

Introduction to Optimization

Peter Richtárik
<https://richtarik.org>



November 2022

Introduction to Optimization

Peter Richtárik



Course Organization

Lecture Outline

- ▶ About Me
- ▶ Teaching Assistants
- ▶ Schedule
- ▶ Brief Course Description
- ▶ Goals and Objectives
- ▶ Knowledge Required
- ▶ Slides
- ▶ Method of Evaluation

About Me I

- ▶ **Name:** Peter Richtárik
- ▶ **Position:** Professor of
 - ▶ Computer Science (CS),
 - ▶ Applied Mathematics and Computational Sciences (AMCS), and
 - ▶ Statistics (STAT)
- at King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia.
- ▶ **Website:** <https://richtarik.org>
- ▶ **Email:** peter.richtarik@kaust.edu.sa
- ▶ **Brief Academic CV:**
 - ▶ 2019–now, Professor, KAUST
 - ▶ 2017–2019, Associate Professor, KAUST
 - ▶ 2009–2017, Assistant and later Associate Professor, University of Edinburgh, Scotland, United Kingdom
 - ▶ 2007–2009, Postdoc, UC Louvain, Belgium
 - ▶ 2002–2007, PhD student, Cornell, USA
 - ▶ 1996–2001, Bc and Mgr student, Comenius University, Slovakia
- ▶ **Nationality:** Slovakia, European Union

About Me II

- ▶ **Research:**
 - ▶ optimization for machine learning
 - ▶ federated learning
 - ▶ convex and nonconvex optimization
 - ▶ stochastic/randomized zero, first and second order methods
 - ▶ distributed centralized and decentralized optimization
 - ▶ deep learning
 - ▶ randomized linear algebra
 - ▶ sampling algorithms
 - ▶ saddle-point problems
 - ▶ operator splitting algorithms
- ▶ **Publishing in:** ICML, NeurIPS, AISTATS, ICLR, JMLR, SIAM Journals, IEEE Journals, Optimization Methods and Software, Mathematical Programming . . .

Teaching Assistants I

Grigory Malinovsky (PhD student at KAUST)



Figure: Grigory Malinovsky

- ▶ Web: <https://grigory-malinovsky.github.io>
- ▶ Email: grigorii.malinovskii@kaust.edu.sa

Teaching Assistants II

Igor Sokolov (PhD student at KAUST)



Figure: Igor Sokolov

- ▶ Web:
<https://cemse.kaust.edu.sa/people/person/igor-sokolov>
- ▶ Email: igor.sokolov.1@kaust.edu.sa

Schedule: The Same for Both Weeks

Identical schedule for Week 1 (Nov 3–6) and Week 2 (Nov 17–20):

Time	Thursday	Friday	Saturday	Sunday
09:00–09:30	Lecture	Lecture	Lecture	Lecture
09:30–09:50	Lecture	Lab	Lab	Lab
09:50–10:00	Break	Break	Break	Break
10:00–10:30	Lecture	Quiz	Lecture	Lecture
10:30–10:50	Lab	Quiz	Lab	Lab
10:50–11:00	Break	Quiz	Break	Break
11:00–11:30	Lecture	Prayer	Quiz	Lecture
11:30–11:50	Lab	Prayer	Quiz	Lab
11:50–12:00	Break	Prayer	Quiz	Break
12:00–13:00	Lunch	Lunch	Lunch	Lunch
13:00–13:30	Lecture	Lecture	Lecture	Exam
13:30–13:50	Lab	Lab	Lab	Exam
13:50–14:00	Break	Break	Break	Exam
14:00–14:30	Lecture	Lecture	Lecture	Exam
14:30–14:50	Lab	Lab	Lab	Exam
14:50–15:00	Break	Break	Break	Exam
15:00–15:30	Lecture	Lecture	Lecture	Exam
15:30–15:50	Lab	Lab	Lab	Exam
15:50–16:00	Break	Break	Break	Exam

Brief Course Description

	Week 1 (Nov 3–6)	Week 2 (Nov 17–20)
Theory	12	0
Applications	6	6
Algorithms	0	12
Labs	17	18
Quizzes	2	2
Exams	1	1

37 Lectures, each lasting 30 minutes:

- ▶ **Organization** – 1 lecture
- ▶ **Theory** – about 12 lectures
- ▶ **Applications** – about 12 lectures
- ▶ **Algorithms** – about 12 lectures

35 Labs, each lasting 20 minutes:

- ▶ **Coding** – Python notebooks will be provided
- ▶ **Exercises** – included in the slides

Goals and Objectives

- ▶ **Theory:** Rigorous foundations of convex optimization (e.g., convex sets, convex functions, Fenchel conjugation, Fenchel duality)
- ▶ **Applications:** Learn about selected applications of optimization (e.g., arbitrage detection via linear programming, portfolio selection via quadratic programming)
- ▶ **Algorithms:** Understand and code up selected key optimization algorithms (e.g., gradient descent, stochastic gradient descent, momentum)

Knowledge Required

- ▶ **Programming:** experience with Python
- ▶ **Linear algebra:** vector spaces, linear independence, basis, linear mappings, quadratic forms, Euclidean spaces, inner product, norm,
...
- ▶ **Matrix theory:** matrices, eigenvalues, ...
- ▶ **Multivariate calculus:** limits, derivatives, gradient, Hessian, chain rule
- ▶ **Probability theory:** random variables, expectation, ...

Slides

- ▶ Slides include all relevant material (e.g., some limited/terse explanation, theorems, algorithms, proofs).
- ▶ However, the course slides are *not* a book.
 - ▶ Listening to the lecture is necessary for the slides to make sense
 - ▶ Asking questions is important and highly desirable for deeper understanding!
- ▶ Slides will be made available in pdf form before each lecture.
- ▶ Recommendation: upload the pdf version of the slides to an iPad, and annotate with an Apple pencil during the lecture.

Method of Evaluation

- ▶ **40% Quizzes**
 - ▶ **4 in-class Quizzes**
 - ▶ Each lasts 60 minutes
 - ▶ **Each worth 10% towards the final grade**
 - ▶ Submission should be done via Blackboard
 - ▶ Submit by the end of the Quiz session
 - ▶ A mix of theoretical and coding problems
- ▶ **60% Exams**
 - ▶ **2 in-class exams**
 - ▶ Each lasts 180 minutes
 - ▶ **Each worth 30% towards the final grade**
 - ▶ Mix of theoretical & coding problems
- ▶ **Extra 2% for each exercise marked with ☕☕**
 - ▶ these are more demanding (e.g., they involve mathematical proofs), and you are not expected to be able to solve them.
- ▶ **Extra 5% for active participation**
 - ▶ Asking questions (lectures are designed in an interactive fashion – there is plenty of room for discussion)

Part I

Theory

Introduction to Optimization

Peter Richtárik



Lecture 1: The Space \mathbb{R}^d

Lecture Outline

- ▶ Linear algebra on \mathbb{R}^d : vectors, matrices, norms & more
- ▶ Topology on \mathbb{R}^d : open, closed and bounded sets & more
- ▶ Analysis on \mathbb{R}^d : functions, gradient, Hessian & more

Linear Algebra on \mathbb{R}^d : Vectors, Matrices, Norms & More

Linear Algebra on \mathbb{R}^d I

- ▶ **Vectors.** \mathbb{R}^d is the **vector space** of d -dimensional (column) vectors

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d$$

composed of real values $x_1, \dots, x_d \in \mathbb{R}$.

- ▶ In order to save space, we will often instead write $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, always keeping in mind that what we mean is a column vector.
- ▶ **Addition of two vectors.** Two vectors $x, y \in \mathbb{R}^d$ can be added, resulting in another vector:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_d + y_d \end{pmatrix} \in \mathbb{R}^d.$$

Linear Algebra on \mathbb{R}^d II

- ▶ **Multiplication of a vector and a scalar.** A vector $x \in \mathbb{R}^d$ can be multiplied by a real scalar $t \in \mathbb{R}$, the result of which is a vector:

$$tx = t \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} tx_1 \\ tx_2 \\ \vdots \\ tx_d \end{pmatrix} \in \mathbb{R}^d.$$

- ▶ **Matrices.** A matrix $A \in \mathbb{R}^{n \times d}$ is a collection of nd real numbers $\{A_{ij} \in \mathbb{R} : i = 1, 2, \dots, n; j = 1, 2, \dots, d\}$ arranged in a table composed of n rows and d columns:

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1d} \\ A_{21} & A_{22} & \dots & A_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nd} \end{pmatrix} \in \mathbb{R}^{n \times d}.$$

The j^{th} column of matrix A will be denoted by $A_{:j} \in \mathbb{R}^n$.

Linear Algebra on \mathbb{R}^d III

- ▶ **Vectors as matrices.** Each vector $x \in \mathbb{R}^d$ can be seen as a matrix with d rows and 1 column, i.e., $x \in \mathbb{R}^{d \times 1}$.
- ▶ **Matrix transpose.** The transpose of a matrix $A \in \mathbb{R}^{n \times d}$ is the matrix $A^\top \in \mathbb{R}^{d \times n}$ given by

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1d} \\ A_{21} & A_{22} & \dots & A_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nd} \end{pmatrix}^\top = \begin{pmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1d} & A_{2d} & \dots & A_{nd} \end{pmatrix} \in \mathbb{R}^{d \times n}.$$

That is, the rows are turned into the columns.

- ▶ **Product of two matrices.** The product of matrices $A = (A_{ij}) \in \mathbb{R}^{n \times d}$ and $B = (B_{jk}) \in \mathbb{R}^{d \times m}$ is the matrix $AB \in \mathbb{R}^{n \times m}$ defined as follows:

$$(AB)_{ik} \stackrel{\text{def}}{=} \sum_{j=1}^d A_{ij} B_{jk}, \quad j = 1, \dots, n; \quad k = 1, \dots, m.$$

Linear Algebra on \mathbb{R}^d IV

- **Matrix-vector product as a linear map from \mathbb{R}^d to \mathbb{R}^n .** The result of the product between a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and a vector $x \in \mathbb{R}^d$ is a vector belonging to \mathbb{R}^n , which arises as a linear combination of the columns of \mathbf{A} with coefficients $\{x_j\}$:

$$\mathbf{A}x = \sum_{j=1}^d x_j \mathbf{A}_{:j} \in \mathbb{R}^n.$$

So, $\mathbf{A} \in \mathbb{R}^{n \times d}$ represents a mapping from \mathbb{R}^d to \mathbb{R}^n . This mapping is linear:

- $\mathbf{A}(tx) = t(\mathbf{A}x)$ for all $t \in \mathbb{R}$ and $x \in \mathbb{R}^d$
- $\mathbf{A}(x + y) = \mathbf{A}x + \mathbf{A}y$ for all $x, y \in \mathbb{R}^d$

Linear Algebra on \mathbb{R}^d V

- ▶ **Standard Euclidean inner product of two vectors:** Given vectors $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ and $y = (y_1, \dots, y_d) \in \mathbb{R}^d$, their inner product is defined by

$$\langle x, y \rangle \stackrel{\text{def}}{=} \sum_{j=1}^d x_j y_j = x^\top y \in \mathbb{R}, \quad (1)$$

where $x^\top y$ correspond to the matrix product of two matrices: $x^\top \in \mathbb{R}^{1 \times d}$ and $y \in \mathbb{R}^{d \times 1}$.

- ▶ **Orthogonality of vectors.** Vectors $x, y \in \mathbb{R}^d$ are **orthogonal** if

$$\langle x, y \rangle = 0.$$

- ▶ **Transpose matrix and inner product.** For every two vectors $x \in \mathbb{R}^d$, $y \in \mathbb{R}^n$ and matrix $A \in \mathbb{R}^{n \times d}$, we have the identity:

$$\langle Ax, y \rangle = \langle x, A^\top y \rangle. \quad (2)$$

Notice that the first inner product is between vectors in \mathbb{R}^n , and the second inner product is between vectors in \mathbb{R}^d .

Linear Algebra on \mathbb{R}^d VI

- ▶ **Standard Euclidean norm of a vector.** The standard Euclidean norm of a vector $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ is defined as

$$\|x\| \stackrel{\text{def}}{=} \sqrt{\langle x, x \rangle} \stackrel{(1)}{=} \sqrt{\sum_{i=1}^d x_i^2}. \quad (3)$$

- ▶ **Positive homogeneity of the norm.**

$$\|tx\| = |t| \|x\|, \quad \forall x \in \mathbb{R}^d, \quad \forall t \in \mathbb{R}. \quad (4)$$

- ▶ **Triangle inequality.**

$$\|x + y\| \leq \|x\| + \|y\|, \quad \forall x, y \in \mathbb{R}^d. \quad (5)$$

- ▶ **Square expansion identity.**

$$\|x + y\|^2 = \|x\|^2 + 2\langle x, y \rangle + \|y\|^2, \quad \forall x, y \in \mathbb{R}^d. \quad (6)$$

Linear Algebra on \mathbb{R}^d VII

- ▶ **Young's inequality:**

$$\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2, \quad \forall x, y \in \mathbb{R}^d. \quad (7)$$

- ▶ **Cauchy-Schwarz inequality:**

$$|\langle x, y \rangle| \leq \|x\| \|y\|, \quad \forall x, y \in \mathbb{R}^d. \quad (8)$$

- ▶ **Ball in \mathbb{R}^d .** A (closed) ball centered at $x \in \mathbb{R}^d$ with radius $r > 0$ is the set of all points in \mathbb{R}^d whose distance from x is at most r :

$$\mathcal{B}(x, r) \stackrel{\text{def}}{=} \{y \in \mathbb{R}^d : \|y - x\| \leq r\}.$$

- ▶ **Nonnegative orthant.** Nonnegative orthant is the subset of vectors from \mathbb{R}^d whose all entries are nonnegative:

$$\mathbb{R}_+^d \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : x_1 \geq 0, \dots, x_d \geq 0\}.$$

Notice that if $x \in \mathbb{R}_+^d$, then $tx \in \mathbb{R}_+^d$ for all $t \geq 0$. This means that \mathbb{R}_+^d is a **cone**.

Linear Algebra on \mathbb{R}^d VIII

- ▶ **Positive orthant.** Positive orthant is the subset of vectors from \mathbb{R}^d whose all entries are positive:

$$\mathbb{R}_{++}^d \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : x_1 > 0, \dots, x_d > 0\}.$$

Notice that if $x \in \mathbb{R}_{++}^d$, then $tx \in \mathbb{R}_{++}^d$ for all $t > 0$.

- ▶ **Positive semidefinite and positive definite matrices.** A square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is positive semidefinite if

$$\langle \mathbf{A}x, x \rangle \geq 0, \quad \forall x \in \mathbb{R}^d.$$

It is positive definite if

$$\langle \mathbf{A}x, x \rangle > 0, \quad \forall x \neq 0.$$

The set of symmetric positive semidefinite (resp. symmetric positive definite) matrices is denoted \mathbb{S}_+^d (resp. \mathbb{S}_{++}^d).

Linear Algebra on \mathbb{R}^d IX

- ▶ **Eigenvectors and eigenvalues.** Given a square symmetric matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, we say that $0 \neq v \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$ are an eigenvalue-eigenvector pair if

$$\mathbf{A}v = \lambda v.$$

If $\mathbf{A} \in \mathbb{S}_+^d$ (resp. $\mathbf{A} \in \mathbb{S}_{++}^d$) all eigenvalues of \mathbf{A} are nonnegative (resp. positive). The largest eigenvalue of \mathbf{A} , denoted $\lambda_{\max}(\mathbf{A})$, satisfies

$$\lambda_{\max}(\mathbf{A}) = \max_{x \neq 0} \frac{x^\top \mathbf{A} x}{x^\top x}.$$

Exercise 1

Find the largest eigenvalue of the matrix aa^\top , where $0 \neq a \in \mathbb{R}^d$.

Exercise 2 (☕☕)

Prove Cauchy-Schwarz inequality.

Topology on \mathbb{R}^d : Open, Closed and Bounded Sets & More

Topology on \mathbb{R}^d |

- ▶ **Complement of a set.** The **complement** of a set $\mathcal{S} \subseteq \mathbb{R}^d$ is the set of all points of \mathbb{R}^d not belonging to \mathcal{S} :

$$\mathcal{S}^c \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : x \notin \mathcal{S}\}.$$

- ▶ **Open sets.** A set $\mathcal{S} \subseteq \mathbb{R}^d$ is called **open** if for all $x \in \mathcal{S}$ there exists $\epsilon > 0$ such that

$$\mathcal{B}(x, \epsilon) \subseteq \mathcal{S}.$$

- ▶ **Interior of a set.** The **interior** of a set $\mathcal{S} \subseteq \mathbb{R}^d$ is the set of all points $x \in \mathcal{S}$ whose neighborhood belongs to \mathcal{S} , i.e.,

$$\text{int}(\mathcal{S}) \stackrel{\text{def}}{=} \{x \in \mathcal{S} : \exists r > 0 \ \mathcal{B}(x, r) \subseteq \mathcal{S}\}.$$

- ▶ **Closed sets.** A set $\mathcal{S} \subseteq \mathbb{R}^d$ is called **closed** if its complement \mathcal{S}^c is open. A closed set \mathcal{S} has the following key property: if $\{x^n\}_{n=1}^\infty$ is a sequence of points $x^n \in \mathcal{S}$ converging to some point $x \in \mathbb{R}^d$, then $x \in \mathcal{S}$.

Topology on \mathbb{R}^d II

- ▶ **Closure of a set.** The **closure** of a set $\mathcal{S} \subseteq \mathbb{R}^d$ is the smallest closed set containing \mathcal{S} . It can be shown that

$$\text{cl}(\mathcal{S}) \stackrel{\text{def}}{=} (\text{int}(\mathcal{S}^c))^c.$$

- ▶ **Bounded sets.** A set $\mathcal{S} \subseteq \mathbb{R}^d$ is called **bounded** if there exists $r > 0$ such that

$$\mathcal{S} \subseteq \mathcal{B}(0, r).$$

- ▶ **Compact sets.** A set $\mathcal{S} \subseteq \mathbb{R}^d$ is called **compact** if it is closed and bounded.

Analysis on \mathbb{R}^d : Functions, Gradient, Hessian & More

Analysis on \mathbb{R}^d I

- ▶ **Gradient of functions on \mathbb{R}^d .** Given a differentiable function

$$f : \mathcal{U} \rightarrow \mathbb{R},$$

where $\mathcal{U} \subseteq \mathbb{R}^d$ is an open set, its **gradient** at point $x \in \mathbb{R}^d$ is the vector of partial derivatives:

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{pmatrix} \in \mathbb{R}^d.$$

- ▶ **First-order Taylor approximation:** Let $\mathcal{U} \subseteq \mathbb{R}^d$ be an open subset of \mathbb{R}^d . Fix any $x \in \mathcal{U}$ and $h \in \mathbb{R}^d$ such that $x + th \in \mathcal{U}$ for all $t \in [0, 1]$. If $f : \mathcal{U} \rightarrow \mathbb{R}$ is **differentiable**, then¹

$$f(x + h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|).$$

Analysis on \mathbb{R}^d II

- **Hessian of functions on \mathbb{R}^d .** Given a twice differentiable function

$$f : \mathcal{U} \rightarrow \mathbb{R},$$

where $\mathcal{U} \subseteq \mathbb{R}^d$ is an open set, its **Hessian** at point $x \in \mathbb{R}^d$ is the matrix of all second partial derivatives:

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_d \partial x_1} & \frac{\partial^2 f(x)}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_d \partial x_d} \end{pmatrix} \in \mathbb{R}^{d \times d}.$$

- **Second-order Taylor approximation:** Let $\mathcal{U} \subseteq \mathbb{R}^d$ be an open subset of \mathbb{R}^d . Fix any $x \in \mathcal{U}$ and $h \in \mathbb{R}^d$ such that $x + th \in \mathcal{U}$ for all $t \in [0, 1]$. If $f : \mathcal{U} \rightarrow \mathbb{R}$ is **twice differentiable**, then

$$f(x + h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \langle \nabla^2 f(x)h, h \rangle + o(\|h\|^2).$$

Analysis on \mathbb{R}^d III

Example 3 (Gradient and Hessian of the quadratic form)

For $\mathbf{A} \in \mathbb{R}^{d \times d}$, let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be the quadratic form defined via

$$f(x) = \frac{1}{2}\langle \mathbf{A}x, x \rangle = \frac{1}{2}x^\top \mathbf{A}x.$$

Then the gradient and Hessian of f are given by

$$\nabla f(x) = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top)x, \quad \nabla^2 f(x) = \frac{1}{2}(\mathbf{A} + \mathbf{A}^\top).$$

In particular, if \mathbf{A} is symmetric (i.e., if $\mathbf{A} = \mathbf{A}^\top$), then

$$\nabla f(x) = \mathbf{A}x, \quad \nabla^2 f(x) = \mathbf{A}.$$

Moreover,

$$f(x + h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\langle \nabla^2 f(x)h, h \rangle, \quad \forall x, h \in \mathbb{R}^d.$$

Analysis on \mathbb{R}^d IV

Example 4 (Gradient and Hessian of linear functions)

For $b \in \mathbb{R}^d$, let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be the linear function defined via

$$f(x) = \langle b, x \rangle = b^\top x.$$

Show that

$$\nabla f(x) = b, \quad \nabla^2 f(x) = 0.$$

Exercise 5

Find the gradient and Hessian of the least-squares function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ defined via

$$f(x) = \frac{1}{2} \|Ax - b\|^2,$$

where $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$.

Analysis on \mathbb{R}^d V

Exercise 6

Is the set $\{x \in \mathbb{R}^d : \|x\| \leq 1\}$ open?

Example 7

$t^2 = o(t)$ since $\lim_{t \rightarrow 0} \frac{t^2}{t} = t = 0$.

¹Landau's little "oh" notation: By $o(t)$ we mean any function $o : \mathbb{R} \rightarrow \mathbb{R}$ such that $\lim_{t \rightarrow 0} \frac{o(t)}{t} = 0$. Informally, these are functions which approach zero faster than "linearly"; i.e., faster than the $t \mapsto t$ function does.

Introduction to Optimization

Peter Richtárik



Lecture 2: Introduction

Lecture Outline

- ▶ Optimization Problems and Solution Concepts
- ▶ Optimization Modelling and Data
- ▶ Optimization: Basic Terminology
- ▶ Optimization Algorithms: Basic Terminology
- ▶ Basic Algorithms
- ▶ Convergence

Optimization Problems and Solution Concepts

Optimization Problems

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && x = (x_1, \dots, x_d) \in \mathcal{S} \subseteq \mathbb{R}^d \end{aligned} \tag{9}$$

- ▶ a set \mathcal{S} (**feasible set / constraint**)
- ▶ a real-valued function $f : \mathcal{S} \rightarrow \mathbb{R}$ (**objective function / cost function / loss function**)

Condition	Problem (9) is called
$\mathcal{S} = \mathbb{R}^d$	unconstrained
$\mathcal{S} \neq \mathbb{R}^d$	constrained
$\mathcal{S} \neq \emptyset$	feasible
$\mathcal{S} = \emptyset$	infeasible
$\exists C \in \mathbb{R}$ such that $f(x) \geq C$ for all $x \in \mathcal{S}$	bounded
$\exists \{x^k\}_{k=0}^{\infty} \subseteq \mathcal{S}$ such that $f(x^k) \rightarrow -\infty$	unbounded

Basic Types of Optimization Problems

The constraint set \mathcal{S} is often described as a solution of a (finite) system of equations and inequalities:

$$\mathcal{S} = \{x \in \mathbb{R}^d : g_i(x) = 0, i \in \mathcal{E}; g_i(x) \leq 0, i \in \mathcal{I}\},$$

where

- ▶ \mathcal{E} is some finite index set, indexing **Equalities**
- ▶ \mathcal{I} is some finite index set, indexing **Inequalities**

Condition	Problem (9) is called
f, g_i are linear	Linear Program (LP)
f not linear	Nonlinear Program (NLP)
f quadratic, g_i linear	Quadratic Program (QP)
f, g_i quadratic	Quadratically Constrained QP (QCQP)
f, \mathcal{S} convex	Convex Program (CP)
f not convex	Nonconvex Program
LP with $x_i \in \mathbb{Z}$ for some i	Integer Program (IP)

We shall deal with **convex problems** and **nonconvex problems** in this course, either without any constraints, or with a convex constraint set \mathcal{S} .



Solution Concepts

Selection of some key solution concepts:

Condition	x is called
any x $x \in \mathcal{S}$ $x \notin \mathcal{S}$	a solution/point feasible solution/point infeasible solution/point
$\exists \delta > 0 : f(x) \leq f(y)$ for all $y \in \mathcal{S} \cap \mathcal{B}(x, \delta)$ $f(x) \leq f(y)$ for all $y \in \mathcal{S}$	local minimizer global minimizer (x^*)
$\exists \delta > 0 : f(x) \leq f(y) + \epsilon$ for all $y \in \mathcal{S} \cap \mathcal{B}(x, \delta)$ $f(x) \leq f(y) + \epsilon$ for all $y \in \mathcal{S}$ $\ x - x^*\ \leq \epsilon$	ϵ -solution (local) ϵ -solution (global) ϵ -solution (global)
$\nabla f(x) = 0$ and $\mathcal{S} = \mathbb{R}^d$ $\ \nabla f(x)\ \leq \epsilon$ and $\mathcal{S} = \mathbb{R}^d$	stationary point ϵ -stationary point

The **δ -neighborhood** of $x \in \mathbb{R}^d$ is the set:

$$\mathcal{B}(x, \delta) \stackrel{\text{def}}{=} \{y \in \mathbb{R}^d : \|y - x\| \leq \delta\}$$

For convex problems we are typically interested in a **global minimizer**, and for nonconvex problems in a **stationary point**.

Optimization Modelling and Data

Optimization Modelling

1. One wants to solve some **problem**
 - ▶ Predict what advert a given user will click on
 - ▶ Deblur an image
 - ▶ Rank a collection webpages based on their relevance to a given query
2. **Represent the solution** of the problem as a mathematical object; typically a vector $x \in \mathbb{R}^d$
 - ▶ x_i represents the probability of clicking on advert i
 - ▶ x_i represents the color of pixel i
 - ▶ x_i represents relevance of website i to the query
3. Choose a set $\mathcal{S} \subseteq \mathbb{R}^d$ of solutions which you are willing to consider. \mathcal{S} is the **feasibility/constraint set**.
 - ▶ $\mathcal{S} = \{x \in \mathbb{R}^d : 0 \leq x_i \leq 1, \sum_i x_i = 1\}$
 - ▶ $\mathcal{S} = \{x \in \mathbb{R}^d : 0 \leq x_i \leq 255\}$
 - ▶ $\mathcal{S} = \mathbb{R}^d$ or $\mathcal{S} = \{x \in \mathbb{R}^d : x \geq 0\}$
4. Choose a **cost/loss/objective function** $f : \mathbb{R}^d \rightarrow \mathbb{R}$ measuring the quality of each potential solution. In particular, $f(x) < f(y)$ should mean that x is a better solution than y .
5. Solve problem (OPT) using a suitable **optimization algorithm**

Optimization and Data

Sources of data:

1. **Description data:** Data forming the initial logical description of the problem
2. **Representation data:** Our choice of representation may depend on some synthetic / computed / collected data
3. **Feasibility data:** Data describing \mathcal{S}
4. **Objective data:** Data describing f

Optimization Algorithms: Basic Terminology

Direct vs Iterative Methods I

Direct methods

- ▶ Perform a **finite number of steps**.
- ▶ Intermediary steps can't be used to produce an approximate solution.
- ▶ The solution found is (typically) **exact**.

Iterative methods

- ▶ Perform an **infinite number of steps** (in principle).
- ▶ Intermediary steps generate a sequence of iterates x^0, x^1, x^2, \dots which get progressively better (in some sense).
- ▶ Any solution in the sequence is **approximate** only

Direct vs Iterative Methods II

Example 8

Solve problem (9) with $f(x) = \frac{1}{2} \|\mathbf{A}x - b\|^2$ and $\mathcal{S} = \mathbb{R}^d$.

- **Direct method:** To find the stationary point, which happens to be the global solution since f is convex, set the gradient to zero:

$$\nabla f(x) = \mathbf{A}^\top(\mathbf{A}x - b) = 0.$$

This results to the linear system $\mathbf{A}^\top \mathbf{A}x = \mathbf{A}^\top b$. Solve the system using any direct solver for linear systems, such as Gaussian elimination. The result is obtained after a finite number of steps.

- **Iterative method:** Choose some initial guess $x^0 \in \mathbb{R}^d$, compute constant L satisfying the bound $L \geq \lambda_{\max}(\mathbf{A}^\top \mathbf{A})$ and perform the following iterative process (“gradient descent”):

$$x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k).$$

It can be shown that $f(x^k) \leq \frac{C}{k}$ for all k and some constant C (which depends on x^0 , \mathbf{A} and b).

We will only talk about iterative methods in this course; these are by far the most prevalent and useful in optimization and machine learning.

Iterative Optimization Methods

All iterative methods can be written in the generic form

$$x^{k+1} = x^k + \alpha_k h^k,$$

where

- ▶ $h^k \in \mathbb{R}^d$ is the **search/descent direction**
- ▶ $\alpha_k \in \mathbb{R}$ is the **stepsize**
- ▶ $\alpha_k h^k \in \mathbb{R}^d$ is the **step**

Iterative optimization methods differ in how:

- ▶ the direction h^k is computed (a major distinguishing factor)
- ▶ the stepsize α_k is determined (a minor distinguishing factor)

Complexity

Two key questions when choosing an optimization method:

- ▶ How costly is it to get from x^k to x^{k+1} ? Typically, we care about time. Some proxies for time:
 - ▶ **computation cost = # of arithmetic operations (additions, multiplications) performed,**
 - ▶ **communication cost = # bits transferred.**
- ▶ How many iterations k do we need to perform to get a solution of acceptable quality?

Definition 9 (Complexity of an Optimization Method)

Let \mathcal{M} be an (iterative) optimization method for solving (9). Let $\epsilon > 0$ be a desired error tolerance with respect to some measure of success. Let W be the cost of performing one step of method \mathcal{M} . Let $k(\epsilon)$ be the number of iterations sufficient for the output of \mathcal{M} to be within the desired tolerance ϵ . The **total complexity** of \mathcal{M} is the function

$$\epsilon \mapsto k(\epsilon) \times W.$$

Remark: Typically, both W and $k(\epsilon)$ depend on the problem at hand (i.e., on f and \mathcal{S}).

Classification of Optimization Methods

Definition 10

We say that an (iterative) optimization algorithm is

- ▶ **feasible** if all iterates x^k are feasible, i.e., if $x^k \in \mathcal{S}$ for all k
- ▶ **infeasible** if it is not feasible
- ▶ **deterministic** if, given x^0 , always the same sequence of iterates $\{x^k\}$ is produced
- ▶ **randomized** if the iterates $\{x^k\}$ form a random process (i.e., if they are random vectors)
- ▶ **zero order** if it is only allowed to compute zero-order information about f (i.e., function values)
- ▶ **first order** if it is only allowed to compute up to first-order information about f (i.e., function values and gradients)
- ▶ **second order** if it is allowed to compute up to second-order information about f (i.e., function values, gradients and Hessian)

Zero vs First vs Second Order Methods

Method	Local information at x about f utilized
zero order	function value $f(x) \in \mathbb{R}$
first order	gradient $\nabla f(x) \in \mathbb{R}^d$
second order	Hessian $\nabla^2 f(x) \in \mathbb{R}^{d \times d}$

Gradient and Hessian

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{pmatrix} \in \mathbb{R}^d; \quad \nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_d \partial x_1} & \frac{\partial^2 f(x)}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_d \partial x_d} \end{pmatrix} \in \mathbb{R}^{d \times d}$$

Standard Algorithms

Three Types of Methods

Direct Search (Zero-order method)

- ▶ Fix a set $\mathcal{D} \stackrel{\text{def}}{=} \{d_1, \dots, d_t\} \subset \mathbb{R}^d$, and constants $\alpha_0 > 0$, $c > 0$
- ▶ If $f(x^k + \alpha_k d_i) < f(x^k) - c\alpha_k^2$ for some i ,
 - ▶ then set $x^{k+1} = x^k + \alpha_k d_i$
 - ▶ otherwise $x^{k+1} = x^k$, $\alpha_{k+1} \leftarrow \alpha_k / 2$

Gradient Descent (First-order method)

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

Newton's Method (Second-order method)

$$x^{k+1} = x^k - \alpha_k (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

Complexity of Standard Methods I

Consider the least-squares function

$$f(x) = \frac{1}{2} \|\mathbf{A}x - b\|^2,$$

where $\mathbf{A} \in \mathbb{R}^{n \times d}$ is a data matrix, $x \in \mathbb{R}^d$ is a vector of parameters to be tuned and $b \in \mathbb{R}^n$ is a data vector. Note that

$$\nabla f(x) = \mathbf{A}^\top(\mathbf{A}x - b), \quad \nabla^2 f(x) = \mathbf{A}^\top \mathbf{A}.$$

Cost of evaluation of $f(x)$:

- ▶ Multiply \mathbf{A} by x : $\mathcal{O}(nd)$
- ▶ Subtract b from the result: $\mathcal{O}(n)$
- ▶ Compute the squared norm of the resulting vector: $\mathcal{O}(n)$
- ▶ Total cost: $\mathcal{O}(nd)$

Cost of evaluation of $\nabla f(x)$:

- ▶ Multiply \mathbf{A} by x : $\mathcal{O}(nd)$

Complexity of Standard Methods II

- ▶ Subtract b from the result: $\mathcal{O}(n)$
- ▶ Multiply the result by \mathbf{A}^\top : $\mathcal{O}(nd)$
- ▶ Total cost: $\mathcal{O}(nd)$

Cost of evaluation of $(\nabla^2 f(x))^{-1} \nabla f(x)$:

- ▶ Form the matrix $\nabla^2 f(x) = \mathbf{A}^\top \mathbf{A}$: $\mathcal{O}(nd^2)$
- ▶ Form the gradient $\nabla f(x)$: $\mathcal{O}(nd)$
- ▶ Solve the $n \times n$ linear system $\nabla^2 f(x)h = \nabla f(x)$ using a direct method (e.g., Gaussian elimination): $\mathcal{O}(d^3)$
- ▶ Total cost: $\mathcal{O}(d^3 + nd^2)$

Convergence

What Quantity do we Want to Converge?

Let x^* be a solution we are interested in (global, local, feasible, . . .).

- ▶ If the method is **deterministic**, we may wish one of the following quantities to converge to zero:
 - ▶ Squared distance to a solution: $a_k \stackrel{\text{def}}{=} \|x^k - x^*\|^2 \rightarrow 0$
 - ▶ Function values: $a_k \stackrel{\text{def}}{=} f(x^k) - f(x^*) \rightarrow 0$
 - ▶ Squared norm of the gradient: $a_k \stackrel{\text{def}}{=} \|\nabla f(x^k)\|^2 \rightarrow 0$
- ▶ If the method is **randomized**, we may wish one of the following quantities to converge to zero:
 - ▶ Expected squared distance to a solution: $a_k \stackrel{\text{def}}{=} \mathbf{E} [\|x^k - x^*\|^2] \rightarrow 0$
 - ▶ Expected function values: $a_k \stackrel{\text{def}}{=} \mathbf{E} [f(x^k) - f(x^*)] \rightarrow 0$
 - ▶ Expected squared gradient norm: $a_k \stackrel{\text{def}}{=} \mathbf{E} [\|\nabla f(x^k)\|^2] \rightarrow 0$

The above are just some of the most common measures of success, several other quantities are studied as well.

The expectation above is with respect to the coin flips / randomness inherent in the randomized algorithm.

Typical Convergence Rates Encountered in Optimization

Bound on a_k	Bound on k implying $a_k \leq \epsilon$	Typical Recursion	Speed
$a_k \leq \frac{C}{\sqrt{k}}$	$k \geq \frac{C}{\epsilon^2}$		slow
$a_k \leq \frac{C}{k}$	$k \geq \frac{C}{\epsilon}$	$a_{k+1} \leq a_k - \frac{a_k^2}{C}$	ok
$a_k \leq \frac{C}{k^2}$	$k \geq \sqrt{\frac{C}{\epsilon}}$		fast
$a_k \leq a_0 e^{-k/C}$	$k \geq C \log \left(\frac{a_0}{\epsilon} \right)$	$a_{k+1} \leq \left(1 - \frac{1}{C}\right) a_k$	very fast

Software

We will work using CVXPY (open source Python-embedded modeling language for convex optimization problems): <https://www.cvxpy.org>

Introduction to Optimization

Peter Richtárik



Lecture 3: Convex Sets - Part 1

Lecture Outline

- ▶ Convex sets: definition
- ▶ Convex sets: examples
- ▶ Convex sets: examples with proofs

Convex Sets: Definition

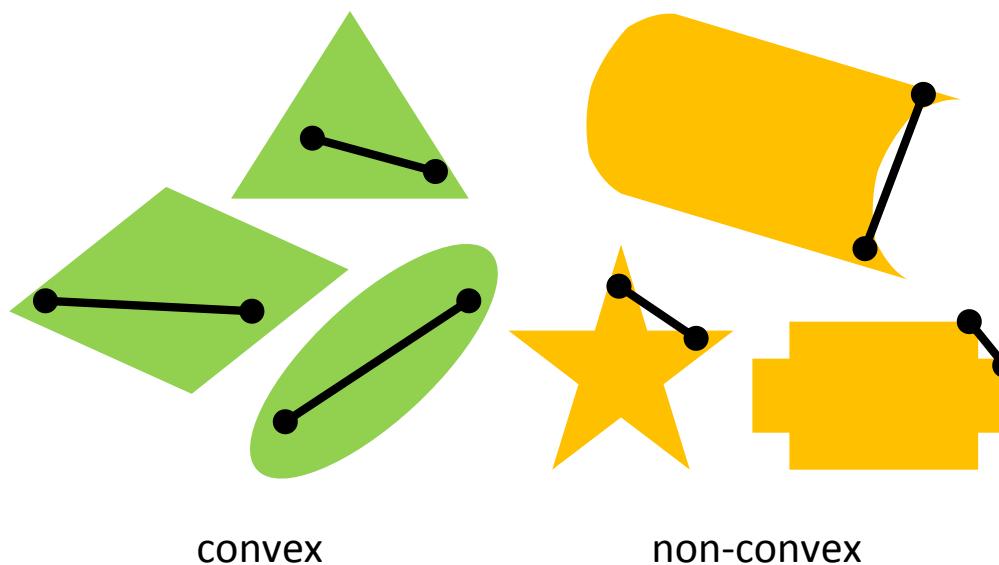
Definition of Convex Sets

Definition 11 (Convex set)

A set $\mathcal{S} \subseteq \mathbb{R}^d$ is **convex** if for any $x, y \in \mathcal{S}$ and any $0 \leq \lambda \leq 1$,

$$\lambda x + (1 - \lambda)y \in \mathcal{S}.$$

That is, \mathcal{S} is convex if with every two points it also contains the line segment joining them.



Convex Sets: Examples

Examples of Convex Sets I

1. **Empty set** \emptyset .
2. **Singleton** $S = \{s\}$, for any $s \in \mathbb{R}^d$.
3. **Line segment** between any two points $x, y \in \mathbb{R}^d$:

$$\mathcal{L} = \{\lambda x + (1 - \lambda)y : \lambda \in [0, 1]\}.$$

One can also consider the **open line segment** between any two points $x, y \in \mathbb{R}^d$:

$$\mathcal{L} = \{\lambda x + (1 - \lambda)y : \lambda \in (0, 1)\}.$$

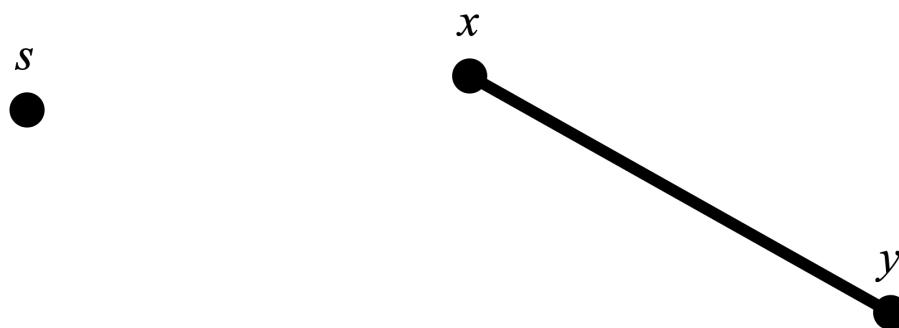


Figure: A singleton (left) and a line segment (right) in \mathbb{R}^2 .

Examples of Convex Sets II

4. **Half line** passing through $a \in \mathbb{R}^d$ in the direction $b \in \mathbb{R}^d$:

$$\mathcal{L} = \{a + tb : t \geq 0\}.$$

One can also consider the **open half line**

$$\mathcal{L} = \{a + tb : t > 0\}.$$

5. **Line** passing through $a \in \mathbb{R}^d$ in the direction $b \in \mathbb{R}^d$:

$$\mathcal{L} = \{a + tb : t \in \mathbb{R}\}.$$

6. **Hyperplane** given by the normal vector $0 \neq a \in \mathbb{R}^d$, passing through vector $b \frac{a}{\|a\|^2}$:

$$\mathcal{H} = \{x \in \mathbb{R}^d : a^\top x = b\}.$$

Examples of Convex Sets III

7. Left halfspace

$$\mathcal{H} = \{x \in \mathbb{R}^d : a^\top x \leq b\}.$$

One can also consider the **open left halfspace**

$$\mathcal{H} = \{x \in \mathbb{R}^d : a^\top x < b\}.$$

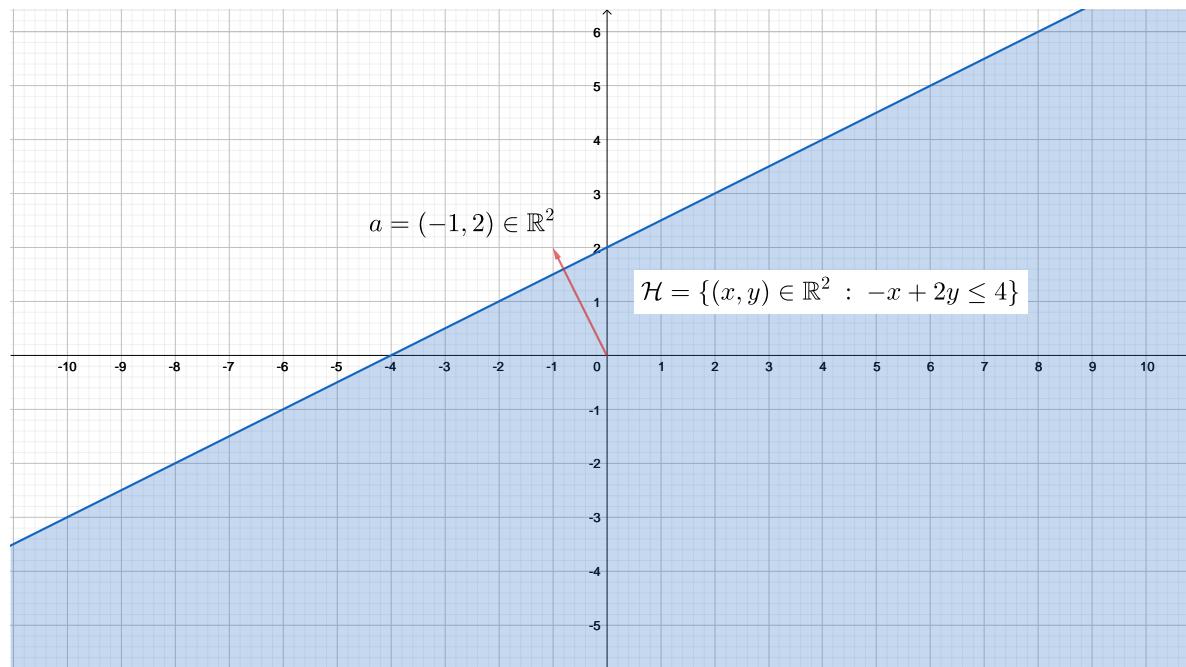


Figure: Left halfspace: $\mathcal{H} = \{x \in \mathbb{R}^2 : a^\top x \leq b\}$, where $a = (-1, 2) \in \mathbb{R}^2$ and $b = 4$.

Examples of Convex Sets IV

8. Right halfspace

$$\mathcal{H} = \{x \in \mathbb{R}^d : a^\top x \geq b\}.$$

One can also consider the **open right halfspace**

$$\mathcal{H} = \{x \in \mathbb{R}^d : a^\top x > b\}.$$

9. Solution set of any system of n linear equations:

$$\mathcal{S} = \{x \in \mathbb{R}^d : \mathbf{A}x = b\},$$

where $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$.

10. Null space (a.k.a. kernel) of a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$:

$$\text{null}(\mathbf{A}) = \{x \in \mathbb{R}^d : \mathbf{A}x = 0\}.$$

11. Range space (a.k.a. image) of a matrix $\mathbf{A}^\top \in \mathbb{R}^{n \times d}$:

$$\text{range}(\mathbf{A}^\top) = \{\mathbf{A}^\top y : y \in \mathbb{R}^n\}.$$

12. Any linear subspace of \mathbb{R}^d .

Examples of Convex Sets V

13. **Ball** with center $a \in \mathbb{R}^d$ and radius $r > 0$:

$$\mathcal{B}(a, r) = \{x \in \mathbb{R}^d : \|x - a\| \leq r\}.$$

One can also consider the **open ball**:

$$\mathcal{B}(a, r) = \{x \in \mathbb{R}^d : \|x - a\| < r\}.$$

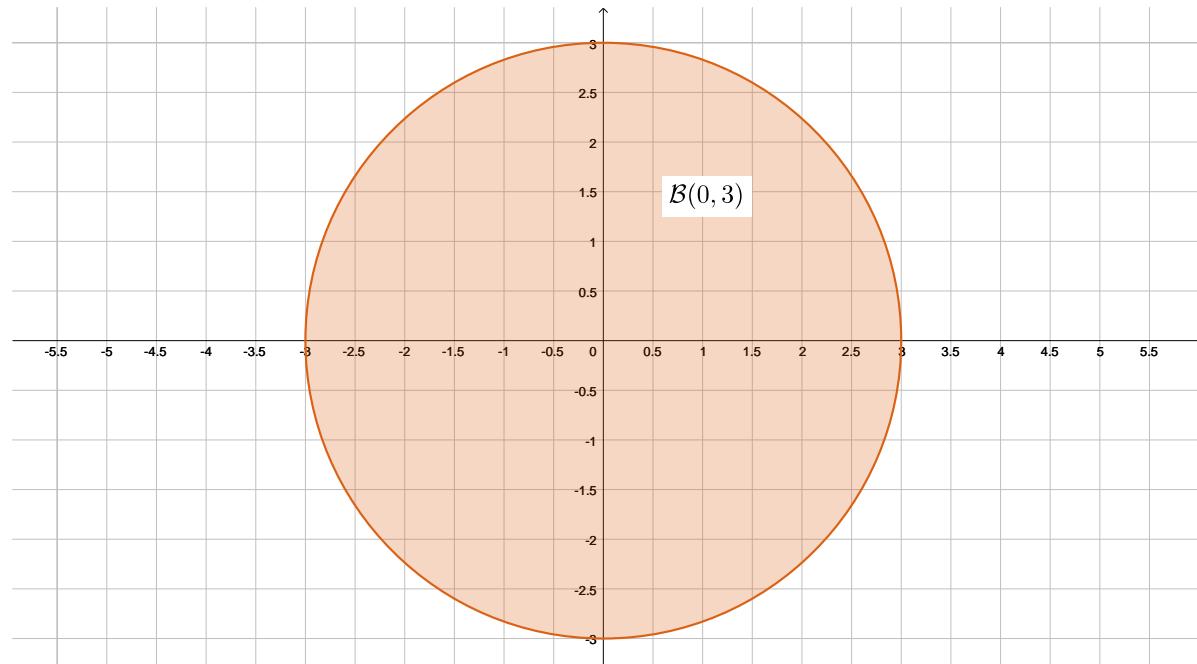


Figure: Ball in \mathbb{R}^2 , centered at $a = (0, 0) \in \mathbb{R}^2$, with radius $r = 3$.

Examples of Convex Sets VI

14. **Ellipsoid** with center $a \in \mathbb{R}^d$ and positive definite “shape” matrix $\mathbf{Q} \in \mathbb{S}_{++}^d$:

$$\mathcal{E} = \{x \in \mathbb{R}^d : (x - a)^\top \mathbf{Q}(x - a) \leq 1\}.$$

Note that the ball $\mathcal{B}(a, r)$ is a special case of this for $\mathbf{Q} = \frac{1}{r^2} \mathbf{I}$, where $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix.

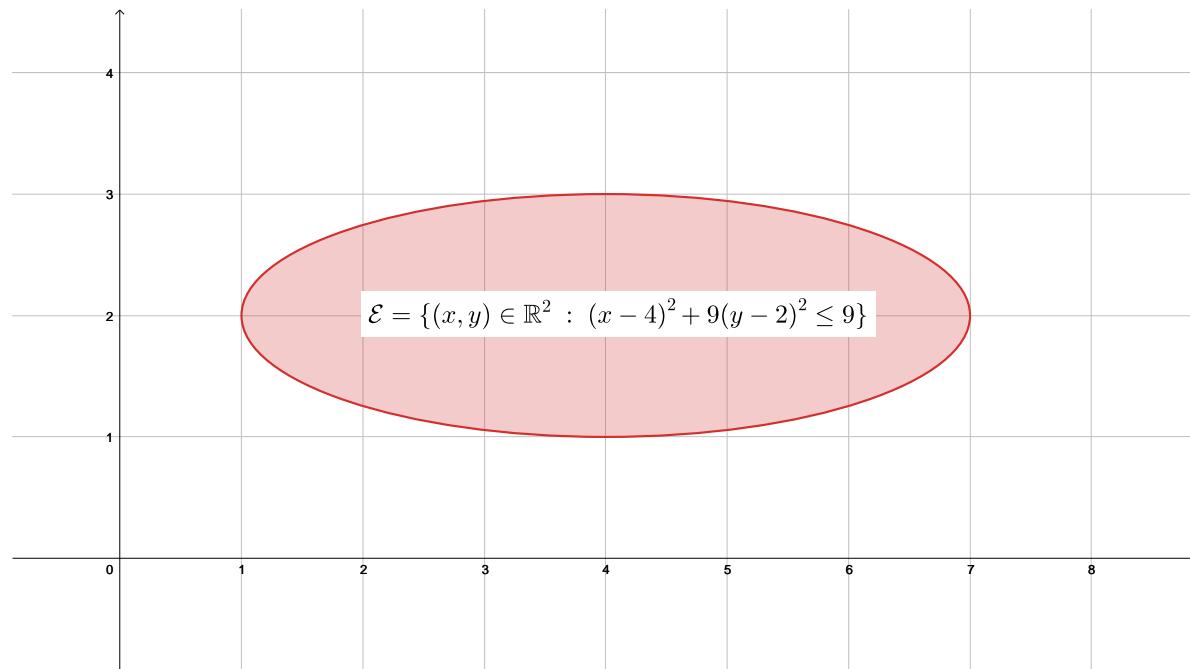


Figure: Ellipsoid in \mathbb{R}^2 with center $a = (4, 2) \in \mathbb{R}^2$ and $\mathbf{Q} = \begin{pmatrix} 1/9 & 0 \\ 0 & 1 \end{pmatrix}$.

Examples of Convex Sets VII

15. **Probability simplex:**

$$\Delta^d \stackrel{\text{def}}{=} \left\{ x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x \geq 0 \right\}.$$

16. **Cone of nonnegative vectors:**

$$\mathbb{R}_+^d \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : x \geq 0\}.$$

17. **Lorentz/second-order cone:**

$$\mathbb{L}^{d+1} = \{(x, t) \in \mathbb{R}^d \times \mathbb{R} : \|x\| \leq t\}.$$

18. **Solution set of any system of linear inequalities:**

$$\mathcal{S} = \{x \in \mathbb{R}^d : \mathbf{A}x \leq b\}.$$

Examples of Convex Sets VIII

19. **Linear hull** of any set $\mathcal{S} \subseteq \mathbb{R}^d$:

$$\text{linear}(\mathcal{S}) \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^n \lambda_i x_i : n \in \mathbb{N}; x_1, \dots, x_n \in \mathcal{S}; \lambda \in \mathbb{R}^n \right\}.$$

20. **Affine hull** of any set $\mathcal{S} \subseteq \mathbb{R}^d$:

$$\text{affine}(\mathcal{S}) \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^n \lambda_i x_i : n \in \mathbb{N}; x_1, \dots, x_n \in \mathcal{S}; \lambda \in \mathbb{R}^n, \sum_{i=1}^n \lambda_i = 1 \right\}.$$

21. **Conic hull** of any set $\mathcal{S} \subseteq \mathbb{R}^d$:

$$\text{cone}(\mathcal{S}) \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^n \lambda_i x_i : n \in \mathbb{N}; x_1, \dots, x_n \in \mathcal{S}; \lambda \in \mathbb{R}_+^n \right\}. \quad (10)$$

22. **Convex hull** of any set $\mathcal{S} \subseteq \mathbb{R}^d$:

$$\text{conv}(\mathcal{S}) \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^n \lambda_i x_i : n \in \mathbb{N}; x_1, \dots, x_n \in \mathcal{S}; \lambda \in \Delta^n \right\}.$$

Examples of Convex Sets IX

23. **Polar cone** of any set $\mathcal{S} \subseteq \mathbb{R}^d$:

$$\mathcal{S}^- \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : \langle s, x \rangle \leq 0 \ \forall s \in \mathcal{S}\}.$$

24. **Normal cone** to a convex set $\mathcal{S} \subseteq \mathbb{R}^d$ at point $x \in \mathcal{S}$:

$$N_{\mathcal{S}}(x) \stackrel{\text{def}}{=} (\mathcal{S} - x)^-,$$

where $\mathcal{S} - x \stackrel{\text{def}}{=} \{s - x : s \in \mathcal{S}\}$.

