Optimization for Data Science
ETH Zürich, FS 2021 261-5110-00L

Lecture 1: Introduction & Convexity

Bernd Gärtner
Niao He
David Steurer

https://www.ti.inf.ethz.ch/ew/courses/ODS21/index.html

February 20, 2021

Course Outline

Part 1: **Optimization** (Bernd Gärtner and Niao He)

- Theory of convex functions
- First-order methods (Gradient Descent and modern algorithms)
- Second-order methods (Newton and Quasi Newton)

Part 2: **Estimation** (David Steurer)

- Regression
- Factorization
- Completion

Focus: provable theoretical guarantees on runtime or estimation quality.

Course Organization

- Lectures (Mon 13-14 and Tue 10-12, **online**); Recordings will be available
- Exercise Classes (Tue 14-16, **online**, starting March 2)
- Non-graded exercises; submission not mandatory but highly recommended, to support the learning experience, and as preparation for special assignments / exam
- Two graded special assignments (March / May; 10% of the grade each)
- Written final exam, 180 minutes, closed book (80% of the grade)
- Material relevant for the exam: **slides, blackboard notes, exercises, special assignments**; lecture notes are available (also contain additional material)

Course page: `https://www.ti.inf.ethz.ch/ew/courses/ODS21/index.html`

All material plus **discussion forums** can be found in the course moodle,
`https://moodle-app2.let.ethz.ch/course/view.php?id=14530`

Optimization

- General optimization problem:

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

  - X: set of **feasible solutions** (if $X = \mathbb{R}^d$: unconstrained optimization)
  - $f : \mathbf{dom}(f) \to \mathbb{R}, X \subseteq \mathbf{dom}(f) \subseteq \mathbb{R}^d$: **objective function**
  - typically assume: f is differentiable (makes efficient methods available)

Why? And How?

Optimization is everywhere

machine learning, big data, statistics, data analysis of all kinds, finance, logistics, planning, control theory, mathematics, search engines, simulations, and many other applications ...

- **Mathematical Modeling**:
  - *defining the optimization problem*
- **Computational Optimization**:
  - *running an (appropriate) optimization algorithm*

Optimization for Machine Learning

- **Mathematical Modeling**:
  - defining the machine learning model
- **Computational Optimization**:
  - training: 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!
  - **Caveat:** Most theoretical results work for convex functions only, but many practical problems are nonconvex (e.g. in deep learning).
  - Still, many relevant problems are convex or can be “convexified”, and there are recent theoretical results on nonconvex functions (some of them covered in this lecture).

Chapter 1

Theory of Convex Functions

Background: The Cauchy-Schwarz inequality

Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. Cauchy-Schwarz inequality (Proof in Section 1.1.2):

$$|\mathbf{u}^\top \mathbf{v}| \leq \|\mathbf{u}\| \, \|\mathbf{v}\| \, .$$

Notation:

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{pmatrix}$$

$$\mathbf{u}^\top = \begin{pmatrix} u_1 & u_2 & \cdots & u_d \end{pmatrix}$$

- $\mathbf{u} = (u_1, \ldots, u_d), \mathbf{v} = (v_1, \ldots, v_d)$, d-dimensional column vectors with real entries
- $\mathbf{u}^\top$, transpose of $\mathbf{u}$, a d-dimensional row vector
- $\mathbf{u}^\top \mathbf{v} = \sum_{i=1}^d u_i v_i$, scalar (or inner) product of $\mathbf{u}$ and $\mathbf{v}$
- $|\mathbf{u}^\top \mathbf{v}|$, absolute value of $\mathbf{u}^\top \mathbf{v}$
- $\|\mathbf{u}\| = \sqrt{\mathbf{u}^\top \mathbf{u}} = \sqrt{\sum_{i=1}^d u_i^2}$, Euclidean norm of $\mathbf{u}$

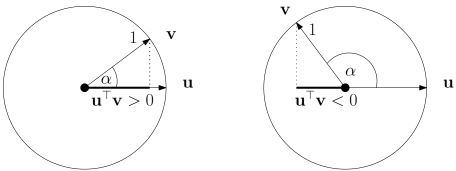
Background: The Cauchy-Schwarz inequality, interpretation

Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. Cauchy-Schwarz inequality: $|\mathbf{u}^\top \mathbf{v}| \leq \|\mathbf{u}\| \, \|\mathbf{v}\|$.

