: If f is convex, then its sublevel set

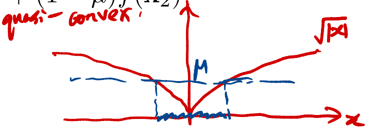
$$\{\mathbf{x} \in \mathrm{dom}(f) : f(\mathbf{x}) \leq \mu\}$$



  is convex for all $\mu \in \mathbb{R}$ (but the converse is not true)

- Jensen's inequality: If f is convex, then

$$f\big(\mu\mathbf{x}_1 + (1-\mu)\mathbf{x}_2\big) \leq \mu f(\mathbf{x}_1) + (1-\mu) f(\mathbf{x}_2)$$



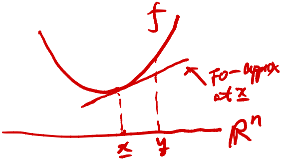
  for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathrm{dom}(f)$ and $0 \leq \mu \leq 1$

Other Important Characterizations of Convex Functions

- First-order characterization: If f is differentiable, then f is convex if and only if $\mathrm{dom}(f)$ is convex, and

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \underbrace{\nabla f^\top(\mathbf{x})}_{= 0}(\mathbf{y} - \mathbf{x})$$

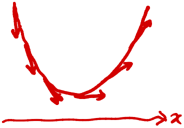
for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$.



- Implying an important consequence: $\underbrace{\nabla f(\mathbf{x}) = 0}_{\Downarrow \; f(y) \geq f(x)} \Longrightarrow \mathbf{x}$ minimizes f

- Second-order characterization: If f is twice differentiable, then f is convex if and only if $\mathrm{dom}(f)$ is convex, and $\mathbf{H}(\mathbf{x}) = \nabla^2 f(\mathbf{x}) \succeq 0$ for all $\mathbf{x} \in \mathrm{dom}(f)$

$$\left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]_{ij} \succeq 0.$$



Important Convexity Notions

- Strictly convex: $f\big(\mu\mathbf{x} + (1-\mu)\mathbf{y}\big) < \mu f(\mathbf{x}) + (1-\mu) f(\mathbf{y})$, i.e., f is convex and has greater curvature than a linear function

- Strongly convex with parameter m: $f(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|^2$ is convex, i.e., f is at least as curvy as a m-parameterized quadratic function

(HW): show: $f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}\|y-x\|^2$

- Note: strongly convex $\Rightarrow$ strictly convex $\Rightarrow$ convex, (converse is not true)

- Similar notions for concave functions

$f(x) = \frac{1}{x} - \frac{m}{2}x^2$ $\qquad$ $f''(x) = 2x^{-3} - m. \geq 0$

$f'(x) = -x^{-2} - mx$

Important Examples of Convex/Concave Functions

- Univariate functions:
  - Exponential functions: e^{ax} is convex for all $a \in \mathbb{R}$
  - Power functions: x^a is convex if $a \in (-\infty, 0] \cup [1, \infty)$ and concave if $a \in [0, 1]$
  - Logarithmic functions: $\log(x)$ is concave for $x > 0$
- Affine function: $\mathbf{a}^\top \mathbf{x} + \mathbf{b}$ is both concave and convex
- Quadratic function: $\frac{1}{2}\mathbf{x}^\top \mathbf{Q} \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$ is convex if $\mathbf{Q} \succeq 0$ (positive semidefinite)
- Least square loss function: $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$ is always convex (since $\mathbf{A}^\top \mathbf{A} \succeq 0$)

  $= (\underline{y} - \underline{A}\underline{x})^T(\underline{y} - \underline{A}\underline{x}) \Rightarrow \underline{Q} = 2\underline{A}^T\underline{A} \leftarrow$ PSD.
- Norm: $\|\mathbf{x}\|$ is always convex for any norm, e.g.,
  - l_p norm: $\|\mathbf{x}\|_p = (\sum_{i=1}^n x_i^p)^{\frac{1}{p}}$ for $p \geq 1$, $\|\mathbf{x}\|_\infty = \max_{i=1,\ldots,n}\{|x_i|\}$
  - Matrix operator (spectral) norm $\|\mathbf{X}\|_{\mathrm{op}} = \sigma_1(\mathbf{X})$
    Matrix trace (nuclear) norm $\|\mathbf{X}\|_{\mathrm{tr}} = \sum_{i=1}^r \sigma_r(\mathbf{X})$, where $\sigma_1(\mathbf{X}) \geq \cdots \geq \sigma_r(\mathbf{X}) \geq 0$ are the singular values of $\mathbf{X}$

More Examples of Convex/Concave Functions

- Indicator function: If $\mathcal{C}$ is convex, then its indicator function

$$\mathbb{1}_{\mathcal{C}}(\mathbf{x}) = \begin{cases} 0 & \mathbf{x} \in \mathcal{C} \\ \infty & \mathbf{x} \notin \mathcal{C} \end{cases}$$

  is convex

- Support function: For any set $\mathcal{C}$ (convex or not), its support function

$$\mathbb{1}^*_{\mathcal{C}}(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{C}} \mathbf{x}^\top \mathbf{y}$$

  is convex

Proof. $\mathbb{1}_C^*(\mu \underline{x}_1 + (1-\mu)\underline{x}_2) = \max_{y \in C} \underbrace{(\mu \underline{x}_1 + (1-\mu)\underline{x}_2)^T y}_{f}$