Examples with Proofs

Examples with Proofs I

Example 12 (Ball is convex)

Prove that the ball $\mathcal{B}(a, r)$, where $a \in \mathbb{R}^d$ and $r > 0$, is a convex set.

Solution: Let $x, y \in \mathcal{B}(a, r)$ and $\lambda \in [0, 1]$. We need to show that $z = \lambda x + (1 - \lambda)y \in \mathcal{B}(a, r)$. Indeed,

$$\begin{aligned}\|z - a\| &= \|\lambda x + (1 - \lambda)y - a\| \\ &= \|\lambda x + (1 - \lambda)y - (\lambda a + (1 - \lambda)a)\| \\ &= \|\lambda(x - a) + (1 - \lambda)(y - a)\| \\ &\stackrel{(5)}{\leq} \|\lambda(x - a)\| + \|(1 - \lambda)(y - a)\| \\ &= \lambda \|x - a\| + (1 - \lambda) \|y - a\| \\ &\leq \lambda r + (1 - \lambda)r \\ &= r,\end{aligned}$$

where the last inequality is due to $\|x - a\| \leq r$ and $\|y - a\| \leq r$, which follows since both x and y belong to $\mathcal{B}(a, r)$.

Examples with Proofs II

Example 13 (Probability simplex is convex)

Show that the probability simplex Δ^d is a convex set.

Solution: Pick any $p, q \in \Delta^d$ and $\lambda \in [0, 1]$. We need to show that

$$r \stackrel{\text{def}}{=} \lambda p + (1 - \lambda)q \in \Delta^d. \quad (11)$$

- ▶ First, we need to argue that $r = (r_1, \dots, r_d) \geq 0$. However, this is obvious since $r_i = \lambda p_i + (1 - \lambda)q_i$, while $p_i \geq 0$ and $q_i \geq 0$.
- ▶ Second, we need to argue that $\sum_{i=1}^d r_i = 1$. Indeed,

$$\begin{aligned} \sum_{i=1}^d r_i &\stackrel{(11)}{=} \sum_{i=1}^d (\lambda p + (1 - \lambda)q)_i = \sum_{i=1}^d (\lambda p_i + (1 - \lambda)q_i) \\ &= \underbrace{\lambda \sum_{i=1}^d p_i}_{=1} + (1 - \lambda) \underbrace{\sum_{i=1}^d q_i}_{=1} = \lambda + (1 - \lambda) = 1. \end{aligned}$$

Cones

Closed Convex Cones I

Definition 14 (Cone)

A set $\mathcal{S} \subseteq \mathbb{R}^d$ is a **cone** if

$$x \in \mathcal{S} \implies tx \in \mathcal{S} \quad \forall t \geq 0.$$

Remarks:

- We will work with closed convex cones, denoted by CCC.

Example 15 (Trivial cone)

The set $\mathcal{S} = \{0\}$ is a CCC.

Example 16 (\mathbb{R}^d)

The entire space $\mathcal{S} = \mathbb{R}^d$ is a CCC.

Example 17 (Linear subspace)

Any linear subspace \mathcal{S} of \mathbb{R}^d is a CCC.

Example 18 (Non-negative orthant)

Closed Convex Cones II

The non-negative orthant

$$\mathcal{S} = \mathbb{R}_+^d \stackrel{\text{def}}{=} \{x = (x_1, \dots, x_d) \in \mathbb{R}^d : x_1, \dots, x_d \geq 0\}$$

is a CCC.

Example 19 (Lorentz/second-order/ice-cream cone)

The set

$$\mathcal{S} = \mathbb{L}^{d+1} \stackrel{\text{def}}{=} \{(x, t) \in \mathbb{R}^d \times \mathbb{R} : \|x\| \leq t\}$$

is a CCC.

Example 20 (Cone of symmetric positive semidefinite matrices)

The set

$$\mathcal{S} = \mathbb{S}_+^d \stackrel{\text{def}}{=} \{\mathbf{A} \in \mathbb{R}^{d \times d} : \mathbf{A} = \mathbf{A}^\top \quad \text{and} \quad x^\top \mathbf{A} x \geq 0 \quad \forall x \in \mathbb{R}^d\}$$

of symmetric positive definite matrices is a CCC.

Exercises I

Exercise 21

Show that $\text{null}(\mathbf{A})$ is a convex set.

Exercise 22

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{B} \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}^m$. Show that the set

$$\mathcal{S} = \{x \in \mathbb{R}^d : \mathbf{A}x = b, \mathbf{B}x \leq c\}$$

is convex.

Exercise 23

Visualize the set

$$\mathcal{S} = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 \leq 1, 0 \leq z \leq 1\}.$$

What does it “look like”? Is this set convex or not? Why?

Exercises II

Exercise 24 (☕)

Show that $(\mathbb{R}_+^d)^- = -\mathbb{R}_+^d$, i.e., show that the polar cone of the nonnegative orthant is equal to $-\mathbb{R}_+^d$.

Exercise 25

Visualize the normal cone $N_{\mathcal{S}}(x)$ to the unit ball $\mathcal{S} = \mathcal{B}(0, 1)$ in \mathbb{R}^2 at point $x = (1, 0) \in \mathbb{R}^2$. Verify that $N_{\mathcal{S}}(x)$ is indeed a convex set.

Introduction to Optimization

Peter Richtárik



Lecture 4: Convex Sets - Part 2

Lecture Outline

- ▶ Manipulating convex sets
- ▶ Topological properties of convex sets

Manipulating Convex Sets

Operations Preserving Convexity

In the next few slides, we will describe the following convexity-preserving operations:

Operation	Theorem
Intersection of convex sets	Theorem 26
Cartesian product of convex sets	Theorem 28
Linear image of a convex set	Theorem 30
Linear combination of convex sets	Theorem 32
Linear pre-image of a convex set	Theorem 34

Table: Convexity preserving operations

They are very useful when checking convexity of sets, including many of the 21 examples we have seen before.

Intersection of Convex Sets

Theorem 26 (Intersection of convex sets)

Let $\mathcal{S}_i, i \in \mathcal{I}$ be any collection of convex sets, where \mathcal{I} is any index set, possibly infinite. Then the intersection of these sets, i.e., the set

$$\mathcal{S} = \cap_{i \in \mathcal{I}} \mathcal{S}_i$$

is a convex set.

Proof.

Assume that $x, y \in \mathcal{S}$, and choose any $\lambda \in [0, 1]$. Then for any $i \in \mathcal{I}$, $x, y \in \mathcal{S}_i$. Since \mathcal{S}_i is convex, $z = \lambda x + (1 - \lambda)y \in \mathcal{S}_i$. Since this is true for any $i \in \mathcal{I}$, this means that $z \in \mathcal{S}$. □

Example 27 (Solution set of a linear system is convex)

Show that the set $\mathcal{S} = \{x \in \mathbb{R}^d : \mathbf{A}x = b\}$, where $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$, is convex.

Solution: Let $a_i = \mathbf{A}_{i:}^\top$ be the i^{th} row of \mathbf{A} , and b_i be the i^{th} element of b . Then $\mathcal{S} = \cap_{i=1}^n \mathcal{S}_i$, where $\mathcal{S}_i = \{x \in \mathbb{R}^d : a_i^\top x = b_i\}$ is a hyperplane. Since each \mathcal{S}_i is convex, then by Theorem 26 we conclude that \mathcal{S} is convex.

Cartesian Product of Convex Sets

Theorem 28 (Cartesian product of convex sets)

Let $\mathcal{S}_1 \in \mathbb{R}^{d_1}, \dots, \mathcal{S}_n \in \mathbb{R}^{d_n}$ be convex sets. Then the set

$$\mathcal{S}_1 \times \cdots \times \mathcal{S}_n \stackrel{\text{def}}{=} \{(x_1, \dots, x_n) \in \mathbb{R}^{d_1 + \cdots + d_n} : x_1 \in \mathcal{S}_1, \dots, x_n \in \mathcal{S}_n\}, \quad (12)$$

is convex.

Example 29 (Hypercube is convex)

Show that the **hypercube** $\mathcal{C} \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : 0 \leq x_i \leq 1, i = 1, \dots, d\}$ is convex.

Solution: The interval $[0, 1] \subset \mathbb{R}$ is convex. Let $\mathcal{C}_i = [0, 1]$ for $i = 1, \dots, d$. By Theorem 28, the Cartesian product of these sets is convex:

$$\begin{aligned} \mathcal{C}_1 \times \cdots \times \mathcal{C}_d &\stackrel{(12)}{=} \{x \in \mathbb{R}^d : x_1 \in \mathcal{C}_1, \dots, x_d \in \mathcal{C}_d\} \\ &= \{x \in \mathbb{R}^d : 0 \leq x_i \leq 1, i = 1, \dots, d\}. \end{aligned}$$

Linear Image of a Convex Set I

Theorem 30 (Linear image of a convex set)

Let $\mathcal{S} \in \mathbb{R}^d$ be a convex set, and let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then the set

$$\mathbf{A}(\mathcal{S}) \stackrel{\text{def}}{=} \{\mathbf{Ax} \in \mathbb{R}^n : x \in \mathcal{S}\}$$

is convex.

Proof.

Let $x', y' \in \mathbf{A}(\mathcal{S})$ and $\lambda \in [0, 1]$. We need to show that

$$\lambda x' + (1 - \lambda)y' \in \mathbf{A}(\mathcal{S}).$$

Note that there exist points $x, y \in \mathcal{S}$ such that $x' = \mathbf{Ax}$ and $y' = \mathbf{Ay}$.

Since \mathcal{S} is convex, we know that $z = \lambda x + (1 - \lambda)y \in \mathcal{S}$. But this means that $\mathbf{Az} \in \mathbf{A}(\mathcal{S})$. However, by linearity,

$$\lambda x' + (1 - \lambda)y' = \lambda \mathbf{Ax} + (1 - \lambda)\mathbf{Ay} = \mathbf{A}(\lambda x + (1 - \lambda)y) = \mathbf{Az} \in \mathbf{A}(\mathcal{S}).$$

Linear Image of a Convex Set II

Example 31 (Conic hull of a finite set is convex)

Show that the conic hull of the finite set $\mathcal{S} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ is convex.

Solution: Let

$$\mathbf{A} = [x_1, \dots, x_n] \in \mathbb{R}^{d \times n}.$$

Since the set \mathbb{R}_+^d is convex, in view of Theorem 30, its image under \mathbf{A} is convex as well. It remains to observe that this image is equal to the conic hull of \mathcal{S} :

$$\mathbf{A}(\mathbb{R}_+^d) = \left\{ \sum_{i=1}^n \lambda_i x_i : \lambda \geq 0 \right\} \stackrel{(10)}{=} \text{cone}(\mathcal{S}).$$

Linear Combination of Convex Sets I

Theorem 32 (Linear combination of convex sets)

Let $\mathcal{S}_1 \in \mathbb{R}^d, \dots, \mathcal{S}_n \in \mathbb{R}^d$ be convex sets, and let $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. Then the set

$$\sum_{i=1}^n \alpha_i \mathcal{S}_i \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^n \alpha_i x_i : x_1 \in \mathcal{S}_1, \dots, x_n \in \mathcal{S}_n \right\}$$

is convex.

Proof.

By Theorem 28, the Cartesian product $\mathcal{S} \stackrel{\text{def}}{=} \mathcal{S}_1 \times \dots \times \mathcal{S}_n \in \mathbb{R}^{d \times \dots \times d}$ is convex. Now let

$$\mathbf{A} = [\alpha_1 \mathbf{I}, \dots, \alpha_n \mathbf{I}],$$

where $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix. By Theorem 30, the image of \mathcal{S} under \mathbf{A} is convex. It remains to notice that this image is precisely the set $\sum_{i=1}^n \alpha_i \mathcal{S}_i$. □

Linear Combination of Convex Sets II

Example 33 (Shift of a convex set is convex)

Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a convex set, and $a \in \mathbb{R}^d$. Then the **shifted set**

$$\mathcal{S} + a \stackrel{\text{def}}{=} \{x + a : x \in \mathcal{S}\}$$

is convex.

Solution: Let $\mathcal{S}_1 = \mathcal{S}$, $\mathcal{S}_2 = \{a\}$, $\alpha_1 = \alpha_2 = 1$. Since both \mathcal{S}_1 and \mathcal{S}_2 are convex, then in view of Theorem 32, the set

$$\alpha_1 \mathcal{S}_1 + \alpha_2 \mathcal{S}_2 = \mathcal{S} + \{a\}$$

is also convex.

Linear Pre-Image of a Convex Set I

Theorem 34 (Linear pre-image of a convex set)

Let $\mathcal{S} \in \mathbb{R}^n$ be a convex set, and let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Then the set²

$$\mathbf{A}^{-1}(\mathcal{S}) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : \mathbf{A}x \in \mathcal{S}\} \quad (13)$$

is convex.

Proof.

Let $x_1, x_2 \in \mathbf{A}^{-1}(\mathcal{S})$ and $\lambda \in [0, 1]$. We want to show that

$$x \stackrel{\text{def}}{=} \lambda x_1 + (1 - \lambda)x_2 \in \mathbf{A}^{-1}(\mathcal{S}).$$

In order to do so, we just need to argue that $\mathbf{A}x \in \mathcal{S}$. Indeed, by linearity

$$\mathbf{A}x = \lambda \mathbf{A}x_1 + (1 - \lambda) \mathbf{A}x_2 \in \mathcal{S},$$

where the last inclusion follows since $\mathbf{A}x_1 \in \mathcal{S}$, $\mathbf{A}x_2 \in \mathcal{S}$ and because \mathcal{S} is convex. □

Linear Pre-Image of a Convex Set II

Example 35 (Convexity of a centered ellipsoid)

Let $\mathbf{Q} \in \mathbb{S}_{++}^d$. Show that the centered ellipsoid

$$\mathcal{E} = \{x \in \mathbb{R}^d : x^\top \mathbf{Q} x \leq 1\}$$

is a convex set.

Solution: Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a matrix such that $\mathbf{Q} = \mathbf{A}^\top \mathbf{A}$. Then

$$\begin{aligned}\mathcal{E} &= \{x \in \mathbb{R}^d : x^\top \mathbf{A}^\top \mathbf{A} x \leq 1\} \\ &= \{x \in \mathbb{R}^d : \|\mathbf{A} x\|^2 \leq 1\} \\ &= \{x \in \mathbb{R}^d : \mathbf{A} x \in \mathcal{B}(0, 1)\} \\ &\stackrel{(13)}{=} \mathbf{A}^{-1}(\mathcal{B}(0, 1)),\end{aligned}$$

where $\mathcal{B}(0, 1)$ is the unit ball in \mathbb{R}^d centered at the origin.

²The notation \mathbf{A}^{-1} does *not* denote the inverse of the matrix! However, if \mathbf{A} is invertible, then $\mathbf{A}^{-1}y = x \Leftrightarrow \mathbf{A}^{-1}(\{y\}) = \{x\}$.

Topological Properties of Convex Sets

Convexity and Topology

Theorem 36 (Closure and Interior of Convex Sets)

- (i) If $\mathcal{S} \subseteq \mathbb{R}^d$ is a convex set, then $\text{cl}(\mathcal{S})$ is convex.
- (ii) Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a convex set with $\text{int}(\mathcal{S}) \neq \emptyset$. Assume $x \in \text{int}(\mathcal{S})$ and $y \in \text{cl}(\mathcal{S})$. Then $\lambda x + (1 - \lambda)y \in \text{int}(\mathcal{S})$ for all $\lambda \in [0, 1]$.
- (iii) If $\mathcal{S} \subseteq \mathbb{R}^d$ is a convex set, then $\text{int}(\mathcal{S})$ is convex.
- (iv) Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a convex set with $\text{int}(\mathcal{S}) \neq \emptyset$. Then $\text{cl}(\text{int}(\mathcal{S})) = \text{cl}(\mathcal{S})$ and $\text{int}(\text{cl}(\mathcal{S})) = \text{int}(\mathcal{S})$.

Theorem 37 (Closedness & Boundedness of Convex/Conic Hulls)

- (v) If $\mathcal{S} \subseteq \mathbb{R}^d$ is a compact set, then $\text{conv}(\mathcal{S})$ is compact.
- (vi) If $\mathcal{S} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$, then $\text{cone}(\mathcal{S})$ is closed.

Exercises I

Exercise 38

Which of the following subsets of \mathbb{R}^2 are convex?

- (i) $\{(x_1, x_2)^\top \in \mathbb{R}^2 : 10 < x_1 \leq 20, x_2 = 5\}$
- (ii) $\{(x_1, x_2)^\top \in \mathbb{R}^2 : x_1^2 \leq 2022\}$
- (iii) $\{(x_1, x_2)^\top \in \mathbb{R}^2 : x_1 \leq x_2\}$
- (iv) $\{(x_1, x_2)^\top \in \mathbb{R}^2 : \max\{x_1, x_2\} \leq -30\}$
- (v) $\{(x_1, x_2)^\top \in \mathbb{R}^2 : \min\{x_1, x_2\} \leq 100\}$
- (vi) $\{(x_1, x_2)^\top \in \mathbb{R}^2 : x_1 + x_2 = 10, x_1 - x_2 = 1\}$
- (vii) $\{(x_1, x_2)^\top \in \mathbb{R}^2 : x_1^4 + x_2^4 \leq 1\}$
- (viii) $\{(x_1, x_2)^\top \in \mathbb{R}^2 : (x_1 - 3x_2)^2 \leq 0\}$

Exercise 39 (☕☕)

Provide an alternative proof of Theorem 32, one that does not rely on Theorem 30.

Exercises II

Exercise 40 (Convex hull of a finite set is convex ☕☕)

Let $\mathcal{S} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$. Show that $\text{conv}(\mathcal{S})$ is convex. Do it in two different ways:

- (i) Use mathematical induction.
- (ii) Use Theorem 30.

Exercise 41 (Convex combination of multiple points ☕☕)

Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a convex set, and $x_1, \dots, x_n \in \mathcal{S}$. Show that

$$\text{conv}(\{x_1, \dots, x_n\}) \subseteq \mathcal{S}.$$

That is, any **convex combination** of points belonging to \mathcal{S} also belongs to \mathcal{S} .

Exercises III

Exercise 42 (Convex hull as the smallest convex set ☕☕)

Let $\mathcal{A} \subset \mathcal{S} \subseteq \mathbb{R}^d$, where \mathcal{S} is convex. Show that then

$$\text{conv}(\mathcal{A}) \subseteq \mathcal{S}.$$

This means that the convex hull of any set \mathcal{A} is the smallest convex set containing \mathcal{A} .

Exercise 43 (Convexity of the ellipsoid ☕☕)

Argue that the ellipsoid

$$\mathcal{E} = \{x \in \mathbb{R}^d : (x - a)^\top Q(x - a) \leq 1\},$$

where $Q \in \mathbb{S}_{++}^d$ is a convex set.

Hint: Combine Example 33 and Example 35.

Introduction to Optimization

Peter Richtárik



Lecture 5: Convex Functions - Part 1

Lecture Outline

- ▶ Definition and examples
- ▶ First-order characterizations of convex functions
- ▶ Second-order characterizations of convex functions

Definition and Examples

Convex Functions I

Definition 44 (Convex function)

A function $f : \mathcal{S} \rightarrow \mathbb{R}$ defined on a convex set $\mathcal{S} \subseteq \mathbb{R}^d$ is **convex** if

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \quad (14)$$

for all $x, y \in \mathcal{S}$ and all $\lambda \in [0, 1]$.

Remarks:

- ▶ If \mathcal{S} is not specified, it is implicitly assumed that $\mathcal{S} = \mathbb{R}^d$.
- ▶ If instead of (14) we insist on the **strict inequality**

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

for all $\lambda \in (0, 1)$, such a function is called **strictly convex**.

- ▶ A function $f : \mathcal{S} \rightarrow \mathbb{R}$ defined on a convex set $\mathcal{S} \subseteq \mathbb{R}^d$ is **concave** if $-f$ is convex.

Convex Functions II

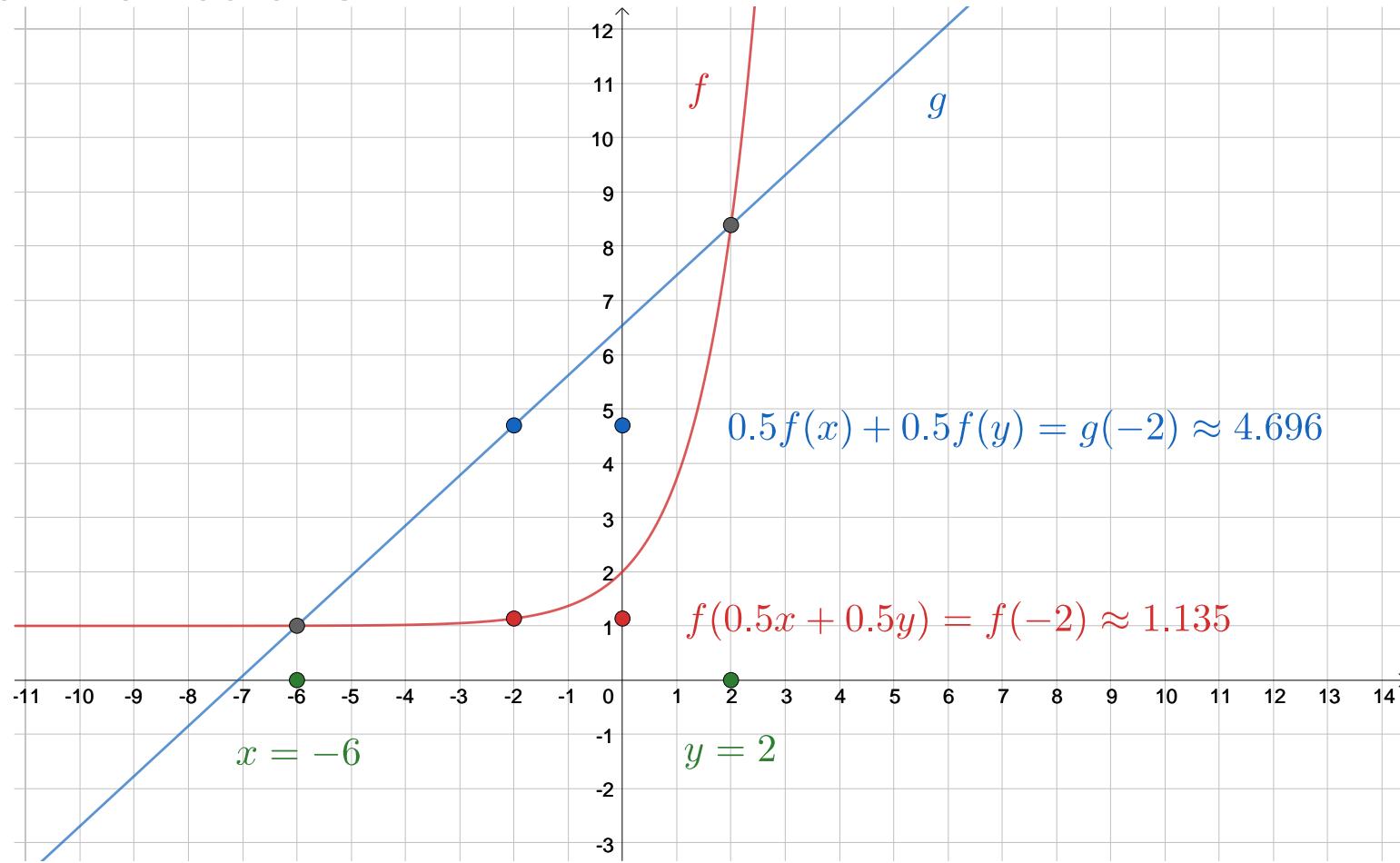


Figure: Geometric illustration of the inequality (14) defining a convex function for $f(x) = 1 + e^x$. We show that $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$ for $x = -6$, $y = 2$ and $\lambda = \frac{1}{2}$.

Convex Functions III

The following inequality generalizes the convexity-defining inequality (14) from 2 points to n points:

Theorem 45 (Jensen's inequality)

Let $f : \mathcal{S} \rightarrow \mathbb{R}$ be a convex function, where $\mathcal{S} \subseteq \mathbb{R}^d$ is a convex set. Then for any collection of vectors $x_1, \dots, x_n \in \mathcal{S}$ and $\lambda \in \Delta^n$, the following inequality holds:

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \leq \sum_{i=1}^n \lambda_i f(x_i).$$

Recall that $\Delta^n \stackrel{\text{def}}{=} \{\lambda \in \mathbb{R}^n : \lambda \geq 0, \sum_{i=1}^n \lambda_i = 1\}$.

Convex Functions IV

Exercise 46

Is there a similar way to visualize Jensen's inequality as was the case for inequality (14)?

Exercise 47

Check numerically that Jensen's inequality holds for the convex function $f(x) = -\sqrt{x}$, $x \in \mathbb{R}_+$.

Exercise 48 (☕)

Prove Jensen's inequality.

Examples of Univariate Convex Functions I

Function	$f(x)$	Condition	Defined on
constant	c	$c \in \mathbb{R}$	\mathbb{R}
linear	bx	$b \in \mathbb{R}$	\mathbb{R}
affine	$bx + c$	$b, c \in \mathbb{R}$	\mathbb{R}
quadratic	$ax^2 + bx + c$	$a \geq 0, b, c \in \mathbb{R}$	\mathbb{R}
reciprocal	$\frac{1}{x}$	—	$(0, \infty)$
exponential	e^x	—	\mathbb{R}
negative logarithmic	$-\log x$	—	\mathbb{R}
absolute value	$ x $	—	\mathbb{R}
simple quartic	x^4	—	\mathbb{R}
power	$ x ^p$	$p \geq 1$	\mathbb{R}
negative square root	$-\sqrt{x}$	—	$(0, \infty)$
logistic loss	$\log(1 + e^{-x})$	—	\mathbb{R}

Table: Examples of convex functions $f : \mathbb{R} \rightarrow \mathbb{R}$.

Examples of Univariate Convex Functions II

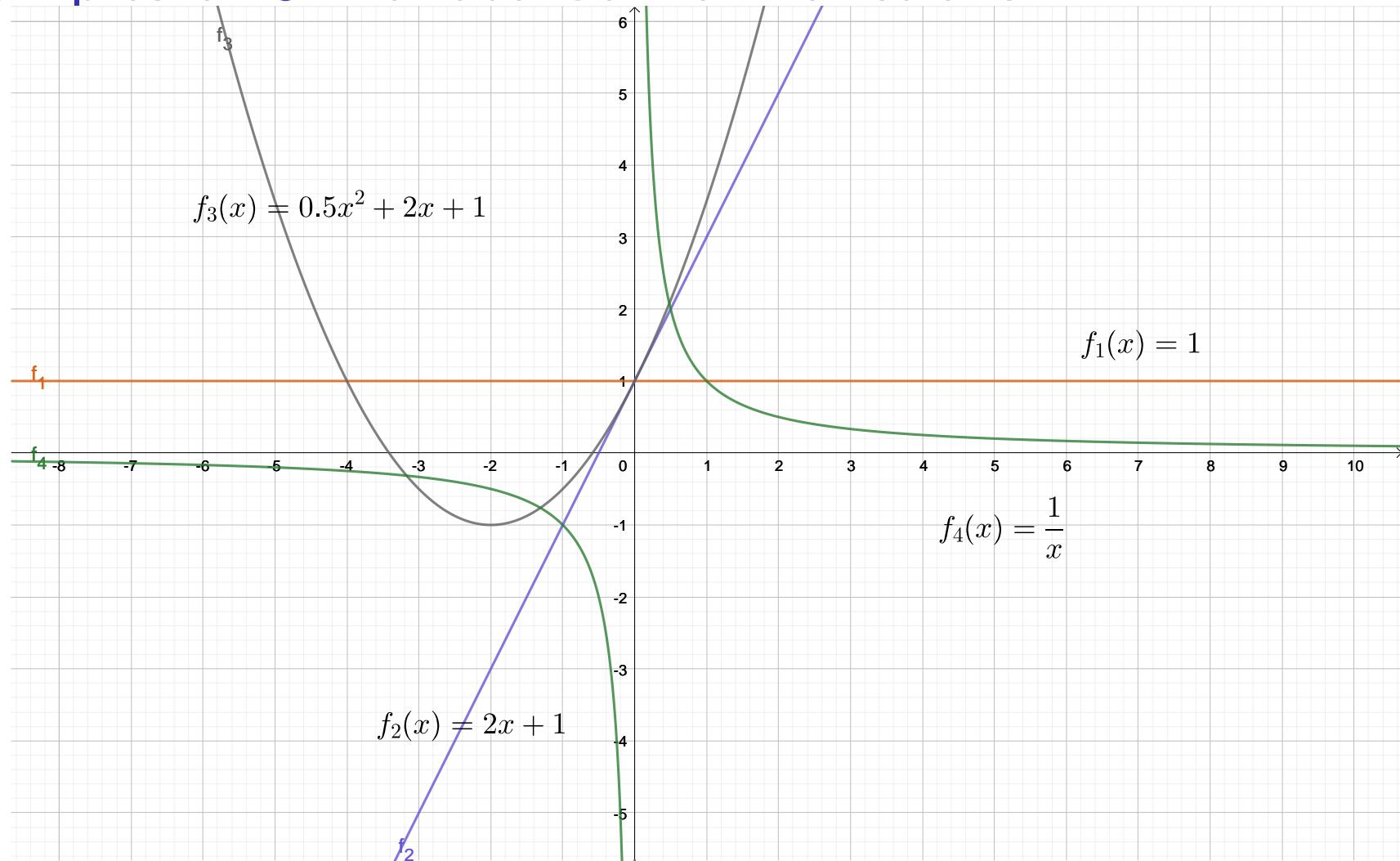


Figure: Examples of univariate convex functions.

Examples of Univariate Convex Functions III

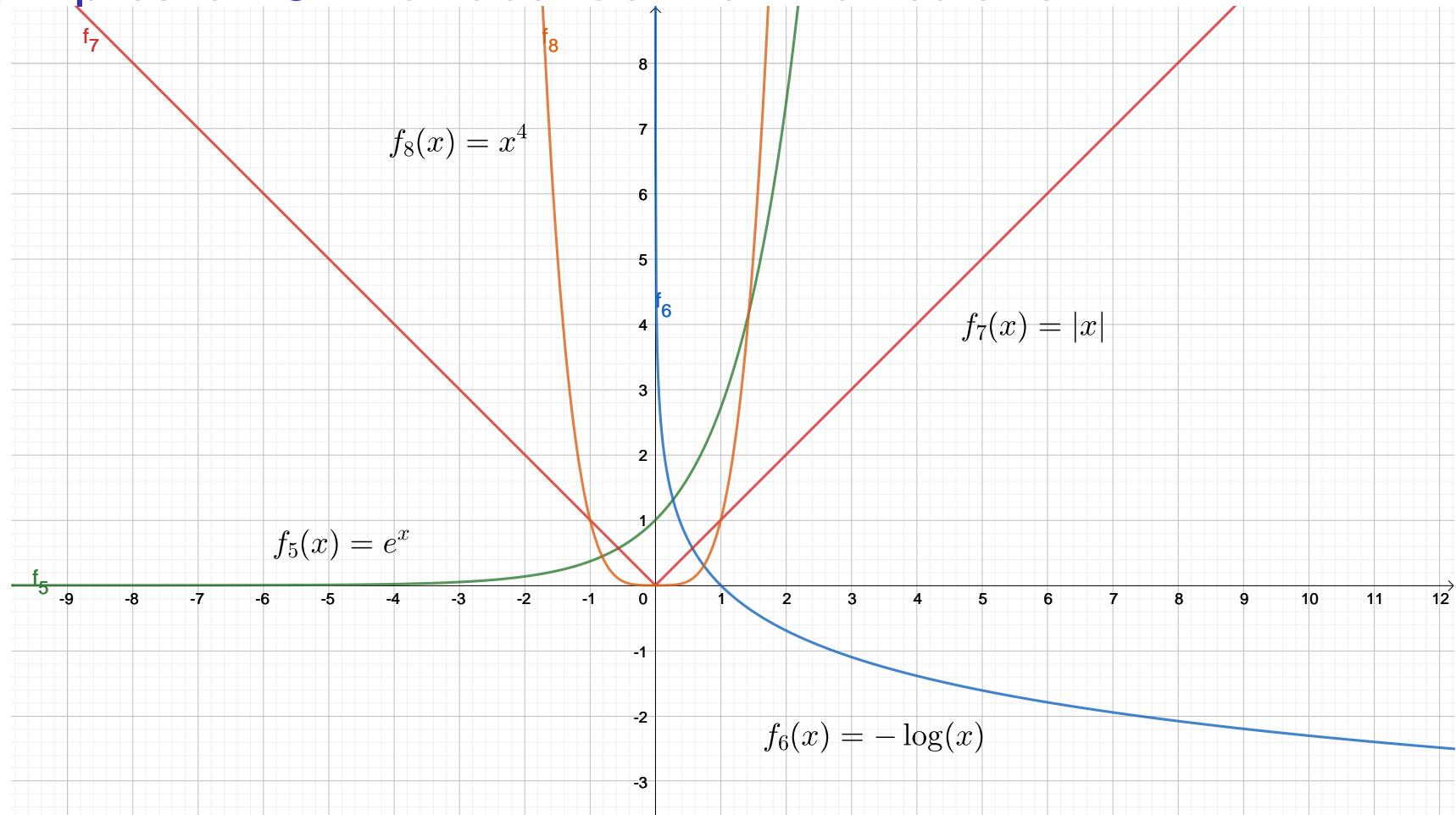


Figure: Examples of univariate convex functions.

Examples of Univariate Convex Functions IV

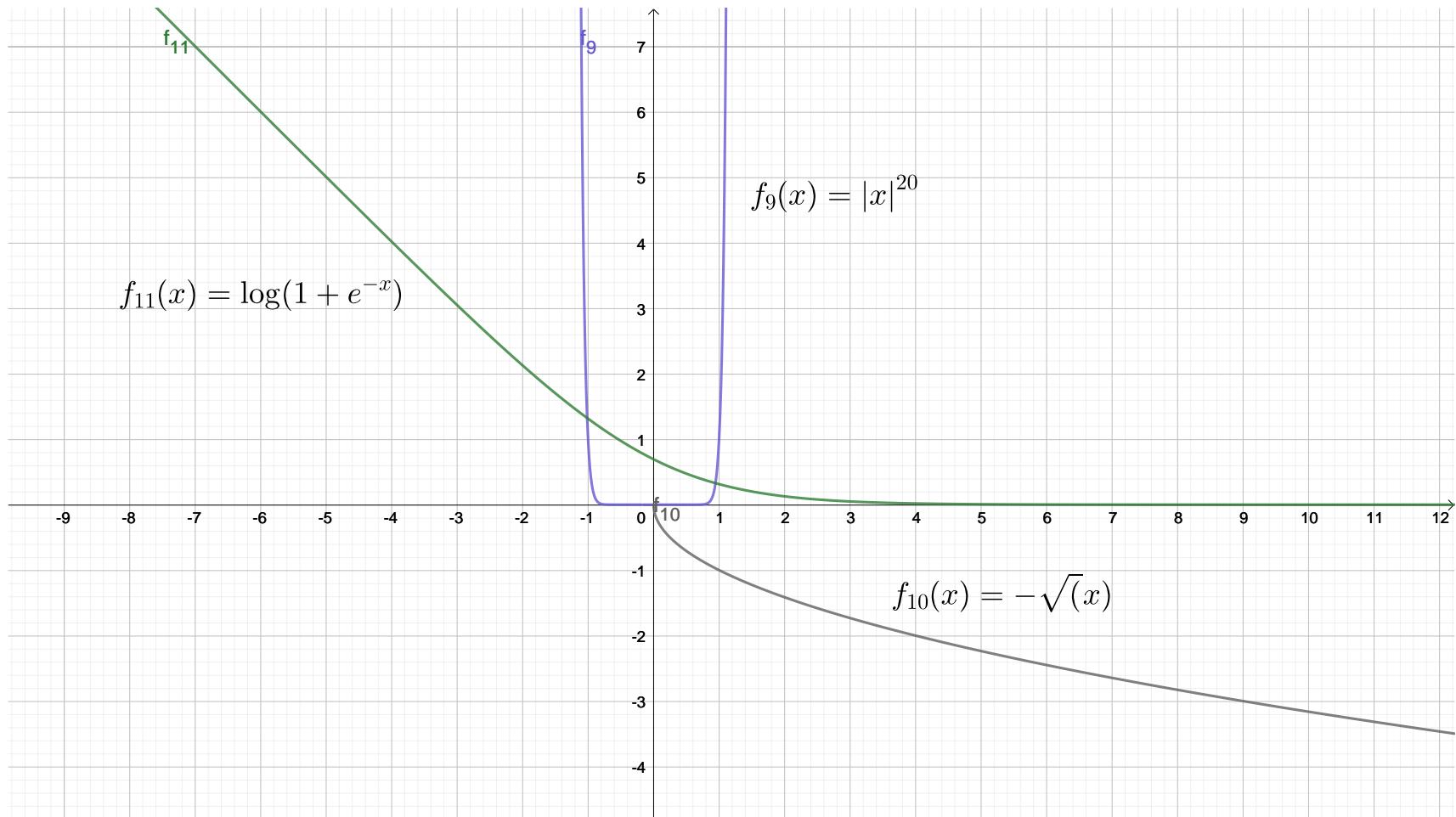


Figure: Examples of univariate convex functions.

Examples with Proofs I

Example 49 (Affine functions are convex)

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be defined as

$$f(x) = a^\top x + b, \quad (15)$$

where $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Choose $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$. Then

$$\begin{aligned} f(\lambda x + (1 - \lambda)y) &\stackrel{(15)}{=} a^\top(\lambda x + (1 - \lambda)y) + b \\ &= \lambda a^\top x + (1 - \lambda)a^\top y + b \\ &= \lambda(a^\top x + b) + (1 - \lambda)(a^\top y + b) \\ &\stackrel{(15)}{=} \lambda f(x) + (1 - \lambda)f(y). \end{aligned}$$

Examples with Proofs II

Example 50 (Norms are convex)

Let

$$f(x) = \|x\| \quad (16)$$

be a norm on \mathbb{R}^d . Choose $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$. Then by the triangle inequality and positive homogeneity of the norm, we get

$$\begin{aligned} f(\lambda x + (1 - \lambda)y) &\stackrel{(16)}{=} \|\lambda x + (1 - \lambda)y\| \\ &\stackrel{(5)}{\leq} \|\lambda x\| + \|(1 - \lambda)y\| \\ &\stackrel{(4)}{=} \lambda \|x\| + (1 - \lambda) \|y\| \\ &\stackrel{(16)}{=} \lambda f(x) + (1 - \lambda)f(y). \end{aligned}$$

Examples with Proofs III

Example 51 (Squared Euclidean norm is convex)

Let

$$f(x) = \|x\|^2 = x^\top x \quad (17)$$

be the standard Euclidean norm on \mathbb{R}^d . Then

$$\begin{aligned} f(\lambda x + (1 - \lambda)y) &\stackrel{(17)}{=} \|\lambda x + (1 - \lambda)y\|^2 \\ &\stackrel{(6)}{=} \|\lambda x\|^2 + 2\langle \lambda x, (1 - \lambda)y \rangle + \|(1 - \lambda)y\|^2 \\ &= \lambda^2 \|x\|^2 + 2\lambda(1 - \lambda)\langle x, y \rangle + (1 - \lambda)^2 \|y\|^2 \\ &= \lambda \|x\|^2 - \lambda(1 - \lambda) \|x\|^2 + 2\lambda(1 - \lambda)\langle x, y \rangle \\ &\quad + (1 - \lambda) \|y\|^2 - (1 - \lambda)(1 - (1 - \lambda)) \|y\|^2 \\ &= \lambda \|x\|^2 - \lambda(1 - \lambda) \underbrace{\left(\|x\|^2 - 2\langle x, y \rangle + \|y\|^2 \right)}_{\stackrel{(6)}{=} \|x - y\|^2} + (1 - \lambda) \|y\|^2 \\ &= \lambda \|x\|^2 - \lambda(1 - \lambda) \|x - y\|^2 + (1 - \lambda) \|y\|^2 \\ &\leq \lambda \|x\|^2 + (1 - \lambda) \|y\|^2 \stackrel{(17)}{=} \lambda f(x) + (1 - \lambda)f(y). \end{aligned}$$

Examples with Proofs IV

Example 52 (Convex quadratic in \mathbb{R}^d – approach 1)

Let

$$f(x) = \frac{1}{2}x^\top \mathbf{A}x,$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a symmetric matrix. Prove that f is convex if and only if \mathbf{A} is positive semidefinite.

Solution: Take $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$. Then

$$\begin{aligned} & \lambda f(x) + (1 - \lambda)f(y) - f(\lambda x + (1 - \lambda)y) \\ = & \lambda(1 - \lambda)(x^\top \mathbf{A}x + y^\top \mathbf{A}y - 2x^\top \mathbf{A}y) \\ = & \lambda(1 - \lambda)(x - y)^\top \mathbf{A}(x - y). \end{aligned}$$

The last expression is nonnegative for all $x, y \in \mathbb{R}^d$ if and only if \mathbf{A} is positive semidefinite. The first expression is nonnegative for all $x, y \in \mathbb{R}^d$ precisely when f is convex.

First Order Characterizations of Convex Functions

First-Order Characterization of Convex Functions I

Theorem 53 (First-order characterization of convex functions)

Let $f : \mathcal{S} \rightarrow \mathbb{R}$ be a continuously differentiable function, where $\mathcal{S} \subseteq \mathbb{R}^d$ is an open convex set. Then the following statements are equivalent:

- (i) f is convex on \mathcal{S} (convexity)
- (ii) $f(y) + \langle \nabla f(y), x - y \rangle \leq f(x)$ for all $x, y \in \mathcal{S}$ (nonnegativity of Bregman divergence)
- (iii) $0 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle$ for all $x, y \in \mathcal{S}$ (gradient monotonicity)

First-Order Characterization of Convex Functions II

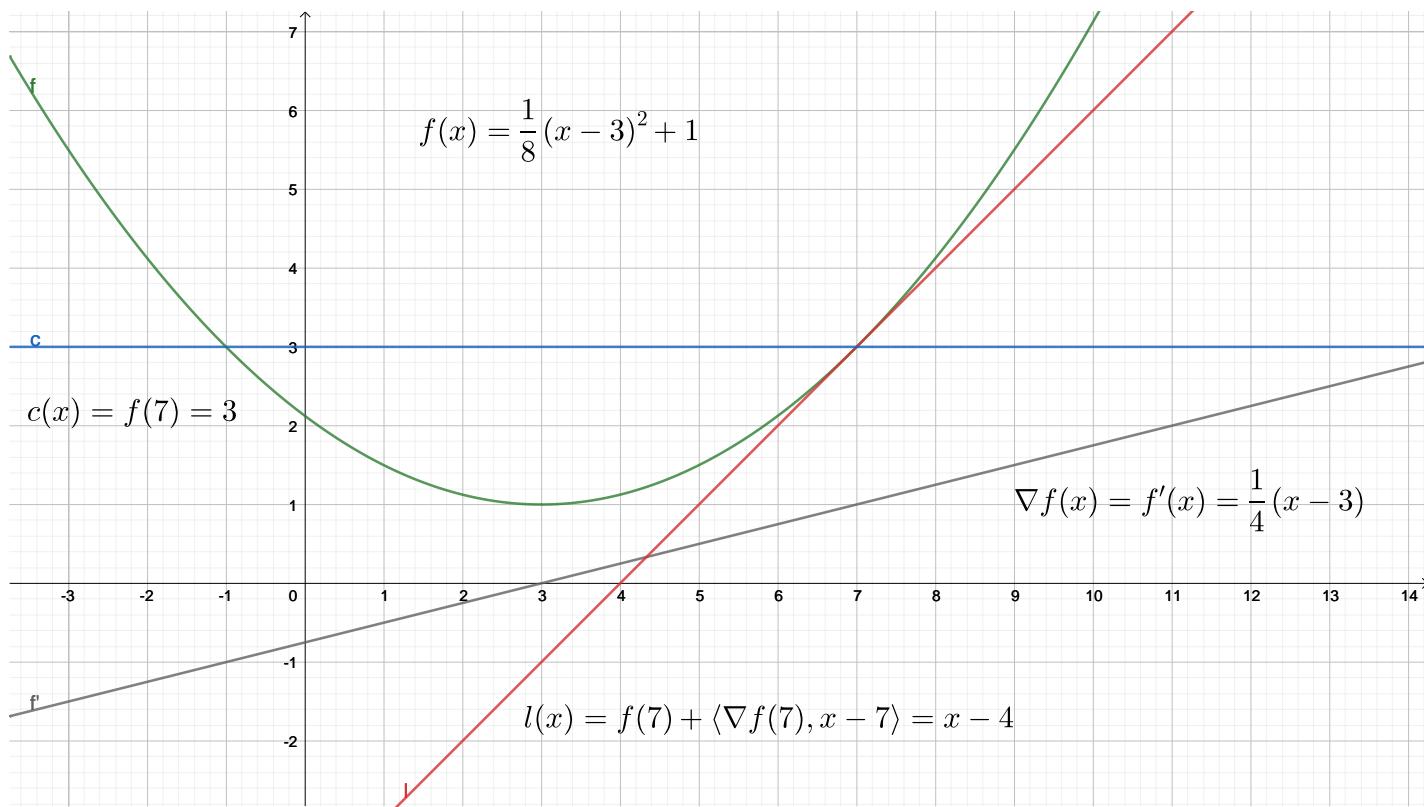


Figure: Convex function $f(x) = \frac{1}{8}(x - 3)^2 + 1$ and its constant ($c(x) = 3$) and linear ($l(x) = x - 4$) approximation at $y = 7$. Notice that the gradient $\nabla f(x) = f'(x) = \frac{1}{4}(x - 3)$ is an increasing function. See Theorem 53.

First-Order Characterization of Convex Functions III

Example 54 (Convex quadratic in \mathbb{R}^d – approach 2)

Let

$$f(x) = \frac{1}{2}x^\top \mathbf{A}x,$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a symmetric matrix. Prove that f is convex if and only if \mathbf{A} is positive semidefinite.

Solution: Clearly, f is continuously differentiable on $\mathcal{S} = \mathbb{R}^d$, which is a convex set. Moreover, its gradient is equal to $\nabla f(x) = \mathbf{A}x$. By Theorem 53, f is convex on \mathbb{R}^d if and only if

$$0 \leq \langle \mathbf{A}x - \mathbf{A}y, x - y \rangle \quad \forall x, y \in \mathcal{S} = \mathbb{R}^d. \quad (18)$$

Since $\langle \mathbf{A}x - \mathbf{A}y, x - y \rangle = (x - y)^\top \mathbf{A}(x - y)$ and $x, y \in \mathbb{R}^d$ are arbitrary, condition (18) is equivalent to requiring $z^\top \mathbf{A}z \geq 0$ for all $z \in \mathbb{R}^d$, which is equivalent to \mathbf{A} being positive semidefinite.

First-order Sufficient Condition for Optimality

Theorem 55 (First-order sufficient condition for optimality)

Let $f : \mathcal{S} \rightarrow \mathbb{R}$ be a continuously differentiable and convex function, where $\mathcal{S} \subseteq \mathbb{R}^d$ is a convex set.

- (i) If $\nabla f(x^*) = 0$ for some $x^* \in \mathcal{S}$, then x^* **minimizes f over \mathcal{S}** , i.e.,

$$f(x^*) \leq f(x) \quad \forall x \in \mathcal{S}.$$

- (ii) Assume $\mathcal{S} = \mathbb{R}^d$. Then $\nabla f(x^*) = 0$ if and only if x^* is a global minimizer of f on \mathbb{R}^d .

Proof.

We only prove (i). Choose $y = x^*$ in Theorem 53 (ii), we get

$$f(x^*) = f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \leq f(x) \quad \forall x \in \mathcal{S}.$$



Second Order Characterizations of Convex Functions

Second-Order Characterization of Convex Functions I

Theorem 56 (Second-order characterization of convex functions)

Let $f : \mathcal{S} \rightarrow \mathbb{R}$ be a twice continuously differentiable function, where $\mathcal{S} \subseteq \mathbb{R}^d$ is an open convex set. Then the following statements are equivalent:

(i) f is convex on \mathcal{S}

(convexity)

(ii) $\nabla^2 f(x) \succeq 0$ for all $x \in \mathcal{S}$

(positive semidefiniteness of the Hessian)

Example 57 (Exponential)

Show that the exponential function $f(x) = e^x$ is convex on \mathbb{R} .

Solution: First, f is infinitely times (and hence also twice) continuously differentiable on \mathbb{R} . It remains to notice that $f''(x) = e^x \geq 0$ for all $x \in \mathbb{R}$, and to apply Theorem 56.

Second-Order Characterization of Convex Functions II

Example 58 (Logistic loss)

Let

$$f(x) = \log(1 + e^{-x})$$

be the logistic loss function. Show that f is convex on \mathbb{R} .

Solution: By chain rule of differentiation, we have

$$f'(x) = -\frac{1}{1 + e^{-x}} e^{-x} = -\frac{1}{e^x + 1}$$

and

$$f''(x) = \frac{e^x}{(e^x + 1)^2} \geq 0$$

for all $x \in \mathbb{R}$. It remains to apply Theorem 56.

Second-Order Characterization of Convex Functions III

Example 59 (Convex quadratic in \mathbb{R}^d – approach 3)

Let

$$f(x) = \frac{1}{2}x^\top \mathbf{A}x,$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a symmetric matrix. Prove that f is convex if and only if \mathbf{A} is positive semidefinite.

Solution: Clearly, f is twice continuously differentiable on $\mathcal{S} = \mathbb{R}^d$, which is a convex set. Moreover, its Hessian is equal to $\nabla^2 f(x) = \mathbf{A}$. By Theorem 56, f is convex on \mathbb{R}^d if and only if \mathbf{A} is positive semidefinite.

Second-Order Characterization of Convex Functions IV

Example 60 (Squared norm)

Show that the function $f(x) = \|x\|^2$ is convex.

Solution: Notice that $f(x) = x^\top \mathbf{I} x$, where $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix. By Examples 54, f is convex if and only if $\nabla^2 f(x) \equiv 2\mathbf{I}$ is positive semidefinite. Since $2\mathbf{I}$ is indeed positive semidefinite, f is convex.

Exercise 61

Show that the function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = x^4$ is convex on \mathbb{R} .

Exercise 62

Show that the function $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ defined by $f(x) = -\sqrt{x}$ is convex on \mathbb{R}_+ .

Exercise 63 (☕☕)

Show that the function $f : \mathbb{R} \times \mathbb{R}_{++} \rightarrow \mathbb{R}$ defined by $f(x, y) = \frac{x^2}{y}$ is convex on $\mathbb{R} \times \mathbb{R}_{++}$.

Second-Order Characterization of Convex Functions V

Exercise 64 (☕☕)

Show that the function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by

$$f(x) = \log \left(\sum_{i=1}^d e^{x_i} \right)$$

is convex on \mathbb{R}^d .

Introduction to Optimization

Peter Richtárik



Lecture 6: Convex Functions - Part 2

Lecture Outline

- ▶ 4 operations preserving convexity
- ▶ Examples

Operations Preserving Convexity

Operation	Convex Function	Condition	Theorem
Multiplication by a nonnegative scalar	$\alpha f(x)$	f is convex; $\alpha \geq 0$	Theorem 65
Lifting	$g(x, y) = f(x)$	f is convex	Theorem 69
Sum	$f_1(x) + \dots + f_n(x)$	f_i are convex	Theorem 71
Maximum	$\max\{f_1(x), \dots, f_n(x)\}$	f_i are convex	Theorem 76
Pre-composition with an affine mapping	$f(\mathbf{A}x + b)$	f is convex	Theorem 80
Post-composition with a nondecreasing convex function	$\phi(f(x))$	f is convex; ϕ is nondecreasing and convex	Theorem 90
Partial minimization	$\min_y f(x, y)$	f is convex; minimum is finite	Theorem 93

Table: Operations preserving convexity.

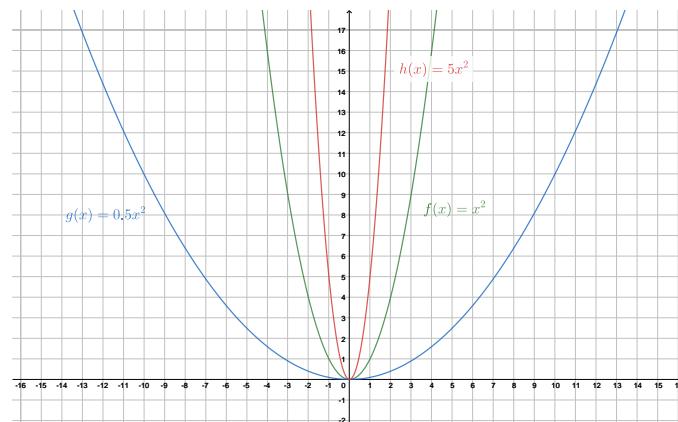
Operations Preserving Convexity: Multiplication by a Nonnegative Scalar I

Theorem 65 (Multiplication by a nonnegative scalar)

Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a convex function, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set. Then

$$g(x) = \alpha f(x) \quad (19)$$

is a convex function on \mathcal{X} provided that $\alpha \geq 0$.



