Optimization for Machine Learning

Lecture 1: Introduction

Sebastian Stich

CISPA – `https://cms.cispa.saarland/optml24/`

April 16, 2024

Quizz

Bachelor / Master / Phd

2 : 10 : 7

Optimization 4

ML ✓

Optimization?

objective funtion $f(x)$ $\ell(w)$

$$\min_{x} f(x)$$

x variable

$x \in C$ constraint

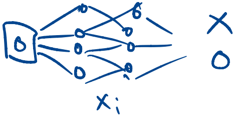
Gradient Descent:

$$x_{t+1} = x_t - \gamma \cdot \overbrace{\nabla f(x_t)}^{\text{Gradient}}$$

"guess" Parameter

Machine Learning Optimization Problems?

model: $x \in \mathbb{R}^d$



"loss" $\qquad \ell_i(x, \overset{\text{input}}{\boxed{\otimes}}, \overset{\text{label}}{"x"})$ $\begin{cases} > 0 & \text{if error} \\ 0 & \text{if prediction is corr.} \end{cases}$

many training samples

$$\ell(x) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(x, \square, " \cdot ")$$ $\leftarrow$ optimizer

General Course Information

Outline

- Prerequisites:
  - Convexity, Linear Algebra
- Main Contents:
  - Gradient Methods, Coordinate Descent, **Stochastic Gradient Descent**
  - **Convergence (proofs)** of SGD on different function classes, impact of of **batch size, momentum, learning late**, etc.
  - variance reduction, adaptive methods
  - Parallel and Distributed Optimization Algorithms, Decentralized and Federated Optimization,
- Advanced Contents:
  - Computational Trade-Offs (Time vs Data vs Accuracy), Lower Bounds, Proximal algorithms, Subgradient Methods, 1–2 recent research papers

Course Organization (Hybrid Format)

- All lectures will be streamed on **zoom**.
- Some lectures can be attended live in **CISPA room 0.01**. (20–30 seats)
- You find **all materials** on the course website https://cms.cispa.saarland/optml24/

Updates 2024:

- **recordings** will be made available on a 'best-effort' basis
- no script, instead we will post additional **reading sources** ($\approx$ 10 pages) each week (there is also a script, but its is very lengthy, so only recommended if you are missing background materials and need to spend additional hours per week to catch up)

Course Organization (6 ETCS = 180hrs)

Lectures	≈ 30hrs
Exercises	≈ 40hrs
Mini-Project	≈ 40hrs
Self-study/exam preparation	≈ 70hrs

I strongly recommend to regularly invest one day a week to study the course material (lecture, slides, reading materials, tutorial, exercises, project, or alternative sources).
15 weeks × 8hrs ≈ 120hrs

Course Organization (Grading)

Final exam criteria:

- pass the **mini-project**
  - groups that fail the mini-project, can submit a revised version

Grading: 25% midterm + 75% final exam.

- (to calculate the weighted average, the points from both exams will be normalized to the same scale)
- **written** exam, closed book
- you can bring one sheet of A4 paper with your own notes (handwritten, or latex, font $\geq$ 10pt)

Exam Date: TBA

only one exam offered this year!

See details on the course webpage.

Lectures & Exercises

Lecture

- 5–10 mins: recap, **quizz**
- 60-75 mins: new materials
- 5-15 mins: discussion
- ! 60-120 mins: self study

Exercises

- a sheet every week
- there are many more exercises available in the lecture notes, or in old exams
- ! Exercises are not mandatory & not graded, but part of the course material.

Course Organization (Exercises & Tutorials)

- **Tutorials/Q&A session**.
  - The assistants are
    - Xiaowen Jiang, <xiaowen.jiang@cispa.de>
    - Yuan Gao, <yuan.gao@cispa.de>
  - 3-4pm on Tuesdays (can adjusted, upon demand)
  - it is highly recommended that to attend the exercise sessions, to either
    - work on the exercises
    - ask questions about the exercises
    - ask **questions about any topic** of the course!
- **Office hours.** You can reach me after class.
- **Please use the forum** for general questions and to discuss the exercises.