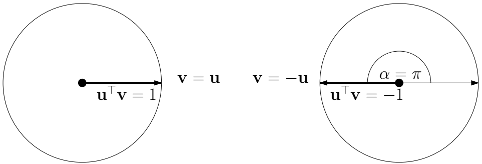
For nonzero vectors, this is equivalent to

$$-1 \leq \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \leq 1.$$

Fraction can be used to define the angle α between $\mathbf{u}$ and $\mathbf{v}$: $\cos(\alpha) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$



Examples for unit vectors ($\|\mathbf{u}\| = \|\mathbf{v}\| = 1$)


$\mathbf{u}^\top \mathbf{v} = 1$
$\mathbf{v} = \mathbf{u}$
$\mathbf{v} = -\mathbf{u}$
$\alpha = \pi$
$\mathbf{u}^\top \mathbf{v} = -1$

Equality in Cauchy-Schwarz if and only if $\mathbf{u} = \mathbf{v}$ or $\mathbf{u} = -\mathbf{v}$.

Background: The spectral norm

Let A be an $(m \times d)$-matrix. Then

$$\|A\| := \max_{\mathbf{v}\in\mathbb{R}^d, \mathbf{v}\neq 0} \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|} = \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\|$$

is the 2-norm (or spectral norm) of A.

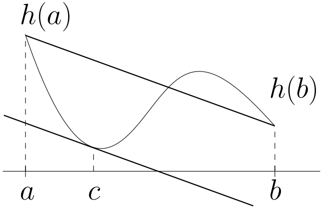
$\|A\|$ is the largest factor by which a vector can be stretched in length under the mapping $\mathbf{v} \to A\mathbf{v}$.

BTW, what is a norm? Read Section 1.1.3 in the notes (or wait a couple of slides)!

Background: The mean value theorem

Let $a < b$ be real numbers, and let $h : [a, b] \to \mathbb{R}$ be a continuous function that is differentiable on (a, b); we denote the derivative by h'. Then there exists $c \in (a, b)$ such that

$$h'(c) = \frac{h(b) - h(a)}{b - a}.$$



Geometric interpretation:

- $(h(b) - h(a))/(b - a)$ is the slope of the line through the two points $(a, h(a))$ and $(b, h(b))$.
- The mean value theorem says that between a and b, we find a tangent to the graph of h that has the same slope.

Background: The fundamental theorem of calculus

Let $a < b$ be real numbers, and let $h : \mathbf{dom}(h) \to \mathbb{R}$ be a differentiable function on an open domain $\mathbf{dom}(h) \supset [a, b]$, and such that h' is continuous on $[a, b]$. Then

$$h(b) - h(a) = \int_a^b h'(t)dt.$$

This theorem is the theoretical underpinning of typical definite integral computations in high school.

For example, to evaluate $\int_2^4 x^2 dx$, we integrate x^2 (giving us $x^3/3$), and then compute

$$\int_2^4 x^2 dx = \frac{4^3}{3} - \frac{2^3}{3} = \frac{56}{3}.$$

BTW, what does differentiable mean? Read Section 1.1.6 in the notes (basics will follow on a later slide).