Operations Preserving Convexity: Multiplication by a Nonnegative Scalar II

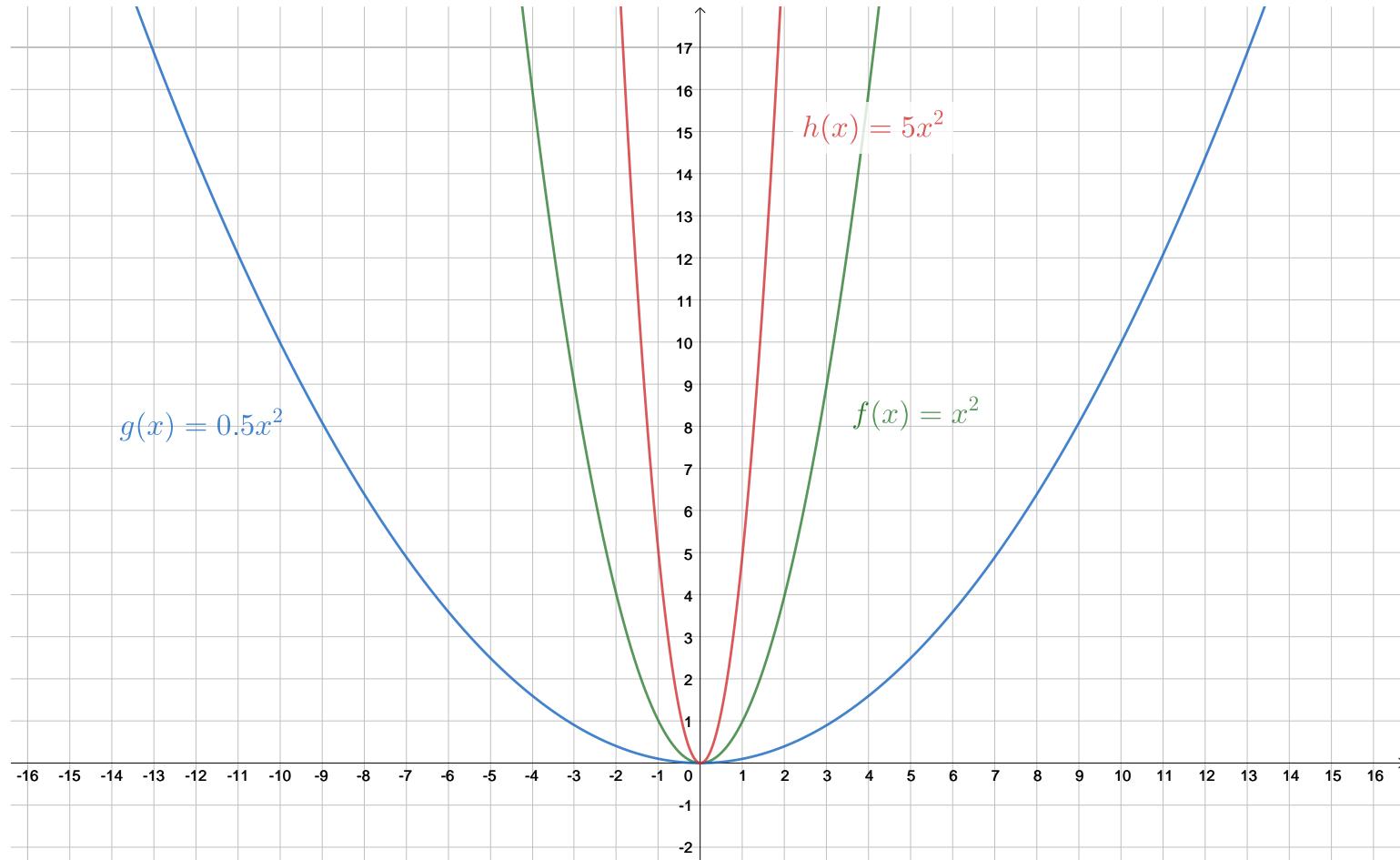


Figure: Illustration of Theorem 65: The function $f(x) = x^2$ is convex, and therefore also the scaled functions $g(x) = 0.5x^2$ and $h(x) = 5x^2$ are convex.

Operations Preserving Convexity: Multiplication by a Nonnegative Scalar III

Proof.

Let $x_1, x_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$. Then

$$\begin{aligned} g(\lambda x_1 + (1 - \lambda)x_2) &\stackrel{(19)}{=} \alpha f(\lambda x_1 + (1 - \lambda)x_2) \\ &\stackrel{(14)}{\leq} \alpha (\lambda f(x_1) + (1 - \lambda)f(x_2)) \\ &= \lambda \alpha f(x_1) + (1 - \lambda)\alpha f(x_2) \\ &\stackrel{(19)}{=} \lambda g(x_1) + (1 - \lambda)g(x_2). \end{aligned}$$

□

Operations Preserving Convexity: Multiplication by a Nonnegative Scalar IV

Example 66

Since $f(x) = \|x\|$ is convex, so is $g(x) = 30\|x\|$.

Example 67

Since $f(x) = \|x\|^2$ is convex, so is $g(x) = 70\|x\|^2$.

Example 68

Since $f(x) = e^x$ is convex, so is $g(x) = 500e^x$.

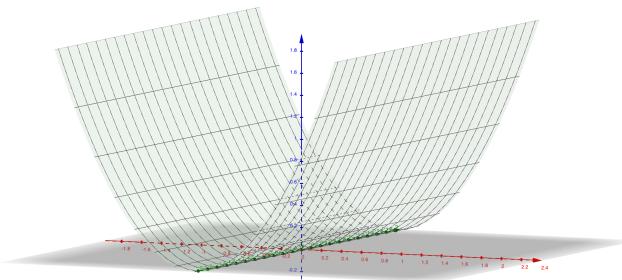
Operations Preserving Convexity: Lifting to a Higher Dimension I

Theorem 69 (Lifting)

Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a convex function, where $\mathcal{X} \subseteq \mathbb{R}^d$. Then the function $g : \mathcal{X} \times \mathbb{R}^n \rightarrow \mathbb{R}$ defined via

$$g(x, y) = f(x) \tag{20}$$

is convex on $\mathcal{X} \times \mathbb{R}^n$.



Example 70

The function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined via $f(x_1) = x_1^2$ is convex on $\mathcal{X} = \mathbb{R}$. So, the function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ defined via $g(x_1, \dots, x_d) = x_1^2$ is convex on \mathbb{R}^d .

Operations Preserving Convexity: Lifting to a Higher Dimension II

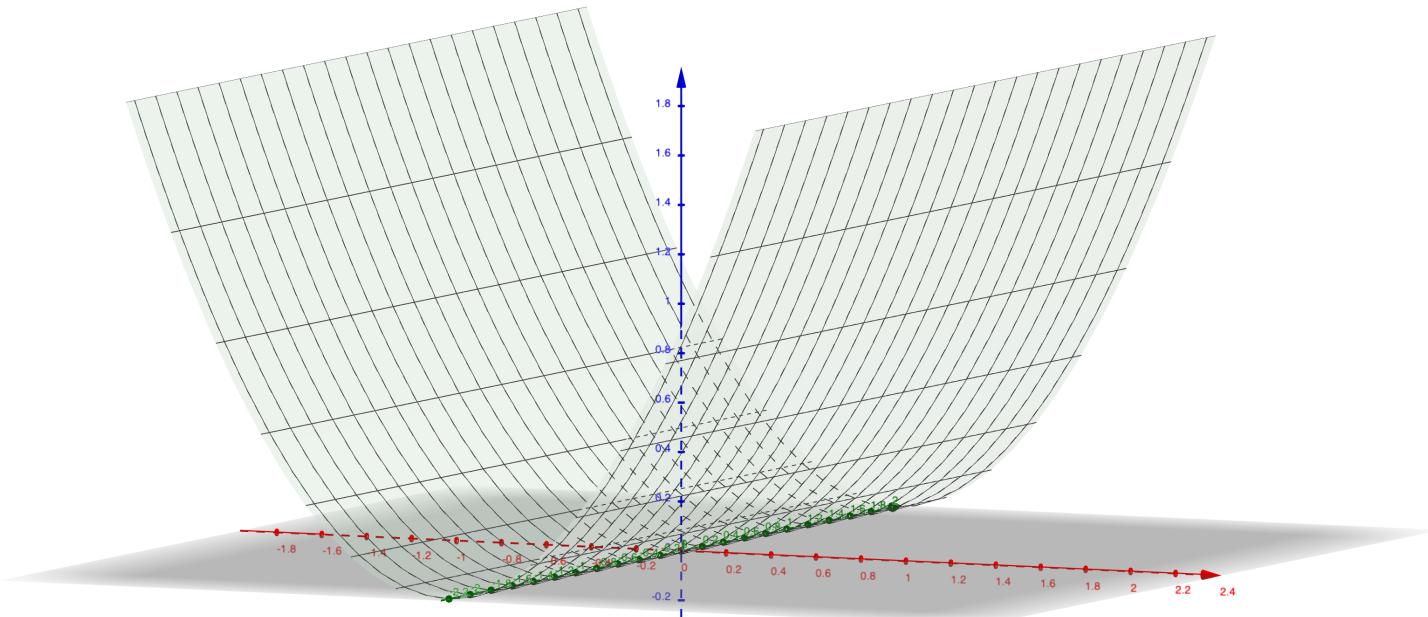


Figure: Illustration of Theorem 69 on Example 70: The function $f(x) = x^2$ (defined on \mathbb{R}) is convex, and therefore also the lifted function $g(x, y) = x^2 + y^2$ (defined on \mathbb{R}^2) is convex.

Operations Preserving Convexity: Lifting to a Higher Dimension III

Remarks:

- ▶ The lifting procedure allows us to “think”, if it is convenient to us, of any convex function defined on some smaller dimensional space \mathbb{R}^d as a function defined on a larger-dimensional space \mathbb{R}^{d+n} , with the extra n dimension being “dummy” in the sense that the function does not depend on them at all.

Operations Preserving Convexity: Lifting to a Higher Dimension IV

Proof.

First, notice that $\mathcal{X} \times \mathbb{R}^n$ is the Cartesian product of two convex sets, \mathcal{X} and \mathbb{R}^n , and hence by Theorem 28, it is a convex set as well. Now let $z_1 = (x_1, y_1) \in \mathcal{X} \times \mathbb{R}^n$, $z_2 = (x_2, y_2) \in \mathcal{X} \times \mathbb{R}^n$, and $\lambda \in [0, 1]$. Then

$$\begin{aligned} g(\lambda z_1 + (1 - \lambda) z_2) &= g(\lambda x_1 + (1 - \lambda) x_2, \lambda y_1 + (1 - \lambda) y_2) \\ &\stackrel{(20)}{=} f(\lambda x_1 + (1 - \lambda) x_2) \\ &\leq \lambda f(x_1) + (1 - \lambda) f(x_2) \\ &\stackrel{(20)}{=} \lambda g(x_1, y_1) + (1 - \lambda) g(x_2, y_2) \\ &= \lambda g(z_1) + (1 - \lambda) g(z_2). \end{aligned}$$



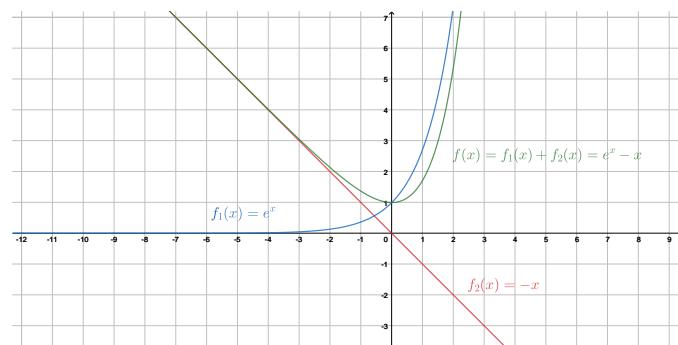
Operations Preserving Convexity: Sum I

Theorem 71 (Sum)

Let $f_1, \dots, f_n : \mathcal{X} \rightarrow \mathbb{R}$ be convex functions, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set. Then their sum

$$g(x) = \sum_{i=1}^n f_i(x) \quad (21)$$

is also a convex function on \mathcal{X} .



Operations Preserving Convexity: Sum II

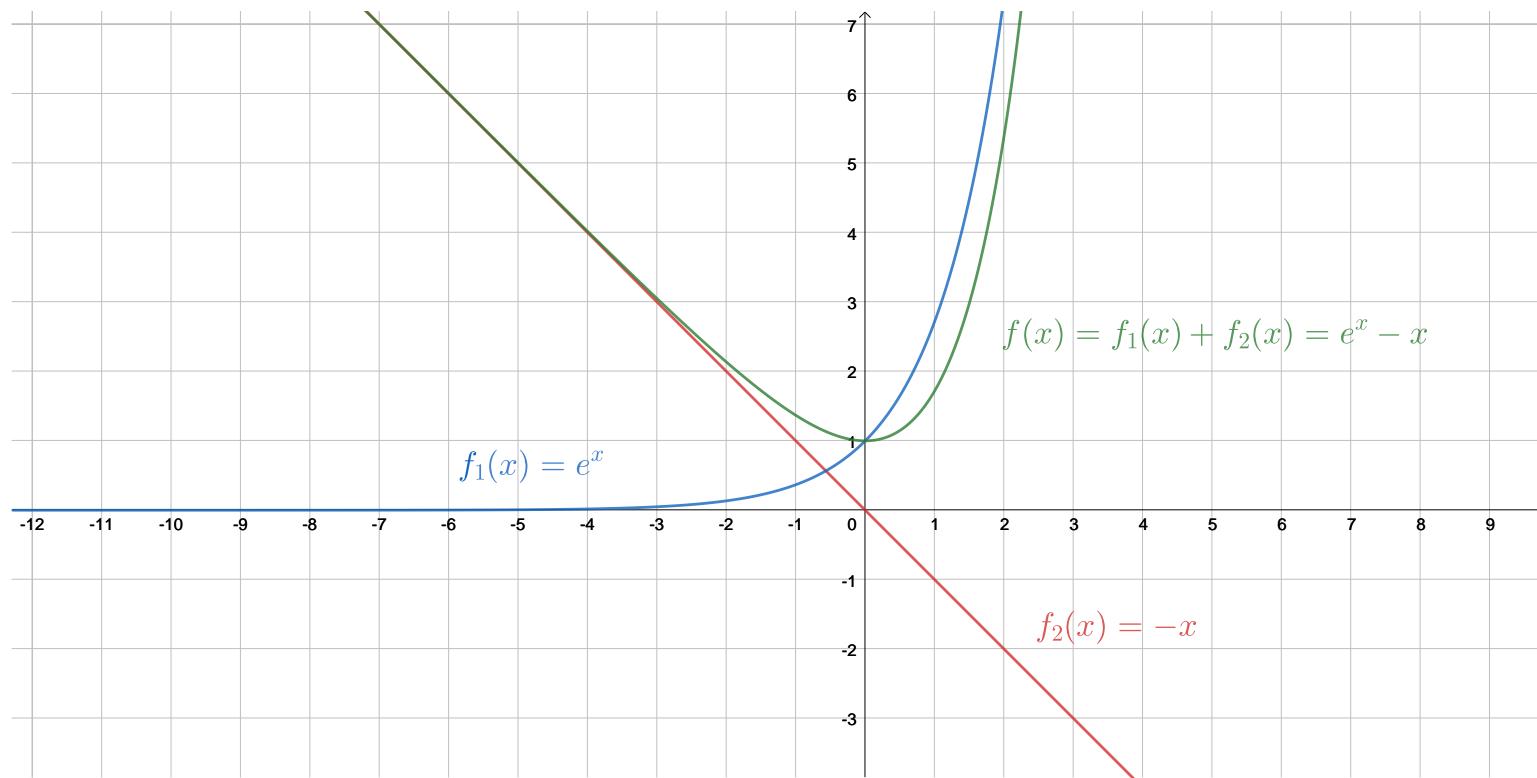


Figure: Illustration of Theorem 71: The functions $f_1(x) = e^x$ and $f_2(x) = -x$ are convex, and therefore also their sum $f(x) = e^x - x$ is convex.

Operations Preserving Convexity: Sum III

Proof.

Let $x_1, x_2 \in \mathcal{X}$ and $\lambda \in [0, 1]$. Since each f_i is convex, we know that

$$f_i(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f_i(x_1) + (1 - \lambda)f_i(x_2). \quad (22)$$

Summing up these inequalities, we get

$$\begin{aligned} g(\lambda x_1 + (1 - \lambda)x_2) &\stackrel{(21)}{=} \sum_{i=1}^n f_i(\lambda x_1 + (1 - \lambda)x_2) \\ &\stackrel{(22)}{\leq} \sum_{i=1}^n (\lambda f_i(x_1) + (1 - \lambda)f_i(x_2)) \\ &= \lambda \sum_{i=1}^n f_i(x_1) + (1 - \lambda) \sum_{i=1}^n f_i(x_2) \\ &\stackrel{(21)}{=} \lambda g(x_1) + (1 - \lambda)g(x_2). \end{aligned}$$



Operations Preserving Convexity: Sum IV

Example 72 (Affine functions from \mathbb{R}^d to \mathbb{R} are convex)

Let $a = (a_1, \dots, a_d) \in \mathbb{R}^d$ and $b \in \mathbb{R}$, and define

$$f(x) = a^\top x - b.$$

This is an affine function from \mathbb{R}^d to \mathbb{R} . Since $f(x) = \sum_{i=0}^d f_i(x)$, where $f_0(x) = -b$ and $f_i(x) = a_i x_i$ for $i = 1, 2, \dots, d$ are constant or linear functions (and hence convex), we conclude that f is convex as well.

Example 73 (Sum of nonconvex functions can be convex or nonconvex)

Let $f_1(x) = \sin(x)$ and $f_2(x) = -\sin(x)$. These are nonconvex functions. However, their sum is 0, which is convex. On the other hand, let $f_1(x) = \sin(x)$ and $f_2(x) = \sin(x)$. These are nonconvex functions, and their sum, $2\sin(x)$, is also nonconvex.

Operations Preserving Convexity: Sum \vee

Exercise 74 (☕☕)

Prove that a conic combination of a finite number of convex functions is a convex function. That is, provided that $f_i : \mathcal{X} \rightarrow \mathbb{R}$ are convex functions for $i = 1, \dots, n$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set, and $\alpha_1, \dots, \alpha_n \geq 0$, show that $f(x) = \sum_{i=1}^n \alpha_i f_i(x)$ is a convex function on \mathcal{X} .

Exercise 75 (☕)

Show that $f(x) = -\log \left(\prod_{i=1}^d x_i \right)$ is convex on \mathbb{R}_{++}^d .

Operations Preserving Convexity: Maximum

Theorem 76 (Maximum)

Let $f_1, \dots, f_n : \mathcal{X} \rightarrow \mathbb{R}$ be convex functions, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set. Then their maximum

$$g(x) = \max\{f_1(x), \dots, f_n(x)\}$$

is also a convex function on \mathcal{X} .

Example 77 (Maximum function)

The function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by $f(x) = \max\{x_1, \dots, x_d\}$ is convex on \mathbb{R}^d .

Example 78 (Infinity norm)

The infinity norm, $\|x\|_\infty \stackrel{\text{def}}{=} \max\{|x_1|, \dots, |x_d|\}$, is convex on \mathbb{R}^d .

Exercise 79 (☕☕)

Prove Theorem 76.

Introduction to Optimization

Peter Richtárik



Lecture 7: Convex Functions - Part 3

Lecture Outline

- ▶ 3 more operations preserving convexity
- ▶ Examples

Operations Preserving Convexity

Operation	Convex Function	Condition	Theorem
Multiplication by a nonnegative scalar	$\alpha f(x)$	f is convex; $\alpha \geq 0$	Theorem 65
Lifting	$g(x, y) = f(x)$	f is convex	Theorem 69
Sum	$f_1(x) + \dots + f_n(x)$	f_i are convex	Theorem 71
Maximum	$\max\{f_1(x), \dots, f_n(x)\}$	f_i are convex	Theorem 76
Pre-composition with an affine mapping	$f(\mathbf{A}x + b)$	f is convex	Theorem 80
Post-composition with a nondecreasing convex function	$\phi(f(x))$	f is convex; ϕ is nondecreasing and convex	Theorem 90
Partial minimization	$\min_y f(x, y)$	f is convex; minimum is finite	Theorem 93

Table: Operations preserving convexity.

Operations Preserving Convexity: Pre-Composition with an Affine Mapping I

Theorem 80 (Pre-composition with an affine mapping)

Let $f : \mathcal{Y} \rightarrow \mathbb{R}$ be a convex function, where $\mathcal{Y} \subseteq \mathbb{R}^n$ is a convex set. Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$. Then the function

$$g(x) = f(\mathbf{A}x + b)$$

is a convex function on $\mathcal{X} = \{x \in \mathbb{R}^d : \mathbf{A}x + b \in \mathcal{Y}\} = \mathbf{A}^{-1}(\mathcal{Y} - b)$.

Example 81 (Least squares)

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$. Show that the least-squares function $g(x) = \frac{1}{2n} \|\mathbf{A}x - b\|^2$ is convex on \mathbb{R}^d .

Solution: Since g is a positive scalar multiple of $\|\mathbf{A}x - b\|^2$, by Theorem 65 it is enough to show that $x \mapsto \|\mathbf{A}x - b\|^2$ is convex. However, this follows from Theorem 80 since $f(x) = \|x\|^2$ is convex.

Operations Preserving Convexity: Pre-Composition with an Affine Mapping II

Exercise 82 (Distance to a point ☕)

Let $a \in \mathbb{R}^d$. Show that the function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by $g(x) = \|x - a\|$ is convex.

Exercise 83 (☕)

Show that the function $g(x) = \|\mathbf{A}x\|$ is convex.

Exercise 84 (Distance between two vectors ☕)

Prove that the function $f : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ defined by $f(x, y) = \|x - y\|$ is convex.

Exercise 85

Is the function $g : \mathbb{R}^3 \rightarrow \mathbb{R}$ defined by $g(x) = \max\{x_1, 5x_2 - 3, x_3 - x_1, 8\}$ convex?

Exercise 86

Is the function $g : \mathbb{R}^3 \rightarrow \mathbb{R}$ defined by $g(x) = \max\{x_1, 2x_2^2, 3x_3^3\}$ convex?



Operations Preserving Convexity: Pre-Composition with an Affine Mapping III

Exercise 87 (Convexity of the cost function in binary classification via logistic regression ☕☕)

The logistic regression problem tries to classify a vector $a \in \mathbb{R}^d$ into one of two classes, $\{-1, +1\}$, via a linear model represented by a vector $x \in \mathbb{R}^d$. This is done as follows. One first collects n training examples of vectors $a_i \in \mathbb{R}^d$ along with their class labels $y_i \in \{-1, +1\}$, for $i = 1, 2, \dots, n$. Next, the following optimization problem is solved to find the optimal parameters x of the linear model:

$$\min_{x \in \mathbb{R}^d} \left[f(x) = \frac{1}{n} \sum_{i=1}^n \log \left(1 + e^{-y_i \langle a_i, x \rangle} \right) \right].$$

Show that the cost function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ defined this way is convex.

Operations Preserving Convexity: Pre-Composition with an Affine Mapping IV

Exercise 88 (☕)

Let $a_1, \dots, a_n \in \mathbb{R}^d$ and $b \in \mathbb{R}^n$. Show that the function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by $g(x) = \max\{a_1^\top x - b_1, \dots, a_n^\top x - b_n\}$ is convex. Do this in three different ways:

- (i) Using the definition of convex functions.
- (ii) Using Theorem 76 and Exercise 72.
- (iii) Using Theorem 76 and Theorem 80.

Exercise 89 (☕☕)

Prove Theorem 80.

Operations Preserving Convexity: Post-Composition with a Non-Decreasing Convex Function I

Theorem 90 (Post-composition with a nondecreasing convex function)

Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a convex function, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set.

Further, let $\phi : \mathcal{Y} \rightarrow \mathbb{R}$ be a non-decreasing convex function, where $\mathcal{Y} \subseteq \mathbb{R}$ is convex. Assume that $f(x) \in \mathcal{Y}$ for all $x \in \mathcal{X}$. Then the function

$$g(x) = \phi(f(x))$$

is convex on \mathcal{X} .

Example 91

If $f : \mathcal{X} \rightarrow \mathbb{R}$ is convex, then $g(x) = e^{f(x)}$ is convex on \mathcal{X} because $g(x) = \phi(f(x))$, where $\phi(t) = e^t$ is convex and increasing.

Exercise 92 (☕☕)

Prove Theorem 90.

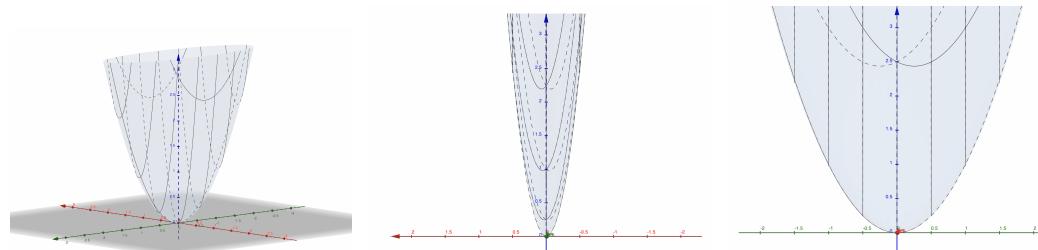
Operations Preserving Convexity: Partial Minimization I

Theorem 93 (Partial minimization)

Let $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ be a convex function, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^n$ are convex sets. Then the functions

$$g(x) \stackrel{\text{def}}{=} \min_{y \in \mathcal{Y}} f(x, y), \quad h(y) \stackrel{\text{def}}{=} \min_{x \in \mathcal{X}} f(x, y)$$

are convex on \mathcal{X} and \mathcal{Y} , respectively, provided that the minima are finite.



Exercise 94 (☕☕)

Prove Theorem 93.

Operations Preserving Convexity: Partial Minimization II

Example 95

Consider the function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ given by

$$f(x, y) = 10x^2 + y^2 - xy.$$

One can verify that $\nabla^2 f(x, y) = \begin{pmatrix} 10 & -1 \\ -1 & 1 \end{pmatrix}$ is a positive semidefinite matrix, and hence f is convex by Theorem 56. It can be calculated that

$$g(x) \stackrel{\text{def}}{=} \min_{y \in \mathbb{R}} f(x, y) = \frac{19}{2}x^2,$$

$$h(y) \stackrel{\text{def}}{=} \min_{x \in \mathbb{R}} f(x, y) = \frac{39}{40}y^2.$$

Clearly, these are both convex functions, as Theorem 93 predicts.

Moreover, the theorem has a nice geometrical interpretation: we “see” these convex functions when looking at the graph of f in the direction of the coordinate axes! See the figure below!

Operations Preserving Convexity: Partial Minimization III

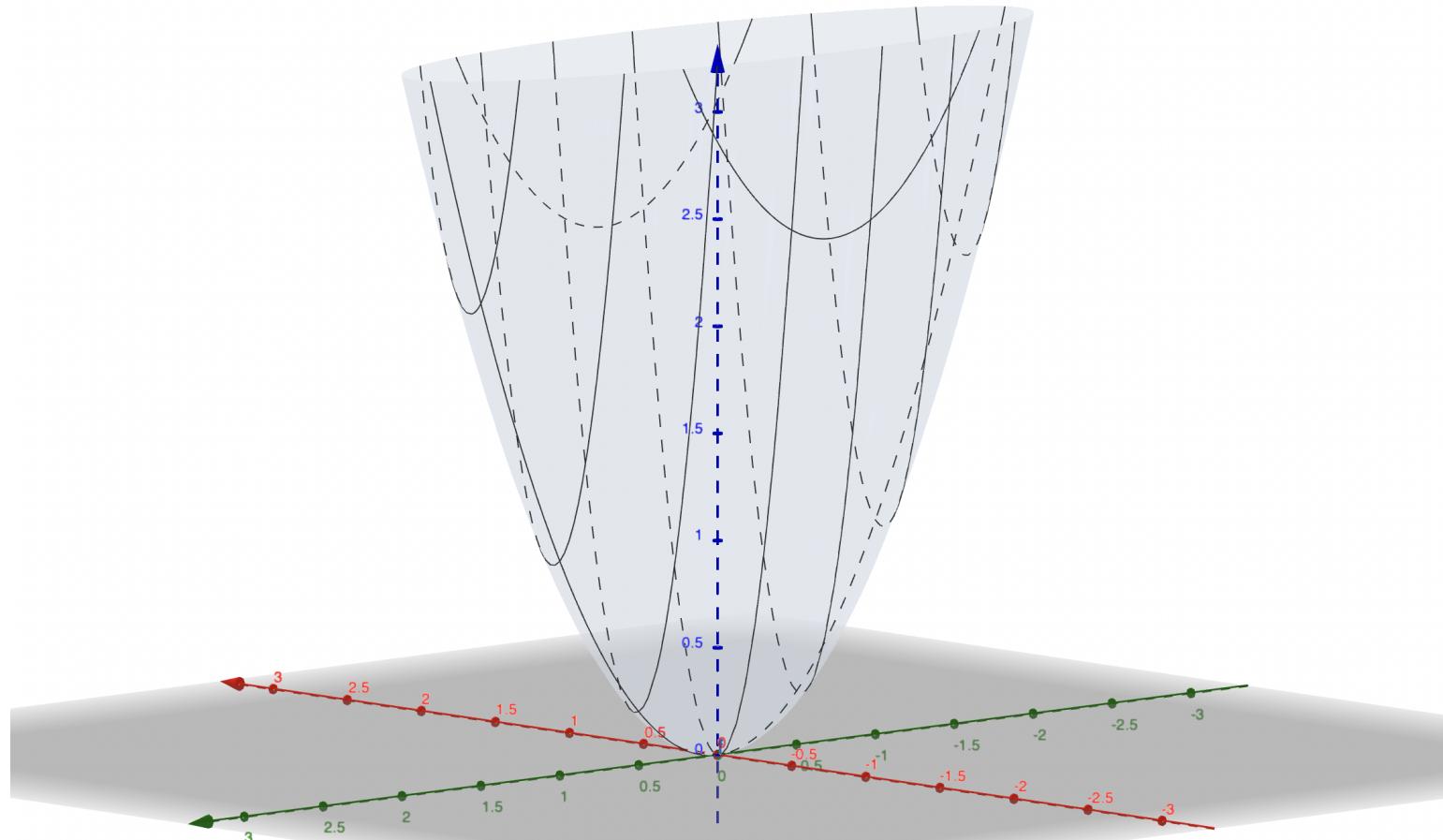


Figure: Convex function $f(x, y) = 10x^2 + y^2 - xy$ on \mathbb{R}^2 .

Operations Preserving Convexity: Partial Minimization IV

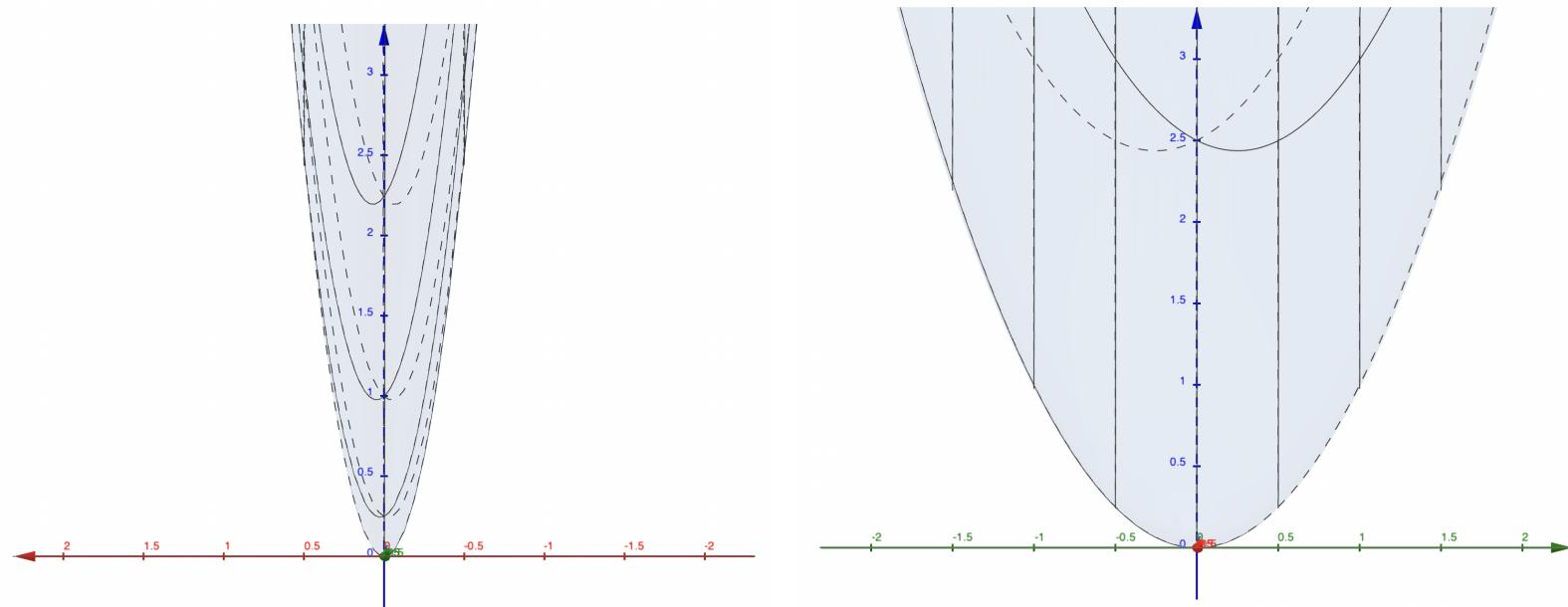


Figure: The two functions obtained by partial minimization:
 $g(x) = \min_{y \in \mathbb{R}} f(x, y) = \frac{19}{2}x^2$ (left) and $h(y) = \min_{x \in \mathbb{R}} f(x, y) = \frac{39}{40}y^2$ (right).

Operations Preserving Convexity: Partial Minimization V

Remarks:

- ▶ It is easy to see that

$$\min_{x \in \mathcal{X}, y \in \mathcal{Y}} f(x, y) = \min_{x \in \mathcal{X}} \left(\min_{y \in \mathcal{Y}} f(x, y) \right) = \min_{x \in \mathcal{X}} g(x)$$

- ▶ So, if we want to minimize a convex function of the form $f(x, y)$, we can first minimize it over $y \in \mathcal{Y}$, obtaining function g (which in view of Theorem 93 turns out to be convex!), and then minimize function g over $x \in \mathcal{X}$.

Operations Preserving Convexity: Partial Minimization VI

- ▶ Note that **this is not the same** as first minimizing $f(x, y)$ over $y \in \mathcal{Y}$ for some **fixed** $\bar{x} \in \mathcal{X}$, thus obtaining

$$y^* = \arg \min_{y \in \mathcal{Y}} f(\bar{x}, y),$$

and then minimizing in x with **fixed** $y = y^*$, thus obtaining

$$x^* = \arg \min_{x \in \mathcal{X}} f(x, y^*).$$

That is to say, the above procedure will **not** in general lead to (x^*, y^*) being the minimizer of $f(x, y)$. These are the first two steps of a method called **Alternating Minimization** or **Block Coordinate Minimization**.

Operations Preserving Convexity: Partial Minimization VII

Example 96 (Separable case)

Let

$$f(x, y) = f_1(x) + f_2(y), \quad (23)$$

where $f_1 : \mathcal{X} \rightarrow \mathbb{R}$ and $f_2 : \mathcal{Y} \rightarrow \mathbb{R}$ are convex functions and $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y} \subseteq \mathbb{R}^n$ are convex sets.

- ▶ We will first argue that f is convex on $\mathcal{X} \times \mathcal{Y}$. Indeed, in view of the “lifting” result (Theorem 69), f_1 is convex on $\mathcal{X} \times \mathbb{R}^n$, f_2 are convex on $\mathbb{R}^d \times \mathcal{Y}$, which means that both f_1 and f_2 are convex on $\mathcal{X} \times \mathcal{Y}$. Since f is the sum of two convex functions, it is convex on $\mathcal{X} \times \mathcal{Y}$ (see Theorem 71).
- ▶ Further, we have

$$g(x) = \min_{y \in \mathcal{Y}} f(x, y) \stackrel{(23)}{=} \min_{y \in \mathcal{Y}} (f_1(x) + f_2(y)) = f_1(x) + \min_{y \in \mathcal{Y}} f_2(y).$$

Since g is the sum of the convex function f_1 and the constant $\min_{y \in \mathcal{Y}} f_2(y)$, it is convex (see Theorem 71).

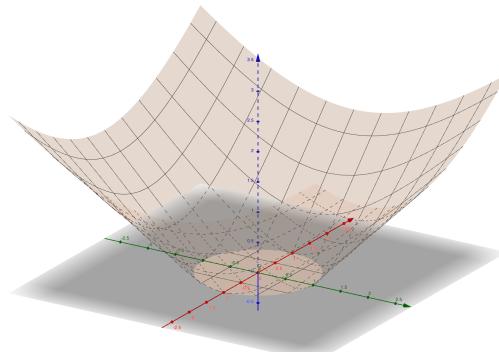
- ▶ So, the assumptions and the claim of Theorem 71 hold in this separable case.

Operations Preserving Convexity: Partial Minimization VIII

Example 97 (Distance to a convex set)

Let $\emptyset \neq \mathcal{Y} \subseteq \mathbb{R}^d$ and let

$$\text{dist}(x, \mathcal{Y}) \stackrel{\text{def}}{=} \min \{ \|x - y\| : y \in \mathcal{Y}\}.$$



This distance function is convex since

$$f(x, y) = \|x - y\|$$

is convex on $\mathbb{R}^d \times \mathbb{R}^d$ (see Example 84) and $\text{dist}(x, \mathcal{Y}) \geq 0$, and we can thus apply Theorem 93.

Operations Preserving Convexity: Partial Minimization IX

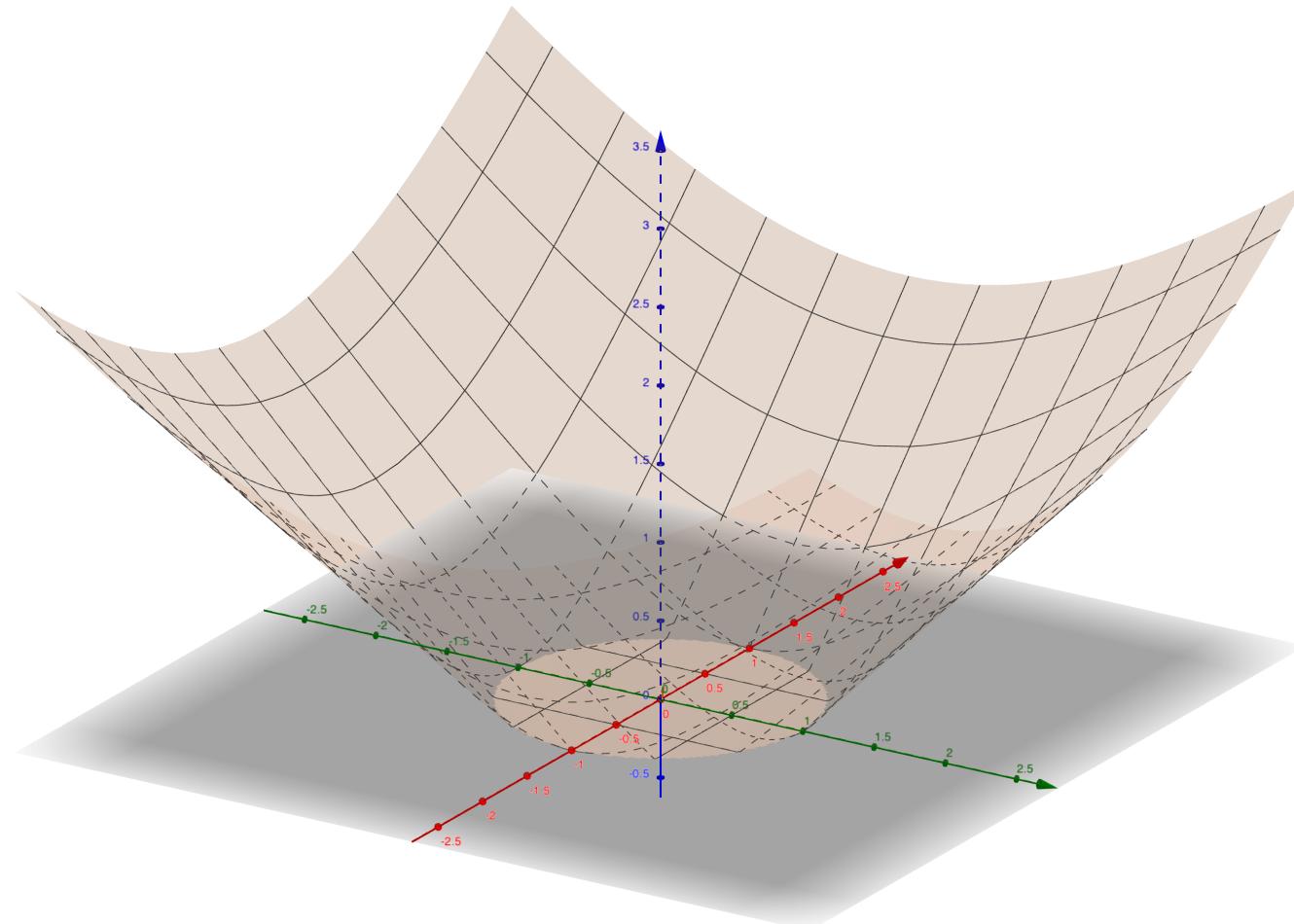


Figure: The graph of the function $\text{dist}(x, \mathcal{Y})$ for $\mathcal{Y} = \mathcal{B}(0, 1) \subseteq \mathbb{R}^2$. That is, the “distance-to-the-unit-2d-ball” function.

Operations Preserving Convexity: Partial Minimization X

Example 98

Let $f(x, y) = x^2 + y^2 - 2xy$. This function is convex on \mathbb{R}^2 (why?). We have

$$g(x) = \min_{y \in \mathbb{R}} f(x, y) = \min_{y \in \mathbb{R}} (x^2 + y^2 - 2xy) = x^2 + \min_{y \in \mathbb{R}} \phi(y), \quad (24)$$

where $\phi(y) = y^2 - 2xy$. Since ϕ is a convex quadratic function in y , its minimum is attained for y^* which satisfies $\phi'(y^*) = 0$, i.e., $2y^* - 2x = 0$. So, $y^* = x$ and therefore, $\phi(y^*) = x^2 - 2x^2 = -x^2$. By plugging this into (24), we get

$$g(x) \stackrel{(24)}{=} x^2 + \phi(y^*) = x^2 - x^2 = 0,$$

which is a constant function, and hence obviously convex.

Example 99

There is a simpler way to solve Example 98. Can you find it?

Operations Preserving Convexity: Partial Minimization XI

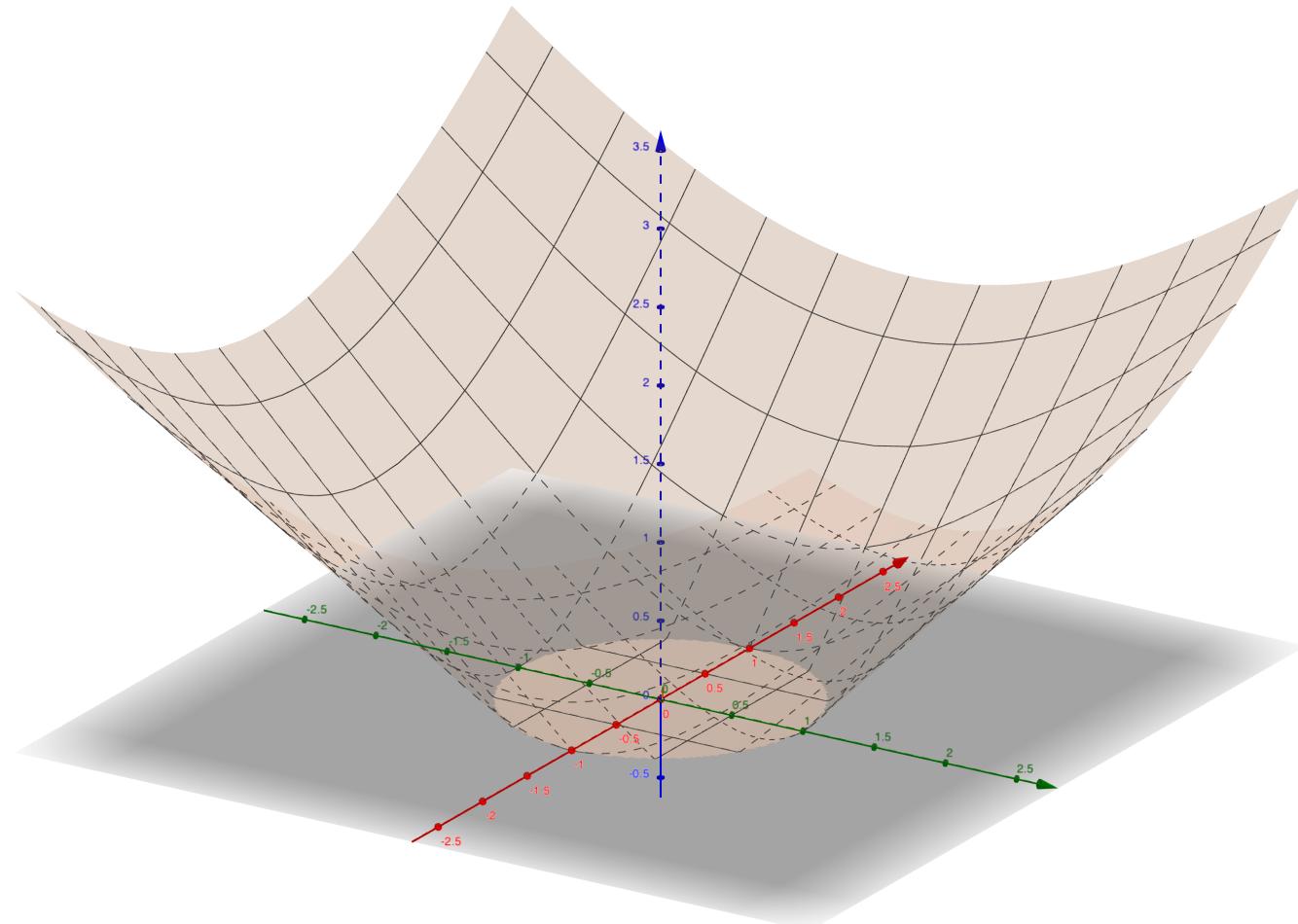


Figure: The graph of the function $\text{dist}(x, \mathcal{Y})$ for $\mathcal{Y} = \mathcal{B}(0, 1) \subseteq \mathbb{R}^2$. That is, the “distance-to-the-unit-2d-ball” function.

Introduction to Optimization

Peter Richtárik



Lecture 8: Convex Functions - Part 4

Lecture Outline

- ▶ Continuity of convex functions
- ▶ Differentiability of convex functions
- ▶ Level sets of convex functions
- ▶ Extended real-valued functions

Continuity of Convex Functions

Continuity of Convex Functions I

As the next example shows, convex functions are nor necessarily continuous everywhere where they are defined.

Example 100 (Convex functions are not necessarily continuous everywhere)

The function $f : [0, +\infty) \rightarrow \mathbb{R}$ defined via

$$f(x) = \begin{cases} 1 & \text{for } x = 0 \\ 0 & \text{for } x > 0 \end{cases} \quad (25)$$

is convex, but not continuous at $x = 0$.



Continuity of Convex Functions II

However, note that f defined in (25) **is continuous on $(0, +\infty)$, which is the interior of $[0, +\infty)$:**

$$\text{int}([0, +\infty)) = (0, +\infty).$$

This is not a coincidence. The next result states that convex functions are continuous in the interior of their domain (of definition).

Theorem 101

If $f : \mathcal{S} \rightarrow \mathbb{R}$ is convex, where $\mathcal{S} \subseteq \mathbb{R}^d$ is a convex set, then f is continuous on $\text{int}(\mathcal{S})$.

Remark: In fact, a stronger property holds: **convex functions are locally Lipschitz continuous in the interior of their domain.**

Differentiability of Convex Functions

Differentiability of Convex Functions I

Convex functions are not necessarily differentiable.

Example 102 (Absolute value)

The function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = |x|$ is convex, but is not differentiable at $x = 0$. It is differentiable everywhere else though:

$$f'(x) = \begin{cases} 1 & \text{for } x > 0 \\ -1 & \text{for } x < 0 \end{cases}.$$

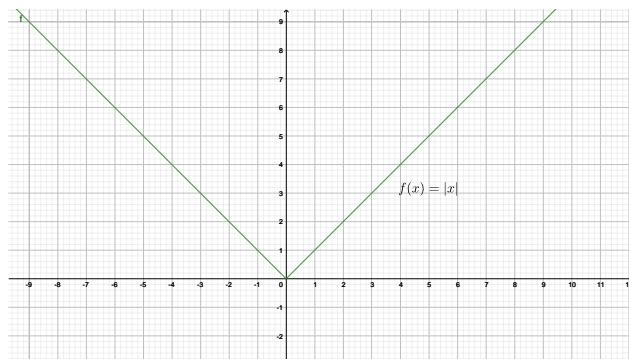


Figure: Convex function $f(x) = |x|$.

Differentiability of Convex Functions II

However, note that despite the fact $f(x) = |x|$ is not differentiable at $x = 0$, the left and right derivatives exist at $x = 0$:

- ▶ The **right derivative** is the **derivative in the direction of the vector $s = +1$** :

$$f'(0, s) \stackrel{\text{def}}{=} \lim_{t \rightarrow 0^+} \frac{f(0 + ts) - f(0)}{t} = \lim_{t \rightarrow 0^+} \frac{t - 0}{t} = 1.$$

- ▶ The **left derivative** is the **derivative in the direction of the vector $s = -1$** :

$$f'(0, s) \stackrel{\text{def}}{=} \lim_{t \rightarrow 0^+} \frac{f(0 + ts) - f(0)}{t} = \lim_{t \rightarrow 0^+} \frac{-t - 0}{t} = -1.$$

Differentiability of Convex Functions III

This is not a coincidence. Indeed, as the next result shows, convex functions have directional derivatives in any direction at points x belonging to the interior of their domains.

Theorem 103 (Existence of directional derivatives)

Let $f : \mathcal{S} \rightarrow \mathbb{R}$ be a convex function, where $\mathcal{S} \subseteq \mathbb{R}^d$ is a convex set. Then for any $x \in \text{int}(\mathcal{S})$ and any $0 \neq s \in \mathbb{R}^d$, the directional derivative of f in direction s at x , defined via

$$f'(x, s) \stackrel{\text{def}}{=} \lim_{t \rightarrow 0^+} \frac{f(x + ts) - f(x)}{t}, \quad (26)$$

exists.

Differentiability of Convex Functions IV

Example 104

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be given by

$$f(x) = \|x\|.$$

We know this is a convex function. We will now calculate the directional derivative of f in the direction $0 \neq s \in \mathbb{R}^d$ at $x = 0$ (note that f is not differentiable at $x = 0$):

$$\begin{aligned} f'(x, s) &\stackrel{(26)}{=} \lim_{t \rightarrow 0^+} \frac{f(x + ts) - f(x)}{t} \\ &= \lim_{t \rightarrow 0^+} \frac{\|0 + ts\| - \|0\|}{t} \\ &= \lim_{t \rightarrow 0^+} \frac{t \|s\|}{t} \\ &= \|s\|. \end{aligned}$$

Differentiability of Convex Functions V

Exercise 105

Find the directional derivatives of the convex function

$$f(x) = \max\{-5x + 3, 2x + 1\}.$$

It will help to first plot the graph of f .

Exercise 106 (☕)

Find a formula for the directional derivatives at $x = 0$ of the convex function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$f(x) = \max\{a_1x, \dots, a_nx\},$$

where $a_1, \dots, a_n \in \mathbb{R}$.

Differentiability of Convex Functions VI

Exercise 107 (☕☕)

Find a formula for the directional derivatives at $x = 0$ of the convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ defined by

$$f(x) = \max\{a_1^\top x, \dots, a_n^\top x\},$$

where $a_1, \dots, a_n \in \mathbb{R}^d$.

Level Sets of Convex Functions

Level Sets of Convex Functions are Convex Sets I

Theorem 108 (Level sets of convex functions)

Let $f : \mathcal{S} \rightarrow \mathbb{R}$ be a convex function, where $\mathcal{S} \subseteq \mathbb{R}^d$ is a convex set. Then for every $t \in \mathbb{R}$, the **level set**

$$\text{level}_f(t) \stackrel{\text{def}}{=} \{x \in \mathcal{S} : f(x) \leq t\} \quad (27)$$

is a convex subset of \mathcal{S} .

Example 109

The level set of $f(x) = x^2 + 1$ (which is defined on $\mathcal{S} = \mathbb{R}$) corresponding to level $t = 5$ is

$$\text{level}_f(5) = \{x \in \mathbb{R} : x^2 + 1 \leq 5\} = [-2, 2].$$

Clearly, this is a convex set, as predicted by Theorem 108.