Mini-Project

- small project with focus on the practical implementation (or deepening of a theoretical aspect)
- can be submitted in **groups of 3 students**
- start: after the midterm exam
- The projects will be graded on a scale of fail, pass, good (top 30%, 0.3 bonus), excellent (top 10%, 0.6 bonus). You are required to pass the project to take part in the exam. If you pass the exam, eventual bonus points from the project will be subtracted to improve your final grade.

See details on the course webpage.

Main Course Materials & Acknowledgment

- The EPFL Opt4ML course, (https://github.com/epfml/OptML_course) gladly shared their
  - Lecture notes
  - Exercises (with solutions)
  - and Python notebooks

Lecture 1

- convexity "what is a convex function"
- minimizer/optimal solution

Optimization

- General optimization problem (**unconstrained minimization**)

$$\begin{array}{rl} \text{minimize} & f(\mathbf{x}) \\ \text{with} & \mathbf{x} \in \mathbb{R}^d \end{array}$$

  - candidate solutions, variables, parameters $\mathbf{x} \in \mathbb{R}^d$
  - objective function $f : \mathbb{R}^d \to \mathbb{R}$
  - typically: technical assumption: f is continuous and differentiable

Optimization for Machine Learning

- **Mathematical Modeling**:
  - defining & and measuring the machine learning model
- **Computational Optimization**:
  - learning the model parameters
- Theory vs. practice:
  - libraries are available, algorithms treated as "black box" by most practitioners
  - **Not here:** we look inside the algorithms and try to understand why and how fast they work!

Optimization Algorithms

- Optimization at large scale: **simplicity** rules!
- Main approaches:
  - **Gradient Descent**
  - **Stochastic Gradient Descent** (SGD)
  - **Coordinate Descent**
- History:
  - 1847: Cauchy proposes gradient descent
  - 1950s: Linear Programs, soon followed by non-linear, SGD
  - 1980s: General optimization, convergence theory
  - 2005-2015: Large scale optimization (mostly convex), convergence of SGD
  - 2015-today: Improved understanding of SGD for deep learning

Chapter 2

Theory of Convex Functions

Convex Sets

A set C is **convex** if the line segment between any two points of C lies in C, i.e., if for any $\mathbf{x}, \mathbf{y} \in C$ and any λ with $0 \leq \lambda \leq 1$, we have

$$\underbrace{\lambda \underline{\mathbf{x}} + (1-\lambda)\underline{\mathbf{y}}}_{\text{line segment}} \in C.$$

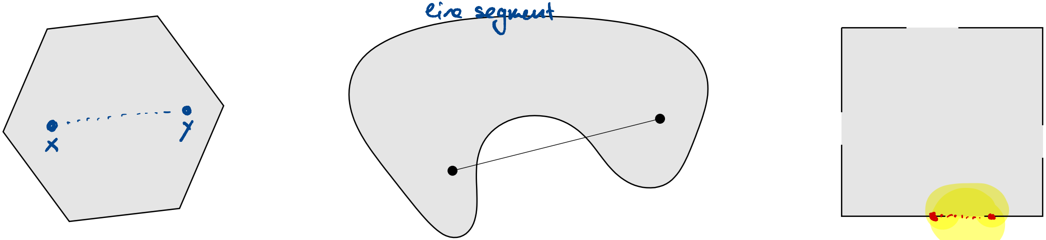


*Figure 2.2 from S. Boyd, L. Vandenberghe

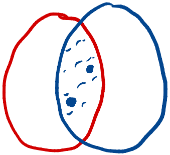
Left Convex.

Middle Not convex, since line segment not in set.

Right Not convex, since some, but not all boundary points are contained in the set.

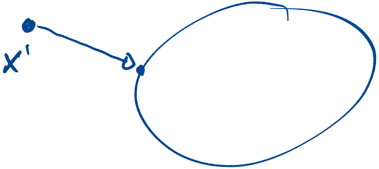
Properties of Convex Sets

- Intersections of convex sets are convex

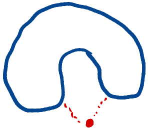


Observation 1.2. Let $C_i, i \in I$ be convex sets, where I is a (possibly infinite) index set. Then $C = \bigcap_{i \in I} C_i$ is a convex set.