Background: Convex Sets

A set $C \subseteq \mathbb{R}^d$ 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

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in C.$$

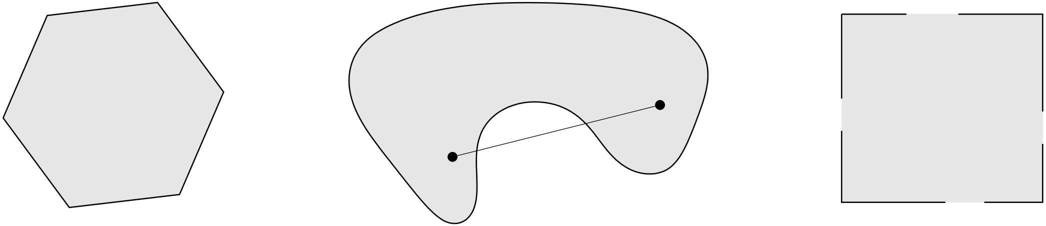


*Figure 2.2 from S. Boyd, L. Vandenberghe

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.

Background: Properties of Convex Sets

Intersections of convex sets are convex.

Observation

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.

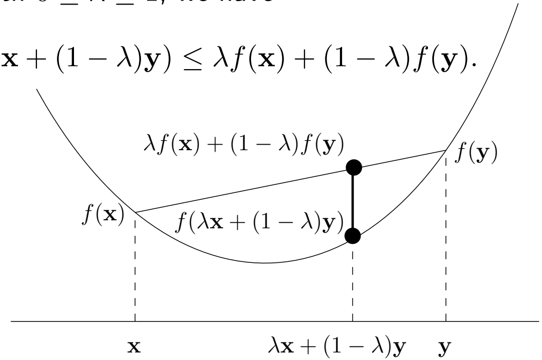
Projections onto convex sets are *unique*, and *often* efficient to compute (we will make use of this later): $P_C(\mathbf{x}) := \mathrm{argmin}_{\mathbf{y} \in C} \|\mathbf{y} - \mathbf{x}\|$

Convex Functions

Definition 1.10

A function $f : \mathbf{dom}(f) \to \mathbb{R}$ 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 $0 \leq \lambda \leq 1$, we have

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



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

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

Motivation: Convex Optimization

Convex Optimization Problems are of the form

$$\min \ f(\mathbf{x}) \qquad \text{such that} \qquad \mathbf{x} \in X \subseteq \mathbb{R}^d,$$

where both

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

Crucial Property of Convex Optimization Problems

- Every local minimum is a **global minimum**, see later...

Motivation: Solving Convex Optimization - Provably

For convex optimization problems, many algorithms compute a sequence $\mathbf{x}_0, \mathbf{x}_1, \ldots$ that does **converge** to a global minimum! (assuming that f is differentiable)

Example Theorem: The **convergence rate** is proportional to $\frac{1}{t}$, i.e.

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{c}{t}$$

(where $\mathbf{x}^\star$ is some optimal solution to the problem.)

Meaning: **Absolute approximation error** converges to 0 over time.

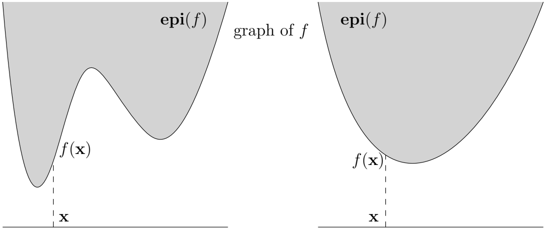
Convex Functions & Sets

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

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



Observation 1.11

f is a convex function if and only if $\mathbf{epi}(f)$ is a convex set.

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 $\|a\mathbf{x}\| = |a| \, \|\mathbf{x}\|$, $a \in \mathbb{R}$:

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

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

Jensen's Inequality

Lemma 1.12

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 2.

Convex Functions are Continuous

Lemma 1.13

Let f be convex and suppose that $\mathbf{dom}(f) \subseteq \mathbb{R}^d$ is open. Then f is continuous.

Not entirely obvious (Exercise 3).

In fact, even linear functions, the simplest convex functions, can in general, meaning over domains of infinite dimension, be discontinuous (Lemma 1.14 gives a classical example).