Level Sets of Convex Functions are Convex Sets II

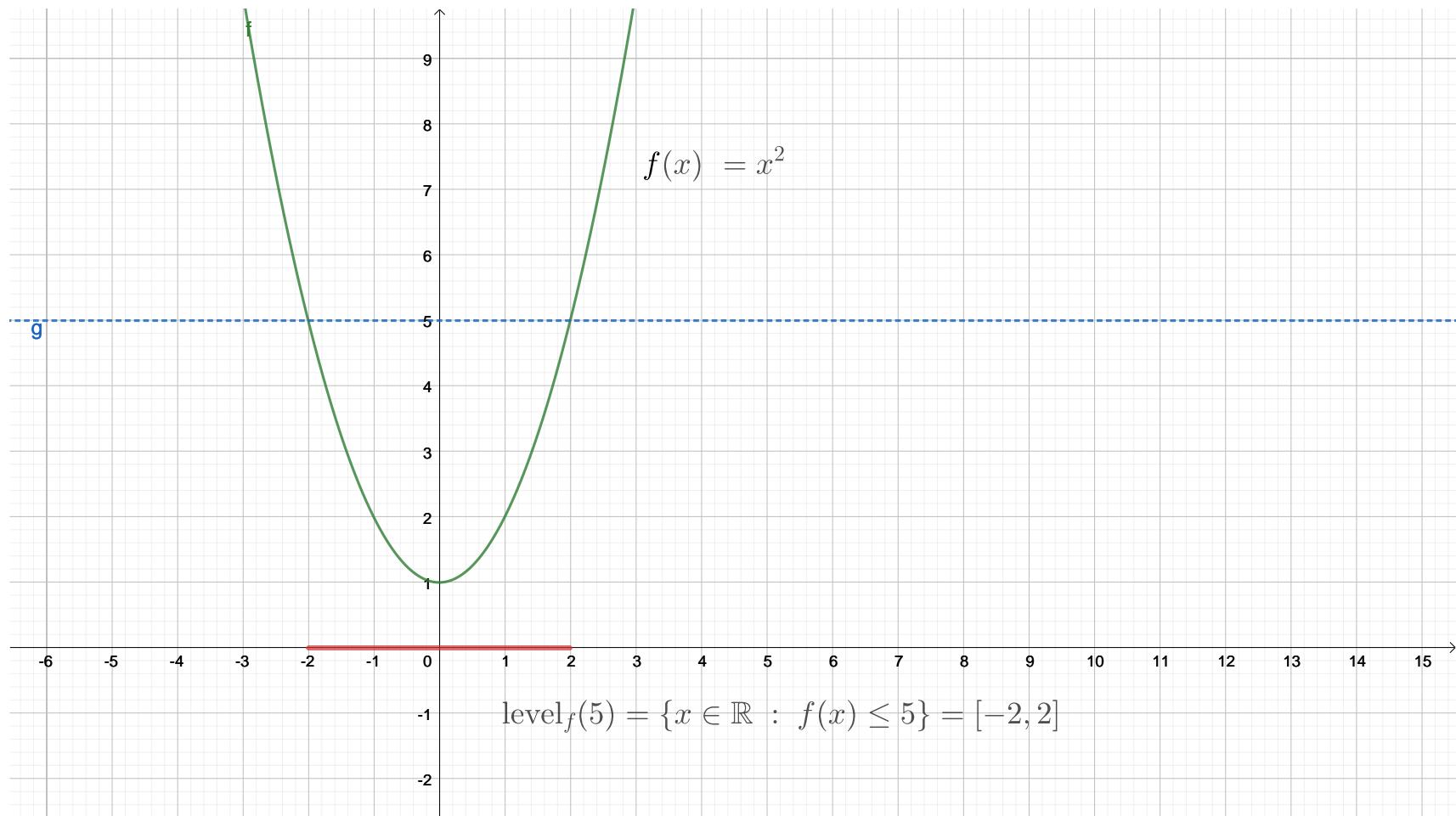


Figure: Level set of $f(x) = x^2 + 1$ at $t = 5$ is the interval $[-2, 2]$.

Level Sets of Convex Functions are Convex Sets III

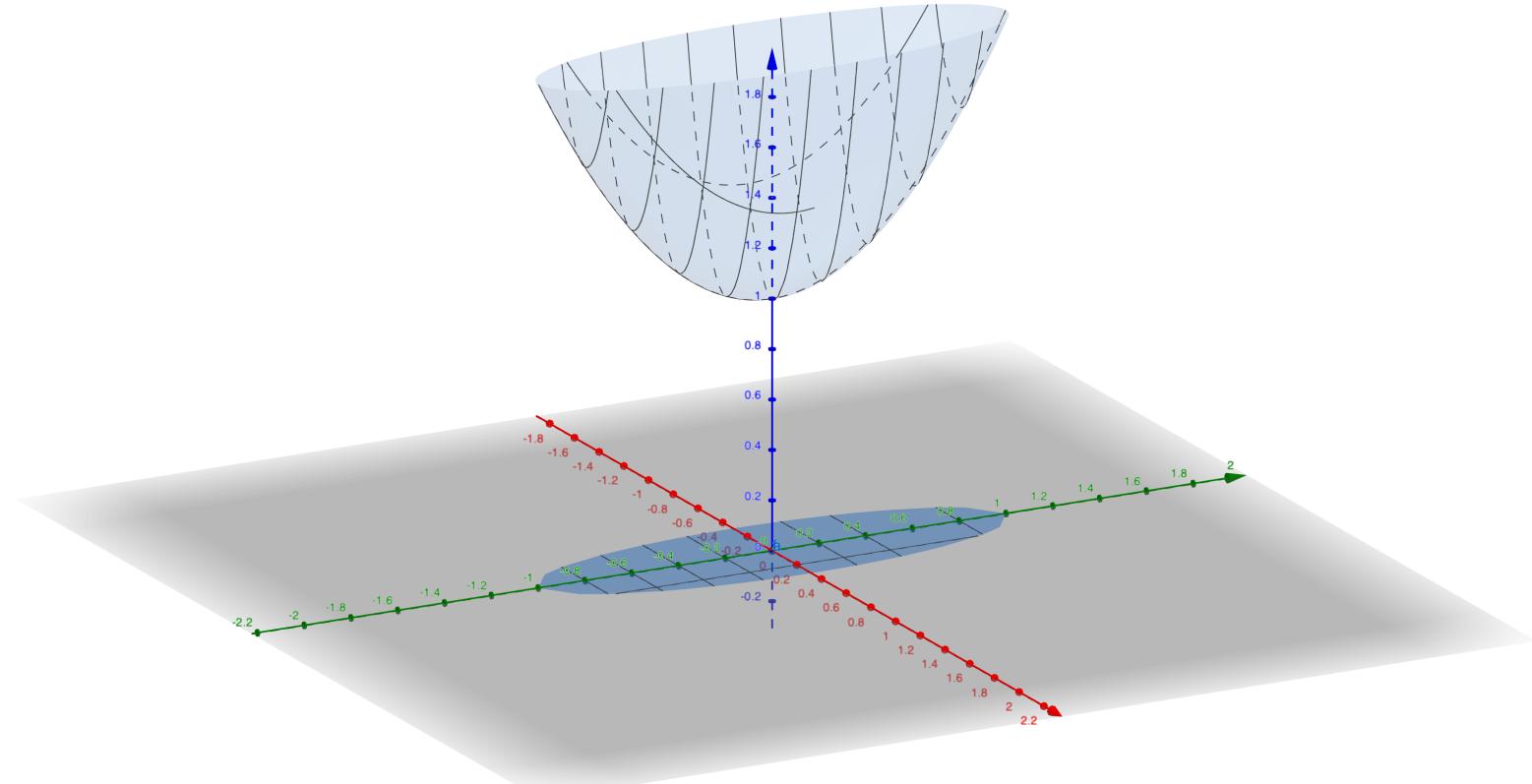


Figure: The level set of $f(x) = 10x^2 + y^2 + 1$ at $t = 2$ is an ellipsoid in \mathbb{R}^2 :
 $\text{level}_f(2) = \{(x, y) \in \mathbb{R}^2 : f(x) \leq 2\} = \{(x, y) \in \mathbb{R}^2 : \begin{pmatrix} x \\ y \end{pmatrix}^\top \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \leq 1\}$.

Level Sets of Convex Functions are Convex Sets IV

Proof. (of Theorem 108)

Let x, y belong to the level set (27), which we shall call $\text{level}_f(t)$, and choose $\lambda \in [0, 1]$. We wish to show that the vector $z = \lambda x + (1 - \lambda)y$ belongs to $\text{level}_f(t)$ as well.

- ▶ First, since $x, y \in \text{level}_f(t)$, we also have $x, y \in \mathcal{S}$. Since \mathcal{S} is convex, we conclude that

$$z \in \mathcal{S}. \quad (28)$$

- ▶ Second, since $x, y \in \text{level}_f(t)$, we know that $f(x) \leq t$ and $f(y) \leq t$. Using convexity of f , we see that

$$\begin{aligned} f(z) &= f(\lambda x + (1 - \lambda)y) \\ &\leq \lambda f(x) + (1 - \lambda)f(y) \\ &\leq \lambda t + (1 - \lambda)t = t. \end{aligned} \quad (29)$$

- ▶ Combining (28) and (29), we conclude that $z \in \text{level}_f(t)$, proving that $\text{level}_f(t)$ is a convex set.



Level Sets of Convex Functions are Convex Sets V

Example 110

The level set of $f(x) = 5x + 1$ (which is defined on $\mathcal{S} = \mathbb{R}$) corresponding to level $t = 11$ is

$$\text{level}_f(11) = \{x \in \mathbb{R} : 5x + 1 \leq 11\} = (-\infty, 2].$$

Clearly, this is a convex set, as predicted by Theorem 108.

Example 111 (Ellipsoids as level sets)

Let $\mathbf{Q} \in \mathbb{S}_{++}^d$. The level set of $f(x) = \frac{1}{2}x^\top \mathbf{Q}x$ (which is defined on $\mathcal{S} = \mathbb{R}^d$) corresponding to level $t = \frac{1}{2}$ is

$$\text{level}_f\left(\frac{1}{2}\right) = \{x \in \mathbb{R}^d : x^\top \mathbf{Q}x \leq 1\}.$$

This is an ellipsoid, and we know that it is convex from Example 43. Applying Theorem 108 is an alternative way of seeing that it is convex.

Level Sets of Convex Functions are Convex Sets VI

Exercise 112

Find all level sets of the convex function $f(x) = \max\{3x - 1, -4x + 5\}$.

Exercise 113

Find all level sets of the convex function $f(x) = -\sqrt{x}$ defined on $[0, \infty)$.

Exercise 114

Describe the level set of the convex function $f(x) = \max\{-x_1, \dots, -x_d\}$ corresponding to the level $t = 0$.

Exercise 115

Describe the level sets of the convex function $f(x) = \|x - a\|$, where $a \in \mathbb{R}^d$.

Exercise 116 (☕)

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Describe the level set of the convex function $f(x) = \|\mathbf{Ax}\|$ corresponding to level $t = 1$.

Level Sets of Convex Functions are Convex Sets VII

Exercise 117 (☕)

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$. Describe the level sets of the convex function

$$f(x) = \|\mathbf{A}x\|_{\infty} = \max \{|\mathbf{A}_{1:x}|, \dots, |\mathbf{A}_{n:x}|\}.$$

Exercise 118 (☕)

Let $\mathbf{A} \in \mathbb{S}_{++}^d$, $b \in \mathbb{R}^d$ and $c \in \mathbb{R}$. Describe the level sets of the convex function $f(x) = \frac{1}{2}x^T \mathbf{A}x + b^T x + c$.

Functions with Convex Level Sets are not Necessarily Convex

While level sets of convex functions are necessarily convex, the converse is not true. That is, functions whose level sets are convex do not necessarily need to be convex.

Exercise 119

Find a nonconvex function $f : \mathbb{R} \rightarrow \mathbb{R}$ whose all level sets are convex.

Extended Real-Valued Functions

Extended Real-Valued Functions I

Definition 120 (Extended real-valued function)

An **extended real-valued function** is a function

$$f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\},$$

i.e., a function whose values are allowed to be $+\infty$.

These functions play an important role in modern **convex optimization** and **convex duality theory**. In particular, we will see two main classes of extended real-valued functions:

- ▶ **Indicator functions** of sets (see Definition 127).
- ▶ **Fenchel conjugates** of functions (see Definition 132).

Extended Real-Valued Functions II

Definition 121 (Domain)

The **domain** of an extended real-valued function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\pm\infty\}$ is the set

$$\text{dom } f \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : f(x) < +\infty\}.$$

Definition 122 (Proper function)

An extended real-valued function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is **proper** if it has nonempty domain. That is, if there exists $x \in \mathbb{R}^d$ such that $f(x) < +\infty$.

Convexity of Extended Real-Valued Functions I

Definition 123 (Epigraph)

The **epigraph** of an extended real-valued function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is the set

$$\text{epi } f \stackrel{\text{def}}{=} \{(x, t) \in \mathbb{R}^d \times \mathbb{R} : f(x) \leq t\}.$$

Remark:

- ▶ Notice that the level set and the epigraph of f are connected via

$$\text{level}_f(t) = \{x \in \mathbb{R}^d : (x, t) \in \text{epi } f\}. \quad (30)$$

Exercise 124

Explain why does the correspondence (30) hold.

Convexity of Extended Real-Valued Functions II

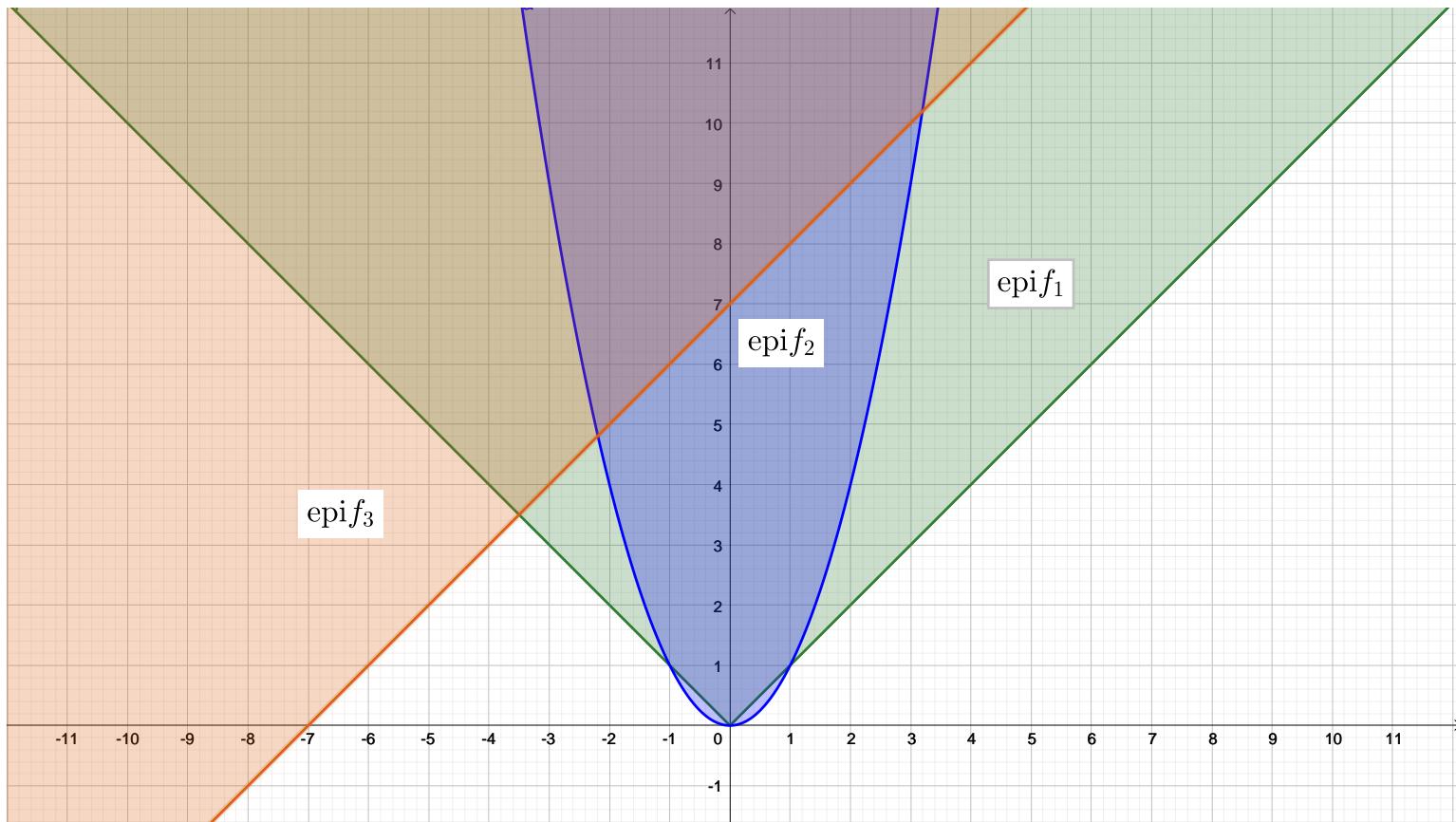


Figure: Epigraph of convex functions are convex sets. The epigraphs of the convex functions $f_1(x) = |x|$, $f_2(x) = x^2$ and $f_3(x) = x - 7$ are depicted by the shaded regions.

Convexity of Extended Real-Valued Functions III

Definition 125 (Convexity of an extended real-valued function)

We say that an extended real-valued function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is **convex** if its epigraph is a convex set.

Remarks:

- ▶ We can transform any real-valued convex function $f : \mathcal{S} \rightarrow \mathbb{R}$, where $\mathcal{S} \subseteq \mathbb{R}^d$, into an “equivalent” extended real-valued function \tilde{f} as follows:

$$\tilde{f}(x) = \begin{cases} f(x) & \text{for } x \in \mathcal{S} \\ +\infty & \text{for } x \notin \mathcal{S} \end{cases}.$$

- ▶ Notice that $\text{dom } \tilde{f} = \mathcal{S}$.
- ▶ It’s possible to check that f is convex (according to Definition 44) if and only if \tilde{f} is convex (according to Definition 125).

Convexity of Extended Real-Valued Functions IV

Exercise 126 (☕☕)

Prove that f is convex (according to Definition 44) if and only if \tilde{f} is convex (according to Definition 125).

Indicator Functions of Sets I

Definition 127 (Indicator function of a set)

Let \mathcal{S} be a subset of \mathbb{R}^d . The **indicator function of \mathcal{S}** is the function

$$\delta_{\mathcal{S}} : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$$

defined as follows:

$$\delta_{\mathcal{S}}(x) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } x \in \mathcal{S} \\ +\infty & \text{if } x \notin \mathcal{S} \end{cases}. \quad (31)$$

Theorem 128 (Properties of $\delta_{\mathcal{S}}$)

- (i) $\text{dom } \delta_{\mathcal{S}} = \mathcal{S}$.
- (ii) $\delta_{\mathcal{S}}$ is proper $\Leftrightarrow \mathcal{S}$ is nonempty.
- (iii) $\delta_{\mathcal{S}}$ is closed $\Leftrightarrow \mathcal{S}$ is closed.
- (iv) $\delta_{\mathcal{S}}$ is convex $\Leftrightarrow \mathcal{S}$ is convex.

Indicator Functions of Sets II

Exercise 129

Let $c \in \mathbb{R}^d$. Show that $\delta_{\{c\}}(x) = \delta_{\{0\}}(x - c)$.

Exercise 130 (☕)

Let $\mathcal{S}_1, \dots, \mathcal{S}_n \subseteq \mathbb{R}^d$. Show that

$$\delta_{\bigcap_{i=1}^n \mathcal{S}_i}(x) = \sum_{i=1}^n \delta_{\mathcal{S}_i}(x).$$

That is, the indicator function of the intersection of sets is the sum of the indicator functions of the individual sets.

Exercise 131 (☕☕)

Prove Theorem 128.

Indicator Functions and Constrained Optimization

Indicator functions are useful since they can encode constraints in optimization problems.

Indeed, the following two optimization problems are equivalent:

$$\min_{x \in \mathcal{S}} f(x) \quad \Leftrightarrow \quad \min_{x \in \mathbb{R}^d} f(x) + \delta_{\mathcal{S}}(x).$$

Remarks:

- ▶ This way, we can turn a constrained optimization problem with constraint \mathcal{S} to an unconstrained optimization problem with includes the extra function $\delta_{\mathcal{S}}$.
- ▶ If f and \mathcal{S} are convex, then $f + \delta_{\mathcal{S}}$ is a convex function. So, a **convex optimization problem with a convex constraint can be turned in an unconstrained convex minimization problem**.
- ▶ This transformation plays an important role in **convex duality theory**.

Introduction to Optimization

Peter Richtárik



Lecture 9: Fenchel Conjugation

Lecture Outline

- ▶ Fenchel conjugation
- ▶ Examples

Fenchel Conjugation: Definition and Basic Properties

Fenchel Conjugate of a Function: Definition

Definition 132 (Fenchel Conjugate)

The **Fenchel conjugate** of a proper function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is the function $f^* : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ defined as

$$f^*(x^*) \stackrel{\text{def}}{=} \sup_{x \in \mathbb{R}^d} \{ \langle x^*, x \rangle - f(x) \}. \quad (32)$$

Remarks:

- ▶ If f is bounded below, then $\text{dom } f^* \neq \emptyset$.
- ▶ This assumption implies that $\inf f \in \mathbb{R}$.
- ▶ Since $f^*(0) = \sup -f = -\inf f$, we see that $f^*(0) \in \mathbb{R}$, which means that $0 \in \text{dom } f^*$, i.e., $\text{dom } f^* \neq \emptyset$.

Fenchel Conjugate of a Function: Properties

Theorem 133 (Properties of Fenchel Conjugation)

Assume $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$.

- (i) $\langle x^*, x \rangle \leq f^*(x^*) + f(x)$ for all $x^* \in \mathbb{R}^d$ and $x \in \text{dom } f$
(Fenchel-Young inequality)
- (ii) If $f \leq g$, then $f^* \geq g^*$
(order-reversing property)
- (iii) f^* is closed and convex
(closedness and convexity)
- (iv) $f^{**} \leq f$
(biconjugate inequality)
- (v) If f is proper, closed and convex, then $f = f^{**}$ *(biconjugate equality)*

Exercise 134 (☕)

Prove parts (i) and (ii) of Theorem 133.

Fenchel Conjugates of Transformed Functions

What are the Fenchel conjugates of functions

$$\phi(bx), \quad \phi(x + c), \quad t\phi(x)$$

if we know the Fenchel conjugate ϕ^* of function ϕ ?

$f(x)$	Assumption	$f^*(x^*)$	Reference
$\phi(bx)$	$b \neq 0$	$\phi^*\left(\frac{x^*}{b}\right)$	
$\phi(x + c)$	—	$\phi^*(x^*) - \langle c, x^* \rangle$	
$t\phi(x)$	$t > 0$	$t\phi^*\left(\frac{x^*}{t}\right)$	

Table: Fenchel conjugates of transformed functions.

Exercise 135 (☕☕)

Prove the formulas for the Fenchel conjugates of transformed functions shown in the table.

Fenchel Conjugation: Examples

Examples: Fenchel Conjugate

$f(x)$	Assumption	$f^*(x^*)$	Reference
$\langle c, x \rangle$	—	$\delta_{\{c\}}(x^*)$	Example 137
$\delta_{\{c\}}(x)$	—	$\langle c, x^* \rangle$	Example 138
$\frac{1}{2} \ x\ ^2$	—	$\frac{1}{2} \ x^*\ ^2$	Example 139
$\frac{1}{2} x^\top A x$	$A \in \mathbb{S}_{++}^d$	$\frac{1}{2} (x^*)^\top A^{-1} x^*$	Example 140
$\delta_{\mathbb{R}_+^d}(x)$	—	$\delta_{-\mathbb{R}_+^d}(x^*)$	Example 141
$\delta_{\mathcal{L}}(x)$	\mathcal{L} is a cone	$\delta_{\mathcal{L}^-}(x^*)$	Example 142
e^x	—	$\begin{cases} x^* \log x^* - x^* & x^* > 0 \\ 0 & x^* = 0 \\ +\infty & x^* < 0 \end{cases}$	

Table: Examples of functions and their Fenchel conjugates.

Exercise 136

Calculate the Fenchel conjugate of $f(x) = e^x$ and show your derivations.

Examples: Fenchel Conjugate

Example 137 (Fenchel conjugate of a linear function)

Let $f(x) = \langle c, x \rangle$, where $c \in \mathbb{R}^d$. Then

$$\begin{aligned} f^*(x^*) &\stackrel{(32)}{=} \sup_{x \in \mathbb{R}^d} \{ \langle x^*, x \rangle - f(x) \} \\ &= \sup_{x \in \mathbb{R}^d} \{ \langle x^* - c, x \rangle \} \\ &\stackrel{(*)}{=} \begin{cases} 0 & \text{if } x^* = c \\ +\infty & \text{if } x^* \neq c \end{cases} \\ &= \begin{cases} 0 & \text{if } x^* \in \{c\} \\ +\infty & \text{if } x^* \notin \{c\} \end{cases} \\ &\stackrel{(31)}{=} \delta_{\{c\}}(x^*). \end{aligned}$$

The Fenchel conjugate of the linear function $f(x) = \langle c, x \rangle$ is $f^*(x^*) = \delta_{\{c\}}(x^*)$.

Examples: Fenchel Conjugate

Example 138 (Fenchel conjugate of an indicator function of a singleton)

Let $f(x) = \delta_{\{c\}}(x)$, where $c \in \mathbb{R}^d$. Then

$$f^*(x^*) \stackrel{(32)}{=} \sup_{x \in \mathbb{R}^d} \{\langle x^*, x \rangle - f(x)\} = \sup_{x \in \mathbb{R}^d} \{\langle x^*, x \rangle - \delta_{\{c\}}(x)\}.$$

Observe that

$$\langle x^*, x \rangle - \delta_{\{c\}}(x) = \begin{cases} \langle x^*, x \rangle - 0 & \text{if } x = c \\ \langle x^*, x \rangle - \infty & \text{if } x \neq c \end{cases} = \begin{cases} \langle x^*, c \rangle & \text{if } x = c \\ -\infty & \text{if } x \neq c \end{cases}.$$

Therefore,

$$f^*(x^*) = \sup_{x \in \mathbb{R}^d} \{\langle x^*, x \rangle - \delta_{\{c\}}(x)\} = \langle x^*, c \rangle. \quad (33)$$

The Fenchel conjugate of $f(x) = \delta_{\{c\}}(x)$ is the linear function $f^*(x^*) = \langle c, x^* \rangle$.

Examples: Fenchel Conjugate

Example 139 (Fenchel conjugate of the squared Euclidean norm)

Let $f(x) = \frac{1}{2} \|x\|^2$. Then

$$f^*(x^*) \stackrel{(32)}{=} \sup_{x \in \mathbb{R}^d} \{\langle x^*, x \rangle - f(x)\} = \sup_{x \in \mathbb{R}^d} \left\{ \langle x^*, x \rangle - \frac{1}{2} \|x\|^2 \right\}. \quad (34)$$

Notice that the function $\phi(x) \stackrel{\text{def}}{=} \langle x^*, x \rangle - \frac{1}{2} \|x\|^2$ is concave and quadratic. We can find the point x at which the maximum is attained by setting the gradient to zero:

$$\nabla \phi(x) = 0 \Leftrightarrow x^* - x = 0. \quad (35)$$

Therefore, by plugging $x = x^*$ into (34), we get

$$f^*(x^*) \stackrel{(35)+(34)}{=} \langle x^*, x^* \rangle - \frac{1}{2} \|x^*\|^2 = \frac{1}{2} \|x^*\|^2.$$

The Fenchel conjugate of $f(x) = \frac{1}{2} \|x\|^2$ is the function f is itself.

Examples: Fenchel Conjugate

Example 140 (Fenchel conjugate of a convex quadratic function)

Let $f(x) = \frac{1}{2}\langle \mathbf{A}x, x \rangle$, where $\mathbf{A} \in \mathbb{S}_{++}^d$. Then

$$f^*(x^*) = \sup_{x \in \mathbb{R}^d} \{ \langle x^*, x \rangle - f(x) \} = \sup_{x \in \mathbb{R}^d} \left\{ \langle x^*, x \rangle - \frac{1}{2} \langle \mathbf{A}x, x \rangle \right\}. \quad (36)$$

The function $\phi(x) \stackrel{\text{def}}{=} \langle x^*, x \rangle - \frac{1}{2} \langle \mathbf{A}x, x \rangle$ is concave and quadratic, and its maximizer can be found by setting its gradient to zero:

$$\nabla \phi(x) = 0 \iff x^* - \mathbf{A}x = 0 \iff x = \mathbf{A}^{-1}x^*. \quad (37)$$

Therefore, by plugging $x = \mathbf{A}^{-1}x^*$ into (36), we get

$$f^*(y) \stackrel{(36)+(37)}{=} \langle x^*, \mathbf{A}^{-1}x^* \rangle - \frac{1}{2} \langle \mathbf{A}\mathbf{A}^{-1}x^*, \mathbf{A}^{-1}x^* \rangle = \frac{1}{2} \langle x^*, \mathbf{A}^{-1}x^* \rangle.$$

The Fenchel conjugate of $f(x) = \frac{1}{2}x^\top \mathbf{A}x$ is the function
 $f^*(x^*) = \frac{1}{2}(x^*)^\top \mathbf{A}^{-1}x^*$.

Examples: Fenchel Conjugate

Example 141 (Fenchel conjugate of the indicator of the nonnegative orthant)

Let $f(x) = \delta_{\mathbb{R}_+^d}(x)$. Then

$$\begin{aligned} f^*(x^*) &= \sup_{x \in \mathbb{R}^d} \{\langle x^*, x \rangle - f(x)\} \\ &= \sup_{x \in \mathbb{R}^d} \left\{ \langle x^*, x \rangle - \delta_{\mathbb{R}_+^d}(x) \right\} \\ &= \sup_{x \in \mathbb{R}_+^d} \langle x^*, x \rangle \\ &= \begin{cases} 0 & \text{if } x^* \in -\mathbb{R}_+^d \\ +\infty & \text{if } x^* \notin -\mathbb{R}_+^d \end{cases} \\ &= \delta_{-\mathbb{R}_+^d}(x^*). \end{aligned}$$

Note also that $\delta_{-\mathbb{R}_+^d}(x^*) = \delta_{(\mathbb{R}_+^d)^-}(x^*) = \delta_{\mathbb{R}_+^d}(-x^*)$.

The Fenchel conjugate of the indicator function of the non-negative orthant \mathbb{R}_+^d is the indicator function of its polar cone $(\mathbb{R}_+^d)^-$.

Examples: Fenchel Conjugate

Example 142 (Fenchel conjugate of the indicator function of a cone)

Let $f(x) = \delta_{\mathcal{L}}(x)$, where \mathcal{L} is a cone. Then

$$f^*(x^*) = \sup_{x \in \mathbb{R}^d} \{\langle x^*, x \rangle - f(x)\} = \sup_{x \in \mathbb{R}^d} \{\langle x^*, x \rangle - \delta_{\mathcal{L}}(x)\} = \sup_{x \in \mathcal{L}} \langle x^*, x \rangle.$$

We now need to calculate the solution to the optimization problem $\sup_{x \in \mathcal{L}} \langle x^*, x \rangle$. We now consider two cases:

- ▶ If there exists $x \in \mathcal{L}$ such that $\langle x^*, x \rangle > 0$, then $\langle x^*, tx \rangle = t\langle x^*, x \rangle \rightarrow \infty$ as $t \rightarrow \infty$, while $tx \in \mathcal{L}$ for all $t > 0$. So, $\sup_{x \in \mathcal{L}} \langle x^*, x \rangle = \infty$.
- ▶ On the other hand, if there does not exist $x \in \mathcal{L}$ such that $\langle x^*, x \rangle > 0$, which can equivalently be written as

$$x^* \in \mathcal{L}^- \stackrel{\text{def}}{=} \{x' : \langle x', x \rangle \leq 0 \ \forall x \in \mathcal{L}\},$$

then $\sup_{x \in \mathcal{L}} \langle x^*, x \rangle \leq 0$. However, since $\langle x^*, 0 \rangle = 0$ and $0 \in \mathcal{L}$, this means that $\sup_{x \in \mathcal{L}} \langle x^*, x \rangle = 0$.

In summary, we have shown that

$$f^*(x^*) = \begin{cases} +\infty & \text{if } x^* \notin \mathcal{L}^- \\ 0 & \text{if } x^* \in \mathcal{L}^- \end{cases} \stackrel{\text{def}}{=} \delta_{\mathcal{L}^-}(x^*).$$

So, the Fenchel conjugate of the indicator function of a cone \mathcal{L} is the indicator function of the polar cone \mathcal{L}^- .

Introduction to Optimization

Peter Richtárik



Lecture 10: Fenchel Duality - Part 1

Lecture Outline

- ▶ Fenchel duality
- ▶ Linear programming duality

Fenchel Duality

Key Tool: Fenchel-Young Inequality

Theorem 143 (Fenchel-Young Inequality)

Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ and let $x \in \text{dom } f$ and $x^* \in \mathbb{R}^d$. Then

$$\langle x^*, x \rangle \leq f^*(x^*) + f(x).$$

Example 144

Let $f(x) = \frac{1}{2} \|x\|^2$. Then $f^*(x^*) = \frac{1}{2} \|x^*\|^2$ and the Fenchel-Young inequality says that

$$\langle x^*, x \rangle \leq \frac{1}{2} \|x^*\|^2 + \frac{1}{2} \|x\|^2.$$

Fenchel Duality: Primal and Dual Problems

Let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$, $g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ and $\mathbf{A} \in \mathbb{R}^{n \times d}$.

Primal problem:

$$OPT_P \stackrel{\text{def}}{=} \inf_{x \in \mathbb{R}^d} \{f(x) + g(\mathbf{A}x)\} \quad (\text{Primal})$$

Dual problem:

$$OPT_D \stackrel{\text{def}}{=} \sup_{y \in \mathbb{R}^n} -f^*(\mathbf{A}^\top y) - g^*(-y) \quad (\text{Dual})$$

We now show that the primal and dual problems are intimately related:

Theorem 145 (Weak Fenchel duality)

The primal and dual optimal values satisfy the inequality

$$OPD_D \leq OPT_P.$$

Proof of Weak Fenchel Duality

This follows from the Fenchel-Young inequality used twice, once for f , obtaining

$$\langle \mathbf{A}^\top \mathbf{y}, \mathbf{x} \rangle \leq f^*(\mathbf{A}^\top \mathbf{y}) + f(\mathbf{x}),$$

and once for g , obtaining

$$\langle -\mathbf{y}, \mathbf{A}\mathbf{x} \rangle \leq g^*(-\mathbf{y}) + g(\mathbf{A}\mathbf{x}).$$

By adding these two inequalities, we get

$$0 \stackrel{(2)}{=} \langle \mathbf{A}^\top \mathbf{y}, \mathbf{x} \rangle + \langle -\mathbf{y}, \mathbf{A}\mathbf{x} \rangle \leq f^*(\mathbf{A}^\top \mathbf{y}) + f(\mathbf{x}) + g^*(-\mathbf{y}) + g(\mathbf{A}\mathbf{x}).$$

By rearranging the inequality, we get

$$-f^*(\mathbf{A}^\top \mathbf{y}) - g^*(-\mathbf{y}) \leq f(\mathbf{x}) + g(\mathbf{A}\mathbf{x}).$$

Since this holds for all \mathbf{x} and \mathbf{y} , the inequality holds if we take supremum in \mathbf{y} on the left-hand side, and infimum in \mathbf{x} on the right-hand side:

$$OPD_D = \sup_{\mathbf{y} \in \mathbb{R}^n} -f^*(\mathbf{A}^\top \mathbf{y}) - g^*(-\mathbf{y}) \leq \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x}) = OPT_P.$$

Strong Fenchel Duality

Definition 146 (Core of a set)

The **core** of a set $\mathcal{S} \subseteq \mathbb{R}^d$ is the set of points $x \in \mathcal{S}$ such that for any $v \in \mathcal{B}(0, 1)$ there exists $\epsilon > 0$ such that $x + tv \in \mathcal{S}$ for all $0 \leq t \leq \epsilon$.

Exercise 147

It turns out that the core always contains the interior: $\text{int}(S) \subseteq \text{core}(S)$. Explain why.

We are now ready to describe the main duality result:

Theorem 148 (Strong Fenchel duality)

If f and g are convex, and

$$0 \in \text{core}(\text{dom } g - \mathbf{A}(\text{dom } f)),$$

then

$$\text{OPT}_P = \text{OPT}_D.$$

Moreover, if OPT_D is finite, then there exists $y \in \mathbb{R}^n$ such that

$$\text{OPT}_D = \sup_{y \in \mathbb{R}^n} -f^*(\mathbf{A}^\top y) - g^*(-y).$$

Linear Programming Duality

Linear Programming Duality I

We now consider the special case with

$$f(x) = c^\top x + \delta_{\mathbb{R}_+^d}(x), \quad c \in \mathbb{R}^d, \quad g(y) = \delta_{\{b\}}(y), \quad b \in \mathbb{R}^n. \quad (38)$$

The **primal problem (Primal)** becomes

$$\begin{aligned} OPT_P &\stackrel{\text{def}}{=} \inf_{x \in \mathbb{R}^d} \{f(x) + g(\mathbf{A}x)\} \\ &\stackrel{(38)}{=} \inf_{x \in \mathbb{R}^d} \left\{ c^\top x + \delta_{\mathbb{R}_+^d}(x) + \delta_{\{b\}}(\mathbf{A}x) \right\} \\ &= \boxed{\inf_{x \in \mathbb{R}^d} \left\{ c^\top x : x \geq 0, \mathbf{A}x = b \right\}} \end{aligned}$$

Linear Programming Duality II

The **dual problem (Dual)** becomes

$$OPT_D \stackrel{\text{def}}{=} \sup_{y \in \mathbb{R}^n} \{-f^*(\mathbf{A}^\top y) - g^*(-y)\}$$

$$\stackrel{(38)}{=} \sup_{y \in \mathbb{R}^n} \{-f^*(\mathbf{A}^\top y) - \delta_b^*(y)\}$$

$$\stackrel{\text{Exercise 138}}{=} \sup_{y \in \mathbb{R}^n} \{-f^*(\mathbf{A}^\top y) - b^\top(-y)\}$$

$$\stackrel{\text{Exercise 149}}{=} \sup_{y \in \mathbb{R}^n} \left\{ -\delta_{\mathbb{R}_+^d}(c - \mathbf{A}^\top y) + b^\top y \right\}$$

$$\stackrel{(138)}{=} \boxed{\sup_{y \in \mathbb{R}^n} \left\{ b^\top y : c - \mathbf{A}^\top y \geq 0 \right\}}$$

Exercise 149

Let $f(x) = c^\top x + \delta_{\mathbb{R}_+^d}(x)$, where $c \in \mathbb{R}^d$. Show that $f^*(y) = \delta_{\mathbb{R}_+^d}(c - y)$.

The LP Problem

minimize $c^\top x$
subject to $a_i^\top x = b_i, \quad i \in \mathcal{E}$ (set of **Equality** constraints)
 $a_i^\top x \leq b_i, \quad i \in \mathcal{I}$ (set of **Inequality** constraints)

where $c, a_i, x \in \mathbb{R}^n$, $b_i \in \mathbb{R}$ for all $i \in \mathcal{E} \cup \mathcal{I}$.

Standard-form LP

minimize $c^\top x$
subject to $\mathbf{A}x = b$
 $x \geq 0$

where $c \in \mathbb{R}^d$, $b \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{n \times d}$.

Primal and Dual Linear Programming Problems on 1 Slide

Primal problem:

$$(P) \quad \begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & \mathbf{A}x = b \\ & x \geq 0 \end{array}$$

Dual problem:

$$(D) \quad \begin{array}{ll} \text{maximize} & b^\top y \\ \text{subject to} & c - \mathbf{A}^\top y \geq 0 \end{array}$$

Theorem 150 (Weak Duality)

If x is feasible for (P) and y is feasible for (D) , then $c^\top x \geq b^\top y$.

Proof.

$$0 \leq (c - \mathbf{A}^\top y)^\top x = c^\top x - y^\top \mathbf{A}x = \underbrace{c^\top x - b^\top y}_{\text{duality gap}}$$

Weak Duality

Corollary 151

- (a) (P) unbounded \Rightarrow (D) infeasible
- (b) (D) unbounded \Rightarrow (P) infeasible
- (c) x feasible for (P) , y feasible for (D) , $c^\top x = b^\top y \implies x, y$ optimal.

Exercise 152 (☕☕)

Prove Corollary 151.

Example: Find the Dual

Example 153

Show that (D') is the dual of the linear program (P') :

$$(P') \quad \begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & \mathbf{A}x \geq b \\ & x \geq 0 \end{array}$$

$$(D') \quad \begin{array}{ll} \text{maximize} & b^\top y \\ \text{subject to} & \mathbf{A}^\top y \leq c \\ & y \geq 0. \end{array}$$

Solution: $c^\top x - b^\top y \geq (\mathbf{A}^\top y)^\top x - (\mathbf{A}x)^\top y = y^\top \mathbf{A}x - x^\top \mathbf{A}^\top y = 0.$

Example: Both (P) and (D) can be Infeasible

Example 154

Give an example of an LP such that it and its dual are both infeasible.

Solution: Standard LP and its dual, with $n = d = 1$, $\mathbf{A} = 0$, $\mathbf{c} = -1$, $\mathbf{b} = 1$.

$$(P) \quad \begin{array}{lll} \text{minimize} & -x \\ \text{subject to} & 0x = 1 \\ & x \geq 0 \end{array}$$

$$(D) \quad \begin{array}{lll} \text{maximize} & 1y \\ \text{subject to} & -1 - 0y \geq 0 \end{array}$$

Strong Duality

Weak duality gives sufficient optimality conditions; that is, conditions sufficient to imply optimality. Indeed, recall that a corollary of weak duality says: x is feasible for (P), y feasible for (D) and $c^\top x = b^\top y$, then x, y are optimal.

What about necessary conditions? We have something even stronger:

Theorem 155 (Strong Duality)

If (P) has optimal solution x , then (D) has optimal solution y and

$$c^\top x = b^\top y.$$

Corollary 156

If both (P) and (D) are feasible, then they have optimal solutions x, y and $c^\top x = b^\top y$.

Proof.

Pick any feasible y' for (D). By weak duality, $c^\top x \geq b^\top y'$ for all feasible x . In particular, (P) is bounded and hence has an optimal solution. It remains to apply strong duality.

Optimality Conditions

Corollary 157 (Optimality Conditions)

Point x is optimal for (P) if and only if

- (i) x is primal feasible (i.e., $\mathbf{A}x = b, x \geq 0$) and
- (ii) there exists dual feasible y (i.e., $c - \mathbf{A}^\top y \geq 0$) such that
- (iii) there is no duality gap (i.e., $c^\top x = b^\top y$).

Condition (iii) can be replaced by

- (iii') $s_i x_i = 0$ for all i , where $s = c - \mathbf{A}^\top y$ (**complementary slackness**)

Goldman-Tucker Theorem

Theorem 158 (Goldman-Tucker)

When both the primal problem (P) and the dual problem (D) are feasible, then they have **optimal solutions** x^* and y^* satisfying **strict complementarity condition**:

$$x^* + c - \mathbf{A}^\top y^* > 0.$$

Remarks:

- ▶ Note that x^* and $s^* \stackrel{\text{def}}{=} c - \mathbf{A}^\top y^*$ are vectors in \mathbb{R}^d .
- ▶ The above condition says that

$$x_i^* + s_i^* > 0, \quad \text{for all } i = 1, 2, \dots, d.$$

Introduction to Optimization

Peter Richtárik



Lecture 11: Fenchel Duality - Part 2

Lecture Outline

- ▶ Quadratic programming duality
- ▶ Von Neumann minimax duality

Quadratic Programming Duality

Quadratic Programming Duality I

We now consider the special case with

$$f(x) = \frac{1}{2}x^\top \mathbf{Q}x + c^\top x + \delta_{\mathbb{R}_+^d}(x), \quad \mathbf{Q} \in \mathbb{S}_{++}^d, \quad c \in \mathbb{R}^d, \quad (39)$$

$$g(y) = \delta_{\{b\}}(y), \quad b \in \mathbb{R}^n. \quad (40)$$

The **primal problem (Primal)** becomes

$$\begin{aligned} OPT_P &\stackrel{\text{def}}{=} \inf_{x \in \mathbb{R}^d} \{f(x) + g(\mathbf{A}x)\} \\ &\stackrel{(39)+(40)}{=} \inf_{x \in \mathbb{R}^d} \left\{ \frac{1}{2}x^\top \mathbf{Q}x + c^\top x + \delta_{\mathbb{R}_+^d}(x) + \delta_{\{b\}}(\mathbf{A}x) \right\} \\ &= \boxed{\inf_{x \in \mathbb{R}^d} \left\{ \frac{1}{2}x^\top \mathbf{Q}x + c^\top x : x \geq 0, \mathbf{A}x = b \right\}} \end{aligned}$$

Quadratic Programming Duality II

Example 159 (☕☕☕)

Let

$$f(x) = \frac{1}{2}x^\top \mathbf{Q}x + c^\top x + \delta_{\mathbb{R}_+^d}(x),$$

where $\mathbf{Q} \in \mathbb{S}_{++}^d$ and $c \in \mathbb{R}^d$. Calculate $f^*(y)$ and then write down the dual problem.

Quadratic Programming: Primal and Dual Problems

Primal problem:

$$(P) \quad \begin{array}{ll} \min_x & \frac{1}{2}x^\top \mathbf{Q}x + c^\top x \\ \text{subject to} & \mathbf{A}x = b \\ & x \geq 0 \end{array}$$

Dual problem:

$$(D) \quad \begin{array}{ll} \max_{x,y,s} & b^\top y - \frac{1}{2}x^\top \mathbf{Q}x \\ \text{subject to} & \mathbf{A}^\top y - \mathbf{Q}x + s = c \\ & x \geq 0, \quad s \geq 0 \end{array}$$

Quadratic Programming: Weak Duality

Theorem 160 (Weak Duality)

If x is feasible for (P) and (x, y, s) is feasible for (D) , then

$$\frac{1}{2}x^\top \mathbf{Q}x + c^\top x \geq b^\top y - \frac{1}{2}x^\top \mathbf{Q}x.$$

Proof.

Let x be primal feasible and (x, y, s) be dual feasible. Then

$$\begin{aligned} & \frac{1}{2}x^\top \mathbf{Q}x + c^\top x - (b^\top y - \frac{1}{2}x^\top \mathbf{Q}x) \\ = & x^\top \mathbf{Q}x + c^\top x - b^\top y \\ = & x^\top \mathbf{Q}x + (\mathbf{A}^\top y - \mathbf{Q}x + s)^\top x - (\mathbf{Ax})^\top y \\ = & \underbrace{(x^\top \mathbf{Q}x - x^\top \mathbf{Q}^\top x)}_{=0} + \underbrace{(y^\top \mathbf{Ax} - x^\top \mathbf{A}^\top y)}_{=0} + \underbrace{s^\top x}_{\geq 0} \geq 0. \end{aligned}$$

We have used the fact that $\mathbf{Q} = \mathbf{Q}^\top$, and that since $\alpha \stackrel{\text{def}}{=} y^\top \mathbf{Ax}$ is a scalar, we must trivially have $\alpha = \alpha^\top = (y^\top \mathbf{Ax})^\top = x^\top \mathbf{A}^\top y$.

Properties of Convex Quadratics

Theorem 161

Let $f(x) = \frac{1}{2}x^\top \mathbf{Q}x + c^\top x$ and assume \mathbf{Q} is symmetric. Then

- (i) $x^\top \mathbf{Q}x < 0$ for some $x \Rightarrow f$ is **unbounded below**
- (ii) $\mathbf{Q} \succeq 0$ & \mathbf{Q} is not positive definite \Rightarrow either f is **unbounded below** or f has **infinitely many minimizers**
- (iii) $\mathbf{Q} \succ 0 \Rightarrow f$ has a **unique minimizer**

Proof.



- (i) Consider $f(tx)$ as $t \rightarrow +\infty$:

$$f(tx) = \frac{1}{2}t^2(\underbrace{x^\top \mathbf{Q}x}_{<0}) + t(c^\top x) \rightarrow -\infty.$$

- (ii) Needs a bit more linear algebra, so we will skip this.
- (iii) \mathbf{Q} is invertible and hence the only solution of $\nabla f(x) = 0$ ($\mathbf{Q}x = -c$) is $x = -\mathbf{Q}^{-1}c$. It is also possible to prove this from the definition directly.

Von Neumann Minimax Duality

Von Neumann Minimax Duality I

We now consider the special case with

$$f(x) = \delta_{\mathcal{X}}(x), \quad g(y) = \delta_{\mathcal{Y}}^*(y), \quad (41)$$

where

- ▶ $\mathcal{X} \subset \mathbb{R}^d$ is a nonempty, convex, compact set,
- ▶ $\mathcal{Y} \subset \mathbb{R}^n$ is a nonempty, convex, compact set.

The **primal problem (P)** becomes

$$\begin{aligned} OPT_P &\stackrel{\text{def}}{=} \inf_{x \in \mathbb{R}^d} \{f(x) + g(\mathbf{A}x)\} \stackrel{(41)}{=} \inf_{x \in \mathbb{R}^d} \{\delta_{\mathcal{X}}(x) + \delta_{\mathcal{Y}}^*(\mathbf{A}x)\} \\ &\stackrel{(32)}{=} \inf_{x \in \mathbb{R}^d} \left\{ \delta_{\mathcal{X}}(x) + \sup_{y \in \mathbb{R}^n} \{ \langle \mathbf{A}x, y \rangle - \delta_{\mathcal{Y}}(y) \} \right\} \\ &= \inf_{x \in \mathbb{R}^d} \sup_{y \in \mathbb{R}^n} \{ \delta_{\mathcal{X}}(x) + \langle \mathbf{A}x, y \rangle - \delta_{\mathcal{Y}}(y) \} \\ &= \inf_{x \in \mathbb{R}^d} \sup_{y \in \mathcal{Y}} \{ \delta_{\mathcal{X}}(x) + \langle \mathbf{A}x, y \rangle \} \\ &= \boxed{\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \langle \mathbf{A}x, y \rangle} \end{aligned}$$

Von Neumann Minimax Duality II

The **dual problem (D)** becomes

$$\begin{aligned}
 OPT_D &\stackrel{\text{def}}{=} \sup_{y \in \mathbb{R}^d} \{-f^*(\mathbf{A}^\top y) - g^*(-y)\} \\
 &\stackrel{(41)}{=} \sup_{y \in \mathbb{R}^d} \{-\delta_{\mathcal{X}}^*(\mathbf{A}^\top y) - \delta_{\mathcal{Y}}^{**}(-y)\} \\
 &\stackrel{(\dagger)}{=} \sup_{y \in \mathbb{R}^d} \{-\delta_{\mathcal{X}}^*(\mathbf{A}^\top y) - \delta_{\mathcal{Y}}(-y)\} \\
 &\stackrel{(32)}{=} \sup_{y \in \mathbb{R}^d} \left\{ - \sup_{x \in \mathbb{R}^n} \{ \langle \mathbf{A}^\top y, x \rangle - \delta_{\mathcal{X}}(x) \} - \delta_{\mathcal{Y}}(-y) \right\} \\
 &= \sup_{y \in \mathbb{R}^d} \left\{ \inf_{x \in \mathbb{R}^n} \{ -\langle \mathbf{A}^\top y, x \rangle + \delta_{\mathcal{X}}(x) \} - \delta_{\mathcal{Y}}(-y) \right\} \\
 &= \sup_{y \in \mathbb{R}^d} \inf_{x \in \mathbb{R}^n} \{ \langle \mathbf{A}^\top(-y), x \rangle + \delta_{\mathcal{X}}(x) - \delta_{\mathcal{Y}}(-y) \} \\
 &= \sup_{y \in \mathbb{R}^d} \inf_{x \in \mathbb{R}^n} \{ \langle \mathbf{A}^\top y, x \rangle + \delta_{\mathcal{X}}(x) - \delta_{\mathcal{Y}}(y) \} \\
 &= \boxed{\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \langle \mathbf{A}^\top y, x \rangle = \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \langle \mathbf{A}x, y \rangle}
 \end{aligned}$$

Von Neumann Minimax Duality III

Theorem 162 (Von Neumann, 1948)

Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}^n$ be nonempty, convex, and compact sets. Then

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \langle \mathbf{A}x, y \rangle = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \langle \mathbf{A}x, y \rangle$$

Exercise 163 (☕)

In step (†) we have used the identity $\delta_{\mathcal{Y}}^{**} = \delta_{\mathcal{Y}}$. Why does it hold?

Two-Player Zero-Sum Games

Von Neumann duality has a key role in **game theory**.

- ▶ Assume there are **2 players**:
 - ▶ player \mathcal{X} , who has **d pure strategies** to choose from (corresponding to the columns of \mathbf{A}), and
 - ▶ player \mathcal{Y} , who has **n pure strategies** to choose from (corresponding to the rows of \mathbf{A})
- ▶ The value A_{ij} is the loss player \mathcal{X} is going to incur (and the profit player \mathcal{Y} will gain) if player \mathcal{X} plays pure strategy j and player \mathcal{Y} plays pure strategy i .
 - ▶ Since what one gains the other one loses, this is a **zero-sum game**.
- ▶ Let $\mathcal{X} = \Delta^d$ and $\mathcal{Y} = \Delta^n$ be the probability simplexes in \mathbb{R}^d and \mathbb{R}^n , respectively. Instead of playing pure strategies, both players are allowed to play **mixed strategies**:
 - ▶ we can think of vector $x \in \mathcal{X} = \Delta^d$ as a **mixed/randomized strategy of player \mathcal{X} (plays column j with probability x_j)** and
 - ▶ we can think of vector $y \in \mathcal{Y} = \Delta^n$ as a **mixed/randomized strategy of player \mathcal{Y} (plays row i with probability y_i)**.
- ▶ The product $\langle \mathbf{Ax}, y \rangle = y^\top \mathbf{Ax}$ is then the **expected loss of player \mathcal{X}** under the mixed strategies $x \in \Delta^d$ and $y \in \Delta^n$.