- (later) Projections onto convex sets are *unique*, and *often* efficient to compute



$P_C(\mathbf{x}') := \operatorname{argmin}_{\mathbf{y} \in C} \|\mathbf{y} - \mathbf{x}'\|$



Convex Functions

Definition

A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if (i) $\mathbf{dom}(f)$ is a convex set and (ii) for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$, and λ with $0 \le \lambda \le 1$, we have

$$f(\underbrace{\lambda \mathbf{x} + (1-\lambda)\mathbf{y}}_{x'}) \le \underbrace{\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})}.$$

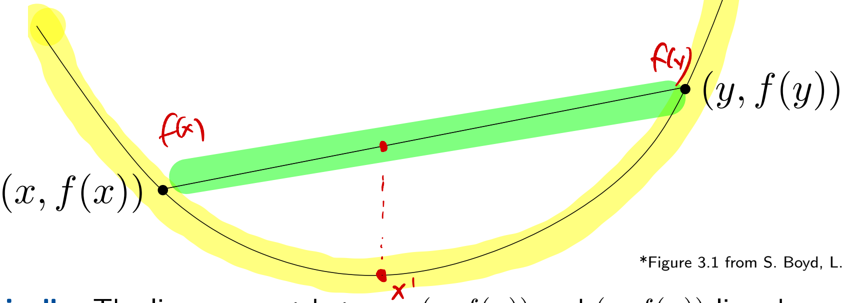


*Figure 3.1 from S. Boyd, L. Vandenberghe

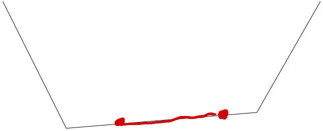
Geometrically: The line segment between $(\mathbf{x}, f(\mathbf{x}))$ and $(\mathbf{y}, f(\mathbf{y}))$ lies above the graph of f.

Strictly Convex Functions

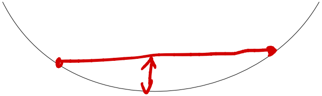
Definition (Lecture-1).1 ([BV04, 3.1.1])

A function $f : \mathbf{dom}(f) \to \mathbb{R}$ is **strictly convex** if (i) $\mathbf{dom}(f)$ is convex and (ii) for all $\mathbf{x} \neq \mathbf{y} \in \mathbf{dom}(f)$ and all $\lambda \in (0,1)$, we have

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}). \tag{1}$$

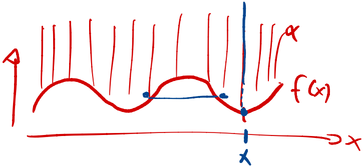


convex, but not strictly convex



strictly convex

Convex Functions & Sets



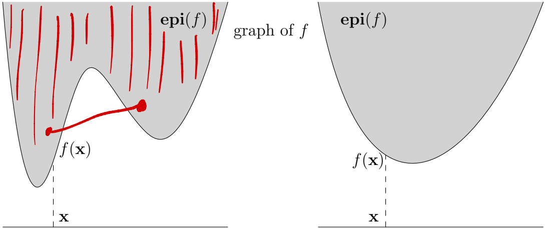
The **graph** of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\{(\mathbf{x}, f(\mathbf{x})) \,|\, \mathbf{x} \in \mathbf{dom}(f)\},$$

The **epigraph** of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathbf{epi}(f) := \{(\mathbf{x}, \underline{\alpha}) \in \mathbb{R}^{d+1} \,|\, \mathbf{x} \in \mathbf{dom}(f), \underline{\alpha} \geq \underline{f(\mathbf{x})}\},$$

Observation 1.4. A function is convex *iff* its epigraph is a convex set.