Differentiable Functions

How to check convexity? Use definition (works, but can be cumbersome).

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

Easier ways exist for differentiable and twice differentiable functions.

Definition 1.5

Let $f : \mathbf{dom}(f) \to \mathbb{R}^m$ where $\mathbf{dom}(f) \subseteq \mathbb{R}^d$ is open. f is called differentiable at $\mathbf{x} \in \mathbf{dom}(f)$ if there exists an $(m \times d)$-matrix A and an error function $r : \mathbb{R}^d \to \mathbb{R}^m$ defined around $\mathbf{0} \in \mathbb{R}^d$ such that for all $\mathbf{y}$ in some neighborhood of $\mathbf{x}$,

$$f(\mathbf{y}) = f(\mathbf{x}) + A(\mathbf{y} - \mathbf{x}) + r(\mathbf{y} - \mathbf{x}),$$

where

$$\lim_{\mathbf{v} \to \mathbf{0}} \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|} = \mathbf{0}.$$

A is unique and called the differential or Jacobian matrix of f at $\mathbf{x}$.

f is locally (around $\mathbf{x}$) well-approximated by a linear function $\ell(\mathbf{y}) = f(\mathbf{x}) + A(\mathbf{y} - \mathbf{x})$.

Differentiable Functions

$$f(\mathbf{y}) = f(\mathbf{x}) + A(\mathbf{y}-\mathbf{x}) + r(\mathbf{y}-\mathbf{x}),$$

$$\lim_{\mathbf{v}\to\mathbf{0}} \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|} = \mathbf{0}.$$

Example: $f(x) = x^2$. We know that derivative $f'(x) = 2x$. Why? For $y = x + v$,

$$\begin{aligned} f(y) = (x+v)^2 &= x^2 + 2vx + v^2 \\ &= f(x) + 2x \cdot v + v^2 \\ &= f(x) + A(y-x) + r(y-x)), \end{aligned}$$

where $A = 2x$, $\quad r(y-x) = r(v) = v^2$. $\qquad \lim_{v\to 0} \frac{|r(v)|}{|v|} = \lim_{v\to 0} |v| = 0.$

Differentiable Functions

$$f(\mathbf{y}) = f(\mathbf{x}) + A(\mathbf{y} - \mathbf{x}) + r(\mathbf{y} - \mathbf{x}),$$

$$\lim_{\mathbf{v} \to \mathbf{0}} \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|} = \mathbf{0}.$$

We denote $Df(\mathbf{x}) := A$. $Df(\mathbf{x})$ is the matrix of partial derivatives at the point $\mathbf{x}$,

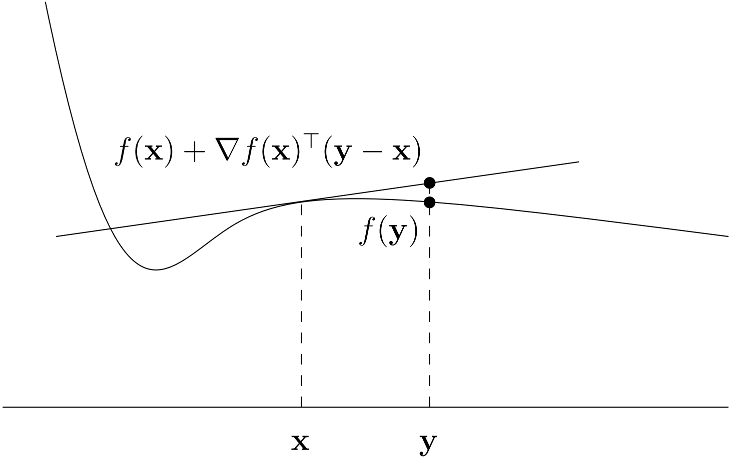
$$Df(\mathbf{x})_{ij} = \frac{\partial f_i}{\partial x_j}(\mathbf{x}).$$

f is called differentiable if f is differentiable at all $\mathbf{x} \in \mathbf{dom}(f)$.