Introduction to Optimization

Peter Richtárik



Lecture 12: Fenchel Duality - Part 3

Lecture Outline

- ▶ From convex optimization to conic programming
- ▶ Conic programming duality

From Convex Optimization to Conic Programming

From Convex Optimization to Conic Programming I

- ▶ **Each convex optimization problem**, i.e., a problem of the form

$$\min_{x \in \mathcal{S}} f(x),$$

where

- ▶ $\mathcal{S} \subseteq \mathbb{R}^d$ is a convex set, and
- ▶ $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a convex function,

can (in principle) be transformed into an equivalent conic optimization problem!

- ▶ Moreover, **there are powerful algorithms and solvers for (reasonably-sized) conic optimization problems** (e.g., interior-point methods and CVXPY).
- ▶ **This makes conic optimization problems very important!**

From Convex Optimization to Conic Programming II

We will now describe the transformation. Given the convex optimization problem

$$(P1) \quad \begin{aligned} & \min_{x \in \mathbb{R}^d} && f(x) \\ & \text{subject to} && x \in \mathcal{S} \end{aligned}$$

we first rewrite it into the form

$$(P2) \quad \begin{aligned} & \min_{x \in \mathbb{R}^d, t \in \mathbb{R}} && t \\ & \text{subject to} && x \in \mathcal{S} \\ & && (x, t) \in \text{epi } f \end{aligned}$$

Introducing the Cartesian product $\mathcal{S} \times \mathbb{R}$, which is convex since both \mathcal{S} and \mathbb{R} are, we can further rewrite the problem into

From Convex Optimization to Conic Programming III

$$(P3) \quad \begin{array}{ll} \min_{x \in \mathbb{R}^d, t \in \mathbb{R}} & t \\ \text{subject to} & (x, t) \in \mathcal{S} \times \mathbb{R} \\ & (x, t) \in \text{epi } f \end{array}$$

and by combining the two constraints on (x, t) , we get

$$(P4) \quad \begin{array}{ll} \min_{x \in \mathbb{R}^d, t \in \mathbb{R}} & t \\ \text{subject to} & (x, t) \in \underbrace{(\mathcal{S} \times \mathbb{R}) \cap \text{epi } f}_{\mathcal{W}} \end{array}$$

We now define the convex cone $\mathcal{Z} \subset \mathbb{R}^{d+2}$ via

$$\mathcal{Z} \stackrel{\text{def}}{=} \text{cone} \left(\{(x, t, 1) \in \mathbb{R}^{d+2} : (x, t) \in \mathcal{W}\} \right),$$

From Convex Optimization to Conic Programming IV

and the affine space

$$\mathcal{L} \stackrel{\text{def}}{=} \{(x, t, u) \in \mathbb{R}^{d+2} : u = 1\},$$

and notice that

$$\mathcal{W} \times \{1\} = \mathcal{Z} \cap \mathcal{L}.$$

So, we have the following equivalences:

$$(x, t) \in \mathcal{W} \Leftrightarrow (x, t, 1) \in \mathcal{W} \times \{1\} \Leftrightarrow (x, t, u) \in \mathcal{Z} \cap \mathcal{L}.$$

Therefore, (P4) can be rewritten into the form

$$(P5) \quad \begin{array}{ll} \min_{x \in \mathbb{R}^d, t \in \mathbb{R}, u \in \mathbb{R}} & t \\ \text{subject to} & (x, t, u) \in \mathcal{Z} \cap \mathcal{L} \end{array}$$

which can in turn be written as

From Convex Optimization to Conic Programming V

$$(P6) \quad \begin{array}{ll} \min_{x \in \mathbb{R}^d, t \in \mathbb{R}, u \in \mathbb{R}} & t \\ \text{subject to} & (x, t, u) \in \mathcal{Z} \\ & (x, t, u) \in \mathcal{L} \end{array}$$

Finally, notice that (P6) has a linear objective function (indeed, $t = 0x + 1t + 0u$), a conic constraint ($z = (x, t, u) \in \mathcal{Z}$), and an affine constraint ($z = (x, t, u) \in \mathcal{L}$) which can be written in the form $(0, 0, 1)z = 1$ and simply means $u = 1$.

Motivated by the universality of optimization problems with a linear objective functions, a conic and an affine constraint, we move on to our next topic: conic programming duality.

Conic Programming Duality

Conic Programming Duality I

We now consider the special case of Fenchel duality with

$$f(x) = \langle c, x \rangle + \delta_{\mathcal{L}}(x), \quad c \in \mathbb{R}^d, \quad \mathcal{L} \subseteq \mathbb{R}^d \text{ is a cone}, \quad (42)$$

$$g(y) = \delta_{\mathcal{M}}(y - b), \quad b \in \mathbb{R}^n, \quad \mathcal{M} \subseteq \mathbb{R}^n \text{ is a cone}. \quad (43)$$

The **primal conic optimization problem (Primal-Conic)** is

$$\begin{aligned} OPT_P &\stackrel{\text{def}}{=} \inf_{x \in \mathbb{R}^d} \{f(x) + g(\mathbf{A}x)\} \\ &\stackrel{(42)+(43)}{=} \inf_{x \in \mathbb{R}^d} \{\langle c, x \rangle + \delta_{\mathcal{L}}(x) + \delta_{\mathcal{M}}(\mathbf{A}x - b)\} \\ &= \inf_{x \in \mathbb{R}^d} \{\langle c, x \rangle : x \in \mathcal{L}, \mathbf{A}x - b \in \mathcal{M}\}. \end{aligned}$$

We can also write the primal problem in this form:

$$\begin{array}{ll} \inf & \langle c, x \rangle \\ \text{subject to} & \mathbf{A}x - b \in \mathcal{M} \\ & x \in \mathcal{L} \end{array} \quad (44)$$

Conic Programming Duality II

The **Fenchel dual** is

$$OPT_D \stackrel{\text{def}}{=} \sup_{y \in \mathbb{R}^n} \{-f^*(\mathbf{A}^\top y) - g^*(-y)\}.$$

We now need to calculate f^* and g^* .

Step 1: Calculating f^* . Since $f(x) = \langle c, x \rangle + \delta_{\mathcal{L}}(x)$, for any $x^* \in \mathbb{R}^d$ we have

$$\begin{aligned} f^*(x^*) &= \sup_{x \in \mathbb{R}^d} \{\langle x^*, x \rangle - f(x)\} &=& \sup_{x \in \mathbb{R}^d} \{\langle x^*, x \rangle - \langle c, x \rangle - \delta_{\mathcal{L}}(x)\} \\ &=& \sup_{x \in \mathbb{R}^d} \{\langle x^* - c, x \rangle - \delta_{\mathcal{L}}(x)\} \\ &=& \delta_{\mathcal{L}}^*(x^* - c) \\ &\stackrel{\text{Example 142}}{=}& \delta_{\mathcal{L}^-}(x^* - c), \end{aligned} \tag{45}$$

where \mathcal{L}^- is the polar cone to \mathcal{L} . So, the Fenchel conjugate of f evaluated at x^* is equal to the indicator function of the polar cone \mathcal{L}^- evaluated at $x^* - c$.

Conic Programming Duality III

Step 2: Calculating g^* . Since $g(y) = \delta_{\mathcal{M}}(y - b)$, for any $y^* \in \mathbb{R}^n$ we have

$$\begin{aligned} g^*(y^*) &= \sup_{y \in \mathbb{R}^n} \{ \langle y^*, y \rangle - g(y) \} \\ &= \sup_{y \in \mathbb{R}^n} \{ \langle y^*, y \rangle - \delta_{\mathcal{M}}(y - b) \} \\ &= \sup_{y \in \mathbb{R}^n} \{ \langle y^*, b \rangle + \langle y^*, y - b \rangle - \delta_{\mathcal{M}}(y - b) \} \\ &= \langle y^*, b \rangle + \sup_{y \in \mathbb{R}^n} \{ \langle y^*, y - b \rangle - \delta_{\mathcal{M}}(y - b) \} \\ &= \langle y^*, b \rangle + \sup_{y' \in \mathbb{R}^n} \{ \langle y^*, y' \rangle - \delta_{\mathcal{M}}(y') \} \\ &= \langle y^*, b \rangle + \delta_{\mathcal{M}}^*(y^*) \\ \text{Example 142} &\stackrel{=}{} \langle y^*, b \rangle + \delta_{\mathcal{M}^-}(y^*). \end{aligned} \tag{46}$$

Conic Programming Duality IV

Step 3. Putting them together. By using the expressions for f^* and g^* , the dual problem can be simplified as follows:

$$\begin{aligned} OPT_D &\stackrel{\text{def}}{=} \sup_{y \in \mathbb{R}^n} \{-f^*(\mathbf{A}^\top y) - g^*(-y)\} \\ &= \sup_{y \in \mathbb{R}^n} \{-\delta_{\mathcal{L}^-}(\mathbf{A}^\top y - c) - (\langle -y, b \rangle + \delta_{\mathcal{M}^-}(-y))\} \\ &\stackrel{(45)+(46)}{=} \sup_{y \in \mathbb{R}^n} \{\langle y, b \rangle - \delta_{\mathcal{L}^-}(\mathbf{A}^\top y - c) - \delta_{\mathcal{M}^-}(-y)\} \\ &= \sup_{y \in \mathbb{R}^n} \{\langle y, b \rangle : \mathbf{A}^\top y - c \in \mathcal{L}^-, \quad -y \in \mathcal{M}^-\} \\ &= \sup_{y \in \mathbb{R}^n} \{\langle y, b \rangle : \mathbf{A}^\top y - c \in \mathcal{L}^-, \quad y \in -\mathcal{M}^-\}. \end{aligned}$$

Conic Programming Duality V

We can also write the dual problem in this form:

$$\begin{array}{ll} \sup & \langle b, y \rangle \\ \text{subject to} & \mathbf{A}^T y - c \in \mathcal{L}^- \\ & y \in -\mathcal{M}^- \end{array} \quad (47)$$

Special Cases of Conic Optimization Problems

Three main classes:

1. **Linear programming** (we have seen this before!)
2. **Second-order cone programming** (this involves the Lorentz/second-order/ice cream cone)
3. **Semidefinite programming** (this involves the cone of positive semidefinite matrices)

Yet Another Duality Pair

Yet Another Duality Pair I

Example 164

Formulate the Fenchel primal and dual problems for the setup where f is a linear function, and g is the indicator function of the nonnegative orthant:

$$f(x) = \langle c, x \rangle, \quad c \in \mathbb{R}^d, \quad g(y) = \delta_{\mathbb{R}_+^n}(y). \quad (48)$$

Solution: The **primal problem (Primal)** becomes

$$\begin{aligned} OPT_P &\stackrel{\text{def}}{=} \inf_{x \in \mathbb{R}^d} \{f(x) + g(\mathbf{A}x)\} \\ &\stackrel{(48)}{=} \inf_{x \in \mathbb{R}^d} \{\langle c, x \rangle + \delta_{\mathbb{R}_+^n}(\mathbf{A}x)\} \\ &= \inf_{x \in \mathbb{R}^d} \{\langle c, x \rangle : \mathbf{A}x \geq 0\}. \end{aligned}$$

Yet Another Duality Pair II

Exercise 165 (☕☕)

The problem $\inf_{x \in \mathbb{R}^d} \{\langle c, x \rangle : Ax \geq 0\}$ can be further simplified. Can you think of a way how?

Yet Another Duality Pair III

The **dual problem (Dual)** becomes

$$\begin{aligned} OPT_D &\stackrel{\text{def}}{=} \sup_{y \in \mathbb{R}^d} \{-f^*(\mathbf{A}^\top y) - g^*(-y)\} \\ &\stackrel{(48)}{=} \sup_{y \in \mathbb{R}^d} \left\{ -\delta_{\{c\}}(\mathbf{A}^\top y) - \delta_{\mathbb{R}_+^n}^*(-y) \right\} \\ &= \sup_{y \in \mathbb{R}^d} \left\{ -\delta_{\mathbb{R}_+^n}^*(-y) : \mathbf{A}^\top y = c \right\} \\ \text{Example 141} &\stackrel{=}{=} \sup_{y \in \mathbb{R}^d} \left\{ -\delta_{\mathbb{R}_-^n}(-y) : \mathbf{A}^\top y = c \right\} \\ &= \sup_{y \in \mathbb{R}^d} \left\{ 0 : \mathbf{A}^\top y = c, y \in \mathbb{R}_+^n \right\} \\ &= \begin{cases} 0 & \text{if } \mathcal{Q} \neq \emptyset \\ -\infty & \text{if } \mathcal{Q} = \emptyset \end{cases}, \end{aligned}$$

where $\mathcal{Q} = \{y \in \mathbb{R}^n : \mathbf{A}^\top y = c, y \geq 0\}$.

Part II

Applications

Introduction to Optimization

Peter Richtárik



Lecture 13: Asset Pricing & Arbitrage Detection via Linear Programming - Part 1

LP Models: Asset Pricing and Arbitrage

Security: asset; something of financial value

Derivative security: a security whose value depends entirely on the value of an **underlying security** (e.g., stock, currency, bond)

Two uses of derivative securities:

- ▶ **Speculation:** Bet on ↑ or ↓ of a security
- ▶ **Hedging:** Reduction of risk in investor's overall position by forming a suitable portfolio of assets that are expected to have opposing risk.

Example 166

Investor holds a share of *XYZ* and is concerned that its price may fall.

Solution: Investor buys a **put option** on *XYZ* and hence protects herself against price falling below the strike price.

Arbitrage

Definition 167 (Type A)

A trading strategy that has

- ▶ **positive initial cash flow** and
- ▶ **no risk of a loss** in the future.

Definition 168 (Type B)

A trading strategy that has

- ▶ **nonnegative initial cash flow**,
- ▶ **no risk of a loss in the future** and
- ▶ **a positive probability of a profit in the future.**

Basic Asset Pricing Model

Model 169 (Asset Pricing)

1. **Securities:** $S^0, S^1, S^2, \dots, S^{n-1}$

2. **Modeling future:**

2.1 We consider **2 time periods**: “now” (time 0) and “future” (time 1)

2.2 States of the world at maturity (time 1) is described by **d scenarios**:

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_d\}$$

2.3 Each scenario occurs with nonzero probability

3. **Prices of securities S^i for $i = 1, \dots, n - 1$:**

3.1 $S_0^i = \text{price of } S^i \text{ at present (time 0)}$

3.2 $S_1^i(\omega_j) = \text{price of } S^i \text{ at maturity under } \omega_j$

4. **Risk-free security:**

4.1 S^0 is the risk-free security (cash)

4.2 $S_0^0 = 1$ (price of risk-free security at present is 1)

4.3 $S_1^0(\omega_j) = 1 + r = R$ for all j (we earn interest rate r on risk-free security; note that it is indeed risk free as its price at time 1 does not depend on ω_j !)

Exercise 170

How realistic is the model? Raise three pieces of criticism.

Type-A Arbitrage and Linear Inequalities

Consider a **portfolio** $y = (y_0, y_1, \dots, y_{n-1})$, where y_i is the amount (whether positive or negative) of security S^i .

- ▶ The **cost of portfolio y at time 0** can be expressed as

$$P_0 = P_0(y) \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} S_0^i y_i$$

- ▶ Both types of arbitrage require: **no risk of a loss in the future**.
This can be written as a system of linear constraints as follows:

$$\sum_{i=0}^{n-1} S_1^i(\omega_j) y_i \geq 0, \quad j = 1, 2, \dots, d. \quad (49)$$

So, **type-A arbitrage exists if and only if** there is portfolio y such that the following **linear inequalities** (in y) hold:

- ▶ $P_0(y) < 0$ and
- ▶ inequalities (49)

Arbitrage-Detecting Linear Program

Consider now the linear program (LP)

$$\begin{aligned} OPT &\stackrel{\text{def}}{=} \text{minimize} & \sum_{i=0}^{n-1} S_0^i y_i \\ &\text{subject to} & \sum_{i=0}^{n-1} S_1^i(\omega_j) y_i \geq 0, \quad j = 1, 2, \dots, d. \end{aligned} \tag{D'}$$

Define $OPT = -\infty$ if the problem is unbounded.

Exercise 171

- ▶ Consider the situation with $OPT = -\infty$.
 - ▶ What does this mean?
 - ▶ What does this say about the model?
- ▶ Would you invest in a portfolio with $P_0(y) > 0$?

Detecting Type-A Arbitrage

The following statement describes the connection between the existence of type-A arbitrage and properties of the linear program (D').

Fact 172

The following statements are equivalent:

- (a) *there is no type-A arbitrage,*
- (b) $OPT = 0,$
- (c) (D') *is bounded.*

Insights:

- ▶ To detect arbitrage: solve (D') and check whether $OPT = 0$ or not.
- ▶ If $OPT \neq 0$, we must have $OPT = -\infty$ (since by Fact 172, (D') must be unbounded!). So, there must exist a feasible portfolio y (i.e., one that does not pose any risk of a future loss) with **negative cost at time 0, i.e.,** $\sum_{i=0}^{n-1} S_0^i y_i < 0$
- ▶ LP solvers detect unboundedness by finding such y . **This y uncovers an arbitrage opportunity (of type-A)!**

Proof of Fact 172

- ▶ Observe that $OPT \leq 0$. Indeed, the empty portfolio ($y = 0$, i.e., we do not own any assets) is feasible and costs nothing at time 0.
- ▶ We now claim that either $OPT = 0$ or $OPT = -\infty$. Let us show this. Assume now that y is a feasible portfolio for which the cost of which at time 0 is negative and finite. Then the portfolio $2y$ is also be feasible, with a smaller cost at time 0. So, OPT cannot be negative *and* finite.
- ▶ In particular, the above observations tell us that $OPT < 0 \Leftrightarrow OPT = -\infty$. On the other hand, by definition, $OPT = -\infty \Leftrightarrow (D')$ is unbounded.
- ▶ Further, note that $OPT < 0$ if and only if there exists type-A arbitrage. Indeed, $OPT < 0$ means that there is a portfolio whose payoff at time 1 is nonnegative but whose cost at present is negative (i.e., it generates positive cashflow at time 0).

We have thus shown that

$$OPT = -\infty \Leftrightarrow \text{type-A arbitrage exists} \Leftrightarrow (D') \text{ is unbounded}$$

The logical negations of each of these three statements must also be equivalent, proving the statement (note we have shown that the logical negation of $OPT = -\infty$ is $OPT = 0$).

Introduction to Optimization

Peter Richtárik



Lecture 14: Asset Pricing & Arbitrage Detection via Linear Programming - Part 2

Type-B Arbitrage and “Linear Inequalities”

By definition, **type-B arbitrage** in our model **exists** if there exists portfolio

$$y = (y_0, \dots, y_{n-1})$$

satisfying the following three conditions:

(i)

$$P_0(y) \equiv \sum_{i=0}^{n-1} S_0^i y_i \leq 0, \quad (50)$$

(we do not lose anything at time 0)

(ii) system of inequalities (49)

(there is no risk of loss at time 1)

(iii)

$$\sum_{i=0}^{n-1} S_1^i(\omega_j) y_i > 0 \text{ for at least one index } j \in \{1, 2, \dots, d\} \quad (51)$$

(positive probability of positive payoff at time 1)

Detecting Type-B Arbitrage

Fact 173

Assume that no type-A arbitrage exists. Then the following statements are equivalent:

- (a) **no type-B arbitrage exists,**
- (b) *for any optimal solution y^* of (D') , all constraints of (D') are tight.* That is,

$$\sum_{i=0}^{n-1} S_1^i(\omega_j) y_i^* = 0 \quad \text{for all } j = 1, 2, \dots, d.$$

Insights:

- ▶ If there **is** type-A arbitrage, Fact 173 does not give us any information
- ▶ If there **isn't** type-A arbitrage (and we know how to detect it from Fact 172!), Fact 173 tells us how to detect type-B arbitrage: solve (D') , obtaining y^* , and test whether all constraints are tight or not.

Exercise 174 (☕☕)

Think about how you would detect type-B arbitrage in the case when there is type-A arbitrage.

Proof of Fact 173

Proof.

- ▶ Since we assume that no type-A arbitrage exists, from Fact 172 we know that $OPT = 0$.
- ▶ Assume (a) holds. This means that there cannot exist y satisfying (49), (50) and (51).
- ▶ Pick any y^* optimal for (D') . Then y^* clearly satisfies (49) (since it is feasible for (D')) and (50) (since $OPT = 0$). Hence, it must be the case that (51) does not hold for y^* .
- ▶ Since (49) holds, this means that (b) must hold.

We have shown that (a) implies (b). It remains to show, using similar arguments, that (b) implies (a).



Exercise 175 (☕☕)

Finish the proof of Fact 173. That is, show that if there is no type-A arbitrage, then (b) implies (a).

Rewriting (D') into Standard Form

If we multiply both the objective function and the constraints of (D') by -1 and observe that minimization of a function is the same as maximization of the negative of the function ($-\min f = \max(-f)$), we immediately observe that

$$\begin{aligned} -OPT &= \max \quad \sum_{i=0}^{n-1} (-S_0^i) y_i \\ \text{subject to} \quad &\sum_{i=0}^{n-1} (-S_1^i(\omega_j)) y_i \leq 0, \quad j = 1, 2, \dots, d. \end{aligned} \tag{D}$$

Note that (D) is a **standard dual problem**. Indeed, if we now let $b = (-S_0^0, -S_0^1, \dots, -S_0^{n-1})^\top \in \mathbb{R}^n$, $c = 0 \in \mathbb{R}^d$ and $\mathbf{A} \in \mathbb{R}^{n \times d}$ with $\mathbf{A}_{ij} = -S_1^i(\omega_j)$, then (D) can be written as

$$\begin{aligned} -OPT &= \max \quad b^\top y \\ \text{subject to} \quad &c - \mathbf{A}^\top y \geq 0. \end{aligned} \tag{D}$$

Problem Dual to (D)

The dual of (D) is the **standard primal problem**:

$$\begin{aligned} -OPT = \min \quad & c^\top x \\ \text{subject to} \quad & \mathbf{A}x = b \\ & x \geq 0, \end{aligned} \tag{P}$$

which, after substitution of $c = 0 \in \mathbb{R}^d$, $b \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{n \times d}$, can be written as

$$\begin{aligned} -OPT = \min \quad & \sum_{j=1}^d 0x_j \\ \text{subject to} \quad & \sum_{j=1}^d -S_1^i(\omega_j)x_j = -S_0^i, \quad i = 0, 1, \dots, n-1 \\ & x_j \geq 0, \quad j = 1, 2, \dots, d. \end{aligned} \tag{P}$$

- ▶ Observe that for any $x \in \mathbb{R}^d$ the value of the objective function of (P) is 0. Hence, all points x feasible for (P) are also optimal!
- ▶ (P) does not necessarily have to be feasible
- ▶ We would like to **understand the meaning of the dual solution x (if it exists!)**

Risk Neutral Probability Measure (RNPM)

Definition 176

A **risk neutral probability measure** on $\Omega = \{\omega_1, \dots, \omega_d\}$ is a vector $p = (p_1, \dots, p_d)^\top \in \mathbb{R}^d$ satisfying

1. $\sum_{j=1}^d p_j = 1; p_j > 0 \forall j$, (i.e., p is a vector of positive probabilities)
2. For all $i = 0, 1, \dots, n - 1$,

$$S_0^i = \frac{1}{R} \sum_{j=1}^d p_j S_1^i(\omega_j) \stackrel{\text{def}}{=} \frac{1}{R} \mathbf{E}[S_1^i]. \quad (52)$$

Insights:

- ▶ Given n, d, R, S_0^i and $S_1^i(\omega_j)$ for all $i = 0, 1, \dots, n - 1$ and $j = 1, \dots, d$, a RNPM does **not have to exist**.
- ▶ However, its existence is related (as we shall see, *equivalent*) to the non-existence of arbitrage.
- ▶ If RNPM exists, it **allows us to see current prices S_0^i as discounted expected values of future prices**, where the expectation is taken wrt the RNPM.

Risk Neutral Probability Measure: Further Insights

- ▶ Note that while p_j is **NOT EQUAL TO** the probability that ω_j happens, yet **it acts as if it was** (in the above sense)
- ▶ It is natural to require that if two securities, say S^i and S^k , have the same future payoffs ($S_1^i(\omega_j) = S_1^k(\omega_j)$ for all j), then they should have the same price at present: $S_0^i = S_0^k$. If RNPM exists, then (52) gives a formula for computing this price.
- ▶ RNPM can be used to **price new securities** using a **replication technique**.
- ▶ If a RNPM exists, as far as we can tell from the above, it **need not be unique**. However, by definition, (52) will hold for all such vectors p , so it does not matter which one is used in the calculation of the present price.

Exercise 177

Construct an example (i.e., find some concrete numbers for n , d , R , S_0^i and $S_1^i(\omega_j)$) such that a RNPM does not exist. Hint: Use $n = 2$, $d = 2$.

First Fundamental Theorem of Asset Pricing

Theorem 178

A *risk neutral probability measure exists if and only if there is no arbitrage.*

Proof.

\Leftarrow : assume that there is no arbitrage.

- ▶ In particular, since there is no type-A arbitrage, applying Fact 172 we know that (D') (and hence (D)) has optimal solution.
- ▶ Using strong duality, we conclude that (P) is feasible (since (D) has optimal solution).
- ▶ In particular, both (P) and (D) are feasible. We can thus apply Goldman-Tucker, which says that there exists strictly complementary optimal solutions x^* and y^* :

$$x^* + (c - \mathbf{A}^\top y^*) > 0. \quad (53)$$

- ▶ Now we utilize the assumption that there is no type-B arbitrage. Applying Fact 173, we get $c - \mathbf{A}^\top y^* = 0$. Plugging this into (53), we conclude that $x^* > 0$.

Proof: (Continued)

We now claim that the vector $p = (p_1, p_2, \dots, p_d)^\top$, where $p_j = Rx_j^*$, is the vector of risk-neutral probabilities. For this we need to check that p satisfied the three properties in the definition.

- ▶ First, clearly $p_j > 0$ for all $j \in \{1, 2, \dots, d\}$ (since $x_j^* > 0$ for all j).
- ▶ Second, note that the first constraint (i.e., the one corresponding to $i = 0$) in (P) looks as follows:

$$\sum_{j=1}^d \underbrace{S_1^0(\omega_j)}_{=R} x_j^* = S_0^0 = 1,$$

which implies that $\sum_{j=1}^d p_j = 1$.

- ▶ Finally, let us look at the i -th constraint in (P) (for any i):

$$\sum_{j=1}^d S_1^i(\omega_j) x_j^* = S_0^i.$$

It can be written as

$$S_0^i = \frac{1}{R} \sum_{j=1}^d S_1^i(\omega_j) Rx_j^* = \frac{1}{R} \sum_{j=1}^d S_1^i(\omega_j) p_j,$$

recovering (52). Hence, p is indeed a RNPM on Ω .

Introduction to Optimization

Peter Richtárik



Lecture 15: Asset Pricing & Arbitrage Detection via Linear Programming - Part 3

Lecture Outline

- ▶ Examples: arbitrage and RNPM
- ▶ Pricing securities via replication

Arbitrage Detection: Example

Example 179 (Arbitrage detection)

A bank accepts deposits yielding an interest rate of 5%. Also, it lends money at the same interest rate. You have an opportunity to invest for 1 year into a government bond that costs \$500, yielding a 10% interest.

1. Is there a type-A arbitrage opportunity?
2. Is there a type-B arbitrage opportunity?
3. Write down a linear program for detecting type-A arbitrage.

Solution: The “situation” is easy: one should “borrow from bank and buy bonds!”. However, this simple problem lets us understand the concepts discussed before. Note that this situation fits Model 169:

- ▶ There are 2 assets only: S^0 (cash/bank) and S^1 (bond); so $n = 2$
- ▶ There are just 2 time periods: now (time 0) and a year from now (time 1)
- ▶ Both bank and government bond positions can be seen as risk-free; and hence we only have one scenario: $\Omega = \{\omega_1\}$; so $d = 1$
- ▶ cash: $S_0^0 = 1$, $R = 1.05$, $S_1^0(\omega_1) = S_0^0 \times R = 1.05$
- ▶ bond: $S_0^1 = 500$, $S_1^1(\omega_1) = S_0^1 \times 1.1 = 550$

We will build a portfolio $y = (y_0, y_1)^\top$.

Strategy 1

Some things we can do:

We can borrow \$1,047.619 from the bank ($\mathbf{y}_0 = -1,047.619$) and use this to buy 2 bonds ($\mathbf{y}_1 = 2$).

- ▶ At time 0, the incurred cost of building our portfolio (i.e., net cash outflow) is

$$P_0(\mathbf{y}) = \sum_{i=0}^{n-1} S_0^i \mathbf{y}_i = S_0^0 \mathbf{y}_0 + S_0^1 \mathbf{y}_1 = 1 \times (-1,047.619) + 500 \times 2 = -47.619$$

- ▶ At time 1, we have only one possible outcome (ω_1), and the cost (=value) of our portfolio (we own it now, so positive value means we are good!) is:

$$\sum_{i=0}^{n-1} S_1^i(\omega_j) \mathbf{y}_i = 1.05 \times (-1,047.619) + 550 \times 2 = 0$$

This is **type-A arbitrage!**

Strategy 2

Do the same as in Strategy 1, except scale your position \mathbf{y} by constant $c > 0$. That is, choose: $\mathbf{y}_0 = c \times (-1, 047.619)$ (borrow c times more cash) and $\mathbf{y}_1 = c \times 2$ (buy c times as many bonds as before).

- ▶ At time 0, the incurred cost of building our portfolio (i.e., net cash outflow) is

$$P_0(\mathbf{y}) = \sum_{i=0}^{n-1} S_0^i \mathbf{y}_i = S_0^0 \mathbf{y}_0 + S_0^1 \mathbf{y}_1 = c \times (-47.619)$$

- ▶ At time 1, we have only one possible outcome (ω_1), and the cost (=value) of our portfolio (we own it now, so positive value means we are good!) is:

$$\sum_{i=0}^{n-1} S_1^i(\omega_1) \mathbf{y}_i = c \times 0$$

So we earn c times as much as with Strategy 1! For $c \rightarrow \infty$, we earn as much as we want. (recall the criticism of Model 169)

Detecting type-A Arbitrage by LP

Recall that the LP (D') for detecting type-A arbitrage has the form:

$$\begin{aligned} \text{minimize } P_0(\mathbf{y}) &\stackrel{\text{def}}{=} \sum_{i=0}^{n-1} S_0^i \mathbf{y}_i \\ \text{subject to } & \sum_{i=0}^{n-1} S_1^i(\omega_j) \mathbf{y}_i \geq 0, \quad j = 1, 2, \dots, d. \end{aligned}$$

In our exercise ($n = 2$, $S_0^0 = 1$, $S_0^1(\omega_1) = 500$, $d = 1$, $S_1^0(\omega_1) = 1.05$ and $S_1^1(\omega_1) = 550$); this takes the form:

$$\begin{aligned} \text{minimize } & \mathbf{y}_0 + 500\mathbf{y}_1 \\ \text{subject to } & 1.05\mathbf{y}_0 + 550\mathbf{y}_1 \geq 0 \end{aligned}$$

- ▶ By solving this LP in Python using CVXPY, we find that it is **unbounded**. Hence, by Fact 172, there is type-A arbitrage.
- ▶ Following the thought process of Strategy 2 we find a feasible portfolio $c\mathbf{y}$ for all $c > 0$ such that $P_0(c\mathbf{y}) \rightarrow -\infty$ as $c \rightarrow \infty$. This is a “proof by hand” that the LP is unbounded (and hence that type-A arbitrage exists).

Strategy 3

We can borrow \$1,000 from the bank ($\mathbf{y}_0 = -1,000$) and use this to buy 2 bonds ($\mathbf{y}_1 = 2$).

- ▶ At time 0, the incurred cost of building our portfolio (i.e., net cash outflow) is

$$P_0(\mathbf{y}) = \sum_{i=0}^{n-1} S_0^i \mathbf{y}_i = S_0^0 \mathbf{y}_0 + S_0^1 \mathbf{y}_1 = 1 \times -1,000 + 500 \times 2 = 0$$

- ▶ At time 1, we have only one possible outcome (ω_1), and the cost (=value) of our portfolio (we own it now, so positive value means we are good!) is:

$$\sum_{i=0}^{n-1} S_1^i(\omega_j) \mathbf{y}_i = 1.05 \times (-1,000) + 550 \times 2 = 50$$

This is **type-B arbitrage!**

Risk Neutral Probability Measure: Example

Example 180 (RNPM)

Show directly from definition that in the previous exercise there is no risk neutral probability measure (RNPM). Then argue this via the theory developed before.

Solution. Since $d = 1$ (we have one scenario only), the only possible probability measure on Ω is $p_1 = 1$. Let us check whether it satisfies the conditions in the definition of a RNPM:

1. $\sum_{j=1}^d p_j = 1$ and $p_j > 0$ for all j trivially holds
2. We further need the following identities to hold:

$$S_0^i = \frac{1}{R} \sum_{j=1}^d p_j S_1^i(\omega_j), \quad i = 0, 1.$$

For $i = 0$ this is $1 = \frac{1}{1.05} p_1 1.05$, which holds. For $i = 1$ this is $500 = \frac{1}{1.05} p_1 550$, which does *not* hold. We conclude that a RNPM does not exist.

Moreover, we reach the same conclusion by applying the Fundamental Theorem of Asset Pricing and the result of the previous Exercise (which says that arbitrage exists).

Pricing of Securities via Replication: Binomial Model

Consider Model 169, with $d = 2$ scenarios ω_1 ("UP"), ω_2 ("DOWN") and 3 securities (i.e., $n = 3$):

- ▶ S^0 : **risk-free security/cash** (yielding interest $R = 1 + r$)
- ▶ S^1 : **underlying security**,
- ▶ S^2 : **derivative security**.

Assume $U > D > 0$ are constants such that

$$S_1^1 = \begin{cases} U \times S_0^1, & \text{under } \omega_1 \text{ ("UP")} \\ D \times S_0^1, & \text{under } \omega_2 \text{ ("DOWN")}. \end{cases}$$

Typically, the price of the **derivative security** at maturity (time 1) is a function of the price of the **underlying security** at maturity:

$$S_1^2 = f(S_1^1).$$

So, if S_1^1 is known, so is S_1^2 . Assuming there is no arbitrage, find a formula for the spot price (i.e., price at time 0) S_0^1 of the **derivative security**.

That is, **price the derivative security**.

Pricing of Securities via Replication (Cont.)

Approach: We will try to form a portfolio $y = (\mathbf{y}_0, \mathbf{y}_1)$ of **cash (S^0)** and the **underlying (S^1)** such that its price at time 1 matches the price of the **derivative security**, under both scenarios:

$$P_1(y, \omega_j) \stackrel{\text{def}}{=} \mathbf{S}_1^0(\omega_j)\mathbf{y}_0 + \mathbf{S}_1^1(\omega_j)\mathbf{y}_1 = \mathbf{S}_1^2(\omega_j), \quad j = 1, 2.$$

- ▶ The above is a system of 2 linear equations in 2 unknowns ($\mathbf{y}_0, \mathbf{y}_1$).
- ▶ The solution is:

$$\mathbf{y}_0 = \frac{\mathbf{S}_1^2(\omega_1) - \mathbf{S}_1^2(\omega_2)}{\mathbf{S}_0^1(U - D)}, \quad \mathbf{y}_1 = \frac{U\mathbf{S}_1^2(\omega_2) - D\mathbf{S}_1^2(\omega_1)}{R(U - D)}, \quad (54)$$

- ▶ We then say that since the portfolio has the same future behavior/payoff/value as the **derivative security S^2** , the **spot price of the derivative security** must be equal to the spot price of the portfolio:

$$\mathbf{S}_0^2 = P_0(y) = \mathbf{S}_0^0\mathbf{y}_0 + \mathbf{S}_0^1\mathbf{y}_1. \quad (55)$$

Pricing of Securities via Replication (Cont.)

- ▶ By plugging (54) into (55), we obtain:

$$S_0^2 = \frac{1}{R} \left[\underbrace{\frac{R - D}{U - D}}_{\stackrel{\text{def}}{=} p_1} S_2^2(\omega_1) + \underbrace{\frac{U - R}{U - D}}_{\stackrel{\text{def}}{=} p_2} S_1^2(\omega_2) \right]. \quad (56)$$

- ▶ If $D < R < U$, then $p = (p_1, p_2)$, where p_1, p_2 are defined in (56), is a **risk neutral probability measure (RNPM)**!
 - ▶ Indeed, then i) $p_1 + p_2 = 1$, $p_1 > 0$, $p_2 > 0$, whereas (56) says that the spot value (S_0^2) of the **derivative security** is equal to its present (i.e., discounted) expected value, where expectation is taken with respect to the RNPM.
 - ▶ However, we still need to check that (56) holds for securities S^0 and S^1 as well.

Exercise 181

Check that (56) holds for securities S^0 and S^1 as well, thus completing the verification of the fact that p is a RNPM.

Pricing of Securities via Replication (Cont.)

- ▶ We have worked with 2 scenarios and 3 securities because we can deal with this by hand.
- ▶ However, the technique can easily be adapted to Model 169 with arbitrary finite number of scenarios ($\omega_1, \dots, \omega_d$) and n securities S^0, \dots, S^{n-1} . In this case, the computations must be done by a computer (e.g., Python via CVXPY).

Introduction to Optimization

Peter Richtárik



Lecture 16: Asset Pricing & Arbitrage Detection via Linear Programming - Part 4

Lecture Outline

Arbitrage detection:

- ▶ Derivative securities written on the same underlying with the same maturity
- ▶ New model without scenarios!
- ▶ Solution via LP
- ▶ Solution without optimization

Detecting Arbitrage in a Portfolio of Derivative Securities

We now introduce a **new model**.

- ▶ We **lift the assumption of a finite state/scenario space**
 $\Omega = \{\omega_1, \dots, \omega_d\}$.
 - ▶ **fidelity challenge:** finite # of scenarios might not describe reality well
→ push for a large # of scenarios
 - ▶ **computational challenge:** the more scenarios one has, the harder it is to solve the LP
 - ▶ **generation challenge:** one has to generate the scenarios somehow in the first place: difficult!
- ▶ The price we pay (for being able to work with an infinite number of scenarios) is that we now only consider a portfolio of
 - ▶ **derivative securities,**
 - ▶ written on the **same underlying**,
 - ▶ with the **same maturity**.

Derivative Securities: Options

An **option** is

- ▶ the right to buy (**call option**) or sell (**put option**) an underlying security (e.g., bond, stock, commodity)
- ▶ at a certain price (**strike**)
 - ▶ any price in $[0, T = \text{maturity}/\text{expiration date}]$ (lookback option)
 - ▶ average price in $[0, T]$ (Asian option)
 - ▶ price at T (European and American options)
- ▶ in a certain time frame
 - ▶ at any point before expiration date: $[0, T]$ (American option)
 - ▶ at **expiration date T (maturity)** (European option)

European Call and Put Options

EU Call Option: The right to **buy** a certain **underlying security** at expiration date at an agreed (strike) price.

$$\underbrace{C_1}_{\text{option payoff at maturity}} = (\underbrace{S_1}_{\text{security value at maturity}} - \text{strike})^+ \stackrel{\text{def}}{=} \max\{S_1 - \text{strike}, 0\}$$

EU Put Option: The right to **sell** a certain **underlying security** at expiration date at an agreed (strike) price.

$$\underbrace{C_1}_{\text{option payoff at maturity}} = (\text{strike} - \underbrace{S_1}_{\text{security value at maturity}})^+ \stackrel{\text{def}}{=} \max\{\text{strike} - S_1, 0\}$$

New Model: Portfolio of Derivative Securities

Model 182 (Portfolio of Derivative Securities)

1. A single **underlying security**:

- ▶ S_0 = price of underlying at time 0 (spot price)
- ▶ S_1 = price of underlying at time 1 (price at maturity)

2. S^1, S^2, \dots, S^n = derivative securities (DS) written on the **underlying security**

- ▶ S_0^i = price of DS S^i at time 0
- ▶ S_1^i = price of DS S^i at time 1

3. Assume

$$S_1^i = \Psi_i(S_1), \quad \text{where}$$

- ▶ Ψ_i is **piece-wise linear**
- ▶ with 2 pieces
- ▶ “joined” at x-axis point K_i

Example 183

$$\Psi_i(S_1) = (S_1 - K_i)^+ = \max\{S_1 - K_i, 0\} \cdots \text{EU call option with strike } K_i$$

$$\Psi_i(S_1) = (K_i - S_1)^+ = \max\{K_i - S_1, 0\} \cdots \text{EU put option with strike } K_i$$

Portfolio of Derivative Securities: Cost, Payoff & Arbitrage

Consider portfolio $x = (x_1, \dots, x_n)$ of derivative securities S^1, \dots, S^n

Cost of portfolio x at time 0:

$$P_0^x \stackrel{\text{def}}{=} \sum_{j=1}^n S_0^j x_j$$

Payoff of portfolio x at maturity (time 1):

$$P_1^x(S_1) \stackrel{\text{def}}{=} \sum_{j=1}^n \psi_j(S_1) x_j$$

To detect type-A arbitrage, we form an **optimization problem** seeking

- ▶ the cheapest portfolio (i.e., we hope for negative cost!, that is, positive profit, at time 0)
- ▶ whose payoff at maturity is **always nonnegative**: $P_1^x(S_1) \geq 0$ for all $S_1 \geq 0$ (i.e., no risk of loss in the future)

Detecting Type-A Arbitrage: 1st Optimization Formulation

We are seeking portfolio $x \in \mathbb{R}^n$ solving the following optimization problem:

$$\text{minimize } P_0^x \quad \text{subject to } P_1^x(S_1) \geq 0 \quad \text{for all } S_1 \geq 0 \quad (57)$$

Comments:

- ▶ Note that $x \rightarrow P_0^x$ and $x \rightarrow P_1^x(S_1)$ are linear functions of x
- ▶ As before, type-A arbitrage corresponds to the problem being unbounded ($OPT = -\infty$). No-type-A arbitrage corresponds to $OPT = 0$
- ▶ **New difficulty: We have infinitely many constraints!** (one constraint for every $S_1 \geq 0$)
- ▶ Due to this difficulty, (57) **is not a linear program!**
- ▶ LP solvers (e.g., CVXPY) are hence unable to solve this in this form

Detecting Type-A Arbitrage: Reformulating Constraints

It turns out that **the feasible set of (57) can be equivalently described by a finite set of linear constraints! So, we can rewrite (57) as an LP after all!**

Theorem 184

Assume the values $\{K_j\}_{j=1}^n$ are ordered as follows:

$0 \leq K_1 \leq K_2 \leq \dots \leq K_n$. The statement

$$P_1^x(S_1) \geq 0 \quad \text{for all } S_1 \geq 0$$

is equivalent to the following three conditions:

1. $P_1^x(0) \geq 0$
2. $P_1^x(K_j) \geq 0$ for all $j = 1, 2, \dots, n$
3. The right-derivative of $s \rightarrow P_1^x(s)$ at $s = K_n$ is nonnegative

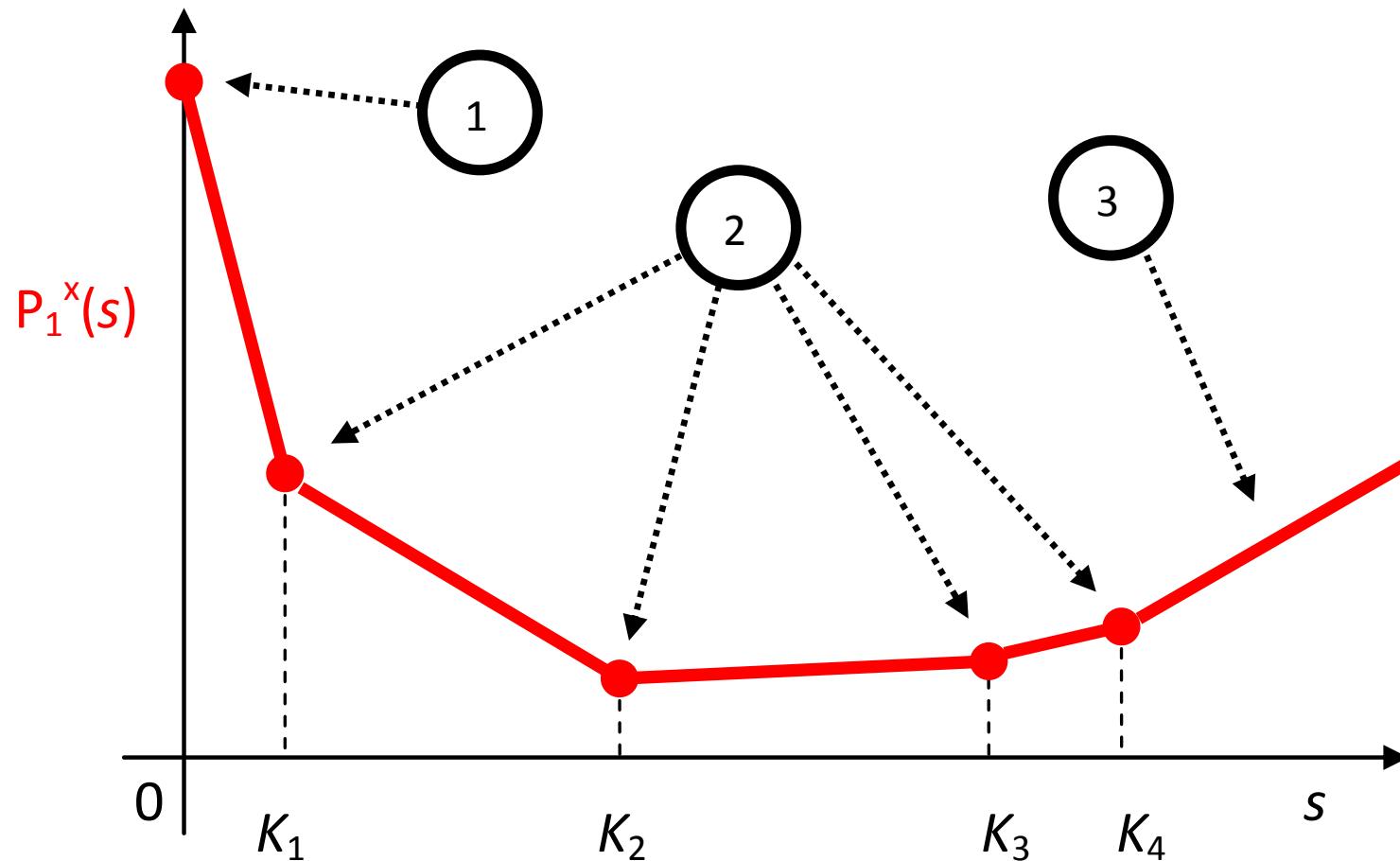
Moreover, the right derivative is equal to $\sum_{j=1}^n (\Psi_j(K_n + 1) - \Psi_j(K_n)) x_j$

Proof.

Note that for fixed x , $s \rightarrow P_1^x(s) = \sum_{j=1}^n \Psi_j(s)x_j$ is a piece-wise linear function, since $s \rightarrow \Psi_j(s)$, $j = 1, \dots, n$, are piece-wise linear functions.

The rest is a “proof by picture” (see next slide: plot of $s \rightarrow P_1^x(s)$ for a single fixed x).

Reformulating Constraints: Proof by Picture



Remark: The picture shows the situation with $n = 4$. Clearly, the argument is general and holds for any n .

Detecting Type-A Arbitrage: 2nd Optim. Formulation

Based on the reformulation, we can write (57) as a **linear program** as follows:

$$\begin{aligned} & \text{minimize} && \sum_{j=1}^n S_0^j x_j \\ & \text{subject to} && \sum_{j=1}^n \Psi_j(0)x_j \geq 0 \\ & && \sum_{j=1}^n \Psi_j(K_i)x_j \geq 0, \quad i = 1, 2, \dots, n \\ & && \sum_{j=1}^n (\Psi_j(K_n + 1) - \Psi_j(K_n)) x_j \geq 0 \end{aligned} \tag{58}$$

Special Case: Detecting Arbitrage in EU Call Options

Assume all derivative securities S^1, \dots, S^n are EU Call Options (on the same underlying, with the same maturity) with strike prices $K_1 < K_2 < \dots < K_n$, respectively. Then

$$\Psi_j(S_1) = (S_1 - K_j)^+ = \max\{S_1 - K_j, 0\}$$

$$\Psi_j(K_i) = (K_i - K_j)^+ = \begin{cases} K_i - K_j, & \text{if } i > j, \\ 0, & \text{otherwise.} \end{cases}$$

Then (58) can be written as

$$\min c^\top x \quad \text{subject to} \quad \mathbf{A}x \geq 0, \quad (59)$$

where $c = (S_0^1, S_0^2, \dots, S_0^n)^\top \in \mathbb{R}^n$ and

$$\mathbf{A} = \begin{pmatrix} K_2 - K_1 & 0 & 0 & \cdots & 0 \\ K_3 - K_1 & K_3 - K_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ K_n - K_1 & K_n - K_2 & K_n - K_3 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

Detecting Arbitrage in EU Call Options: Main Theorem

Theorem 185

Consider a collection of EU Call Options S^1, \dots, S^n with strike prices $K_1 < K_2 < \dots < K_n$, respectively, written on the same underlying security, with the same maturity.

Then there is no arbitrage if and only if the following three conditions hold:

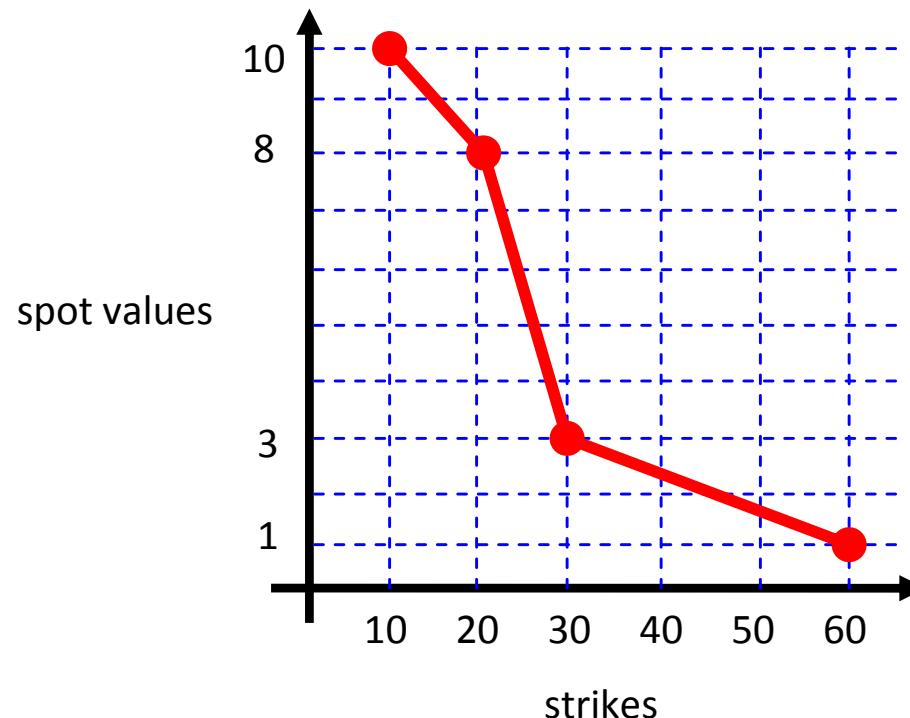
1. $S_0^j > 0$ for all $j = 1, 2, \dots, n$
2. $S_0^j > S_0^{j+1}$ for all $j = 1, 2, \dots, n - 1$
3. The piece-wise linear function mapping K_j to S_0^j , for $j = 1, \dots, n$, is **strictly convex**. That is, the slopes of its linear pieces are **strictly increasing**.

Example

Example 186

Let $n = 4$ and $S_0^1 = 10$, $S_0^2 = 8$, $S_0^3 = 3$, $S_0^4 = 1$ and $K_1 = 10$, $K_2 = 20$, $K_3 = 30$, $K_4 = 60$. Does there exist arbitrage?

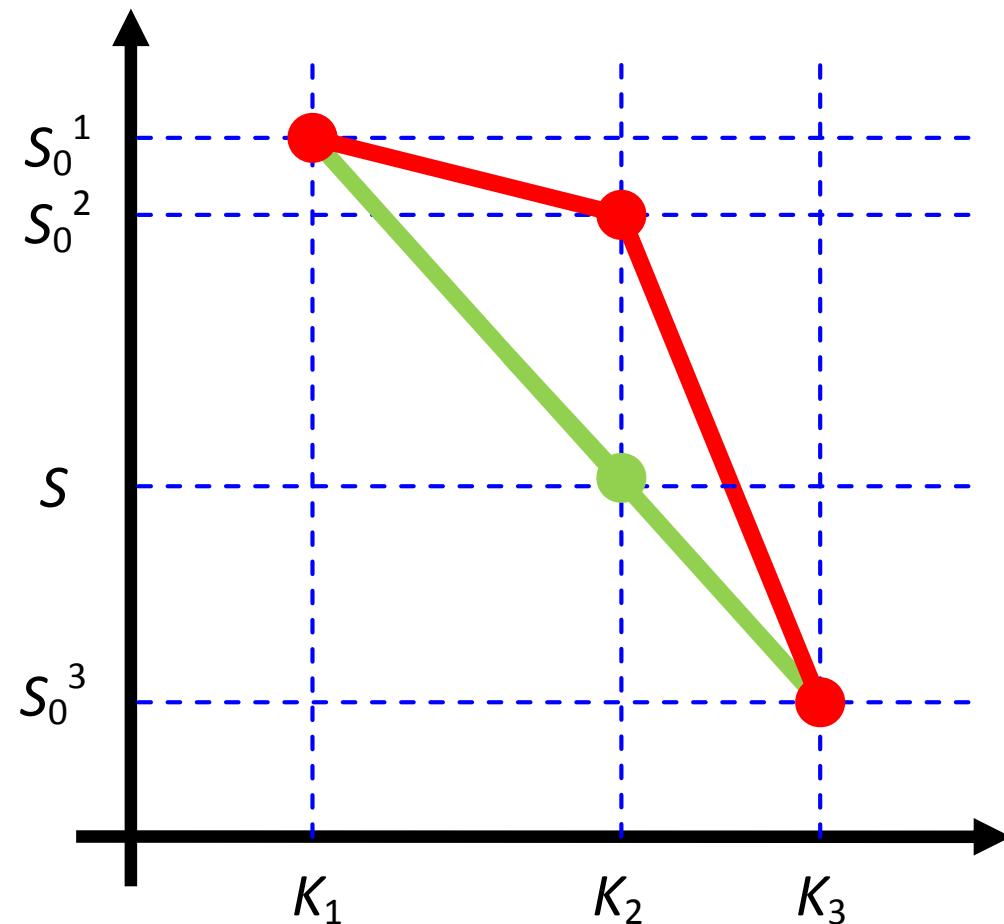
Solution: Properties 1) and 2) are satisfied. However, property 3) is not; since the slopes are: -2 , -5 and $-2/3$, but it is not true that $-2 < -5 < -2/3$, and hence there is arbitrage:



How to find portfolio x detecting arbitrage opportunity?