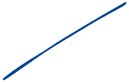
Convex Functions & Sets

Proof: recall $\mathbf{epi}(f) := \{(\mathbf{x}, \alpha) \in \mathbb{R}^{d+1} \,|\, \mathbf{x} \in \mathbf{dom}(f), \alpha \geq f(\mathbf{x})\}$

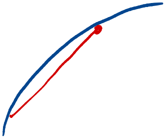
Examples

- $f(x) = x^2$

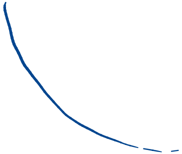
- $f(x) = x$



- $\log(x)$



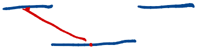
$-\log(x)$



- indicator function of a set C

$$\begin{cases} \cancel{1}\,\infty & x \notin C \\ 0 & x \in C \end{cases}$$

$C = [0, 1]$



$C = [0, 1] \cup [2, 3]$

Gradients and Differentiable Functions

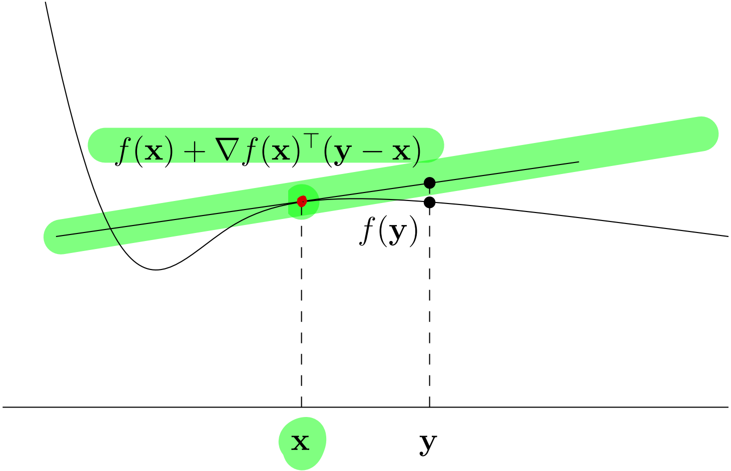
Gradient vector

For a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$,

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \\ \\ \frac{\partial f(x)}{\partial x_d} \end{bmatrix}$$

Differentiable Functions

Graph of the affine function $f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$ is a tangent hyperplane to the graph of f at $(\mathbf{x}, f(\mathbf{x}))$.


$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})$
$f(\mathbf{y})$
$\mathbf{x}$
$\mathbf{y}$

linear approximation

First-order Characterization of Convexity

Lemma (Lecture-1).2 ([BV04, 3.1.3])
Suppose that $\mathbf{dom}(f)$ *is open and that* f *is differentiable; in particular, the* **gradient** *(vector of partial derivatives)*

$$\nabla f(\mathbf{x}) := \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_d}(\mathbf{x})\right)$$

exists at every point $\mathbf{x} \in \mathbf{dom}(f)$*. Then* f *is convex if and only if* $\mathbf{dom}(f)$ *is convex and*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \tag{2}$$

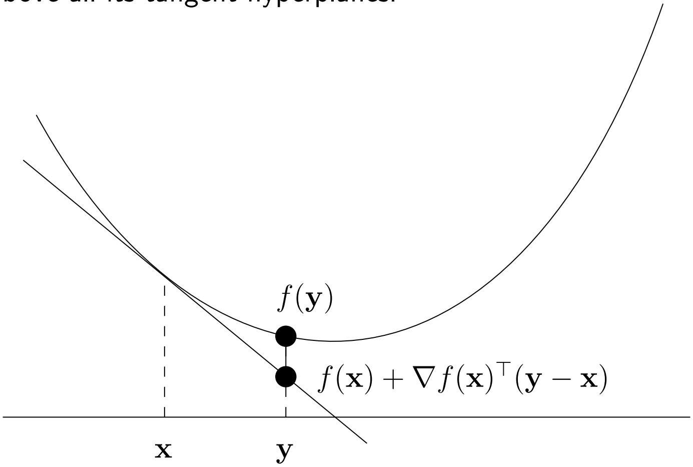
holds for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$*.*

First-order Characterization of Convexity

linear approximation

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y}-\mathbf{x}), \quad \mathbf{x},\mathbf{y} \in \mathbf{dom}(f).$$

Graph of f is above all its tangent hyperplanes.