If $f : \mathbf{dom}(f) \to \mathbb{R}$, $Df(\mathbf{x})$ is a row vector typically denoted by $\underbrace{\nabla f(\mathbf{x})}_{\text{gradient}}{}^{\top}$.

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 $(\mathbf{x}, f(\mathbf{x}))$.


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

First-order Characterization of Convexity

Lemma 1.15

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

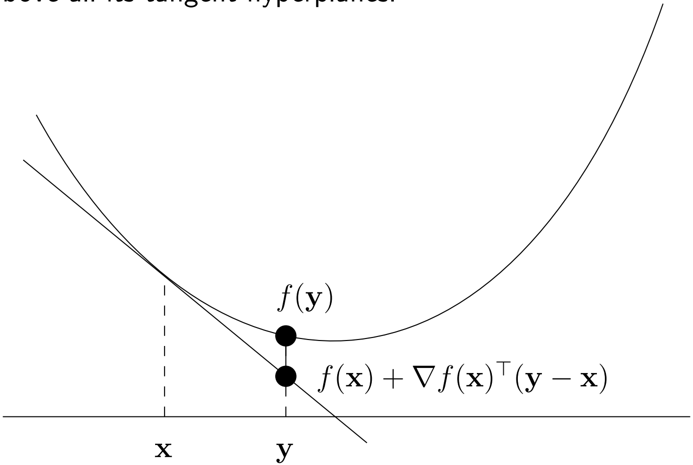
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \qquad (1)$$

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

First-order Characterization of Convexity

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

Graph of f is above all its tangent hyperplanes.


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

First-order Characterization of Convexity: Proof

f convex iff $\mathbf{dom}(f)$ convex and

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

$\Rightarrow$: suppose f is convex.

Then, for all $t \in (0,1)$,

$$f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) = f((1-t)\mathbf{x} + t\mathbf{y}) \leq (1-t)f(\mathbf{x}) + tf(\mathbf{y}) = f(\mathbf{x}) + t(f(\mathbf{y}) - f(\mathbf{x})).$$

Subtracting $f(\mathbf{x})$ on both sides, dividing by t, and using differentiability at $\mathbf{x}$:

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \frac{f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{t} \\ &= f(\mathbf{x}) + \frac{\nabla f(\mathbf{x})^\top t(\mathbf{y} - \mathbf{x}) + r(t(\mathbf{y} - \mathbf{x}))}{t} \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \underbrace{\frac{r(t(\mathbf{y} - \mathbf{x}))}{t}}_{\to 0 \text{ for } t \to 0} \end{aligned}$$

First-order Characterization of Convexity: Proof

f convex iff $\mathbf{dom}(f)$ convex and

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

$\Leftarrow$: suppose the inequality holds.

$\mathbf{z} := \lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in \mathbf{dom}(f)$ for $\lambda \in [0,1]$ (by convexity of $\mathbf{dom}(f)$)

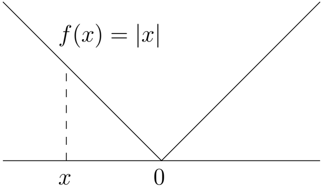
$$\begin{array}{rcll} f(\mathbf{x}) & \geq & f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z}), & \cdot \lambda \\ f(\mathbf{y}) & \geq & f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z}) & \cdot (1-\lambda) \end{array}$$

Adding up, gradient terms cancel:

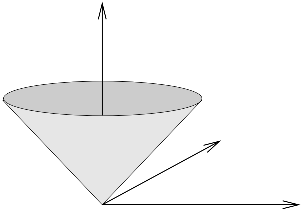
$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \geq f(\mathbf{z}) = f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}).$$

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:



Second-order Characterization of Convexity

Lemma 1.17

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

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

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

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

Operations that Preserve Convexity

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 $\mathbf{dom}(f \circ g) := \{\mathbf{x} \in \mathbb{R}^m : g(\mathbf{x}) \in \mathbf{dom}(f)\}$.