$= \mu \underline{x}_1^T \hat{y} + (1-\mu)\underline{x}_2^T \hat{y} \le \mu \max_{y\in C} \underline{x}_1^T y + (1-\mu) \max_{y \in C} \underline{x}_2^T y$

$\hat{y} \in \arg\max_{y \in C} f$

$= \mu \mathbb{1}_C^*(\underline{x}_1) + (1-\mu)\mathbb{1}_C^*(\underline{x}_2).$

- Max function: $f(\mathbf{x}) = \max\{x_1, \ldots, x_n\}$ is convex

Min fn is concave

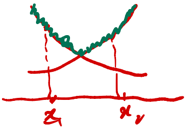
Operations That Preserve Convexity of Functions

- Nonnegative linear combinations: $f_1, \dots, f_m$ being convex implies $\mu_1 f_1 + \cdots + \mu_m f_m$ is convex for any $\mu_1, \dots, \mu_m \geq 0$

- Pointwise maximization: If f_i is convex for any index $i \in \mathcal{I}$, then

  e.g., $\mathbb{1}_{\mathcal{C}}^*(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{C}} \mathbf{x}^T \mathbf{y}$

  $$f(\mathbf{x}) = \max_{i \in \mathcal{I}} f_i(\mathbf{x})$$

  is convex. Note that the index set $\mathcal{I}$ can be infinite

- Partial minimization: If $g(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}, \mathbf{y}$ and $\mathcal{C}$ is convex, then

  $$f(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{C}} g(\mathbf{x}, \mathbf{y})$$

  fixed.

  is convex (the basis for ADMM, coordinate descent, ...)

Examples of Composite Operations to Prove Convexity

Example 1: Let $\mathcal{C}$ be an arbitrary set. Show that maximum distance to $\mathcal{C}$ under an arbitrary norm $\|\cdot\|$, i.e., $f(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{C}} \|\mathbf{x}-\mathbf{y}\|$ is convex.

Proof.

- Note that $f_{\mathbf{y}}(\mathbf{x}) = \|\mathbf{x}-\mathbf{y}\|$ is convex in $\mathbf{x}$ for any fixed $\mathbf{y}$.
- By pointwise maximization rule, f is convex. □

Example 2: Let $\mathcal{C}$ be a convex set. Show that minimum distance to $\mathcal{C}$ under an arbitrary norm $\|\cdot\|$, i.e., $f(\mathbf{x}) = \min_{\mathbf{y}\in\mathcal{C}} \|\mathbf{x}-\mathbf{y}\|$ is also convex.

Proof.

- Note that $f(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$ is convex in both $\mathbf{x}$ and $\mathbf{y}$.
- $\mathcal{C}$ is convex by assumption.
- By partial minimization rule, f is convex. □

More Operations That Preserve Convexity of Functions

- Affine composition: f is convex $\Longrightarrow g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$ is convex

- General composition: Suppose $f = h \circ g$, where $g : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$. Then:
  - f is convex if h is convex & nondecreasing, g is convex
  - f is convex if h is convex & nonincreasing, g is concave
  - f is concave if h is concave & nondecreasing, g is concave
  - f is concave if h is concave & nonincreasing, g is convex

  How to remember these? Think of the chain rule when $n = 1$

$$f''(x) = \underbrace{h''(g(x))}_{\geq 0}\underbrace{g'(x)^2}_{\geq 0} + \underbrace{h'(g(x))}_{\geq 0}\underbrace{g''(x)}_{\geq 0} \quad \geq 0.$$

Generalization

- Vector-valued composition: Suppose that

$$f(\mathbf{x}) = h(\mathbf{g}(\mathbf{x})) = h(g_1(\mathbf{x}), \ldots, g_k(\mathbf{x}))$$

where $g : \mathbb{R}^n \to \mathbb{R}^k$, $h : \mathbb{R}^k \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$. Then:

  - f is convex if h is convex & nondecreasing in each argument, g is convex
  - f is convex if h is convex & nonincreasing in each argument, g is concave
  - f is concave if h is concave & nondecreasing in each argument, g is concave
  - f is concave if h is concave & nonincreasing in each argument, g is convex

Example of Composite Operations to Prove Convexity

Log-sum-exp function: Show that $g(\mathbf{x}) = \log(\sum_{i=1}^{k} \exp(\mathbf{a}_i^\top \mathbf{x} + b_i))$ is convex, where $\mathbf{a}_i, b_i$, $i = 1, \ldots, k$ are fixed parameters (often called "soft max" in ML literature since it smoothly approximates $\max_{i=1,\ldots,k}(\mathbf{a}_i^\top \mathbf{x} + b_i)$.

lin. op. preserve convexity

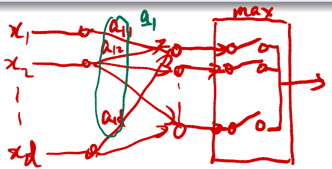
Proof.

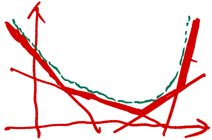
- Note that it suffices to prove $f(\mathbf{x}) = \log(\sum_{i=1}^{n} \exp(x_i))$ is convex (Why?)
- According to second-order characterization, compute the Hessian to obtain:

$$\nabla^2 f(\mathbf{x}) = \mathrm{Diag}\{\mathbf{z}\} - \mathbf{z}\mathbf{z}^\top$$

where $(\mathbf{z})_i = e^{x_i}/(\sum_{l=1}^{n} e^{x_l})$. This matrix is diagonally dominant $\Rightarrow$ PSD. $\square$

NN:





Next Class

Gradient Descent