$f(\mathbf{y})$
$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y}-\mathbf{x})$
$\mathbf{x}$
$\mathbf{y}$

Second-order Characterization of Convexity

Lemma (Lecture-1).3 ([BV04, 3.1.4])

Suppose that $\mathbf{dom}(f)$ *is open and that* f *is twice differentiable; in particular, the* **Hessian** *(matrix of second partial derivatives)*

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(\mathbf{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d}(\mathbf{x}) \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_d \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_d}(\mathbf{x}) \end{pmatrix}$$

exists at every point $\mathbf{x} \in \mathbf{dom}(f)$ *and is symmetric. Then* f *is convex if and only if* $\mathbf{dom}(f)$ *is convex, and for all* $\mathbf{x} \in \mathbf{dom}(f)$*, we have*

$$\nabla^2 f(\mathbf{x}) \succeq 0 \quad \textit{(i.e. } \nabla^2 f(\mathbf{x}) \textit{ is positive semidefinite).}$$

(A symmetric matrix M *is positive semidefinite if* $\mathbf{x}^\top M \mathbf{x} \geq 0$ *for all* $\mathbf{x}$*, and positive definite if* $\mathbf{x}^\top M \mathbf{x} > 0$ *for all* $\mathbf{x} \neq \mathbf{0}$*.)*

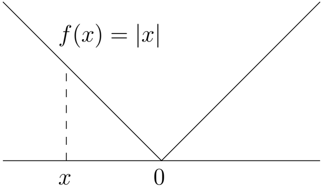
Second-order Characterization of Convexity

Example: $f(x_1, x_2) = x_1^2 + x_2^2$.

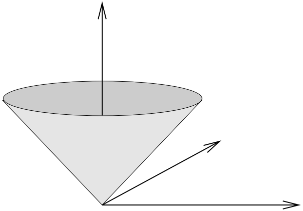
$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \succeq 0.$$

Nondifferentiable Functions. . .

are also relevant in practice.



More generally, $f(\mathbf{x}) = \|\mathbf{x}\|$ (Euclidean norm). For $d = 2$, graph is the ice cream cone:



Convex Optimization Problems

Motivation: Convex Optimization

Convex Optimization Problems are of the form

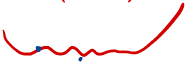
$$\min \ f(\mathbf{x}) \qquad \text{s.t.} \qquad \mathbf{x} \in X$$

where both

- f is a convex function
- $X \subseteq \mathbf{dom}(f)$ is a convex set (note: $\mathbb{R}^d$ is convex)

Local Minima and Critical Points

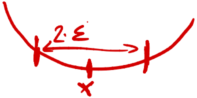
$\min f(x) = f(x^*)$



Definition (Lecture-1).4

A **local minimum** of $f : \mathbf{dom}(f) \to \mathbb{R}$ is a point $\mathbf{x}$ such that there exists $\varepsilon > 0$ with

$$f(\mathbf{x}) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathbf{dom}(f) \text{ satisfying } \|\mathbf{y} - \mathbf{x}\| < \varepsilon.$$



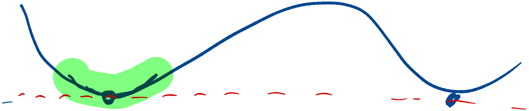
Definition (Lecture-1).5

A **critical point** of a differentiable function $f : \mathbf{dom}(f) \to \mathbb{R}$ is a point $\mathbf{x}$ such that

$$\nabla f(\mathbf{x}) = 0.$$

Note: $\|\nabla f(x)\| \leq \varepsilon$ "approximate critical point"

Local Minima are Global Minima



Lemma (Lecture-1).6

Let $\mathbf{x}^\star$ be a **local minimum** *of a convex function $f : \mathbf{dom}(f) \to \mathbb{R}$. Then $\mathbf{x}^\star$* **is a global minimum**, *meaning that $f(\mathbf{x}^\star) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in \mathbf{dom}(f)$.*

Proof.