Local Minima are Global Minima

Definition 1.19

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

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

Lemma 1.20

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 1.21

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 the first-order characterization of convexity, we have

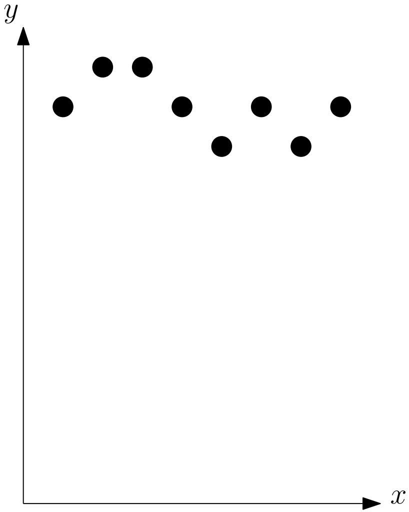
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) = f(\mathbf{x})$$

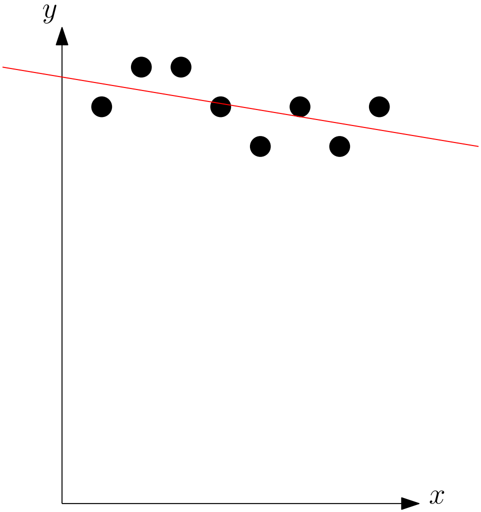
for all $\mathbf{y} \in \mathbf{dom}(f)$, so $\mathbf{x}$ is a global minimum. □

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

Example: Least Squares

Problem: Fit a line to a set of points





Example: Least Squares

Points $(x_i, y_i), i = 1, \ldots, 8$:

$$(1, 10), (2, 11), (3, 11), (4, 10), (5, 9), (6, 10), (7, 9), (8, 10),$$

Line: $y = w_0 + w_1 x$

Fitting error (sum of squared vertical distances of points to the line):

$$\begin{aligned} f(w_0, w_1) &= \sum_{i=1}^{8} (w_1 x_i + w_0 - y_i)^2 \\ &= 204w_1^2 + 72w_1w_0 - 706w_1 + 8w_0^2 - 160w_0 + 804 \end{aligned}$$

Function is convex:

$$\nabla^2(w_0, w_1) = \begin{pmatrix} 16 & 72 \\ 72 & 408 \end{pmatrix} \succeq 0.$$

Another proof of convexity: f is the sum of (more easily seen to be) convex functions $(w_1 x_i + w_0 - y_i)^2, i = 1, \ldots, 8.$

Example: Least Squares

Global minimum (solve for critical point):

$$\nabla f(w_0, w_1) = (72w_1 + 16w_0 - 160, 408w_1 + 72w_0 - 706) = (0, 0).$$

System of linear equations.

$$(w_0^\star, w_1^\star) = \left(\frac{43}{4}, -\frac{1}{6}\right).$$

Hence, the optimal line is

$$y = -\frac{1}{6}x + \frac{43}{4}.$$

Fact

Convex quadratic functions can be minimized by solving a system of linear equations, no need to run any optimization algorithm.

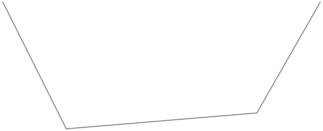
Solving for a critical point is always a system of equations, but these are typically nonlinear and therefore hard to solve analytically → optimization!