How to Find Arbitrage-Detecting Portfolio x ? I

Consider a situation like on this picture, i.e., successive slopes are **not strictly increasing**:



How to Find Arbitrage-Detecting Portfolio x ? II

Let $0 < \alpha < 1$ be chosen so that

$$K_2 = \alpha K_1 + (1 - \alpha) K_3. \quad (60)$$

Consider a portfolio of $x_1 = \alpha$ units of S^1 and $x_3 = 1 - \alpha$ units of S^3 .

- ▶ For this portfolio we pay at time 0 less than for 1 unit of S^2 :

$$\underbrace{x_1 S_0^1 + x_3 S_0^3}_{\text{cost of portfolio } x \text{ at time 0}} \stackrel{(60)}{=} S < \underbrace{S_0^2}_{\text{cost of } S^2 \text{ at time 0}}$$

- ▶ Recall that for a scalar $s \in \mathbb{R}$, we write $s^+ \stackrel{\text{def}}{=} \max\{0, s\}$. It can be easily verified that for $t \geq 0$, and $a, b \in \mathbb{R}$ we have:

$$t(a^+) = (ta)^+, \quad a^+ + b^+ \geq (a + b)^+.$$

How to Find Arbitrage-Detecting Portfolio x ? III

- ▶ Therefore:

$$\begin{aligned} P_1^x(S_1) &= (S_1 - K_1)^+ x_1 + (S_1 - K_3)^+ x_3 \\ &= (x_1(S_1 - K_1))^+ + (x_3(S_1 - K_3))^+ \\ &\geq (x_1(S_1 - K_1) + x_3(S_1 - K_3))^+ \stackrel{(60)}{=} (S_1 - K_2)^+. \end{aligned}$$

So, the payoff of the portfolio cannot be worse than the payoff of S^2 .

- ▶ This means, using standard arguments, that there is an arbitrage opportunity: **short-sell S^2 and invest the cash into the portfolio!**

Introduction to Optimization

Peter Richtárik



Part 17: Image Manipulation

Lecture Outline

- ▶ Image inpainting
- ▶ Image denoising

Image Inpainting

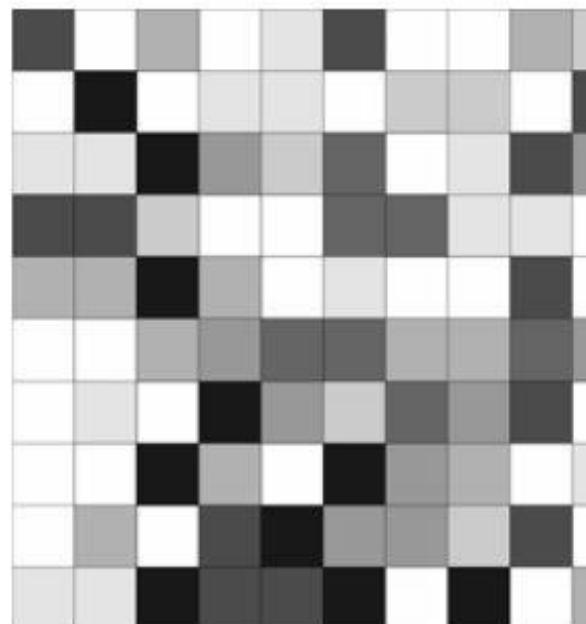
Images as Matrices

An 8-bit grayscale image can be represented as a matrix

$$\mathbf{X} \in \mathbb{R}^{m \times n}$$

whose entries between 0 (black) and 255 (white), i.e.,

$$X_{ij} \in \{0, 1, \dots, 255\} \quad \text{for all } (i, j) \in \mathcal{K}^{\text{all}} \stackrel{\text{def}}{=} \{1, \dots, m\} \times \{1, \dots, n\}.$$



254	107
255	165

Image Inpainting II

Original Image



Recovered Image



Corrupted Image

Difference Image

Image Inpainting II

Image inpainting

We work with an 8-bit grayscale image $\mathbf{X}^{\text{original}} \in \mathbb{R}^{m \times n}$ under the following conditions:

- ▶ The pixels $(i, j) \in \mathcal{K}^{\text{all}}$ are partitioned into two disjoint parts:

$$\mathcal{K}^{\text{all}} = \mathcal{K}^{\text{original}} \cup \mathcal{K}^{\text{corrupted}}, \quad \mathcal{K}^{\text{original}} \cap \mathcal{K}^{\text{corrupted}} = \emptyset.$$

- ▶ The pixels $(i, j) \in \mathcal{K}^{\text{corrupted}}$ are **corrupted** (e.g., not known at all, overwritten by text, damaged via scratches), i.e., we do not know their true values.
- ▶ The pixels $(i, j) \in \mathcal{K}^{\text{original}}$ are **known** (i.e., uncorrupted).

The goal of image inpainting is to **reconstruct the original image** by finding the values of the corrupted pixels, i.e., those with indices (i, j) in $\mathcal{K}^{\text{corrupted}}$.

Total Variation I

Definition 187 (Total variation of a matrix)

The total variation of a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ is the quantity

$$\begin{aligned}\text{TV}(\mathbf{X}) &\stackrel{\text{def}}{=} \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \left\| \begin{pmatrix} \mathbf{X}_{i+1,j} - \mathbf{X}_{i,j} \\ \mathbf{X}_{i,j+1} - \mathbf{X}_{i,j} \end{pmatrix} \right\| \\ &\stackrel{(3)}{=} \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \sqrt{(\mathbf{X}_{i+1,j} - \mathbf{X}_{i,j})^2 + (\mathbf{X}_{i,j+1} - \mathbf{X}_{i,j})^2}.\end{aligned}$$

Remarks:

- ▶ **TV(\mathbf{X}) does not depend on $\mathbf{X}_{m,n}$.**
- ▶ $\text{TV}(\mathbf{X}) = 0$ if and only if all entries of \mathbf{X} are identical, with the exception of $\mathbf{X}_{m,n}$, which is allowed to have arbitrary value.
- ▶ **The function $\mathbf{X} \mapsto \text{TV}(\mathbf{X})$ is convex.**

Exercise 188

Explain why $\mathbf{X} \mapsto \text{TV}(\mathbf{X})$ is a convex function.

Total Variation II

Example 189 (Total variation)

Let $m = 2$ and $n = 3$, and

$$\mathbf{X} = \begin{pmatrix} 5 & 1 & 3 \\ 2 & 0 & 4 \end{pmatrix}.$$

Then

$$\begin{aligned}\text{TV}(\mathbf{X}) &= \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \left\| \begin{pmatrix} \mathbf{x}_{i+1,j} - \mathbf{x}_{i,j} \\ \mathbf{x}_{i,j+1} - \mathbf{x}_{i,j} \end{pmatrix} \right\| = \sum_{i=1}^1 \sum_{j=1}^2 \left\| \begin{pmatrix} \mathbf{x}_{i+1,j} - \mathbf{x}_{i,j} \\ \mathbf{x}_{i,j+1} - \mathbf{x}_{i,j} \end{pmatrix} \right\| \\ &= \sum_{j=1}^2 \left\| \begin{pmatrix} \mathbf{x}_{2,j} - \mathbf{x}_{1,j} \\ \mathbf{x}_{1,j+1} - \mathbf{x}_{1,j} \end{pmatrix} \right\| = \left\| \begin{pmatrix} \mathbf{x}_{2,1} - \mathbf{x}_{1,1} \\ \mathbf{x}_{1,2} - \mathbf{x}_{1,1} \end{pmatrix} \right\| + \left\| \begin{pmatrix} \mathbf{x}_{2,2} - \mathbf{x}_{1,2} \\ \mathbf{x}_{1,3} - \mathbf{x}_{1,2} \end{pmatrix} \right\| \\ &= \left\| \begin{pmatrix} 2 - 5 \\ 1 - 5 \end{pmatrix} \right\| + \left\| \begin{pmatrix} 0 - 1 \\ 3 - 1 \end{pmatrix} \right\| = \left\| \begin{pmatrix} -3 \\ -4 \end{pmatrix} \right\| + \left\| \begin{pmatrix} -1 \\ 2 \end{pmatrix} \right\| \\ &\stackrel{(3)}{=} \sqrt{(-3)^2 + (-4)^2} + \sqrt{(-1)^2 + 2^2} = \sqrt{25} + \sqrt{5} = 5 + \sqrt{5}.\end{aligned}$$

Total Variation III

Example 190 (Total variation)

Let $m = 2$ and $n = 3$, and

$$\mathbf{X} = \begin{pmatrix} 6 & 6 & 6 \\ 6 & 6 & 2^{100} \end{pmatrix}.$$

Then

$$\begin{aligned} \text{TV}(\mathbf{X}) &= \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \left\| \begin{pmatrix} \mathbf{x}_{i+1,j} - \mathbf{x}_{i,j} \\ \mathbf{x}_{i,j+1} - \mathbf{x}_{i,j} \end{pmatrix} \right\| = \sum_{i=1}^1 \sum_{j=1}^2 \left\| \begin{pmatrix} \mathbf{x}_{i+1,j} - \mathbf{x}_{i,j} \\ \mathbf{x}_{i,j+1} - \mathbf{x}_{i,j} \end{pmatrix} \right\| \\ &= \sum_{j=1}^2 \left\| \begin{pmatrix} \mathbf{x}_{2,j} - \mathbf{x}_{1,j} \\ \mathbf{x}_{1,j+1} - \mathbf{x}_{1,j} \end{pmatrix} \right\| = \left\| \begin{pmatrix} \mathbf{x}_{2,1} - \mathbf{x}_{1,1} \\ \mathbf{x}_{1,2} - \mathbf{x}_{1,1} \end{pmatrix} \right\| + \left\| \begin{pmatrix} \mathbf{x}_{2,2} - \mathbf{x}_{1,2} \\ \mathbf{x}_{1,3} - \mathbf{x}_{1,2} \end{pmatrix} \right\| \\ &= \left\| \begin{pmatrix} 6 - 6 \\ 6 - 6 \end{pmatrix} \right\| + \left\| \begin{pmatrix} 6 - 6 \\ 6 - 6 \end{pmatrix} \right\| = \left\| \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\| + \left\| \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\| \\ &= 0. \end{aligned}$$

Total Variation Inpainting

It turns out that a surprisingly good reconstruction of the original image $\mathbf{X}^{\text{original}}$ can be found by searching for a matrix \mathbf{X} whose total variation is as small as possible, subject to the constraint that the known pixel values are preserved:

$$\begin{aligned} & \min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \quad \text{TV}(\mathbf{X}) \\ \text{subject to} \quad & \mathbf{X}_{i,j} = \mathbf{X}_{i,j}^{\text{original}} \quad \text{for all } (i,j) \in \mathcal{K}^{\text{original}} \end{aligned}$$

This is an **optimization problem involving a nonsmooth convex objective function subject to linear constraints!**

Image Denoising

Image Denoising I

Image denoising

We have access to a **noisy version** of **image** $\mathbf{X}^{\text{original}} \in \mathbb{R}^{m \times n}$:

$$\mathbf{x}^{\text{noisy}} = \mathbf{X}^{\text{original}} + \Sigma,$$

where $\Sigma \in \mathbb{R}^{m \times n}$ is Gaussian noise.

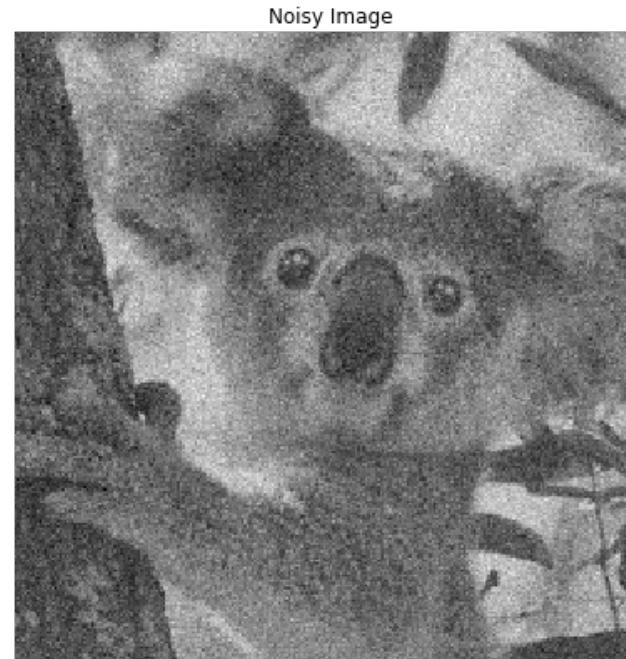


Image Denoising II

Besides being useful to perform image inpainting, total variation can be used for image denoising as well.

Total variation image denoising

Solve the optimization problem

$$\min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \|\mathbf{X} - \mathbf{X}^{\text{noisy}}\|_F^2 + \lambda \cdot \text{TV}(\mathbf{X}),$$

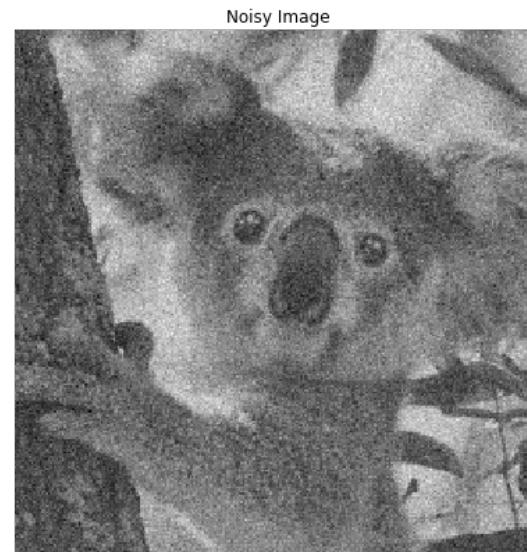
where

- ▶ $\lambda > 0$ is a positive regularization constant controlling the strength of denoising, and
- ▶ $\|\cdot\|_F$ denotes the **Frobenius matrix norm** defined via

$$\|\mathbf{X}\|_F \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^m \sum_{j=1}^n \mathbf{X}_{i,j}^2}.$$

Image Denoising III

Result of total variation image denoising with $\lambda = 10^{-3}$:



Introduction to Optimization

Peter Richtárik



Part 18: Truss Topology Design - Part 1

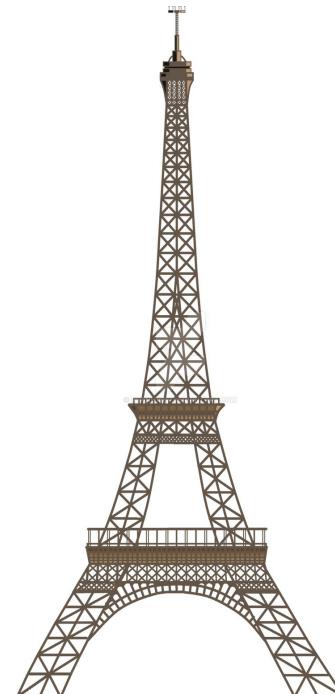
Lecture Outline

- ▶ truss
- ▶ truss topology design (TTD)
- ▶ the physics of TTD

Truss I

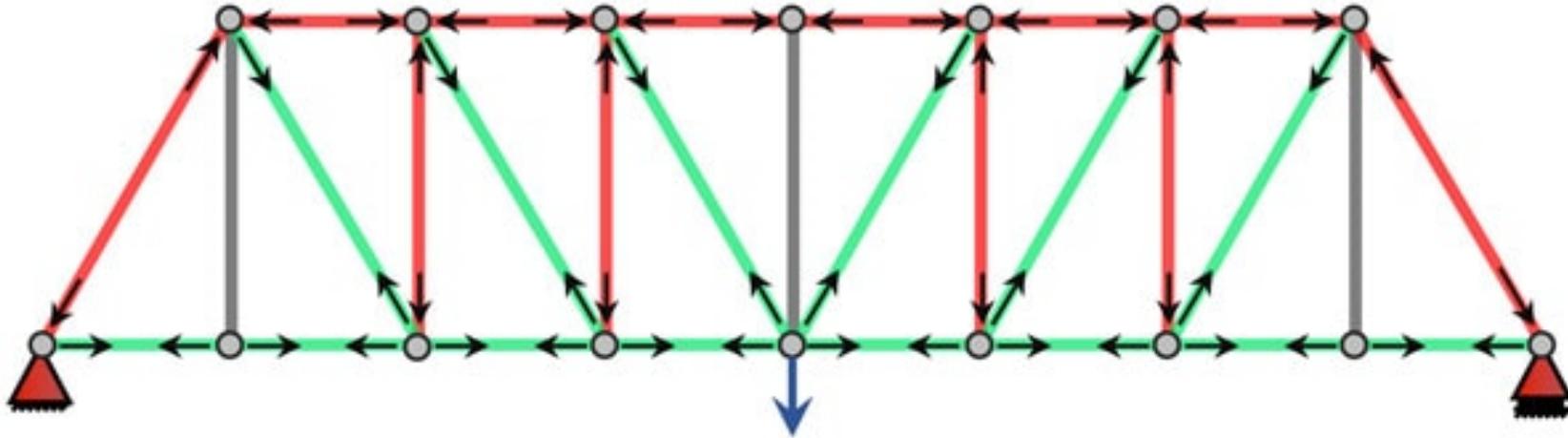
Definition 191 (Truss)

A **truss** is a mechanical construction comprising of thin **elastic bars** linked to each other at **nodes**.



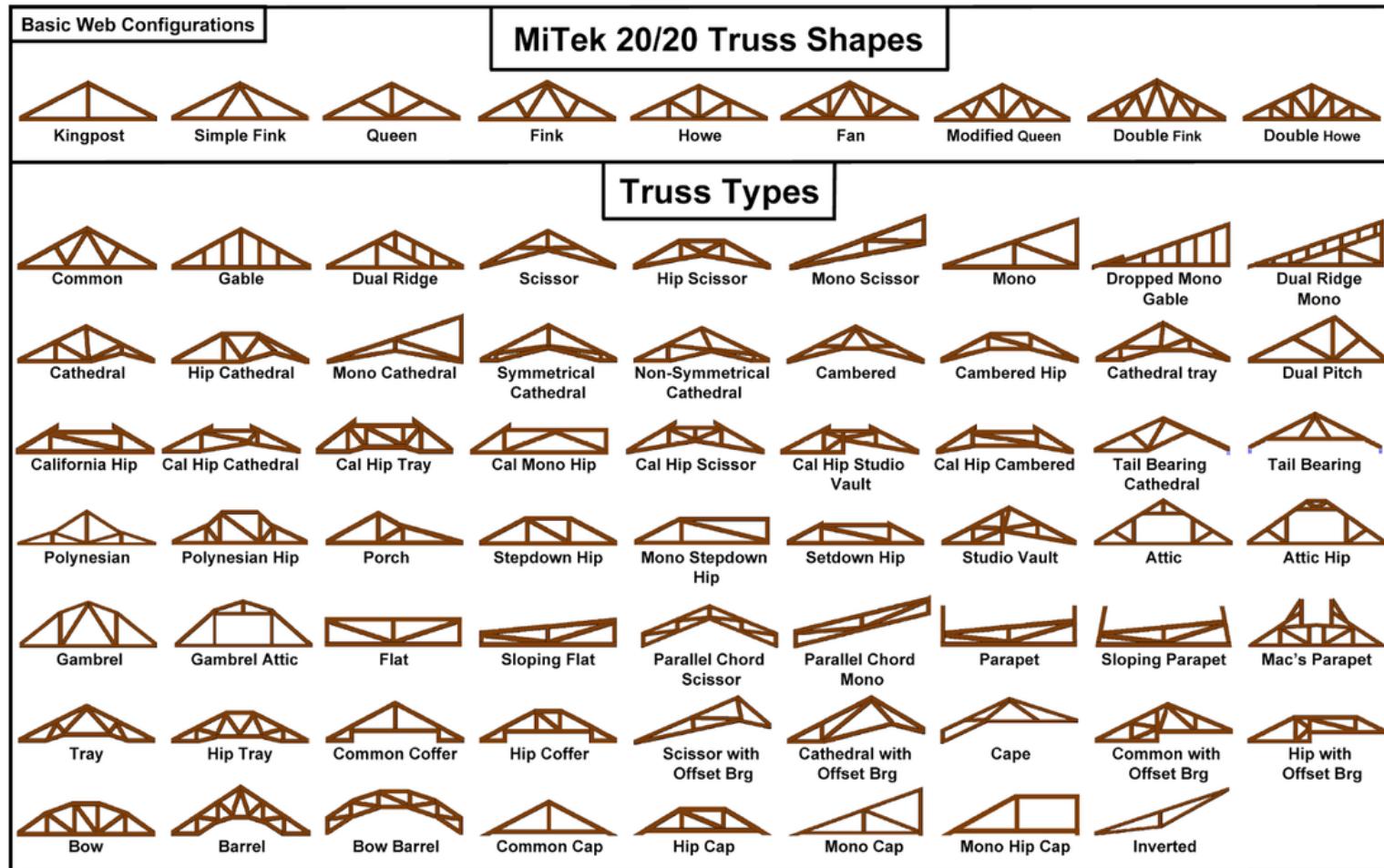
Examples of trusses: electric mast, Eiffel tower, railroad bridge, ...

Truss II



Bridge truss under a load.

Truss III



TTD: Truss Topology Design I

Truss under a load

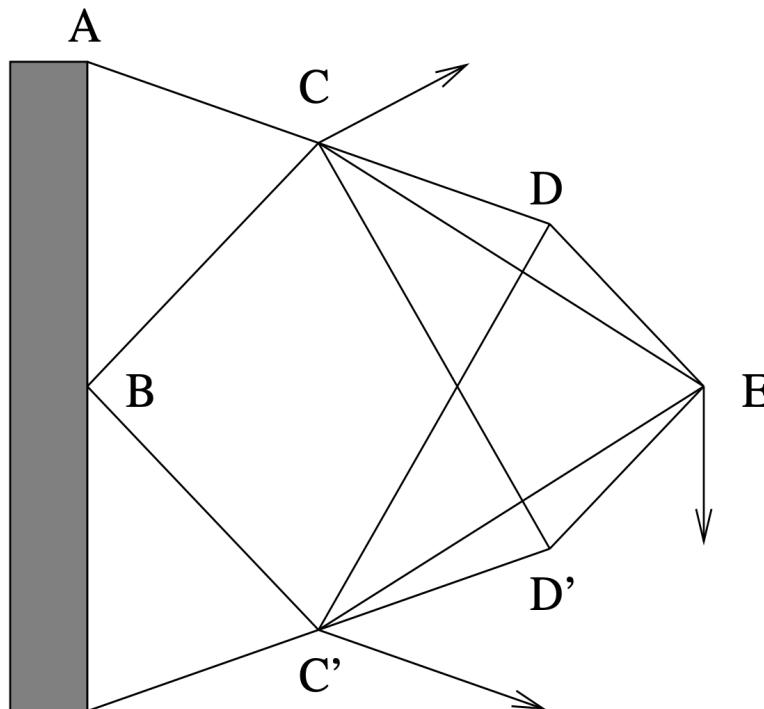
- ▶ A truss can be subjected to an **external load** – a collection of simultaneous forces acting at the nodes.
- ▶ Under a load, the construction **deforms** a bit, until the tensions caused by the deformation compensate the external forces.
- ▶ When deformed, the **truss stores certain potential energy**; this energy is called the **compliance** of the truss with respect to the load.
- ▶ **The less the compliance, the more rigid the truss with respect to the load in question.**

TTD: Truss Topology Design II

Input of a TTD problem

- ▶ a set of **nodes**, which is a finite set of points on the plane \mathbb{R}^2 or in the space \mathbb{R}^3 where the bars of the truss to be designed can be linked,
- ▶ **boundary conditions** specifying the nodes that are supported and cannot move (like nodes A , B , A' on the wall in figure below)
- ▶ a **load**, which is a collection of external forces acting at the nodes.

TTD: Truss Topology Design III



A'

nodes: A,A',B,C,C',D,D',E

bars: AC,A'C',BC,BC',CD,CD',C'D',C'D,CE,C'E,DE,D'E

forces: arrows

TTD: Truss Topology Design IV

The goal of truss topology design is to

- ▶ **design a truss of a given total weight**
- ▶ **best able to withstand the given load**, i.e., to link some pairs of the nodes by bars of appropriate sizes, not exceeding a given total weight,
- ▶ in such a way that the **compliance of the resulting truss with respect to the load of interest will be as small as possible.**

An attractive feature of the TTD problem is that although it seems to deal with the size (weights) of the bars only, **it finds the geometric shape (layout) of the truss as well!**

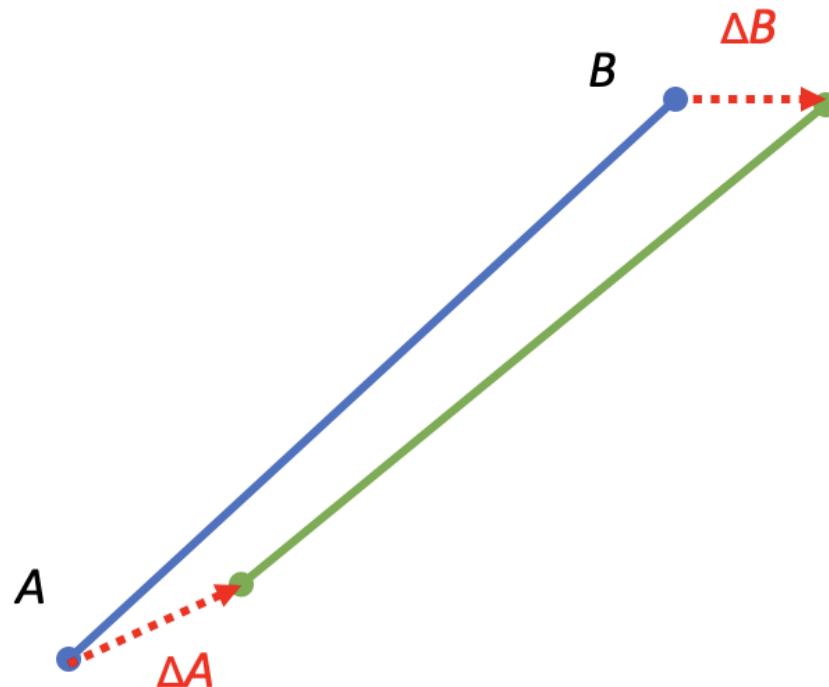
Indeed, we may start with a dense grid of nodes and allow all pairs of nodes to be connected by bars. In the optimal truss, yielded by the optimization process, some of the bars (typically the majority of them) will get zero weights. In other words, the optimization problem will by itself decide which nodes to use and how to link them, i.e., it will find both the optimal pattern (topology) of the construction and the optimal sizing.

The Physics of Truss Topology Design

The Physics of Truss Topology Design (TTD) I

To pose the TTD problem as an optimization problem, let us look in more detail at **what happens with a truss under a load:**

- ▶ Consider a particular bar AB in the unloaded truss.
- ▶ After the load is applied, the nodes A and B move a little bit, as shown in the figure below. This movement is called **displacement**.



The Physics of Truss Topology Design (TTD) II

Assuming the nodal displacements $\Delta A \in \mathbb{R}^d$ and $\Delta B \in \mathbb{R}^d$ (where $d \in \{2, 3\}$) to be small and neglecting the second order terms, the **elongation ℓ_{AB} of the bar AB under the load** is equal to the projection of the vector $\Delta B - \Delta A$ on the direction of the bar:

$$\ell_{AB} \stackrel{\text{def}}{=} \frac{\langle \Delta B - \Delta A, B - A \rangle}{\|B - A\|} \in \mathbb{R}. \quad (61)$$

By **Hooke's law³**, the **tension** (i.e., the magnitude of the reaction force) caused by this elongation, which we will denote τ_{AB} , satisfies

$$\tau_{AB} = \kappa \frac{\ell_{AB} \times S_{AB}}{\|B - A\|} = \kappa \frac{\ell_{AB} \times t_{AB}}{\|B - A\|^2}, \quad (62)$$

where

- ▶ κ is a **characteristic of the material** (Young's modulus⁴),
- ▶ S_{AB} is the **cross-sectional area of the bar AB** , and
- ▶ $t_{AB} = S_{AB} \|B - A\|$ is the **volume of the bar AB** .

The Physics of Truss Topology Design (TTD) III

By combining (62) and (61), the tension equals

$$\tau_{AB} \stackrel{(62)+(61)}{=} \kappa t_{AB} \frac{\langle \Delta B - \Delta A, B - A \rangle}{\|B - A\|^3}. \quad (63)$$

The **reaction force at point B associated with the tension is the vector $F_B \in \mathbb{R}^d$** (where $d \in \{2, 3\}$) given by

$$\begin{aligned} F_B &\stackrel{\text{def}}{=} -\tau_{AB} \frac{B - A}{\|B - A\|} \\ &\stackrel{(63)}{=} -\kappa t_{AB} \frac{\langle \Delta B - \Delta A, B - A \rangle}{\|B - A\|^4} (B - A) \\ &= -t_{AB} \langle \Delta B - \Delta A, \beta_{AB} \rangle \beta_{AB}, \end{aligned} \quad (64)$$

where

$$\beta_{AB} \stackrel{\text{def}}{=} \sqrt{\kappa} \frac{B - A}{\|B - A\|^2} \in \mathbb{R}^d. \quad (65)$$

Note that the vector β_{AB} depends on the positions of the nodes linked by the bar and is independent of the load and of the design.

The Physics of Truss Topology Design (TTD) IV

Now let us look at the **potential energy stored by our bar AB as a result of its elongation.**

Mechanics says that this energy is the half-product of the tension and the elongation, i.e., it is

$$E_{AB} = \frac{\tau_{AB} \times \ell_{AB}}{2}$$
$$\stackrel{(61)+(63)+(65)}{=} \frac{t_{AB}}{2} \langle \Delta B - \Delta A, \beta_{AB} \rangle^2. \quad (66)$$

³Hooke's law is the law of elasticity discovered by the English scientist Robert Hooke in 1660, which states that, for relatively small deformations of an object, the displacement or size of the deformation is directly proportional to the deforming force or load. Under these conditions the object returns to its original shape and size upon removal of the load.

⁴Young's modulus is a property of the material that tells us how easily it can stretch and deform and is defined as the ratio of tensile stress to tensile strain. Stress is the amount of force applied per unit area and strain is extension per unit length.

Introduction to Optimization

Peter Richtárik



Part 19: Truss Topology Design - Part 2

Lecture Outline

- ▶ TTD: mathematical model
- ▶ TTD via convex optimization

Truss Topology Design: Mathematical Model

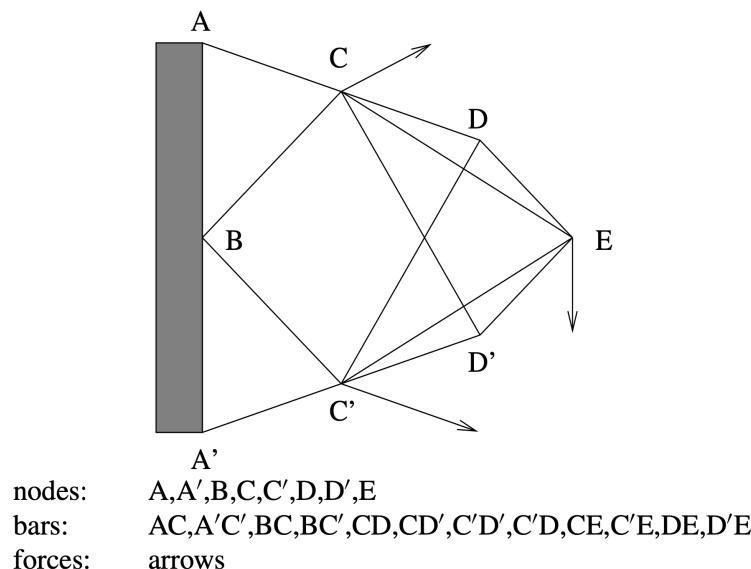
TTD Model I

Here is how the mathematical model of truss topology design (TTD) is built:

1. **The number m of free nodes:** Let M be the number of nodes in the grid of nodes and M_f be the number of the **free nodes** — those that are not fixed by the **boundary conditions**.

Example 192

In the picture below, all 8 nodes are free except for A , B and A' , which are attached to a wall. So, $M = 8$ and $M_f = 5$.



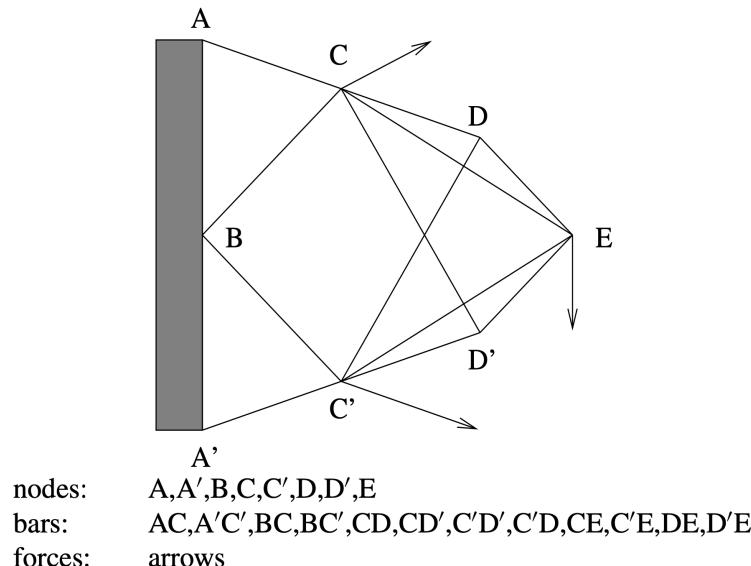
TTD Model II

2. **Tentative bars:** Let n be the number of tentative bars (i.e., pair connections between distinct nodes from the grid, at least one node in the pair being free).

Example 193

In the picture below, there are $n = 12$ tentative bars:

$AC, CD, DE, ED', D'C', C'A', BC, BC', CE, C'E, CD', C'D$. Notice that we do not consider the bar AD' , for example.



TTD Model III

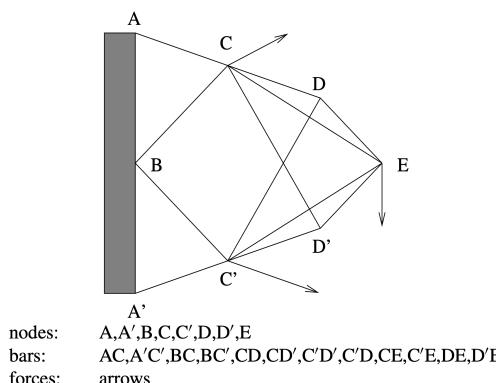
3. **Space \mathbb{R}^m of virtual displacements:** We define the space \mathbb{R}^m of **virtual displacements** of the construction as the product of the spaces of displacements of the free nodes:

$$\mathbb{R}^m = \underbrace{\mathbb{R}^d \times \cdots \times \mathbb{R}^d}_{M_f \text{ times}}.$$

Thus, $m = dM_f$, where $d = 2$ or $d = 3$, depending on whether we are speaking about planar (i.e., $d = 2$) or spatial (i.e., $d = 3$) trusses.

Example 194

In the picture below, we have a planar (i.e., $d = 2$) truss, and the five nodes C, C', D, D', E are free. So, $M_f = 5$ and $m = 2M_f = 10$.



TTD Model IV

4. **Load $f \in \mathbb{R}^m$ of external forces acting at the free nodes:** a collection of external forces acting at the free nodes. This collection can be represented by a vector $f \in \mathbb{R}^m$. For every free node $j \in \{1, \dots, M_f\}$, the corresponding subvector $f[j] \in \mathbb{R}^d$ of f is the external force acting at node j , where $d \in \{2, 3\}$. Recall that $m = dM_f$.
5. **Displacement vector $v \in \mathbb{R}^m$:** The vector v from \mathbb{R}^m represents a displacement of the grid of free nodes: the subvector $v[j] \in \mathbb{R}^d$ corresponding to a free node $j \in \{1, \dots, M_f\}$ represents the “physical” d -dimensional ($d = 2$ in case of a planar truss and $d = 3$ in case of a spatial truss) displacement of the node j . So, $v[j] \in \mathbb{R}^d$ where $d \in \{2, 3\}$.
6. **Volume of the tentative bar i :** A particular truss can be identified with a nonnegative vector

$$t = (t_1, \dots, t_n) \in \mathbb{R}_+^n,$$

where t_i is the volume of bar i in the truss. Alternatively, we write $t_i = t_{A_i B_i}$, where A_i, B_i are the nodes identifying bar i .

TTD Model V

7. **Vector $b_i \in \mathbb{R}^m$ associated with a tentative bar i :** Let us somehow order all our n tentative bars and consider the i^{th} of them. This bar links two nodes $A_i \in \mathbb{R}^d$ and $B_i \in \mathbb{R}^d$ where $d \in \{2, 3\}$. Let us associate with tentative bar i a vector $b_i \in \mathbb{R}^m$ as follows:

$$b_i[j] \stackrel{\text{def}}{=} \begin{cases} \beta_{A_i; B_i} \in \mathbb{R}^d & j = B_i \text{ and } j \text{ is free} \\ -\beta_{A_i; B_i} \in \mathbb{R}^d & j = A_i \text{ and } j \text{ is free} \\ 0 \in \mathbb{R}^d & \text{otherwise,} \end{cases} \quad (67)$$

where $d \in \{2, 3\}$.

Remark: Note that **vectors b_1, \dots, b_n are independent of the load $f \in \mathbb{R}^m$ and the truss volumes $t \in \mathbb{R}_+^n$** . They merely depend on the discretization of the plane/space and on the partition of the nodes into those that are free and those that are fixed (i.e., bound by boundary conditions).

TTD Model VI

8. **Reaction forces:** Consider a truss $t \in \mathbb{R}_+^m$, and let us look at the reaction forces caused by a displacement $v \in \mathbb{R}^m$ of the n nodes of the truss.

8.1 From (64) and (67) it follows that for every free node $j \in \{1, \dots, M_f\}$, the component of the reaction force caused, under the displacement v , by the i^{th} bar at the node j is

$$-t_i(b_i^\top v)b_i[j].$$

8.2 Consequently, the total reaction force at the node $j \in \{1, \dots, M_f\}$ is

$$F_{\text{react}}[j] \stackrel{\text{def}}{=} -\sum_{i=1}^n t_i(b_i^\top v)b_i[j]$$

and the vector of all reaction forces at all nodes is

$$F_{\text{react}} \stackrel{\text{def}}{=} -\sum_{i=1}^n t_i(b_i^\top v)b_i = -\underbrace{\left(\sum_{i=1}^n t_i b_i b_i^\top \right)}_{A(t)} v.$$

TTD Model VII

9. **Stiffness matrix:** The matrix

$$\mathbf{A}(t) \stackrel{\text{def}}{=} \sum_{i=1}^n t_i b_i b_i^\top$$

is called the **stiffness matrix**. It is an $m \times m$ symmetric matrix which depends linearly on the design variables — the volumes t_i of tentative bars.

10. **No rigid body motion assumption:** We'll need a technical assumption: The vectors $b_1, \dots, b_n \in \mathbb{R}^m$ span the space \mathbb{R}^m , i.e.,

$$\text{linear}(b_1, \dots, b_n) = \mathbb{R}^m.$$

This assumption prevents rigid body motions of the set of nodes.

TTD Model VIII

11. **Equilibrium:** At equilibrium, the reaction forces must compensate the external ones, which gives us a system of linear equations determining the displacement of the truss under an external load f :

$$\mathbf{A}(t)\mathbf{v} = \mathbf{f}. \quad (68)$$

If the linear system (68) does not have a solution, then the truss is not strong enough to carry the load. It will break.

TTD Model IX

12. **Compliance:** The **compliance of truss** $t \in \mathbb{R}^n$ **under load** $f \in \mathbb{R}^m$ is the total potential energy stored in the bars at equilibrium after deformation. It can be computed as follows:

$$\begin{aligned} E &\stackrel{\text{def}}{=} \sum_{i=1}^n E_{A_i B_i} \stackrel{(66)}{=} \sum_{i=1}^n \frac{\tau_{A_i B_i} \times \ell_{A_i B_i}}{2} \stackrel{(66)}{=} \sum_{i=1}^n \frac{t_{A_i B_i}}{2} \langle \Delta B_i - \Delta A_i, \beta_{A_i B_i} \rangle^2 \\ &\stackrel{(67)+(*)}{=} \frac{1}{2} \sum_{i=1}^n t_i (\nu^\top b_i)^2 \\ &= \frac{1}{2} \nu^\top \left(\sum_{i=1}^n t_i b_i b_i^\top \right) \nu \\ &= \frac{1}{2} \nu^\top \mathbf{A}(t) \nu \stackrel{(68)}{=} \frac{1}{2} f^\top \nu. \quad (69) \end{aligned}$$

Remarks:

- ▶ In step (*) we write t_i (instead of $t_{A_i B_i}$) to denote the volume of bar $A_i B_i$, which we also refer to as bar i .
- ▶ Compliance is one half of the mechanical work performed by the external load on the displacement of the truss until equilibrium.

Truss Topology Design: Optimization Model

TTD: Optimization Model I

Model 195 (TTD: Optimization Model)

Given

- ▶ *n (number of tentative bars),*
- ▶ *m (number of free nodes),*
- ▶ *vectors $b_1, \dots, b_n \in \mathbb{R}^m$,*
- ▶ *nonzero load $f \in \mathbb{R}^m$ (force acting on the free nodes), and*
- ▶ *a budget for the total volume of all bars $w > 0$,*

find truss $t \in \mathbb{R}_+^n$ (i.e., find volumes of all tentative bars) which minimizes the compliance with respect to the load f , and satisfies the total volume budget/constraint

$$\sum_{i=1}^n t_i \leq w.$$

TTD: Optimization Model II

That is,

$$\begin{array}{ll}\min_{t \in \mathbb{R}^n, v \in \mathbb{R}^m} & \frac{1}{2} v^\top \mathbf{A}(t) v \\ \text{subject to} & \mathbf{A}(t)v = f \\ & \sum_{i=1}^n t_i \leq w \\ & t \geq 0\end{array}$$

- ▶ Note that compliance is a function of both $t \in \mathbb{R}^n$ (the volumes of the n tentative bars) and $v \in \mathbb{R}^m$ the displacements of the nodes.
Neither t nor v is known a-priori, the optimization problem seeks to find both of these vectors! The objective function is convex in $t \in \mathbb{R}^n$ alone, and it is convex in $v \in \mathbb{R}^m$ alone, but it is not jointly convex in $(t, v) \in \mathbb{R}^n \times \mathbb{R}^m$.

TTD: Optimization Model III

- ▶ The constraints

$$\sum_{i=1}^n t_i \leq w, \quad t \geq 0$$

are linear in t .

- ▶ The equilibrium constraint

$$\mathbf{A}(t)v = f$$

is linear in t alone, and it is linear in v alone, but it is not linear jointly in (t, v) .

- ▶ As a result, **this looks like a difficult nonconvex problem.**

Truss Topology Design via Convex Optimization

TTD as a Convex Optimization Problem I

Solve the optimization problem

$$\begin{aligned} \min_{q \in \mathbb{R}^n} \quad & \sum_{i=1}^n |q_i| \\ \text{subject to} \quad & \sum_{i=1}^n q_i b_i = f, \end{aligned} \tag{70}$$

where $b_1, \dots, b_n, f \in \mathbb{R}^m$ are from Model 195 of the truss topology design problem.

Remarks:

- ▶ The objective function is convex; it is equal to the L_1 norm of vector $q \in \mathbb{R}^n$: $\|q\|_1 \stackrel{\text{def}}{=} \sum_{i=1}^n |q_i|$.
- ▶ The constraints are linear. Since $f \in \mathbb{R}^m$, there are m constraints,
- ▶ So, this is a convex optimization problem.

TTD as a Convex Optimization Problem II

The key theorem of this lecture:

Theorem 196 (Computing the optimal truss)

Assume the load $f \in \mathbb{R}^m$ to be nonzero. Given a solution $q^* \in \mathbb{R}^n$ of the convex optimization problem (70), the **optimal truss**

$$t^* = (t_1^*, \dots, t_n^*) \in \mathbb{R}_+^m$$

w.r.t. Model 195 (i.e., a truss minimizing the compliance $\frac{1}{2}f^\top v$ whose total volume does not exceed w) is given by

$$t_i^* = \frac{w}{w^*} |q_i^*|, \quad i = 1, 2, \dots, n,$$

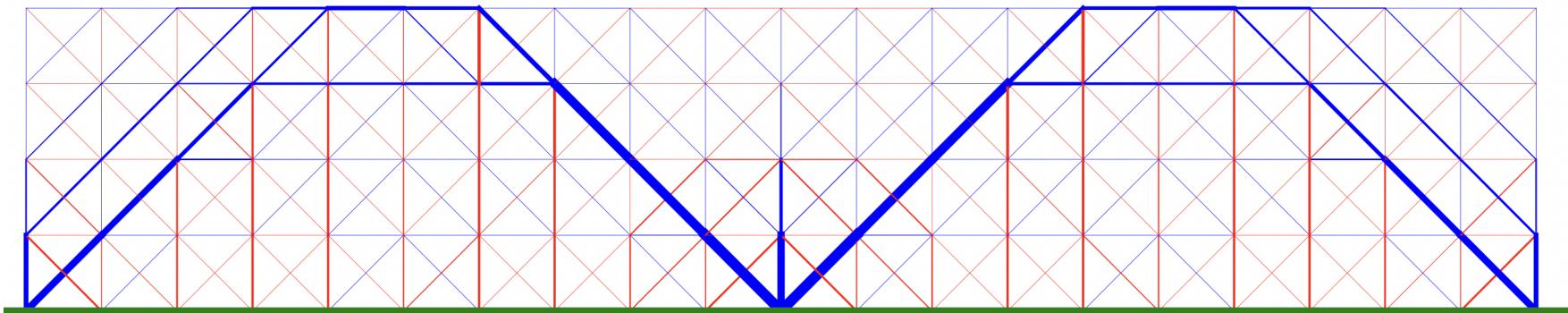
where

$$w^* \stackrel{\text{def}}{=} \sum_{i=1}^n |q_i^*| > 0.$$

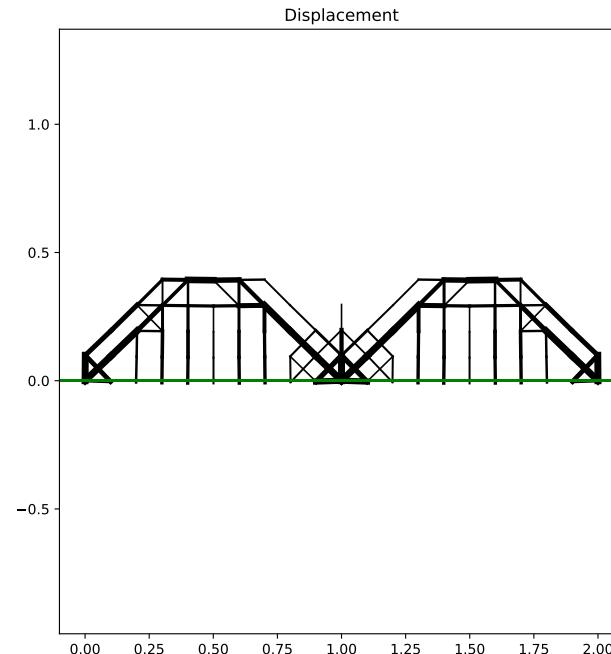
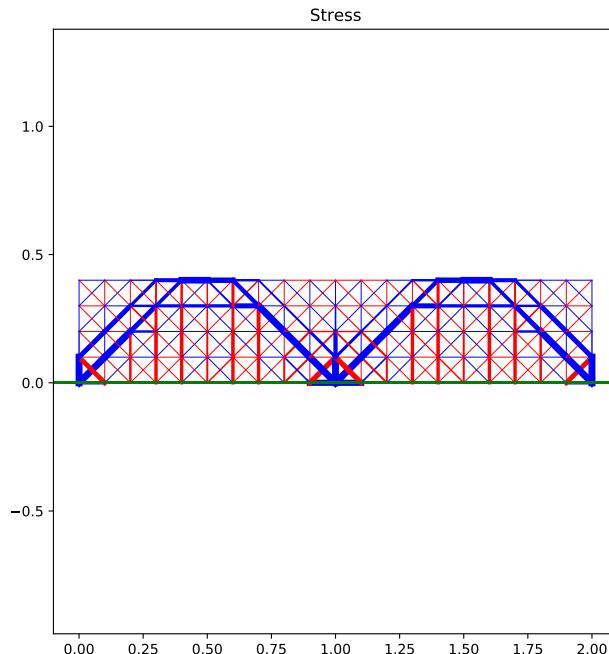
TTD as a Convex Optimization Problem III

An optimal truss computed via optimization on a 5×21 point grid composed of 105 nodes.

- ▶ Number of tentative bars $n = 344$
- ▶ Number of free nodes $m = 102$ (since there are three fixed nodes)



TTD as a Convex Optimization Problem IV



TTD as a Convex Optimization Problem V

Exercise 197 (☕☕)

Show that the dual problem to (70) is the problem⁵

$$\begin{aligned} \min_{v \in \mathbb{R}^m} \quad & f^\top v \\ \text{subject to} \quad & |b_i^\top v| \leq 1, \quad i = 1, 2, \dots, n. \end{aligned} \tag{71}$$

⁵For a loaded truss, a stress in a bar is the absolute value of the corresponding tension (i.e., the magnitude of the reaction force caused by bar's deformation) divided by the cross-sectional area of the bar; the larger this quantity, the worse the conditions the material is working in. The stress in bar i is (up to a constant factor) a simple function of the displacement vector v : $s_i = |b_i^\top v|$. The mechanical interpretation of this problem is: find a displacement v that maximizes the work $f^\top v$ of the load under the constraint that all stresses s_i are at most 1.

Introduction to Optimization

Peter Richtárik



Part 20: Smoothness and Strong Convexity

Lecture Outline

- ▶ Bregman divergence
- ▶ μ -convexity of functions
- ▶ L -smoothness of functions
- ▶ Functions that are both μ -convex and L -smooth

Bregman Divergence of a Function

Bregman Divergence of a Function I

Definition 198 (Bregman divergence)

The **Bregman divergence** of a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is the mapping $D_f : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ defined by

$$D_f(x, y) \stackrel{\text{def}}{=} f(x) - (f(y) + \langle \nabla f(y), x - y \rangle).$$

Remarks:

- ▶ Notice that the Bregman divergence is the difference between the function f and its linear approximation computed at y .
- ▶ Notice that we have shown before that a continuously differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex if and only if $D_f(x, y) \geq 0$ for all $x, y \in \mathbb{R}^d$. Indeed, this is what the equivalence of statements (i) and (ii) in Theorem 53 means!

Bregman Divergence of a Function II

Example 199

Let $f(x) = x^2$. Then

$$\begin{aligned}D_f(x, y) &= x^2 - (y^2 + \langle 2y, x - y \rangle) \\&= x^2 - y^2 - 2xy + 2y^2 \\&= x^2 + y^2 - 2xy \\&= (x - y)^2.\end{aligned}$$

Clearly, the Bregman divergence is nonnegative for all x, y , and hence by Theorem 53, f is convex. But we already knew this.

Bregman Divergence of a Function III

Example 200

Let $f(x) = \frac{1}{2}x^\top \mathbf{A}x$, where $\mathbf{A} \in \mathbb{S}_+^d$. Then since $\nabla f(y) = \mathbf{A}y$, we get

$$\begin{aligned} D_f(x, y) &= \frac{1}{2}x^\top \mathbf{A}x - \left(\frac{1}{2}y^\top \mathbf{A}y + \langle \mathbf{A}y, x - y \rangle \right) \\ &= \frac{1}{2}x^\top \mathbf{A}x - \frac{1}{2}y^\top \mathbf{A}y - x^\top \mathbf{A}y + y^\top \mathbf{A}y \\ &= \frac{1}{2}x^\top \mathbf{A}x + \frac{1}{2}y^\top \mathbf{A}y - x^\top \mathbf{A}y \\ &= \frac{1}{2}(x - y)^\top \mathbf{A}(x - y). \end{aligned}$$

Since \mathbf{A} is positive semidefinite, the Bregman divergence is nonnegative for all x, y , and hence by Theorem 53, f is convex. But we already knew this as well.