Suppose there exists $\mathbf{y} \in \mathbf{dom}(f)$ such that $f(\mathbf{y}) < f(\mathbf{x}^\star)$.
Define $\mathbf{y}' := \lambda\mathbf{x}^\star + (1-\lambda)\mathbf{y}$ for $\lambda \in (0,1)$.
From convexity, we get that that $f(\mathbf{y}') < f(\mathbf{x}^\star)$. Choosing λ so close to 1 that $\|\mathbf{y}' - \mathbf{x}^\star\| < \varepsilon$ yields a contradiction to $\mathbf{x}^\star$ being a local minimum. □

Critical Points are Global Minima

Lemma (Lecture-1).7

Suppose that f is convex and differentiable over an open domain $\mathbf{dom}(f)$. Let $\mathbf{x} \in \mathbf{dom}(f)$. If $\nabla f(\mathbf{x}) = \mathbf{0}$ (**critical point**)*, then $\mathbf{x}$ is a* **global minimum***.*

Proof.

Suppose that $\nabla f(\mathbf{x}) = \mathbf{0}$. According to our Lemma on the first-order characterization of convexity, we have

$$f(y) \geq f(x) + \langle \underbrace{\nabla f(x)}_{=0}, y - x \rangle$$

□

Geometrically, tangent hyperplane is horizontal at $\mathbf{x}$.

Useful inequalities

Convex Functions

Examples of convex functions

- Linear functions: $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$
- Affine functions: $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$
- Exponential: $f(x) = e^{\alpha x}$
- Norms. Every norm on $\mathbb{R}^d$ is convex.

Convexity of a norm $\|\mathbf{x}\|$

By the triangle inequality $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ and homogeneity of a norm $\|a\mathbf{x}\| = |a| \, \|\mathbf{x}\|$, a scalar:

$$\|\lambda \mathbf{x} + (1-\lambda)\mathbf{y}\| \leq \|\lambda \mathbf{x}\| + \|(1-\lambda)\mathbf{y}\| = \lambda \|\mathbf{x}\| + (1-\lambda) \|\mathbf{y}\| .$$

We used the triangle inequality for the inequality and homogeneity for the equality.

Jensen's Inequality

Lemma (Lecture-1).8 (Jensen's inequality)

Let f be convex, $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbf{dom}(f)$, $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$ such that $\sum_{i=1}^m \lambda_i = 1$. Then

$$f\left(\sum_{i=1}^m \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^m \lambda_i f(\mathbf{x}_i).$$

For $m = 2$, this is convexity. The proof of the general case is Exercise 7.

$$f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)$$

Operations that Preserve Convexity

Lemma (Lecture-1).9 (Exercise 5)

(i) *Let* $f_1, f_2, \ldots, f_m$ *be convex functions,* $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$. *Then* $f := \sum_{i=1}^m \lambda_i f_i$ *is convex on* $\mathbf{dom}(f) := \bigcap_{i=1}^m \mathbf{dom}(f_i)$.

(ii) *Let* f *be a convex function with* $\mathbf{dom}(f) \subseteq \mathbb{R}^d$, $g : \mathbb{R}^m \to \mathbb{R}^d$ *an affine function, meaning that* $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, *for some matrix* $A \in \mathbb{R}^{d \times m}$ *and some vector* $\mathbf{b} \in \mathbb{R}^d$. *Then the function* $f \circ g$ *(that maps* $\mathbf{x}$ *to* $f(A\mathbf{x} + \mathbf{b})$*) is convex on* $\mathbf{dom}(f \circ g) := \{\mathbf{x} \in \mathbb{R}^m : g(\mathbf{x}) \in \mathbf{dom}(f)\}$.

Lecture 1 Recap

- General Course Information
- Convex Sets & Convex Functions
  - We have seen the definition and different characterizations of convex functions.
  - We have seen the definition and different characterizations of a minimizer.
  - In the next lecture, we will study the convergence of **Gradient Descent** on convex functions.

Bibliography

Sébastien Bubeck.
Convex Optimization: Algorithms and Complexity.
Foundations and Trends in Machine Learning, 8(3-4):231–357, 2015.

Stephen Boyd and Lieven Vandenberghe.
Convex Optimization.
Cambridge University Press, New York, NY, USA, 2004.
`https://web.stanford.edu/~boyd/cvxbook/`.