Strictly Convex Functions

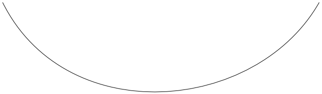
Definition 1.23

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

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



convex, but not strictly convex



strictly convex

Lemma 1.25

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be strictly convex. Then f has at most one global minimum.

Constrained Minimization

Definition 1.26

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex and let $X \subseteq \mathbf{dom}(f)$ be a convex set. A point $\mathbf{x} \in X$ is a minimizer of f over X if

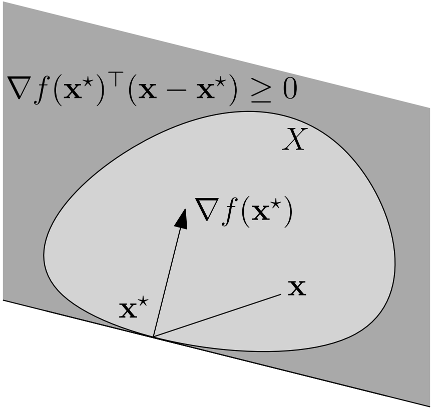
$$f(\mathbf{x}) \leq f(\mathbf{y}) \quad \forall \mathbf{y} \in X.$$

Lemma ([BV04, 4.2.3])

Suppose that $f : \mathbf{dom}(f) \to \mathbb{R}$ is convex and differentiable over an open domain $\mathbf{dom}(f) \subseteq \mathbb{R}^d$, and let $X \subseteq \mathbf{dom}(f)$ be a convex set. Point $\mathbf{x}^\star \in X$ is a minimizer of f over X if and only if

$$\nabla f(\mathbf{x}^\star)^\top (\mathbf{x} - \mathbf{x}^\star) \geq 0 \quad \forall \mathbf{x} \in X.$$

Constrained Minimization


$\nabla f(\mathbf{x}^\star)^\top(\mathbf{x}-\mathbf{x}^\star) \geq 0$
X
$\nabla f(\mathbf{x}^\star)$
$\mathbf{x}$
$\mathbf{x}^\star$

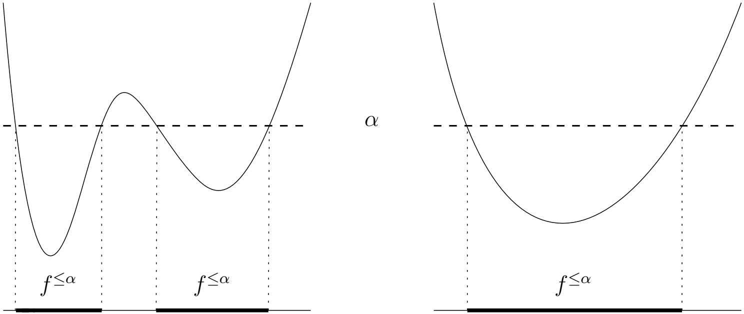
Existence of a minimizer

How do we know that a global minimum exists?

Not necessarily the case, even if f bounded from below ($f(x) = e^x$)

Definition 1.28

$f : \mathbb{R}^d \to \mathbb{R}$, $\alpha \in \mathbb{R}$. The set $f^{\leq \alpha} := \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \leq \alpha\}$ is the α-sublevel set of f.



The Weierstrass Theorem

Theorem 1.29

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, and suppose there is a nonempty and bounded sublevel set $f^{\leq \alpha}$. Then f has a global minimum.