μ -Convex Functions

μ -Convex Functions I

Definition 201 (μ -convexity)

We say that a (continuously differentiable) function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is **μ -convex** if

$$f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2 \leq f(x), \quad \text{for all } x, y \in \mathbb{R}^d.$$

Note that this can be equivalently written in the form

$$\frac{\mu}{2} \|x - y\|^2 \leq D_f(x, y), \quad \text{for all } x, y \in \mathbb{R}^d. \quad (72)$$

When $\mu > 0$, we say that f is **strongly convex**. Note that $\mu = 0$ reduces to standard **convexity**.

Example 202

The function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = \frac{a}{2}x^2$, where $a \geq 0$, is **μ -convex** with $\mu = a$.

μ -Convex Functions II

Example 203

Consider the function $f(x) = \frac{1}{2}x^\top \mathbf{A}x$, where $\mathbf{A} \in \mathbb{S}_+^d$. We know from Example 200 that the Bregman divergence of this function is given by

$$D_f(x, y) = \frac{1}{2}(x - y)^\top \mathbf{A}(x - y).$$

Is f μ -convex? If so, what is μ ? We need to check whether (72) holds:

$$\frac{\mu}{2}(x - y)^\top (x - y) \leq \frac{1}{2}(x - y)^\top \mathbf{A}(x - y), \quad \text{for all } x, y \in \mathbb{R}^d.$$

Basic linear algebra says that this holds for μ being the smallest eigenvalue of \mathbf{A} :

$$\mu = \lambda_{\min}(\mathbf{A}).$$

L -Smooth Functions

L -Smooth Functions I

Definition 204 (L -smoothness)

We say that a (continuously differentiable) function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth if

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2, \quad \text{for all } x, y \in \mathbb{R}^d.$$

Example 205

The function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = \frac{a}{2}x^2$, where $a \geq 0$, is L -convex with $L = a$.

L -Smooth Functions II

Example 206

Consider the function $f(x) = \frac{1}{2}x^\top \mathbf{A}x$, where $\mathbf{A} \in \mathbb{S}_+^d$. We know from Example 200 that the Bregman divergence of this function is given by

$$D_f(x, y) = \frac{1}{2}(x - y)^\top \mathbf{A}(x - y).$$

Is f **L -smooth?** If so, what is L ? We need to check whether (72) holds:

$$\frac{L}{2}(x - y)^\top (x - y) \leq \frac{1}{2}(x - y)^\top \mathbf{A}(x - y), \quad \text{for all } x, y \in \mathbb{R}^d.$$

Basic linear algebra says that this holds for L being the smallest eigenvalue of \mathbf{A} :

$$L = \lambda_{\min}(\mathbf{A}).$$

Functions that are Both L -Smooth μ -Convex

Properties of L -Smooth and μ -Convex Functions I

As the next exercise states, the ridge regression objective function is both L -smooth and μ -convex. Such functions are “very nice” from an optimization point of view.

Exercise 207 (Ridge regression objective ☕)

The ridge regression objective has the form

$$f(x) = \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|^2,$$

where $\lambda > 0$. Show that f is L -smooth and μ -convex, where

$$L = \lambda_{\max}(A^\top A) + \lambda,$$

and

$$\mu = \lambda_{\min}(A^\top A) + \lambda.$$

Properties of L -Smooth and μ -Convex Functions II

The next theorem summarizes some very useful properties of L -smooth and μ -convex functions. These properties are often used to develop convergence theory for gradient-type optimization algorithms such as gradient descent (GD) and stochastic gradient descent (SGD).

Theorem 208

If f is both L -smooth and μ -convex, the following inequalities hold:

$$\mu \|x - y\|^2 \leq 2D_f(x, y) \leq L \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d, \quad (73)$$

$$\frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq 2D_f(x, y) \leq \frac{1}{\mu} \|\nabla f(x) - \nabla f(y)\|^2, \quad \forall x, y \in \mathbb{R}^d. \quad (74)$$

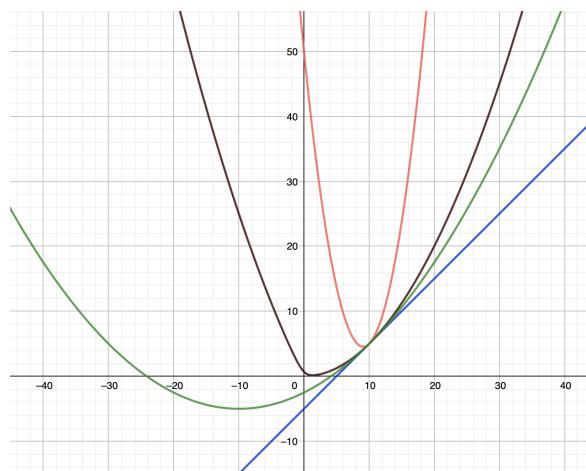
Properties of L -Smooth and μ -Convex Functions III

Example 209

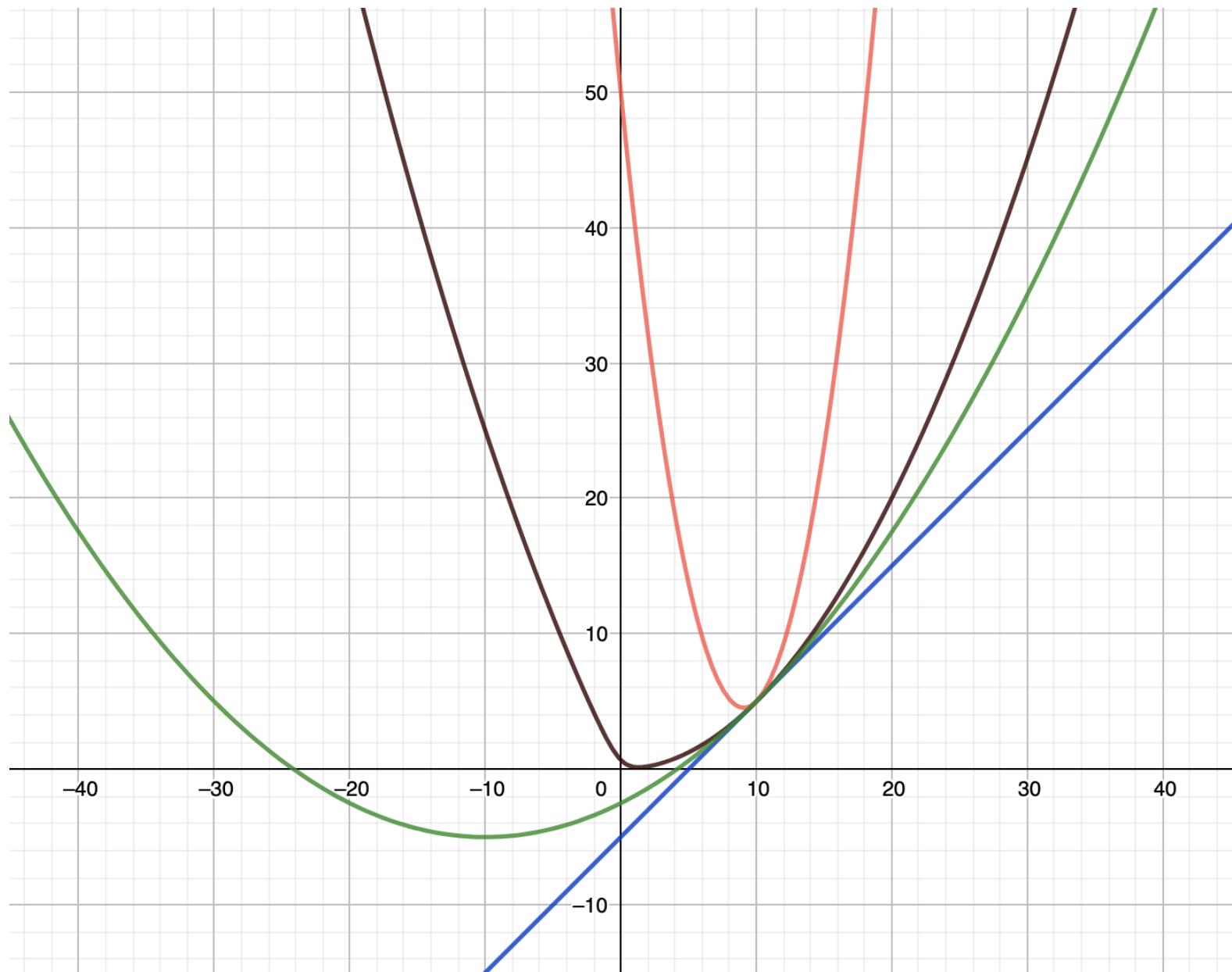
The function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$f(x) = \ln(1 + e^{-\theta x}) + \frac{\zeta}{2}x^2$$

is L -smooth for any $L \geq \frac{\theta^2}{4} + \zeta$ and μ -convex for any $0 \leq \mu \leq \zeta$. In the below plots we choose $\theta = 2$ and $\zeta = 0.1$.



Properties of L -Smooth and μ -Convex Functions IV



Properties of L -Smooth and μ -Convex Functions V

The plot depicts the functions $f(x)$, and three approximations of $f(x)$ defined as follows:

$$C_f(x, y) = f(x) + \langle \nabla f(y), x - y \rangle$$

$$U_f(x, y) = f(x) + \langle \nabla f(y), x - y \rangle + \frac{L}{2} \|x - y\|^2$$

$$L_f(x, y) = f(x) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2$$

for $y = 10$, $L = 1.01$ and $\mu = 0.05$.

Introduction to Optimization

Peter Richtárik



Part 21: SGD - Part 1

SGD in Generic Form

Consider the optimization problem

$$\min_{x \in \mathbb{R}^d} f(x).$$

We will now talk about the following **generic SGD method:**

Algorithm 1 SGD (generic)

- 1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: Construct an **unbiased estimator** $g^k \stackrel{\text{def}}{=} g(x^k)$ of the gradient
 - 4: $x^{k+1} = x^k - \gamma g^k$
-

Unbiasedness

Assumption 210 (Unbiasedness)

For all $k \geq 0$, $g^k = g(x^k)$ is an **unbiased estimator** of the gradient $\nabla f(x^k)$. That is,

$$\mathbb{E} [g^k | x^k] = \nabla f(x^k). \quad (75)$$

Remarks:

- ▶ Note that in the above assumption we only need to construct an unbiased stochastic gradient along the path of the iterates only.
- ▶ In practice, we often construct a random mapping $x \mapsto g(x)$ satisfying the property

$$\mathbb{E} [g(x)] = \nabla f(x) \quad \text{for all } x \in \mathbb{R}^d$$

and simply apply it to the iterates.

- ▶ This is a somewhat stronger assumption than what we need for analysis of SGD, but it has the advantage of being independent of the iterates of the algorithm.

Basic Facts Used in the Analysis

Basic Facts - I

- **Young's inequality.** For all $a, b \in \mathbb{R}^d$ and $t > 0$ the following inequalities holds:

$$\langle a, b \rangle \leq \frac{\|a\|^2}{2t} + \frac{t \|b\|^2}{2}, \quad (76)$$

$$\|a + b\|^2 \leq 2 \|a\|^2 + 2 \|b\|^2, \quad (77)$$

and

$$\frac{1}{2} \|a\|^2 - \|b\|^2 \leq \|a + b\|^2. \quad (78)$$

- **Variance decomposition.** For a random vector $X \in \mathbb{R}^d$ (with finite second moment) and any $c \in \mathbb{R}^d$ the variance can be decomposed as

$$\mathbf{E} [\|X - \mathbf{E}[X]\|^2] = \mathbf{E} [\|X - c\|^2] - \|\mathbf{E}[X] - c\|^2. \quad (79)$$

Some consequences:

Basic Facts - II

- ▶ The solution of the optimization problem

$$\min_{c \in \mathbb{R}^d} \mathbf{E} [\|X - c\|^2]$$

is $c = \mathbf{E}[X]$.

- ▶ For $c = 0$ we get the identity

$$\mathbf{Var}[X] \stackrel{\text{def}}{=} \mathbf{E} [\|X - \mathbf{E}[X]\|^2] = \mathbf{E} [\|X\|^2] - \|\mathbf{E}[X]\|^2. \quad (80)$$

- ▶ Since $f(x) = \|x\|^2$ is **2-convex** and **2-smooth**, the **strong** and **reverse Jensen's inequalities** applied to f say that

$$\mathbf{Var}[X] = \frac{\mu}{2} \mathbf{Var}[X] \leq \mathbf{E} [\|X\|^2] - \|\mathbf{E}[X]\|^2 \leq \frac{L}{2} \mathbf{Var}[X] = \mathbf{Var}[X].$$

- ▶ This can be seen as an alternative proof of (80).
- ▶ Perhaps more interestingly, this shows that the **strong and reverse Jensen's inequalities are tight** for $f(x) = \|x\|^2$.

Basic Facts - III

Identity (80) immediately leads to the well known inequality

$$\mathbb{E} [\|X\|^2] \geq \|\mathbb{E}[X]\|^2. \quad (81)$$

- **Markov's inequality.** For any nonnegative random variable X and any $t > 0$,

$$\text{Prob}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}. \quad (82)$$

Tower Property

Lemma 211 (Tower Property / Iterated Expectation)

For any random variables X and Y , we have

$$\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X]. \quad (83)$$

Proof.

We shall only prove this for discrete random variables; the proof is more technical in the continuous case.

$$\begin{aligned}\mathbb{E}[X] &= \sum_x x \text{Prob}(X = x) &= \sum_x x \sum_y \text{Prob}(X = x \& Y = y) \\ &= \sum_y \sum_x x \text{Prob}(X = x \& Y = y) \\ &= \sum_y \sum_x \text{Prob}(Y = y) x \frac{\text{Prob}(X = x \& Y = y)}{\text{Prob}(Y = y)} \\ &= \sum_y \text{Prob}(Y = y) \underbrace{\sum_x x \text{Prob}(X = x | Y = y)}_{\mathbb{E}[X | Y=y]} \\ &= \mathbb{E}[\mathbb{E}[X | Y]].\end{aligned}$$

Tower Property of Expectations: Intuition via Example

Example 212

Consider discrete random variables \mathbf{X} and \mathbf{Y} :

- ▶ \mathbf{X} has 2 outcomes: x_1 and x_2
- ▶ \mathbf{Y} has 3 outcomes: y_1 , y_2 and y_3

Their joint probability mass function is given in this table:

	y_1	y_2	y_3	
x_1	0.05	0.20	0.03	0.28
x_2	0.25	0.30	0.17	0.72
	0.30	0.50	0.20	1

Obviously, $E[\mathbf{X}] = 0.28x_1 + 0.72x_2$. But we can also write:

$$\begin{aligned} E[\mathbf{X}] &= (0.05x_1 + 0.25x_2) + (0.20x_1 + 0.30x_2) + (0.03x_1 + 0.17x_2) \\ &= \underbrace{0.30}_{\text{Prob}(\mathbf{Y}=y_1)} \underbrace{\left(\frac{0.05}{0.30}x_1 + \frac{0.25}{0.30}x_2 \right)}_{E[\mathbf{X} | \mathbf{Y}=y_1]} + \underbrace{0.50}_{\text{Prob}(\mathbf{Y}=y_2)} \left(\frac{0.20}{0.50}x_1 + \frac{0.30}{0.50}x_2 \right) \\ &\quad + \underbrace{0.20}_{\text{Prob}(\mathbf{Y}=y_3)} \left(\frac{0.03}{0.20}x_1 + \frac{0.17}{0.20}x_2 \right) \\ &= E[E[\mathbf{X} | \mathbf{Y}]]. \end{aligned}$$

SGD Analysis

Analysis of Generic SGD

Lemma 213

Consider Algorithm 1. If f is μ -convex and Assumption 210 (unbiasedness) is satisfied, then for all $\gamma \geq 0$ and $k \geq 0$ we have

$$\begin{aligned} \mathbb{E} \left[\|r^{k+1}\|^2 \mid x^k \right] &\leq (1 - \gamma \mu) \|r^k\|^2 - 2\gamma D_f(x^k, x^*) \\ &\quad + \gamma^2 \underbrace{\mathbb{E} \left[\|g^k - \nabla f(x^*)\|^2 \mid x^k \right]}_{G^k}. \end{aligned} \quad (84)$$

Proof of Lemma 213 - I

- ▶ Let $r^k \stackrel{\text{def}}{=} x^k - x^*$. Then

$$\begin{aligned}
\|r^{k+1}\|^2 &= \left\| \left(x^k - \gamma g^k \right) - x^* \right\|^2 \\
&= \left\| \left(x^k - \gamma g^k \right) - \left(x^* - \gamma \nabla f(x^*) \right) \right\|^2 \\
&= \left\| x^k - x^* - \gamma \left(g^k - \nabla f(x^*) \right) \right\|^2 \\
&= \|r^k\|^2 - 2\gamma \langle r^k, g^k - \nabla f(x^*) \rangle + \gamma^2 \|g^k - \nabla f(x^*)\|^2.
\end{aligned}$$

- ▶ We now compute expectation of both sides of the above inequality, conditional on x^k . This gives

$$\begin{aligned}
\mathbb{E} [\|r^{k+1}\|^2 \mid x^k] &\leq \|r^k\|^2 - 2\gamma \mathbb{E} [\langle r^k, g^k - \nabla f(x^*) \rangle \mid x^k] + \gamma^2 \mathbb{E} [\|g^k - \nabla f(x^*)\|^2 \mid x^k] \\
&= \|r^k\|^2 - 2\gamma \langle r^k, \mathbb{E} [g^k \mid x^k] - \nabla f(x^*) \rangle + \gamma^2 \mathbb{E} [\|g^k - \nabla f(x^*)\|^2 \mid x^k] \\
&\stackrel{(75)}{=} \|r^k\|^2 - 2\gamma \langle r^k, \nabla f(x^k) - \nabla f(x^*) \rangle + \gamma^2 \mathbb{E} [\|g^k - \nabla f(x^*)\|^2 \mid x^k],
\end{aligned}$$

where in the second step we used linearity of expectation and the fact that r^k is a constant vector conditioned on x^k , and in the last step we used the unbiasedness assumption.

Proof of Lemma 213 - II

- We now use μ -convexity to bound

$$\left\langle r^k, \nabla f(x^k) - \nabla f(x^*) \right\rangle \stackrel{(\text{Theorem 208})}{\geq} D_f(x^k, x^*) + \frac{\mu}{2} \|x^k - x^*\|^2,$$

which leads to

$$E \left[\|r^{k+1}\|^2 \mid x^k \right] \leq (1 - \gamma \mu) \|r^k\|^2 - 2\gamma D_f(x^k, x^*) + \gamma^2 E \left[\|g^k - \nabla f(x^*)\|^2 \mid x^k \right],$$

which is what we wanted to prove.

Bounding The Third Term - I

To analyze SGD using Lemma 213, we will need to come up with some appropriate upper bound on

$$G^k \stackrel{\text{def}}{=} \mathbf{E} \left[\|g^k - \nabla f(x^*)\|^2 \mid x^k \right].$$

- ▶ Recall that in GD, we used the bound

$$G^k = \|\nabla f(x^k) - \nabla f(x^*)\|^2 \leq 2LD_f(x^k, x^*),$$

which happens to be the consequence of *L-smoothness* and convexity.

- ▶ Stepsize γ was then chosen so that the last two terms in (84) can be neglected (i.e., upper bounded by 0).
- ▶ This led to the recursion $\|r^{k+1}\|^2 \leq (1 - \gamma\mu) \|r^k\|^2$.
- ▶ Motivated by this, it looks tempting to use an analogous inequality in the SGD case as well. Specifically, we require that there exists $A \geq 0$ such that

$$G^k \leq 2AD_f(x^k, x^*)$$

holds for all $k \geq 0$.

Bounding The Third Term - II

- ▶ However, this inequality does not hold for many typical choices of the stochastic gradient g^k unless the problem $\min f + R$ has some very special properties.
- ▶ Fortunately, there is a simple trick that fixes the issue: we require the inequality to hold up to some additive error $C \geq 0$:

$$G^k \leq 2AD_f(x^k, x^*) + C.$$

Key Assumption

Assumption 214 (AC)

There exist constants $A \geq 0$ and $C \geq 0$ such that for all $k \geq 0$,

$$G^k \stackrel{\text{def}}{=} \mathbf{E} \left[\|g^k - \nabla f(x^*)\|^2 \mid x^k \right] \leq 2AD_f(x^k, x^*) + C. \quad (85)$$

Remarks:

- ▶ In subsequent lectures, we will give two examples of practical stochastic gradient estimators g^k satisfying the above “AC” assumption/inequality.
- ▶ In a somewhat more distant future, we will further generalize this inequality in several ways, which will enable us to also consider
 - ▶ “variance reduced” SGD methods,
 - ▶ nonconvex f .

Convergence of SGD

Convergence Theory for SGD

Theorem 215 (Convergence of SGD – General Result)

Assume that f is μ -convex, g^k is unbiased (Assumption 210) and that the AC assumption (Assumption 214) is satisfied. Choose a stepsize satisfying

$$0 < \gamma \leq \frac{1}{A}. \quad (86)$$

Then the iterates $\{x^k\}_{k \geq 0}$ of SGD (Algorithm 1) satisfy

$$\mathbb{E} \left[\|x^k - x^*\|^2 \right] \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2 + \frac{\gamma C}{\mu}. \quad (87)$$

As we shall see, the above theorem **covers many useful variants of SGD methods**, and describes a **linear convergence rate up to a certain oscillation radius** controlled by the constant C .

Convergence Theory: Commentary I

Gradient Descent (GD) convergence result. In the case of GD, $g(x) = \nabla f(x)$ is trivially unbiased, and provided that f is convex and L -smooth, the AC assumption is satisfied with $A = L$ and $C = 0$. Moreover, the sequence of iterates $\{x^k\}_{k \geq 0}$ is deterministic.

- ▶ So, if f is μ -convex, then by Theorem 215 for all $0 < \gamma \leq \frac{1}{L}$ we get

$$\|x^k - x^*\|^2 \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2.$$

Convergence of the first term. The first term

$$T_1 \stackrel{\text{def}}{=} (1 - \gamma\mu)^k \|x^0 - x^*\|^2$$

appearing on the right hand side of (128) is smaller (=better) if

- ▶ the number of iterations k is larger (we have this under control),
- ▶ the stepsize γ is larger (we have this under control),
- ▶ the modulus of strong convexity μ is larger (typically, we do not have this under control),

Convergence Theory: Commentary II

- ▶ the starting point x^0 is closer to the optimal point x^* (we can control this in principle, but this is hard to do as we do not know what x^* is).

Typically, k (knowing when to stop) and γ (knowing what stepsize to choose) are the hyperparameters we directly control. Note that

- ▶ $k \rightarrow \infty \Rightarrow T_1 \rightarrow 0$
- ▶ More precisely, for any $\varepsilon > 0$,

$$k \geq \frac{1}{\gamma \mu} \log \frac{1}{\varepsilon} \Rightarrow T_1 \leq \varepsilon \|x^0 - x^*\|^2. \quad (88)$$

- ▶ Focusing on T_1 only and ignoring the effect of the stepsize on the second term (which does not matter in the lucky/favorable case when $C = 0$), the optimal choice of the stepsize is $\gamma = \frac{1}{A}$.

Convergence of the second term. The second term

$$T_2 \stackrel{\text{def}}{=} \frac{\gamma C}{\mu}$$

appearing on the right hand side of (128) is smaller (=better) when

Convergence Theory: Commentary III

- ▶ the stepsize γ is smaller
- ▶ the strong convexity modulus μ is larger,
- ▶ C is smaller
 - ▶ The favorable case $C = 0$ is special, as then the second term vanishes, and so do the associated convergence complications.
 - ▶ In the general $C > 0$ case, we can decrease the size of T_1 by decreasing the stepsize γ .

We have

$$\gamma \leq \frac{\mu \varepsilon \|x^0 - x^*\|^2}{C} \Rightarrow T_2 \leq \varepsilon \|x^0 - x^*\|^2. \quad (89)$$

Driving $T_1 + T_2$ to zero in the $C > 0$ case.

By combining (88) and (89), we get

$$E \left[\|x^k - x^*\|^2 \right] \leq T_1 + T_2 \leq 2\varepsilon \|x^0 - x^*\|^2 \quad (90)$$

as long as $\gamma = \min \left\{ \frac{1}{A}, \frac{\mu \varepsilon \|x^0 - x^*\|^2}{C} \right\}$ and $k \geq \frac{1}{\gamma \mu} \log \frac{1}{\varepsilon}$.

Convergence Theory: Commentary IV

- ▶ Note that

$$k = \mathcal{O}\left(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right),$$

which is **much worse** than $\mathcal{O}(\log \frac{1}{\varepsilon})$ (the latter was sufficient in the favorable $C = 0$ case).

- ▶ Indeed, the $C > 0$ case **forces us to forgo fast (linear) convergence rate and settle for slow (sublinear) convergence rate.**
- ▶ Later on in the course, we will talk about algorithmic strategies (commonly referred to as **variance reduction**) which enable us to maintain the fast (linear) rate even in the $C > 0$ case. This will require more sophisticated algorithmic and convergence analysis tools.

Proof of Theorem 215 - I

By combining Lemma 213 with Assumption 214, we get

$$\begin{aligned} \mathbf{E} \left[\|r^{k+1}\|^2 \mid x^k \right] &\stackrel{(84)}{\leq} (1 - \gamma \mu) \|r^k\|^2 - 2\gamma D_f(x^k, x^*) + \gamma^2 \mathbf{E} \left[\|g^k - \nabla f(x^*)\|^2 \mid x^k \right] \\ &\stackrel{(214)}{\leq} (1 - \gamma \mu) \|r^k\|^2 - 2\gamma D_f(x^k, x^*) + 2\gamma^2 \mathbf{A} D_f(x^k, x^*) + \gamma^2 \mathbf{C} \\ &= (1 - \gamma \mu) \|r^k\|^2 - 2\gamma(1 - \gamma \mathbf{A}) D_f(x^k, x^*) + \gamma^2 \mathbf{C} \\ &\leq (1 - \gamma \mu) \|r^k\|^2 + \gamma^2 \mathbf{C}, \end{aligned}$$

where in the last step we dropped the Bregman divergence term because $\gamma \leq \frac{1}{\mathbf{A}}$ and because Bregman divergence of a convex function is nonnegative.

Proof of Theorem 215 - II

We now take expectation again on both sides, and use the tower property

$$\mathbf{E} \left[\|r^{k+1}\|^2 \right] \stackrel{(83)}{=} \mathbf{E} \left[\mathbf{E} \left[\|r^{k+1}\|^2 \mid x^k \right] \right],$$

arriving at

$$\mathbf{E} \left[\|r^{k+1}\|^2 \right] \leq (1 - \gamma \mu) \mathbf{E} \left[\|r^k\|^2 \right] + \gamma^2 C.$$

Unrolling the recurrence and noting that $\mathbf{E} \left[\|r^0\|^2 \right] = \|r^0\|^2$ gives us

$$\begin{aligned} \mathbf{E} \left[\|r^k\|^2 \right] &\leq (1 - \gamma \mu)^k \|r^0\|^2 + \gamma^2 C \sum_{i=0}^{k-1} (1 - \gamma \mu)^i \\ &\leq (1 - \gamma \mu)^k \|r^0\|^2 + \frac{\gamma C}{\mu}. \end{aligned}$$

More on the AC Assumption

When is the AC Assumption Satisfied?

Theorem 215 depends on the AC assumption (Assumption 215). When is it satisfied?

- ▶ It will be useful (and notationally cleaner) to “decouple” the AC assumption from the iterates of SGD and cast it in a more abstract setting.
- ▶ In particular, let $g(x)$ be an unbiased estimator of $\nabla f(x)$. That is, $\mathbf{E}[g(x)] = \nabla f(x)$. We define

$$G(x, y) \stackrel{\text{def}}{=} \mathbf{E} \left[\|g(x) - \nabla f(y)\|^2 \right]. \quad (91)$$

- ▶ Using the above notation, the AC assumption can be written in the form

$$G(x^k, x^*) \leq 2AD_f(x^k, x^*) + C.$$

- ▶ Typically, providing an upper bound $G(x, y)$ for any x and y is equally complicated as providing a bound for $x = x^k$ and $y = y^*$. So, in order to establish the AC assumption, we will typically seek to establish an inequality of the form

$$G(x, y) \leq 2AD_f(x, y) + C(y), \quad x, y \in \mathbb{R}^d, \quad (92)$$

where $C(y)$ is allowed to depend on y , after which we specialize it to $x = x^k$ and $y = y^*$.

Bounding $G(x, y)$: First Approach

Theorem 216

Let $g(x)$ be an unbiased estimator of $\nabla f(x)$ and assume that $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and **L -smooth**. Then

$$G(x, y) \leq 2LD_f(x, y) + \text{Var}[g(x)]. \quad (93)$$

If, moreover, there exist constants A' and $C'(y)$ such that

$$\text{Var}[g(x)] \leq 2A'D_f(x, y) + C'(y), \quad (94)$$

then

$$G(x, y) \leq 2(L + A')D_f(x, y) + C'(y). \quad (95)$$

Proof of Theorem 216

By applying the variance decomposition identity (79) (with $X = g(x)$ and $c = \nabla f(y)$) and because $\nabla f(x) = \mathbf{E}[g(x)]$, we can write

$$\begin{aligned} G(x, y) &\stackrel{(91)}{=} \mathbf{E} \left[\|g(x) - \nabla f(y)\|^2 \right] \\ &\stackrel{(79)}{=} \mathbf{E} \left[\|g(x) - \mathbf{E}[g(x)]\|^2 \right] + \|\mathbf{E}[g(x)] - \nabla f(y)\|^2 \\ &= \mathbf{Var}[g(x)] + \|\nabla f(x) - \nabla f(y)\|^2. \end{aligned}$$

It remains to apply the inequality $\|\nabla f(x) - \nabla f(y)\|^2 \leq 2L D_f(x, y)$, which holds since f is assumed to be convex and **L -smooth** (see Proposition ??(iii)).

Example: Stochastic Gradient with Bounded Variance

Example 217

The most classic assumption on the stochastic gradient beyond unbiasedness is **boundedness of its variance**:

$$\text{Var}[g(x)] \leq \sigma^2 \quad \text{for all } x \in \mathbb{R}^d.$$

This arises in practice in the situation when a zero-mean bounded-variance noise is added to the true gradient:

$$g(x) \stackrel{\text{def}}{=} \nabla f(x) + \xi, \quad \mathbb{E}[\xi] = 0, \quad \text{Var}[\xi] \leq \sigma^2.$$

Since $\text{Var}[g(x)] = \text{Var}[\xi]$, inequality (93) then implies that

$$G(x, y) \leq 2LD_f(x, y) + \sigma^2.$$

Theorem 215 now says that as long as $\gamma \leq \frac{1}{L}$, we have

$$\mathbb{E}[\|x^k - x^*\|^2] \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2 + \frac{\gamma\sigma^2}{\mu}.$$

Bounding $G(x, y)$: Second Approach

Theorem 218

Let $g(x)$ be an unbiased estimator of $\nabla f(x)$ and assume that

$$\mathbf{E} \left[\|g(x) - g(y)\|^2 \right] \leq 2\mathbf{A}'' D_f(x, y) + \mathbf{C}''(y). \quad (96)$$

Then

$$G(x, y) \leq 2(2\mathbf{A}'') D_f(x, y) + 2(\mathbf{Var}[g(y)] + \mathbf{C}''(y)). \quad (97)$$

Proof.

$$\begin{aligned} G(x, y) &\stackrel{(91)}{=} \mathbf{E} \left[\|g(x) - \nabla f(y)\|^2 \right] \\ &= \mathbf{E} \left[\|g(x) - g(y) + g(y) - \nabla f(y)\|^2 \right] \\ &\stackrel{(77)}{\leq} \mathbf{E} \left[2 \|g(x) - g(y)\|^2 + 2 \|g(y) - \nabla f(y)\|^2 \right] \\ &\stackrel{(96)}{\leq} 2(2\mathbf{A}'') D_f(x, y) + 2(\mathbf{Var}[g(y)] + \mathbf{C}''(y)). \end{aligned}$$

Bounding $G(x, y)$: Commentary

Theorem 216

- ▶ The AC inequality holds with

$$A = L + A', \quad C = C'(x^*).$$

- ▶ **Hard part:** Establishing (94)

Theorem 218

- ▶ The AC inequality holds with

$$A = 2A'', \quad C = 2(\text{Var}[g(x^*)] + C''(x^*)).$$

- ▶ **Hard part:** Establishing (96)

Exercises

Exercises I

Exercise 219 (Young's inequality ☕)

Prove inequalities (76), (77) and (78).

Exercise 220 (Variance decomposition ☕)

Provide two different proofs of the variance decomposition identity (79).

Exercise 221 (Chebyshev's inequality ☕)

Use Markov's inequality to prove that for a random vector $X \in \mathbb{R}^d$ and any $t > 0$, we have

$$\text{Prob}(\|X - \mathbf{E}[X]\| \geq t) \leq \frac{\text{Var}[X]}{t^2},$$

Exercises II

Exercise 222 (Convergence in mean square ☕☕)

We say that a sequence $\{X^k\}_{k \geq 0}$ of random vectors **converges in mean square** to random vector $X \in \mathbb{R}^d$ if the second moments $E[\|X^k\|^2]$ and $E[\|X\|^2]$ exist, and if

$$\lim_{k \rightarrow \infty} E[\|X^k - X\|^2] = 0.$$

Under the assumptions of Theorem 215, show that the sequence of random vectors $\{x^k\}_{k \geq 0}$ produced by SGD converges in mean square to x^* if the AC assumption holds with $C = 0$.

Exercises III

Exercise 223 (Convergence in probability ☕☕)

We say that a sequence $\{X^k\}_{k \geq 0}$ of random vectors **converges in probability** to random vector $X \in \mathbb{R}^d$ if for all $\varepsilon > 0$

$$\lim_{k \rightarrow \infty} \text{Prob} (\|X^k - X\| > \varepsilon) = 0.$$

Under the assumptions of Theorem 215, show that the sequence of random vectors $\{x^k\}_{k \geq 0}$ produced by SGD converges in probability to x^* if the AC assumption holds with $C = 0$.

Exercises IV

Exercise 224 (☕☕)

Let $f = \frac{1}{n} \sum_{i=1}^n f_i$, and assume that each f_i is convex and L_i -smooth.

Define a gradient estimator by setting $g(x) = \nabla f_i(x)$, where the index i is chosen uniformly at random. Establish a bound of the form

$$G(x, y) \leq 2AD_f(x, y) + C,$$

and specify what A and C are. Use this to derive a convergence result for SGD using this gradient estimator.

Exercise 225 (Code ☕)

Code up the SGD method described in the previous example on an optimization problem of your choice. Perform a sequence of numerical experiments testing various parts of the established theory.

Introduction to Optimization

Peter Richtárik



Part 22: SGD - Part 2

Introduction

We will now consider the “finite-sum” optimization problem

$$\min_{x \in \mathbb{R}^d} \left[f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n f_i(x) \right]. \quad (98)$$

- ▶ We now consider a non-uniform sampling variant of SGD which uses the gradient estimator

$$g(x) = \frac{\nabla f_i(x)}{np_i}, \quad (99)$$

where i is chosen with probability $p_i > 0$.

- ▶ Note that the gradient estimator is unbiased:

$$\mathbf{E}[g(x)] = \sum_{i=1}^n p_i \frac{\nabla f_i(x)}{np_i} = \sum_{i=1}^n \frac{\nabla f_i(x)}{n} = \nabla f(x). \quad (100)$$

SGD-NS: the Algorithm

For the record, here is the formal algorithm:

Algorithm 2 SGD-NS

- 1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, probabilities p_1, \dots, p_n summing up to one
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: Sample $i^k = i \in \{1, 2, \dots, n\}$ with probability $p_i > 0$
 - 4: $g^k = g(x^k) = \frac{\nabla f_i(x^k)}{np_i}$
 - 5: $x^{k+1} = x^k - \gamma g^k$
-

Two notable special cases:

- ▶ SGD-US (SGD with uniform sampling): $p_i = \frac{1}{n}$
- ▶ SGD-IS (SGD with importance sampling): $p_i = \frac{L_i}{\sum_j L_j}$

Convergence Theory

Applying Theorem 215: What's Missing?

In order to analyze SGD-NS, we will use Theorem 215, which requires the following assumptions (written here in the “decoupled” form):

- ▶ μ -convexity of f ,
- ▶ unbiasedness: $E[g(x)] = \nabla f(x)$ for all $x \in \mathbb{R}^d$,
- ▶ AC assumption: $E\left[\|g(x) - \nabla f(y)\|^2\right] \leq 2AD_f(x, y) + C(y)$ for all $x, y \in \mathbb{R}^d$.

We will continue to assume that f is μ -convex, and we already checked that the stochastic gradient estimator (99) used in SGD-NS is unbiased; see (100). **So, in order to be able to apply Theorem 215, we only need to establish the AC inequality.**

However, the AC inequality will not hold without us imposing further assumptions on the functions f_1, f_2, \dots, f_n . Here is an assumption that works.

Assumption 226

Each f_i is convex and L_i -smooth.

Computing the AC Constants

If each f_i is L_i -smooth and 0-convex (as required by Assumption 226), then in view of Theorem 208, the following inequalities hold for all $i \in [n]$:

$$0 \leq 2D_{f_i}(x, y) \leq L_i \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d, \quad (101)$$

$$\frac{1}{L_i} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2D_{f_i}(x, y), \quad \forall x, y \in \mathbb{R}^d. \quad (102)$$

Lemma 227

Let Assumption 226 hold (convexity and L_i -smoothness of f_i for all i). Let the gradient estimator g be defined as in (99). Then

$$\mathbf{E} [\|g(x) - g(y)\|^2] \leq 2A'' D_f(x, y), \quad \text{for all } x, y \in \mathbb{R}^d,$$

where $A'' = \max_i \frac{L_i}{np_i}$. Moreover,

$$G(x, y) \leq 2(2A'')D_f(x, y) + 2\mathbf{Var}[g(y)] \quad \text{for all } x, y \in \mathbb{R}^d. \quad (103)$$

Proof of Lemma 227

Fix any $x, y \in \mathbb{R}^d$. Then

$$\mathbb{E} \left[\|g(x) - g(y)\|^2 \right] \stackrel{(99)}{=} \sum_{i=1}^n \textcolor{red}{p}_i \left\| \frac{\nabla f_i(x)}{np_i} - \frac{\nabla f_i(y)}{np_i} \right\|^2 \quad (104)$$

$$= \frac{1}{n} \sum_{i=1}^n \frac{1}{np_i} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \quad (105)$$

$$\begin{aligned} &\stackrel{(102)}{\leq} \frac{1}{n} \sum_{i=1}^n \frac{1}{np_i} 2\textcolor{red}{L}_i D_{f_i}(x, y) \\ &\leq 2 \left(\max_i \frac{L_i}{np_i} \right) \frac{1}{n} \sum_{i=1}^n D_{f_i}(x, y) \\ &= 2 \left(\max_i \frac{L_i}{np_i} \right) D_f(x, y). \end{aligned} \quad (106)$$

Finally, (103) follows by applying Theorem 218.

Complexity of SGD-NS

The following result is a direct consequence of Theorem 215 and Lemma 227.

Theorem 228

Assume that

- ▶ f is μ -convex,
- ▶ Each f_i is convex and L_i -smooth (Assumption 226)

Then SGD-NS with stepsize $0 < \gamma \leq \frac{1}{2A''}$, where $A'' = \max_i \frac{L_i}{np_i}$, satisfies

$$\mathbb{E} \left[\|x^k - x^*\|^2 \right] \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2 + \frac{2\gamma\sigma_*^2}{\mu}, \quad (107)$$

where

$$\begin{aligned} \sigma_*^2 &\stackrel{\text{def}}{=} \text{Var}[g(x^*)] \stackrel{(80)}{=} \mathbb{E} \left[\|g(x^*)\|^2 \right] - \|\mathbb{E}[g(x^*)]\|^2 \\ &= \left(\frac{1}{n^2} \sum_{i=1}^n \frac{\|\nabla f_i(x^*)\|^2}{p_i} \right) - \|\nabla f(x^*)\|^2. \end{aligned} \quad (108)$$

Observation 1: Interpolation Regime

Zero Variance

The regime in which the variance of the stochastic gradient evaluated at x^* vanishes (i.e., $\sigma_*^2 = 0$), is particularly favorable, and it makes sense to spend a bit of time to think about it. Main reason why it is special:

Theorem 229 (Factor of 2 Improvement)

Assume that $\sigma_*^2 = 0$ and let $A'' = \max_i \frac{L_i}{np_i}$. Then

- (i) *The upper bound in (103) can be improved by a factor of 2:*

$$G(x, x^*) \leq 2A'' D_f(x, x^*). \quad (109)$$

- (ii) *Theorem 228 has the following stronger form: SGD-NS allows for a twice larger stepsize $0 < \gamma \leq \frac{1}{A''}$, and satisfies*

$$\mathbf{E} \left[\|x^k - x^*\|^2 \right] \leq (1 - \gamma \mu)^k \|x^0 - x^*\|^2. \quad (110)$$

In particular, SGD-NS in this case converges to the solution x^ (in mean square) rather than to a neighborhood thereof only.*

Interpolation Regime: Definition

Definition 230 (Interpolation regime)

Assume that there is no regularizer (i.e., $R \equiv 0$). We say that problem (115) is in the **interpolation regime** if

$$\nabla f_i(x^*) = 0 \quad \text{for all } i = 1, 2, \dots, n.$$

Interpretation:

- ▶ If f_i is convex and $f_i(x) = \ell(h_x(a_i), b_i)$, where ℓ is a loss function and (a_i, b_i) is a data example, then $\nabla f_i(x^*) = 0$ means that **model h_{x^*} incurs minimum (typically zero) loss on example (a_i, b_i)** .
- ▶ If this holds for all i , this means that model h_{x^*} perfectly fits (i.e., **interpolates**) all training data.

Interpolation Regime: Example

Example 231 (Linear system)

Assume the function $h^* : \mathcal{A} \rightarrow \mathcal{B}$ we are learning is given by

$$h^*(a) = \langle x^*, a \rangle,$$

where $\mathcal{A} = \mathbb{R}^d$ and $\mathcal{B} = \mathbb{R}$. Given an input a_i , the true label is given by $b_i = h^*(a_i) = \langle x^*, a_i \rangle$. Assume we use model

$$h_x(a) = \langle x, a \rangle.$$

If we use quadratic loss

$$\ell(b, b') = \frac{1}{2}(b - b')^2,$$

then

$$f_i(x) = \ell(h_x(a_i), b_i) = \frac{1}{2}(\langle a_i, x \rangle - b_i)^2.$$

Notice that $f_i(x^*) = 0$, and $\nabla f_i(x^*) = 0$ for all i . Hence, in this case we are in the interpolation regime.

Observation 2: Uniform Sampling

Uniform Sampling: SGD-US

For the record, here is the formal algorithm:

Algorithm 3 SGD-US

- ```

1: Parameters: learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$
2: for $k = 0, 1, 2, \dots$ do
3: Sample $i^k = i \in \{1, 2, \dots, n\}$ with probability $\frac{1}{n}$
4: $g^k = \nabla f_i(x^k)$ obtain a stochastic gradient
5: $x^{k+1} = x^k - \gamma g^k$

```

# Comparing GD and SGD-US in the Interpolation Regime

Let us compare the convergence of GD and SGD-US:

- ▶ GD
  - ▶ Using the stepsize  $\gamma = \frac{1}{L}$ , GD achieves iteration complexity
$$k \geq \frac{L}{\mu} \log \frac{1}{\varepsilon} \Rightarrow \|x^k - x^*\|^2 \leq \varepsilon \|x^0 - x^*\|^2$$
  - ▶ Cost of 1 iteration:  $n$  gradient evaluations
  - ▶ **Total complexity:**  $\tilde{\mathcal{O}}\left(\frac{nL}{\mu}\right)$
- ▶ SGD-US
  - ▶ By Theorem 229, using the uniform probabilities  $p_i = \frac{1}{n}$ , stepsize  $\gamma = \frac{1}{A''} = \frac{1}{\max_i L_i}$ , SGD-US achieves iteration complexity
$$k \geq \frac{\max_i L_i}{\mu} \log \frac{1}{\varepsilon} \Rightarrow \mathbb{E} \left[ \|x^k - x^*\|^2 \right] \leq \varepsilon \|x^0 - x^*\|^2$$
  - ▶ Cost of 1 iteration: 1 gradient evaluation
  - ▶ **Total complexity:**  $\tilde{\mathcal{O}}\left(\frac{\max_i L_i}{\mu}\right)$

**Key Question:** How does  $\max_i L_i$  compare to  $nL$ ?

## Example: Large $L_i$

Example 232 (One can have  $L_i = nL$  for all  $i$ )

Choose  $\theta > 0$  and let

$$f_i(x) = \frac{\theta}{2}x_i^2$$

for  $i = 1, 2, \dots, n$ . We have  $f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_i f_i(x) = \frac{\theta/n}{2} \|x\|^2$ .

► Since

$$\frac{\theta}{2} \|x\|^2 - f_i(x) = \frac{\theta}{2} \sum_{j \neq i} x_j^2$$

is convex,  $f_i$  is  $L_i$ -smooth with  $L_i = \theta$ . Moreover, the constant  $L_i$  can't be improved further.

► Since

$$\frac{\theta/n}{2} \|x\|^2 - f(x) = 0$$

is convex,  $f$  is  $L$ -smooth with  $L = \frac{\theta}{n}$ . Moreover, the constant  $L$  can't be improved further.

Note that

$$L_i = nL \quad \text{for all } i = 1, 2, \dots, n.$$

## Example: Small $L_i$

Example 233 (One can have  $L_i = L$  for all  $i$ )

Choose  $\theta > 0$  and let

$$f_i(x) = \frac{\theta}{2} \|x\|^2$$

for  $i = 1, 2, \dots, n$ . We have  $f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_i f_i(x) = \frac{\theta}{2} \|x\|^2$ . Clearly, all functions  $f_1, \dots, f_n$  and  $f$  are  $\theta$ -smooth, and the smoothness constant can't be further improved. Note that

$$L_i = L \quad \text{for all } i = 1, 2, \dots, n.$$

# Generalizing the Examples

It turns out that the two examples above are extreme cases. In particular, we will show in Lemma 234 that (in a certain precise sense)

$$L \leq \max_i L_i \leq nL.$$

- ▶ Because of the inequality  $\max_i L_i \leq nL$ , we concluded that in the interpolation regime, **SGD-US is not worse than GD in terms of total complexity!**
- ▶ In situations when  $\max_i L_i \ll nL$  (recall that there exists a scenario in which  $\max_i L_i = L$ ), in the interpolation regime, **SGD-US can be up to  $n$  times faster than GD!**

# Smoothness Constants of Functions $f_i$ and $f$

## Lemma 234

Let  $f_1, f_2, \dots, f_n : \mathbb{R}^d \rightarrow \mathbb{R}$  be differentiable functions and define  $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$ .

- (i) Assume  $f_i$  is  $L_i$ -smooth for all  $i$ . Then  $f$  is  $L$ -smooth with

$$L \leq \frac{1}{n} \sum_{i=1}^n L_i.$$

- (ii) Assume that  $f$  is  $L$ -smooth and  $f_i$  is convex for all  $i$ . Then each  $f_i$  is  $L_i$ -smooth with

$$L_i \leq nL.$$

- (iii) Assume that  $f$  is  $L'$ -smooth,  $f_i$  is convex and  $L'_i$ -smooth for all  $i$ . Then there exist constants  $L \leq L'$  and  $L'_i \leq L_i$  for all  $i$  such that  $f$  is  $L$ -smooth,  $f_i$  is  $L_i$ -smooth for all  $i$ , and

$$L \leq \max_i L_i \leq nL.$$

# Proof of Lemma 234

(i) Since  $f_i$  is  $L_i$ -smooth, the function

$$\phi_i(x) \stackrel{\text{def}}{=} \frac{L_i}{2} \|x\|^2 - f_i(x)$$

is convex. Hence, the average of these functions,

$$\frac{1}{n} \sum_{i=1}^n \phi_i(x) = \frac{\frac{1}{n} \sum_{i=1}^n L_i}{2} \|x\|^2 - f(x),$$

is also convex. This means that  $f$  is  $(\frac{1}{n} \sum_i L_i)$ -smooth.

(ii) Since  $f$  is  $L$ -smooth, the function

$$\phi(x) \stackrel{\text{def}}{=} \frac{L}{2} \|x\|^2 - f(x)$$

is convex, and hence the function

$$n\phi(x) = \frac{nL}{2} \|x\|^2 - \sum_{i=1}^n f_i(x)$$

is also convex. Now, since all  $f_i$ 's are convex, also  $\sum_{i:i \neq j} f_i$  is convex, and hence

$$\frac{nL}{2} \|x\|^2 - \sum_{i=1}^n f_i(x) + \sum_{i:i \neq j} f_i = \frac{nL}{2} \|x\|^2 - f_j(x)$$

is convex. This means that  $f_j$  is  $L_j$ -smooth with  $L_j \leq nL$ .

(iii) Left as an exercise.

## Observation 3: Importance Sampling

# The Many Faces of Importance Sampling I

The (open) **standard simplex** in  $\mathbb{R}^n$  is the set

$$\Delta_n \stackrel{\text{def}}{=} \left\{ \mathbf{p} = (p_1, \dots, p_n) : \sum_{i=1}^n p_i = 1 \quad \text{and} \quad p_i > 0 \text{ for all } i \right\}.$$

1. Optimizing the linear convergence rate without worrying about the neighborhood:

$$\begin{aligned} \arg \min_{\mathbf{p} \in \Delta_n, \gamma} \left\{ (1 - \gamma \mu)^k : 0 < \gamma \leq \frac{1}{2A''} \right\} &= \arg \min_{\mathbf{p} \in \Delta_n} \left( 1 - \frac{\mu}{2 \max_i \frac{L_i}{n p_i}} \right)^k \\ &= \arg \min_{\mathbf{p} \in \Delta_n} \max_i \frac{L_i}{n p_i}. \end{aligned}$$

The solution of this problem leads to the optimal stepsize  $\gamma = \frac{1}{2\bar{L}}$  and importance sampling probabilities

$$p_i = \frac{L_i}{\sum_j L_j} \quad \Rightarrow \quad A'' = \frac{\sum_i L_i}{n} \stackrel{\text{def}}{=} \bar{L}.$$

# The Many Faces of Importance Sampling II

We will refer to SGD with these probabilities as SGD-IS.

## Algorithm 4 SGD-IS

- ```

1: Parameters: learning rate  $\gamma > 0$ , starting point  $x^0 \in \mathbb{R}^d$ 
2: for  $k = 0, 1, 2, \dots$  do
3:   Sample  $i^k = i \in \{1, 2, \dots, n\}$  with probability  $p_i = \frac{L_i}{\sum_j L_j}$ 
4:    $g^k = \frac{\nabla f_i(x^k)}{np_i}$                                 obtain a stochastic gradient
5:    $x^{k+1} = x^k - \gamma g^k$ 

```

On the next slide we compare the convergence of GD and SGD-IS in the **interpolation regime**.

The Many Faces of Importance Sampling III

- ▶ GD
 - ▶ We know that using stepsize $\gamma = \frac{1}{L}$, GD achieves iteration complexity

$$k \geq \frac{L}{\mu} \log \frac{1}{\varepsilon} \Rightarrow \|x^k - x^*\|^2 \leq \varepsilon \|x^0 - x^*\|^2$$

- ▶ Cost of 1 iteration: n gradient evaluations
- ▶ **Total complexity:** $\tilde{\mathcal{O}}\left(\frac{nL}{\mu}\right)$

▶ SGD-IS

- ▶ By Theorem 229, using the **importance sampling probabilities** $p_i = \frac{L_i}{\sum_j L_j}$, stepsize $\gamma = \frac{1}{A''} = \frac{1}{\frac{1}{n} \sum_i L_i}$, SGD-IS achieves iteration complexity

$$k \geq \frac{\frac{1}{n} \sum_i L_i}{\mu} \log \frac{1}{\varepsilon} \Rightarrow \mathbb{E} \left[\|x^k - x^*\|^2 \right] \leq \varepsilon \|x^0 - x^*\|^2$$

- ▶ Cost of 1 iteration: 1 gradient evaluation
- ▶ **Total complexity:** $\tilde{\mathcal{O}}\left(\frac{\frac{1}{n} \sum_i L_i}{\mu}\right)$

The Many Faces of Importance Sampling IV

Key Question: How does $\frac{1}{n} \sum_i L_i$ compare to nL ?

By Lemma 234,

$$L \leq \frac{1}{n} \sum_{i=1}^n L_i \leq nL.$$

- ▶ Because of the inequality $\frac{1}{n} \sum_i L_i \leq \max_i L_i \leq nL$, we conclude that in the interpolation regime, SGD-IS is better than SGD-US, which in turn is better than GD (in terms of total complexity!)
- ▶ In situations when $\frac{1}{n} \sum_i L_i \ll \max_i L_i$, in the interpolation regime, SGD-IS can be much faster than SGD-US!

The Many Faces of Importance Sampling V

2. Optimizing the neighborhood size without worrying about convergence speed:

$$\arg \min_{p \in \Delta_n, \gamma} \left\{ \frac{2\gamma \sigma_*^2}{\mu} : 0 < \gamma \leq \frac{1}{2A''} \right\}.$$

Notice we can push the above objective arbitrarily close to 0 for any fixed p by choosing γ sufficiently small. So, there is no useful notion of importance sampling in this case.