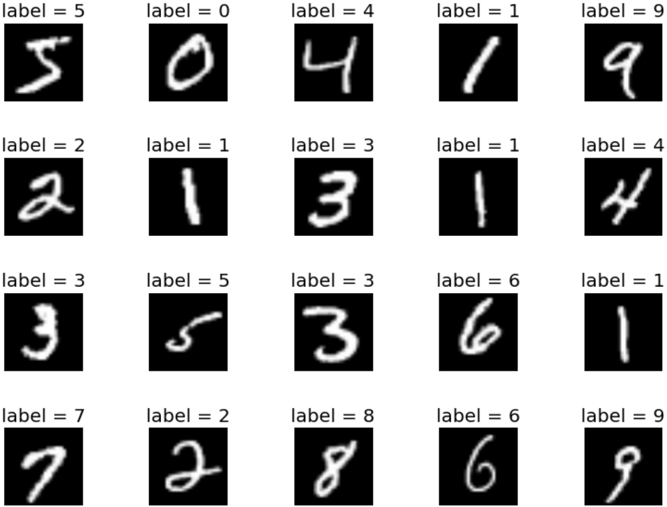
Proof:

We know that f—as a continuous function—attains a minimum over the closed and bounded ($=$ compact) set $f^{\leq \alpha}$ at some $\mathbf{x}^\star$. This $\mathbf{x}^\star$ is also a global minimum as it has value $f(\mathbf{x}^\star) \leq \alpha$, while any $\mathbf{x} \notin f^{\leq \alpha}$ has value $f(\mathbf{x}) > \alpha \geq f(\mathbf{x}^\star)$.

Generalizes to suitable domains $\mathbf{dom}(f) \neq \mathbb{R}^d$.

Example: Handwritten Digit Recognition (MNIST database)

Task: recognize handwritten decimal digits $0, 1, \ldots, 9$



Example: Handwritten Digit Recognition

Training data:

- set P of grayscale images (28×28 pixels)
- for each $\mathbf{x} \in P$, the correct digit $d(\mathbf{x}) \in \{0, \ldots, 9\}$

Approach:

- represent image as feature vector $\mathbf{x} \in \mathbb{R}^{28 \cdot 28} = \mathbb{R}^{784}$, where x_i is the gray value of the i-th pixel
- fit a matrix $W \in \mathbb{R}^{10 \times 784}$ to the training data
- use vector $\mathbf{y} = W\mathbf{x} \in \mathbb{R}^{10}$ to predict the digit seen in an arbitrary image $\mathbf{x}$
- idea: $y_j, j = 0, \ldots, 9$ should tell us the probability of the digit being j
- For example, use probabilities $z_j = z_j(\mathbf{y}) = \frac{e^{y_j}}{\sum_{k=0}^{9} e^{y_k}}$

Example: Handwritten Digit Recognition

Matrix W should minimize the recognition error on the training data.

Measure recognition error by a loss function (there are many choices).

For example,

$$\ell(W) = -\sum_{\mathbf{x}\in P} \ln\left(z_{d(\mathbf{x})}(W\mathbf{x})\right) = \sum_{\mathbf{x}\in P}\left(\ln\left(\sum_{k=0}^{9} e^{(W\mathbf{x})_k}\right) - (W\mathbf{x})_{d(\mathbf{x})}\right).$$

- $z_{d(\mathbf{x})}(W\mathbf{x}) \in (0,1)$: probability of predicting the correct digit $d(\mathbf{x})$ on training image $\mathbf{x}$
- $-\ln\left(z_{d(\mathbf{x})}(W\mathbf{x})\right) > 0$
- tends to ∞ for probability tending to 0 (punishes small probability)
- tends to 0 for probability tending to 1 (rewards large probability)

Example: Handwritten Digit Recognition

Exercise 7

The function $\ell : \mathbb{R}^{10\cdot 784} \to \mathbb{R}$ given by

$$\ell(W) = -\sum_{\mathbf{x}\in P} \ln\left(z_{d(\mathbf{x})}(W\mathbf{x})\right) = \sum_{\mathbf{x}\in P}\left(\ln\left(\sum_{k=0}^{9} e^{(W\mathbf{x})_k}\right) - (W\mathbf{x})_{d(\mathbf{x})}\right)$$

is convex.

The function ℓ does not necessarily have a global minimum, but one can characterize the training sets for which it does. This needs material on weakly coercive functions, see notes (Exercise 8).

Bibliography

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