3. Optimizing the neighborhood size while making sure the convergence speed is good enough.

We choose $\gamma_{\min} > 0$ (can't be too big) and require that $\gamma \geq \gamma_{\min}$. This ensures that

$$(1 - \gamma\mu)^k \leq (1 - \gamma_{\min}\mu)^k.$$

The Many Faces of Importance Sampling VI

Using this as a constraint, we can now pose the problem

$$\begin{aligned} \arg \min_{\mathbf{p} \in \Delta_n, \gamma} & \left\{ \frac{2\gamma\sigma_*^2}{\mu} : 0 < \gamma \leq \frac{1}{2A''}, \gamma \geq \gamma_{\min} \right\} \\ = \arg \min_{\mathbf{p} \in \Delta_n} & \left\{ \frac{2\gamma_{\min}\sigma_*^2}{\mu} : \gamma_{\min} \leq \frac{1}{2A''} \right\} \\ = \arg \min_{\mathbf{p} \in \Delta_n} & \left\{ \sigma_*^2 : \gamma_{\min} \leq \min_i \frac{n p_i}{2L_i} \right\}. \end{aligned}$$

We'll stop here and not worry about computing the optimal probabilities in this setting.

4. Optimizing the variance σ_*^2 . Since

$$\sigma_*^2 \stackrel{(108)}{=} \left(\frac{1}{n^2} \sum_{i=1}^n \frac{\|\nabla f_i(x^*)\|^2}{p_i} \right) - \|\nabla f(x^*)\|^2,$$

The Many Faces of Importance Sampling VII

the optimal probabilities are the solution of

$$\min_{\mathbf{p} \in \Delta_n} \sum_{i=1}^n \frac{\|\nabla f_i(x^*)\|^2}{p_i}. \quad (111)$$

It turns out the optimal solution of (111) is given by the **importance sampling probabilities**

$$p_i = \frac{\|\nabla f_i(x^*)\|}{\sum_{j=1}^n \|\nabla f_j(x^*)\|}. \quad (112)$$

Using these optimal probabilities, the variance is equal to

$$\sigma_*^2 = \left(\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*)\| \right)^2 - \|\nabla f(x^*)\|^2. \quad (113)$$

Exercises

Exercises I

Exercise 235 (Decomposition of Bregman divergence)

Let $f = \frac{1}{n} \sum_{i=1}^n f_i$. Prove that $D_f(x, y) = \frac{1}{n} \sum_{i=1}^n D_{f_i}(x, y)$.

Exercise 236

Write down the complete proof of Theorem 228.

Exercise 237

Prove that $\sigma_*^2 = 0$ if and only if $\nabla f_i(x^*) = np_i \nabla f(x^*)$ for all i .

Exercise 238

Prove part (iii) of Lemma 234.

Exercise 239

Give an example of L_i -smooth functions f_i for which $\frac{1}{n} \sum_i L_i \ll \max_i L_i$.

Exercises II

Exercise 240 (Importance sampling 1)

Let $a_1, \dots, a_n > 0$. Prove that the solution to the optimization problem

$$\min_{p \in \Delta_n} \max_i \frac{a_i}{p_i}$$

is given by $p_i = \frac{a_i}{\sum_j a_j}$ for all $i = 1, 2, \dots, n$.

Exercise 241 (Importance sampling 2 - simple version)

Let $a, b > 0$. Prove that the solution to the optimization problem

$$\min_{0 < p < 1} \frac{a}{p} + \frac{b}{1-p}$$

is given by $p = \frac{\sqrt{a}}{\sqrt{a} + \sqrt{b}}$.

Exercises III

Exercise 242 (Importance sampling 2)

Let $a_1, \dots, a_n > 0$. Prove that the solution to the optimization problem

$$\min_{p \in \Delta_n} \sum_{i=1}^n \frac{a_i}{p_i}$$

is given by $p_i = \frac{\sqrt{a_i}}{\sum_j \sqrt{a_j}}$ for all $i = 1, 2, \dots, n$.

Exercises IV

Exercise 243

Let $a_1, a_2, \dots, a_n \in \mathbb{R}^d$ and let $a = \frac{1}{n} \sum_{i=1}^n a_i$. Show that

$$\left(\frac{1}{n} \sum_{i=1}^n \|a_i\| \right)^2 \geq \|a\|^2. \quad (114)$$

- (i) Deduce that the expression in (113) is nonnegative.⁶
- (ii) Is it true that (114) holds as an equality if and only if there exists $c \in \mathbb{R}^d$ such that for each i we have $a_i = t_i c$ for some $t_i \geq 0$?

⁶Note that we knew this was nonnegative already through other means: it is equal to variance of $g(x^*)$ wrt the probabilities given in (112).

Introduction to Optimization

Peter Richtárik



Part 23: SGD - Part 3

Introduction |

Recall we consider the **finite-sum** problem

$$\min_{x \in \mathbb{R}^d} \left[f(x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n f_i(x) \right]. \quad (115)$$

We will describe **three families of SGD methods** which allow us to use multiple functions f_i , chosen at random, in the formation of the stochastic gradient. If we assume that each function f_i corresponds to a single training data point only, then this means that **a random collection of all training data points (a “minibatch”) is processed in each iteration.**

- ▶ SGD-NS in each iteration samples & processes a single training data point only
- ▶ GD in each iteration samples & processes all the training data points
- ▶ In some sense, these methods we will **interpolate between SGD-NS (or a variant thereof) and GD**. The methods are also known as **minibatch SGD** methods.

Introduction II

To do define and analyze each method, we need to:

- ▶ Describe the **unbiased** stochastic gradient estimator g
- ▶ Compute **expected smoothness** constant $A'' \geq 0$ such that

$$\mathbb{E} \left[\|g(x) - g(y)\|^2 \right] \leq 2A'' D_f(x, y)$$

Sampling without Replacement (Nice Sampling)

Sampling without Replacement: Nice Sampling

Fix a minibatch size $\tau \in \{1, 2, \dots, n\}$ and let S be a random subset of $\{1, 2, \dots, n\}$ of size τ , chosen uniformly at random.⁷ Define the gradient estimator via

$$g(x) \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{i \in S} \nabla f_i(x). \quad (116)$$

This estimator leads to the following SGD algorithm:

Algorithm 5 SGD-NICE

- 1: **Parameters:** learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, minibatch size $\tau \in \{1, 2, \dots, n\}$
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: Sample $S^k \subseteq \{1, \dots, n\}$ uniformly from all subsets of cardinality τ
 - 4: $g^k = \frac{1}{\tau} \sum_{i \in S^k} \nabla f_i(x^k)$ obtain a stochastic gradient
 - 5: $x^{k+1} = x^k - \gamma g^k$
-

⁷That is, we choose a single subset from the $\binom{n}{\tau}$ subsets of $\{1, 2, \dots, n\}$ of cardinality τ , each with probability $1/\binom{n}{\tau}$. Such a random set S is also known in the literature under the name τ -nice sampling.

SGD-NICE: Unbiasedness and Expected Smoothness

Lemma 244

The gradient estimator g defined in (116) is unbiased. If we further assume that $n \geq 2$, f_i is convex and L_i -smooth for all i , and f is L -smooth, then

$$\mathbb{E} \left[\|g(x) - g(y)\|^2 \right] \leq 2A'' D_f(x, y),$$

where

$$A'' = \frac{n - \tau}{\tau(n - 1)} \max_i L_i + \frac{n(\tau - 1)}{\tau(n - 1)} L. \quad (117)$$

Finally,

$$\mathbb{V}\text{ar}[g(y)] = \frac{n - \tau}{\tau(n - 1)} \left(\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(y)\|^2 - \|\nabla f(y)\|^2 \right). \quad (118)$$

Commentary - I

Let

$$a(\tau) \stackrel{\text{def}}{=} \frac{n - \tau}{\tau(n - 1)}, \quad b(\tau) \stackrel{\text{def}}{=} \frac{n(\tau - 1)}{\tau(n - 1)}.$$

Notice that

- ▶ $a(\tau) + b(\tau) = 1$ for all $\tau \in \{0, 1, \dots, n\}$
- ▶ a is decreasing, with $a(1) = 1$, $a(n) = 0$
- ▶ b is increasing, with $b(1) = 0$, $b(n) = 1$

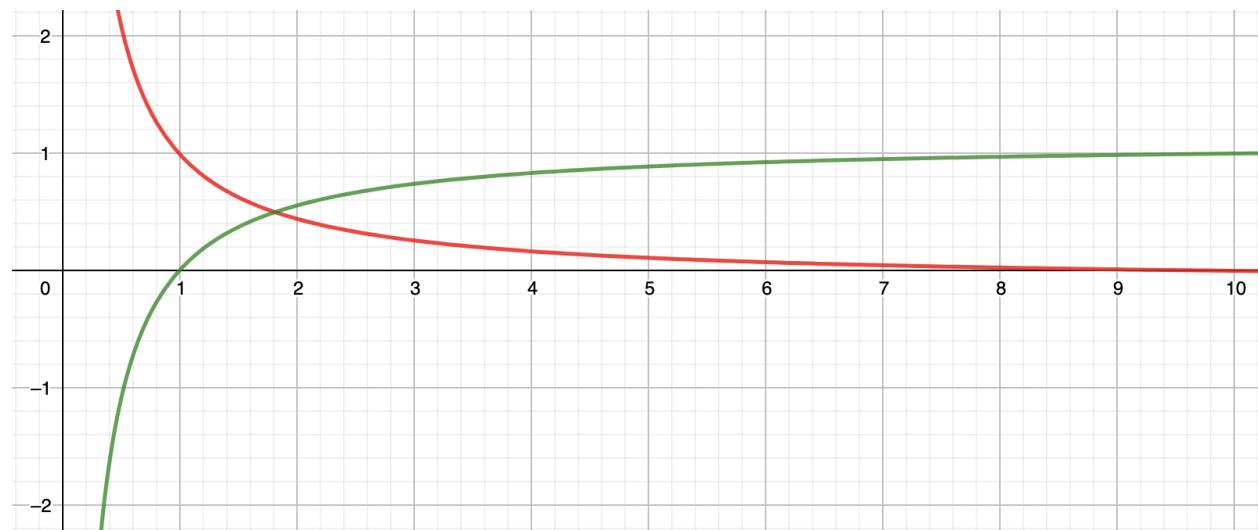


Figure: Functions a and b plotted for $1 \leq \tau \leq n$ and $n = 10$.

Commentary - II

Summary table:

τ	$a(\tau)$	$b(\tau)$	A''	Algorithm
1	1	0	$\max_i L_i$	SGD-US
τ	$\frac{n-\tau}{\tau(n-1)}$	$\frac{n(\tau-1)}{\tau(n-1)}$	$\frac{n-\tau}{\tau(n-1)} \max_i L_i + \frac{n(\tau-1)}{\tau(n-1)} L$	SGD-NICE
n	0	1	L	GD

Key insights:

- ▶ For $\tau = 1$, we recover SGD-US, its maximum stepsize, and hence also its rate
- ▶ For $\tau = n$, we recover GD, its maximum stepsize, and hence also its rate
- ▶ SGD-NICE is therefore a minibatch SGD method that interpolates between SGD-US and GD as τ moves from 1 to n .

Proof of Lemma 244 - I

Unbiasedness. Let χ_i be the random variable defined by

$$\chi_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases}.$$

It is easy to show that

$$\mathbb{E}[\chi_i] = \text{Prob}(i \in S) = \frac{\tau}{n}. \quad (119)$$

Unbiasedness of $g(x)$ now follows via direct computation:

$$\begin{aligned} \mathbb{E}[g(x)] &\stackrel{(116)}{=} \mathbb{E}\left[\frac{1}{\tau} \sum_{i \in S} \nabla f_i(x)\right] = \mathbb{E}\left[\frac{1}{\tau} \sum_{i=1}^n \chi_i \nabla f_i(x)\right] \\ &= \frac{1}{\tau} \sum_{i=1}^n \mathbb{E}[\chi_i] \nabla f_i(x) \\ &\stackrel{(119)}{=} \frac{1}{\tau} \sum_{i=1}^n \text{Prob}(i \in S) \nabla f_i(x) \\ &\stackrel{(119)}{=} \frac{1}{\tau} \sum_{i=1}^n \frac{\tau}{n} \nabla f_i(x) \\ &= \nabla f(x). \end{aligned}$$

Proof of Lemma 244 - II

Expected smoothness (i.e., computing constant A''). Fix $x, y \in \mathbb{R}^d$ and let

$$a_i \stackrel{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y). \quad (120)$$

Let χ_{ij} be the random variable defined by

$$\chi_{ij} = \begin{cases} 1 & i \in S \text{ and } j \in S \\ 0 & \text{otherwise} \end{cases}.$$

Note that

$$\chi_{ij} = \chi_i \chi_j. \quad (121)$$

Further, it is easy to show that

$$\mathbf{E} [\chi_{ij}] = \text{Prob}(i \in S, j \in S) = \frac{\tau(\tau-1)}{n(n-1)}. \quad (122)$$

It is easy to check that for any vectors $b_1, \dots, b_n \in \mathbb{R}^d$ we have the identity

$$\left\| \sum_{i=1}^n b_i \right\|^2 - \sum_{i=1}^n \|b_i\|^2 = \sum_{i \neq j} \langle b_i, b_j \rangle. \quad (123)$$

Proof of Lemma 244 - III

We will use this identity twice in what follows:

$$\begin{aligned}
 \mathbf{E} [\|g(x) - g(y)\|^2] &\stackrel{(116)}{=} \mathbf{E} \left[\left\| \frac{1}{\tau} \sum_{i \in S} \nabla f_i(x) - \frac{1}{\tau} \sum_{i \in S} \nabla f_i(y) \right\|^2 \right] \\
 &\stackrel{(120)}{=} \frac{1}{\tau^2} \mathbf{E} \left[\left\| \sum_{i \in S} a_i \right\|^2 \right] \\
 &= \frac{1}{\tau^2} \mathbf{E} \left[\left\| \sum_{i=1}^n \chi_i a_i \right\|^2 \right] \\
 &\stackrel{(123)}{=} \frac{1}{\tau^2} \mathbf{E} \left[\sum_{i=1}^n \|\chi_i a_i\|^2 + \sum_{i \neq j} \langle \chi_i a_i, \chi_j a_j \rangle \right] \\
 &\stackrel{(121)}{=} \frac{1}{\tau^2} \mathbf{E} \left[\sum_{i=1}^n \|\chi_i a_i\|^2 + \sum_{i \neq j} \chi_{ij} \langle a_i, a_j \rangle \right] \\
 &= \frac{1}{\tau^2} \sum_{i=1}^n \mathbf{E} [\chi_i] \|a_i\|^2 + \sum_{i \neq j} \mathbf{E} [\chi_{ij}] \langle a_i, a_j \rangle. \quad (124)
 \end{aligned}$$

Proof of Lemma 244 - IV

Using the formulas (119) and (122) and the decomposition identity (123) again, we can continue:

$$\begin{aligned}
 \mathbb{E} [\|g(x) - g(y)\|^2] &\stackrel{(124)}{=} \frac{1}{\tau^2} \left(\frac{\tau}{n} \sum_{i=1}^n \|a_i\|^2 + \frac{\tau(\tau-1)}{n(n-1)} \sum_{i \neq j} \langle a_i, a_j \rangle \right) \\
 &= \frac{1}{\tau n} \sum_{i=1}^n \|a_i\|^2 + \frac{\tau-1}{\tau n(n-1)} \sum_{i \neq j} \langle a_i, a_j \rangle \\
 &\stackrel{(123)}{=} \frac{1}{\tau n} \sum_{i=1}^n \|a_i\|^2 + \frac{\tau-1}{\tau n(n-1)} \left(\left\| \sum_{i=1}^n a_i \right\|^2 - \sum_{i=1}^n \|a_i\|^2 \right) \\
 &= \frac{n-\tau}{\tau(n-1)} \frac{1}{n} \sum_{i=1}^n \|a_i\|^2 + \frac{n(\tau-1)}{\tau(n-1)} \left\| \frac{1}{n} \sum_{i=1}^n a_i \right\|^2. \quad (125)
 \end{aligned}$$

Since f_i is convex and L_i -smooth, we know that

$$\|a_i\|^2 \stackrel{(120)}{=} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L_i D_{f_i}(x, y).$$

Since f is convex and L -smooth, we know that

$$\left\| \frac{1}{n} \sum_{i=1}^n a_i \right\|^2 \stackrel{(120)}{=} \|\nabla f(x) - \nabla f(y)\|^2 \leq 2L D_f(x, y).$$

Proof of Lemma 244 - V

It only remains to plug these bounds to (125), apply the bound $L_i \leq \max_i L_i$ and use the identity $D_f(x, y) = \frac{1}{n} \sum_{i=1}^n D_{f_i}(x, y)$:

$$\begin{aligned}\mathbb{E} [\|g(x) - g(y)\|^2] &\stackrel{(125)}{\leq} \frac{n - \tau}{\tau(n - 1)} \frac{1}{n} \sum_{i=1}^n 2L_i D_{f_i}(x, y) + \frac{n(\tau - 1)}{\tau(n - 1)} 2LD_f(x, y) \\ &\leq 2 \frac{n - \tau}{\tau(n - 1)} \max_i L_i \frac{1}{n} \sum_{i=1}^n D_{f_i}(x, y) + 2 \frac{n(\tau - 1)}{\tau(n - 1)} LD_f(x, y) \\ &= 2 \frac{n - \tau}{\tau(n - 1)} \max_i L_i D_f(x, y) + 2 \frac{n(\tau - 1)}{\tau(n - 1)} LD_f(x, y) \\ &= 2 \left(\frac{n - \tau}{\tau(n - 1)} \max_i L_i + \frac{n(\tau - 1)}{\tau(n - 1)} L \right) D_f(x, y).\end{aligned}$$

Variance. Using the variance decomposition and then rewriting the second moment using the exact same steps that led to (125), but with $a'_i = \nabla f_i(y)$ instead of a_i , we

Proof of Lemma 244 - VI

arrive at the identity

$$\begin{aligned}\text{Var}[g(y)] &= \mathbb{E} \left[\|g(y)\|^2 \right] - \|\mathbb{E}[g(y)]\|^2 \\ &= \frac{n - \tau}{\tau(n-1)} \frac{1}{n} \sum_{i=1}^n \|a'_i\|^2 + \frac{n(\tau-1)}{\tau(n-1)} \left\| \frac{1}{n} \sum_{i=1}^n a'_i \right\|^2 - \left\| \frac{1}{n} \sum_{i=1}^n a'_i \right\|^2 \\ &= \frac{n - \tau}{\tau(n-1)} \left(\frac{1}{n} \sum_{i=1}^n \|a'_i\|^2 - \left\| \frac{1}{n} \sum_{i=1}^n a'_i \right\|^2 \right) \\ &= \frac{n - \tau}{\tau(n-1)} \left(\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(y)\|^2 - \|\nabla f(y)\|^2 \right).\end{aligned}$$

Convergence of SGD-NICE - I

By combining

- ▶ Theorem 215 (main convergence theorem for SGD under the AC assumption),
- ▶ Theorem 218 (result that reduced checking the AC assumption to checking expected smoothness: $A = 2A''$ and $C = 2\text{Var}[g(x^*)]$) and
- ▶ Lemma 244 (computation of expected smoothness constant A''),

we arrive at Theorem 245.

Convergence of SGD-NICE - II

Theorem 245

Consider the SGD-NICE method (Algorithm 5) with minibatch size $\tau \in \{1, 2, \dots, n\}$. Then the variance of the gradient estimator satisfies

$$\text{Var}[g(x^*)] = \frac{n - \tau}{\tau(n - 1)} \left(\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*)\|^2 - \|\nabla f(x^*)\|^2 \right). \quad (126)$$

Further, assume that f_i is convex and L_i -smooth for all i , and that f is μ -convex and L -smooth. Let

$$A'' = \frac{n - \tau}{\tau(n - 1)} \max_i L_i + \frac{n(\tau - 1)}{\tau(n - 1)} L,$$

$A = 2A''$ and $C = 2\text{Var}[g(x^*)]$. Finally, choose a stepsize satisfying

$$0 < \gamma \leq \frac{1}{A}. \quad (127)$$

Then the iterates $\{x^k\}_{k \geq 0}$ of SGD-NICE (Algorithm 5) satisfy

$$\mathbb{E} \left[\|x^k - x^*\|^2 \right] \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2 + \frac{\gamma C}{\mu}. \quad (128)$$

Convergence of SGD-NICE - III

Let's now choose a small enough stepsize so that the method would converge to an ε -accurate solution.

Choose any relative error tolerance $0 < \delta < 1$ and stepsize

$$\gamma = \min \left\{ \frac{1}{2A''}, \frac{\mu \delta \|x^0 - x^*\|^2}{2\text{Var}[g(x^*)]} \right\}.$$

Then

$$k \geq \max \left\{ \frac{2A''}{\mu}, \frac{2\text{Var}[g(x^*)]}{\delta \mu^2 \|x^0 - x^*\|^2} \right\} \log \left(\frac{1}{\delta} \right) \Rightarrow \mathbf{E} [\|x^k - x^*\|^2] \leq 2\delta \|x^0 - x^*\|^2.$$

Moving from relative error δ to absolute error $\varepsilon = 2\delta \|x^0 - x^*\|^2$, the above translates to

$$k \geq \max \left\{ \frac{2A''}{\mu}, \frac{4\text{Var}[g(x^*)]}{\varepsilon \mu^2} \right\} \log \left(\frac{2 \|x^0 - x^*\|^2}{\varepsilon} \right) \Rightarrow \mathbf{E} [\|x^k - x^*\|^2] \leq \varepsilon.$$

Optimal Minibatch Size - I

Given the above convergence result, we may wish to ask which minibatch size is optimal with respect to the **total complexity** of SGD-NICE, defined as the **product of the number of iterations and the cost of one iteration**. Since

- ▶ the number of iterations is $\max \left\{ \frac{2A''}{\mu}, \frac{4\text{Var}[g(x^*)]}{\varepsilon\mu^2} \right\}$ (we ignore the logarithmic factor which does not depend on τ), and
- ▶ cost of each iteration is τ ,

we arrive at the following **total complexity minimization** problem:

$$\min_{1 \leq \tau \leq n} \mathcal{C}(\tau),$$

where

$$\mathcal{C}(\tau) \stackrel{\text{def}}{=} \frac{2}{\mu(n-1)} \max \left\{ \underbrace{(n-\tau) \max_i L_i + n(\tau-1)L}_{\text{increasing linear}}, \underbrace{(n-\tau) \frac{2\sigma_*^2}{\varepsilon\mu}}_{\text{decreasing linear}} \right\}.$$

Optimal Minibatch Size - II

Observations:

- ▶ If $\sigma_*^2 = 0$ (e.g., in the interpolation regime), then the “decreasing” part is equal to zero, and hence $\mathcal{C}(\tau)$ is an increasing function. So, the optimal minibatch size is

$$\tau^* = 1.$$

- ▶ Further, if σ_*^2 is not too large, then the increasing linear function dominates the decrasing linear function on the interval $[1, n]$, and hence the optimal minibatch size is again

$$\tau^* = 1.$$

This happens for

$$\sigma_*^2 \leq \frac{\varepsilon \mu \max_i L_i}{2}.$$

Optimal Minibatch Size - III

- ▶ Otherwise, the increasing linear and the decreasing linear lines intersect on $(1, n)$, and the optimal minibatch size can be found by computing the intersection:

$$\tau^* = \frac{n(\theta + L - \max_i L_i)}{\theta + nL - \max_i L_i},$$

where $\theta = \frac{2\sigma_*^2}{\varepsilon\mu}$.

- ▶ Notice that

$$\sigma_*^2 = \frac{\varepsilon\mu \max_i L_i}{2} \Rightarrow \theta = \max_i L_i \Rightarrow \tau^* = \frac{n(\max_i L_i + L - \max_i L_i)}{\max_i L_i + nL - \max_i L_i} = 1.$$

- ▶ Notice that

$$\sigma_*^2 \rightarrow \infty \Rightarrow \tau^* \rightarrow n.$$

Key takeaway: If we care about the total complexity of SGD-NICE, the larger the quantity $\sigma_*^2 = \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*)\|^2 - \|\nabla f(x^*)\|^2$ is, the larger minibatch size should be chosen. As this quantity grows, the optimal minibatch size approaches n . On the other hand, in the “small” σ_*^2 regime (and in particular if $\sigma_*^2 = 0$), the optimal minibatch size is $\tau^* = 1$.

Sampling without Replacement (Independent Sampling)

Sampling without Replacement: Independent Sampling

- ▶ Let p_1, p_2, \dots, p_n be probabilities ($0 < p_i \leq 1$ for all i).
- ▶ We do **not** require these probabilities to add up to 1! So, $\sum_i p_i$ can be anywhere in the interval $(0, n]$.

For each i define a random set as follows:

$$\textcolor{teal}{S}_i \stackrel{\text{def}}{=} \begin{cases} \{i\} & \text{with probability } p_i \\ \emptyset & \text{with probability } 1 - p_i \end{cases}.$$

We now define a random subset $\textcolor{teal}{S} \subseteq \{1, 2, \dots, n\}$ by taking the union of these simple sets

$$\textcolor{teal}{S} \stackrel{\text{def}}{=} \bigcup_{i=1}^n \textcolor{teal}{S}_i. \tag{129}$$

Define the gradient estimator via

$$g(x) \stackrel{\text{def}}{=} \sum_{i \in S} \frac{1}{np_i} \nabla f_i(x). \tag{130}$$

SGD-IND: The Algorithm

Gradient estimator (130) leads to the following new variant of SGD:

Algorithm 6 SGD-IND

- ```

1: Parameters: learning rate $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, probabilities
 $0 < p_i \leq 1$ for $i = 1, 2, \dots, n$
2: for $k = 0, 1, 2, \dots$ do
3: Sample set $S^k = \cup_{i=1}^n S_i^k$, where $S_i^k = \{i\}$ with probability p_i ;
4: $g^k = \sum_{i \in S^k} \frac{1}{np_i} \nabla f_i(x^k)$ obtain a stochastic gradient
5: $x^{k+1} = x^k - \gamma g^k$

```

# Minibatch Size

Note that  $S$  has a random size/cardinality. In such cases, we will use the word **minibatch size** to refer to the **expected cardinality**:

$$\tau = \mathbb{E}[|S|].$$

Note that

$$\mathbb{E}[|S|] = \mathbb{E} \left[ \sum_{i=1}^n |S_i| \right] = \sum_{i=1}^n \mathbb{E}[|S_i|] = \sum_{i=1}^n 1p_i + 0(1-p_i) = \sum_{i=1}^n p_i. \quad (131)$$

- ▶ If we choose  $\tau = n$ , we must necessarily have  $p_i = 1$  for all  $i$ ,  $S \equiv \{1, 2, \dots, n\}$  and hence we recover the gradient estimator used by GD:  $g(x) = \nabla f(x)$ .
- ▶ If we choose  $\tau = 1$ , we **do not** recover the gradient estimator used by SGD-NS:

| Algorithm | Gradient estimator $g(x)$                     | Minibatch size $\tau$ |
|-----------|-----------------------------------------------|-----------------------|
| SGD-NS    | $\frac{1}{np_i} \nabla f_i(x)$                | 1 (deterministically) |
| SGD-IND   | $\sum_{i \in S} \frac{1}{np_i} \nabla f_i(x)$ | 1 (in expectation)    |

# SGD-IND: Unbiasedness and Expected Smoothness

## Lemma 246

The gradient estimator  $\hat{g}$  defined in (130) is unbiased. If we further assume that  $f_i$  is convex and  $L_i$ -smooth for all  $i$ , and  $f$  is  $L$ -smooth, then

$$\mathbb{E} \left[ \|\hat{g}(x) - \hat{g}(y)\|^2 \right] \leq 2A'' D_f(x, y),$$

where

$$A'' = \frac{\max_i \left( \frac{1}{p_i} - 1 \right) L_i}{n} + L. \quad (132)$$

# Commentary - I

**Minibatch size = 1:**

| Algorithm / probabilities | uniform ( $p_i = 1/n$ )        | nonuniform                                                |
|---------------------------|--------------------------------|-----------------------------------------------------------|
| SGD-NS                    | $\max_i L_i$                   | $\max_i \frac{L_i}{np_i}$                                 |
| SGD-IND                   | $L + \frac{n-1}{n} \max_i L_i$ | $\frac{\max_i \left(\frac{1}{p_i} - 1\right) L_i}{n} + L$ |

Table: The value of  $A''$  for two variants of SGD under uniform and nonuniform probabilities.

## Commentary - II

Note that in the  $\tau = 1$  case with uniform probabilities, we have

$$\begin{aligned} A''_{\text{SGD-IND}} &= \textcolor{red}{L} + \frac{n-1}{n} \max_i \textcolor{red}{L}_i \\ &= \frac{1}{n} n \textcolor{red}{L} + \left(1 - \frac{1}{n}\right) \max_i \textcolor{red}{L}_i \\ &\geq \frac{1}{n} \max_i \textcolor{red}{L}_i + \left(1 - \frac{1}{n}\right) \max_i \textcolor{red}{L}_i \\ &= \max_i \textcolor{red}{L}_i \\ &= A''_{\text{SGD-US}} \end{aligned}$$

where the inequality follows from Lemma 234(ii). So, SGD-IND has a worse  $A''$  constant than SGD-US, which means its rate is worse.

**Minibatch size =  $n$ :** In the  $\tau = n$  case, we recover GD, its maximum stepsize, and hence also its rate since  $A'' = \textcolor{red}{L}$

# Commentary - III

## More insights:

- ▶ The estimator (130) thus leads to a minibatch SGD method that interpolates between something “similar” to SGD-NS and GD as  $\tau$  moves from 1 to  $n$ .
- ▶ Unlike in the case of  $\tau$ -nice sampling, here we can make use of nonuniform probabilities, which means we can think about constructing **importance sampling for minibatches**.

# Proof of Lemma 246 - I

**Unbiasedness.** As before, let  $\mathbf{x}_i$  be the random variable defined by

$$\mathbf{x}_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases}.$$

It is easy to show that

$$\mathbf{E}[\mathbf{x}_i] = \text{Prob}(i \in S) = \text{Prob}(i \in S_i) = p_i. \quad (133)$$

Unbiasedness of  $g(x)$  now follows via direct computation:

$$\begin{aligned} \mathbf{E}[g(x)] &\stackrel{(130)}{=} \mathbf{E}\left[\sum_{i \in S} \frac{1}{np_i} \nabla f_i(x)\right] \\ &= \mathbf{E}\left[\sum_{i=1}^n \mathbf{x}_i \frac{1}{np_i} \nabla f_i(x)\right] \\ &= \sum_{i=1}^n \mathbf{E}[\mathbf{x}_i] \frac{1}{np_i} \nabla f_i(x) \\ &\stackrel{(133)}{=} \sum_{i=1}^n \frac{1}{n} \nabla f_i(x) \\ &= \nabla f(x). \end{aligned}$$

## Proof of Lemma 246 - II

**Expected smoothness (i.e., computing  $A''$ ).** Fix  $x, y \in \mathbb{R}^d$  and let

$$a_i \stackrel{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y). \quad (134)$$

We will use this identity twice in what follows:

$$\begin{aligned} \mathbf{E} [\|g(x) - g(y)\|^2] &\stackrel{(130)}{=} \mathbf{E} \left[ \left\| \sum_{i \in S} \frac{1}{np_i} \nabla f_i(x) - \sum_{i \in S} \frac{1}{np_i} \nabla f_i(y) \right\|^2 \right] \\ &\stackrel{(134)}{=} \mathbf{E} \left[ \left\| \sum_{i \in S} \frac{a_i}{np_i} \right\|^2 \right] \\ &= \mathbf{E} \left[ \left\| \sum_{i=1}^n \chi_i \frac{a_i}{np_i} \right\|^2 \right] \\ &\stackrel{(123)}{=} \mathbf{E} \left[ \sum_{i=1}^n \left\| \chi_i \frac{a_i}{np_i} \right\|^2 + \sum_{i \neq j} \left\langle \chi_i \frac{a_i}{np_i}, \chi_j \frac{a_j}{np_j} \right\rangle \right] \\ &= \mathbf{E} \left[ \sum_{i=1}^n \chi_i \left\| \frac{a_i}{np_i} \right\|^2 + \sum_{i \neq j} \chi_i \chi_j \left\langle \frac{a_i}{np_i}, \frac{a_j}{np_j} \right\rangle \right] \\ &= \sum_{i=1}^n \mathbf{E} [\chi_i] \left\| \frac{a_i}{np_i} \right\|^2 + \sum_{i \neq j} \mathbf{E} [\chi_i \chi_j] \left\langle \frac{a_i}{np_i}, \frac{a_j}{np_j} \right\rangle. \end{aligned} \quad (135)$$

## Proof of Lemma 246 - III

Since  $E[\mathbf{x}_i] = p_i$ , and since by independence we have  $E[\mathbf{x}_i \mathbf{x}_j] = E[\mathbf{x}_i] E[\mathbf{x}_j] = p_i p_j$ , we can further write

$$\begin{aligned}
E\left[\|g(x) - g(y)\|^2\right] &= \frac{1}{n^2} \sum_{i=1}^n \frac{1}{p_i} \|a_i\|^2 + \sum_{i \neq j} \left\langle \frac{a_i}{n}, \frac{a_j}{n} \right\rangle \\
&\stackrel{(123)}{=} \frac{1}{n^2} \sum_{i=1}^n \frac{1}{p_i} \|a_i\|^2 + \left( \left\| \sum_{i=1}^n \frac{a_i}{n} \right\|^2 - \sum_{i=1}^n \left\| \frac{a_i}{n} \right\|^2 \right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \left( \frac{1}{p_i} - 1 \right) \|a_i\|^2 + \left\| \sum_{i=1}^n \frac{a_i}{n} \right\|^2. \tag{136}
\end{aligned}$$

Since  $f_i$  is convex and  $L_i$ -smooth, we know that

$$\|a_i\|^2 \stackrel{(134)}{=} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L_i D_{f_i}(x, y).$$

Since  $f$  is convex and  $L$ -smooth, we know that

$$\left\| \frac{1}{n} \sum_{i=1}^n a_i \right\|^2 \stackrel{(134)}{=} \|\nabla f(x) - \nabla f(y)\|^2 \leq 2L D_f(x, y).$$

Plugging these estimates into (136), and using the identity

$$D_f(x, y) = \frac{1}{n} \sum_{i=1}^n D_{f_i}(x, y),$$

## Proof of Lemma 246 - IV

we finally get

$$\begin{aligned}\mathbf{E} [\|g(x) - g(y)\|^2] &\leq \frac{1}{n^2} \sum_{i=1}^n \left( \frac{1}{p_i} - 1 \right) 2\textcolor{red}{L}_i D_{f_i}(x, y) + 2\textcolor{red}{L} D_f(x, y) \\ &\leq 2 \frac{\max_i \left( \frac{1}{p_i} - 1 \right) \textcolor{red}{L}_i}{n} \frac{1}{n} \sum_{i=1}^n D_{f_i}(x, y) + 2\textcolor{red}{L} D_f(x, y) \\ &= 2 \frac{\max_i \left( \frac{1}{p_i} - 1 \right) \textcolor{red}{L}_i}{n} D_f(x, y) + 2\textcolor{red}{L} D_f(x, y) \\ &= 2 \left( \frac{\max_i \left( \frac{1}{p_i} - 1 \right) \textcolor{red}{L}_i}{n} + \textcolor{red}{L} \right) D_f(x, y).\end{aligned}$$

# Sampling with Replacement

# Sampling with Replacement: Multisampling

Let  $q_1, q_2, \dots, q_n$  be probabilities summing up to 1 and let  $s$  be the random variable equal to  $i$  with probability  $q_i$ . Fix a minibatch size  $\tau \in \{1, 2, \dots\}$  and let  $s_1, s_2, \dots, s_\tau$  be independent copies of  $s$ . Define the gradient estimator via

$$g(x) \stackrel{\text{def}}{=} \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{nq_{s_t}} \nabla f_{s_t}(x). \quad (137)$$

Gradient estimator (137) leads to the following new variant of SGD:

---

## Algorithm 7 SGD-MULTI

---

- 1: **Parameters:** learning rate  $\gamma > 0$ , starting point  $x^0 \in \mathbb{R}^d$ , positive probabilities  $q_1, \dots, q_n$  summing up to 1, minibatch size  $\tau \in \{1, 2, \dots\}$
  - 2: **for**  $k = 0, 1, 2, \dots$  **do**
  - 3:     Sample  $\tau$  i.i.d. random variables  $s_1^k, \dots, s_\tau^k$ , where each is equal to  $i$  with probability  $q_i$
  - 4:     
$$g^k = \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{nq_{s_t^k}} \nabla f_{s_t^k}(x^k)$$
 obtain a stochastic gradient
  - 5:     
$$x^{k+1} = x^k - \gamma g^k$$
-

# SGD-MULTI: Unbiasedness and Expected Smoothness

## Lemma 247

The gradient estimator  $\hat{g}$  defined in (137) is unbiased. If we further assume that  $f_i$  is convex and  $L_i$ -smooth for all  $i$ , and  $f$  is  $L$ -smooth, then

$$\mathbb{E} \left[ \|\hat{g}(x) - \hat{g}(y)\|^2 \right] \leq 2A'' D_f(x, y),$$

where

$$A'' = \frac{1}{\tau} \left( \max_i \frac{L_i}{nq_i} \right) + \left( 1 - \frac{1}{\tau} \right) L. \quad (138)$$

# Commentary - I

Let

$$a(\tau) \stackrel{\text{def}}{=} \frac{1}{\tau}, \quad b(\tau) \stackrel{\text{def}}{=} 1 - \frac{1}{\tau}.$$

Notice that

- ▶  $a(\tau) + b(\tau) = 1$  for all  $\tau \in \{0, 1, \dots\}$
- ▶  $a$  is decreasing, with  $a(1) = 1$ ,  $a(+\infty) = 0$
- ▶  $b$  is increasing, with  $b(1) = 0$ ,  $b(+\infty) = 1$

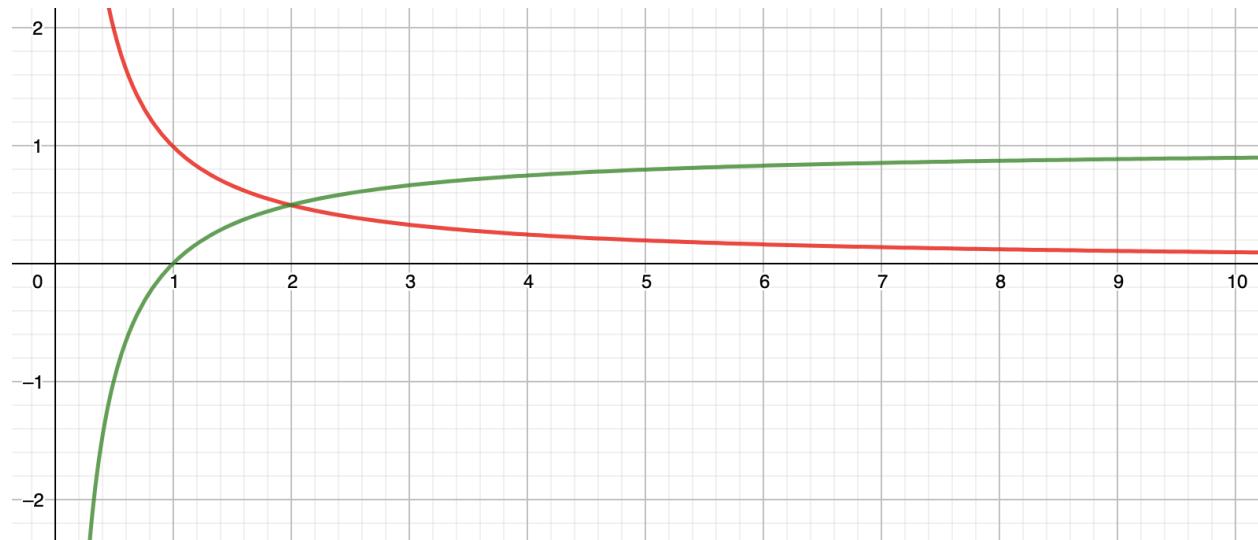


Figure: Functions  $a$  and  $b$  plotted for  $1 \leq \tau \leq +\infty$ .

## Commentary - II

Summary table:

| $\tau$    | $a(\tau)$        | $b(\tau)$            | $A''$                                                                                       | Algorithm |
|-----------|------------------|----------------------|---------------------------------------------------------------------------------------------|-----------|
| 1         | 1                | 0                    | $\max_i \frac{L_i}{nq_i}$                                                                   | SGD-NS    |
| $\tau$    | $\frac{1}{\tau}$ | $1 - \frac{1}{\tau}$ | $\frac{1}{\tau} \left( \max_i \frac{L_i}{nq_i} \right) + \left(1 - \frac{1}{\tau}\right) L$ | SGD-MULTI |
| $+\infty$ | 0                | 1                    | $L$                                                                                         | GD        |

### Key insights:

- ▶ For  $\tau = 1$ , we recover SGD-NS, its maximum stepsize, and hence also its rate
- ▶ For  $\tau = +\infty$ , we recover GD, its maximum stepsize, and hence also its rate
- ▶ The estimator (116) thus leads to a minibatch SGD method that interpolates between SGD-NS and GD as  $\tau$  moves from 1 to  $+\infty$ .

# Proof of Lemma 247 - I

**Unbiasedness.** Unbiasedness of  $g(x)$  follows via direct computation:

$$\begin{aligned}\mathbf{E}[g(x)] &= \mathbf{E} \left[ \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{nq_{st}} \nabla f_{st}(x) \right] \\ &= \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbf{E} \left[ \frac{1}{nq_{st}} \nabla f_{st}(x) \right] \\ &= \frac{1}{\tau} \sum_{t=1}^{\tau} \sum_{i=1}^n q_i \frac{1}{nq_i} \nabla f_i(x) \\ &= \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{n} \sum_{i=1}^n \nabla f_i(x) \\ &= \frac{1}{\tau} \sum_{t=1}^{\tau} \nabla f(x) \\ &= \nabla f(x).\end{aligned}$$

**Expected smoothness (i.e., computing  $A''$ ).** Fix  $x, y \in \mathbb{R}^d$  and let

$$a_i \stackrel{\text{def}}{=} \nabla f_i(x) - \nabla f_i(y). \quad (139)$$

# Proof of Lemma 247 - II

Then

$$\begin{aligned}
 \mathbf{E} [\|g(x) - g(y)\|^2] &= \mathbf{E} \left[ \left\| \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{nq_{s_t}} \nabla f_{s_t}(x) - \frac{1}{\tau} \sum_{t=1}^{\tau} \frac{1}{nq_{s_t}} \nabla f_{s_t}(y) \right\|^2 \right] \\
 &\stackrel{(139)}{=} \frac{1}{\tau^2} \mathbf{E} \left[ \left\| \sum_{t=1}^{\tau} \frac{a_{s_t}}{nq_{s_t}} \right\|^2 \right] \\
 &= \frac{1}{\tau^2} \mathbf{E} \left[ \sum_{t=1}^{\tau} \left\| \frac{a_{s_t}}{nq_{s_t}} \right\|^2 + \sum_{t \neq u}^{\tau} \left\langle \frac{a_{s_t}}{nq_{s_t}}, \frac{a_{s_u}}{nq_{s_u}} \right\rangle \right] \\
 &= \frac{1}{\tau^2} \sum_{t=1}^{\tau} \mathbf{E} \left[ \left\| \frac{a_{s_t}}{nq_{s_t}} \right\|^2 \right] + \frac{1}{\tau^2} \sum_{t \neq u}^{\tau} \mathbf{E} \left[ \left\langle \frac{a_{s_t}}{nq_{s_t}}, \frac{a_{s_u}}{nq_{s_u}} \right\rangle \right] \quad (140)
 \end{aligned}$$

## Proof of Lemma 247 - III

We now separately bound the two terms in (140) by a multiple of the Bregman divergence  $D_f(x, y)$ . First, we estimate

$$\begin{aligned} \mathbf{E} \left[ \left\| \frac{\mathbf{a}_{\textcolor{teal}{s}_t}}{nq_{\textcolor{teal}{s}_t}} \right\|^2 \right] &= \sum_{i=1}^n q_i \left\| \frac{\mathbf{a}_i}{nq_i} \right\|^2 \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{nq_i} \|\mathbf{a}_i\|^2 \\ &\stackrel{(139)}{=} \frac{1}{n} \sum_{i=1}^n \frac{1}{nq_i} \|\nabla f_i(x) - \nabla f_i(y)\|^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \frac{2L_i}{nq_i} D_{f_i}(x, y) \\ &\leq 2 \left( \max_i \frac{L_i}{nq_i} \right) \frac{1}{n} \sum_{i=1}^n D_{f_i}(x, y) \\ &= 2 \left( \max_i \frac{L_i}{nq_i} \right) D_f(x, y). \end{aligned} \tag{141}$$

The first inequality follows from Proposition ?? (iii) since  $f_i$  is convex and  $L_i$ -smooth, the second inequality follows by bounding  $\frac{L_i}{nq_i} \leq \max_i \frac{L_i}{nq_i}$ , and the last identity was the subject of an exercise.

# Proof of Lemma 247 - IV

Next, since  $s_t$  and  $s_u$  are independent for  $t \neq u$ , we have

$$\begin{aligned}
\mathbb{E} \left[ \left\langle \frac{a_{s_t}}{nq_{s_t}}, \frac{a_{s_u}}{nq_{s_u}} \right\rangle \right] &= \left\langle \mathbb{E} \left[ \frac{a_{s_t}}{nq_{s_t}} \right], \mathbb{E} \left[ \frac{a_{s_u}}{nq_{s_u}} \right] \right\rangle \\
&= \left\langle \sum_{i=1}^n q_i \frac{a_i}{nq_i}, \sum_{i=1}^n q_i \frac{a_i}{nq_i} \right\rangle \\
&= \left\| \frac{1}{n} \sum_{i=1}^n a_i \right\|^2 \\
&\stackrel{(139)}{=} \|\nabla f(x) - \nabla f(y)\|^2 \\
&\leq 2LD_f(x, y). \tag{142}
\end{aligned}$$

Finally, plugging (141) and (142) into (140), we obtain

$$\begin{aligned}
\mathbb{E} [\|g(x) - g(y)\|^2] &\leq \frac{1}{\tau^2} \sum_{t=1}^{\tau} 2 \left( \max_i \frac{L_i}{nq_i} \right) D_f(x, y) + \frac{1}{\tau^2} \sum_{t \neq u} 2LD_f(x, y) \\
&= 2 \left( \frac{1}{\tau} \left( \max_i \frac{L_i}{nq_i} \right) + \left( 1 - \frac{1}{\tau} \right) L \right) D_f(x, y),
\end{aligned}$$

as desired.

# Exercises

# Exercises I

## Exercise 248

Show that if  $S$  is a  $\tau$ -nice sampling (i.e., a random subset of  $\{1, 2, \dots, n\}$  of cardinality  $\tau$  chosen uniformly at random), then

$$\text{Prob}(i \in S) = \frac{\tau}{n}.$$

## Exercise 249

Show that if  $S$  is a  $\tau$ -nice sampling (i.e., a random subset of  $\{1, 2, \dots, n\}$  of cardinality  $\tau$  chosen uniformly at random), then

$$\text{Prob}(i \in S, j \in S) = \frac{\tau(\tau-1)}{n(n-1)}.$$

## Exercises II

### Exercise 250

Prove identity (123). That is, prove that for any vectors  $b_1, \dots, b_n \in \mathbb{R}^d$ ,

$$\left\| \sum_{i=1}^n b_i \right\|^2 - \sum_{i=1}^n \|b_i\|^2 = \sum_{i \neq j} \langle b_i, b_j \rangle.$$

### Exercise 251 (Second moment of SGD-IND)

Compute  $\mathbf{E} [\|g(y)\|^2]$  for the gradient estimator  $g(x)$  defined in (130).

# Introduction to Optimization

Peter Richtárik



## Lecture 24: Markowitz Mean-Variance Portfolio via Quadratic Programming

# Lecture Outline

- ▶ In the lectures devoted to detecting arbitrage via linear programming, we attempted to form a “good” portfolio with **zero risk**
- ▶ Instead, we will now:
  - ▶ Describe a model (Markowitz model) for designing a “good” portfolio in situations when there is **risk**
  - ▶ In particular, we will design a portfolio trading **risk** for **expected return** (hence the term “mean-variance portfolio”)

# Markowitz Portfolio Model

## Model 252 (Markowitz Portfolio)

1. **two time periods:** now (time 0) and future (time 1)
2.  $S^1, \dots, S^n$ : **assets** with random returns
3.  $r_i$ : **random return** of asset  $S^i$

$$r = (r_1, \dots, r_n)^\top \in \mathbb{R}^n \quad \text{vector of random returns}$$

4.  $\mu_i = \mathbf{E}[r_i]$  : **expected return** of asset  $S^i$

$$\mu = (\mu_1, \dots, \mu_n)^\top \in \mathbb{R}^n \quad \text{vector of expected returns}$$

5.  $\Sigma$  : **covariance matrix** for the returns ( $n \times n$  **positive definite**)

$$\Sigma = \mathbf{E}[(r - \mu)(r - \mu)^\top] \in \mathbb{R}^{n \times n}, \quad \Sigma_{ij} = \mathbf{E}[(r_i - \mu_i)(r_j - \mu_j)] \in \mathbb{R}$$

6.  $x_i$  : funds invested in asset  $S^i$

$$x = (x_1, \dots, x_n)^\top \in \mathbb{R}^n \quad \text{vector representing a **portfolio** of assets}$$

# Portfolio: Expected Return and Variance

Random return of portfolio  $x$ :

$$r(x) = \sum_{i=1}^n r_i x_i = r^\top x$$

Expected return of portfolio  $x$ :

$$\mathbb{E}[r(x)] = \sum_{i=1}^n \mathbb{E}[r_i] x_i = \sum_{i=1}^n \mu_i x_i = \mu^\top x$$

Variance of the return of portfolio  $x$ :

$$\begin{aligned}\sigma^2(x) &\stackrel{\text{def}}{=} \mathbb{V}\text{ar}[r(x)] = \mathbb{E}[(r(x) - \mathbb{E}[r(x)])^2] \\&= \mathbb{E}[(r^\top x - \mu^\top x)^2] \\&= \mathbb{E}[x^\top (r - \mu)(r - \mu)^\top x] \\&= x^\top \underbrace{\mathbb{E}[(r - \mu)(r - \mu)^\top]}_{\Sigma} x = x^\top \Sigma x.\end{aligned}$$

# Example: Positive Semidefiniteness Assumption

## Example 253

Recall that we assume that  $\Sigma$  is positive definite.

- (i) However, each covariance matrix is necessarily positive semidefinite. Show this.
- (ii) So, the assumption on positive definiteness is really equivalent to requiring that  $x^\top \Sigma x \neq 0$  for all  $x \neq 0$ . What does this *mean* in the language of our model?

### Solution:

- (i) Indeed, for any  $x \in \mathbb{R}^n$  we have

$$\begin{aligned} x^\top \Sigma x &= x^\top \mathbf{E} [(r - \mu)(r - \mu)^\top] x \\ &= \mathbf{E} [x^\top (r - \mu)(r - \mu)^\top x] = \mathbf{E} [(x^\top (r - \mu))^2] \geq 0. \end{aligned}$$

- (ii) In other words, the question asks: what would it mean if there was a nonzero portfolio  $x$  for whose **variance**,  $x^\top \Sigma x$ , was 0? Intuitively, since its variance is 0, it means that there are some “redundancies” in the model: the returns of some assets depend deterministically on the return of others. We can keep removing these assets from the model until the assumption is satisfied.

# Set of Admissible Portfolios $\mathcal{X} \subset \mathbb{R}^n$

The set of **admissible portfolios**, denoted by  $\mathcal{X}$ , is the set of portfolios the modeler is interested in. This will typically be a much smaller set than  $\mathbb{R}^n$  as the modeler wants to impose various constraints.

- ▶ We typically assume that  $\mathcal{X}$  is **convex** for tractability purposes.
  - ▶ Sometimes we will assume that the set is **polyhedral**, i.e., described by a finite number of linear equations and inequalities:

$$\mathcal{X} = \{x \in \mathbb{R}^n : \mathbf{A}x = b, \mathbf{C}x \geq d\}.$$

Advantage of this: we will get **convex quadratic programming**

- ▶ Set  $\mathcal{X}$  can be constructed to model **practical constraints**, such as
  - ▶ **budget constraint:** We often assume that

$$\sum_{i=1}^n x_i = 1 \quad \text{for all } x \in \mathcal{X}$$

(i.e., we have a unit budget to form the portfolio)

- ▶ **short selling limitations:**  $x_i \geq -\delta_i, i = 1, 2, \dots, n$
- ▶ **no short selling:**  $x_i \geq 0$  for all  $i$
- ▶ **diversification:**  $x_i \leq m$  for all  $i$
- ▶ **transaction costs**

# Efficient Portfolios

## Definition 254 (Efficient portfolios)

We say that a portfolio  $x$  is **efficient** if it satisfies one of the following conditions:

- (1) it has the **maximum expected return** among all admissible portfolios of the **same (or smaller) variance**:

$$(EP1) \quad \begin{aligned} & \max_{x \in \mathbb{R}^n} \quad \mu^\top x \\ & \text{subject to} \quad x^\top \Sigma x \leq \sigma^2 \\ & \quad \quad \quad x \in \mathcal{X} \end{aligned}$$

- (2) it has the **minimum variance** among all admissible portfolios with the **same (or larger) expected return**:

$$(EP2) \quad \begin{aligned} & \min_{x \in \mathbb{R}^n} \quad x^\top \Sigma x \\ & \text{subject to} \quad \mu^\top x \geq R \\ & \quad \quad \quad x \in \mathcal{X} \end{aligned}$$

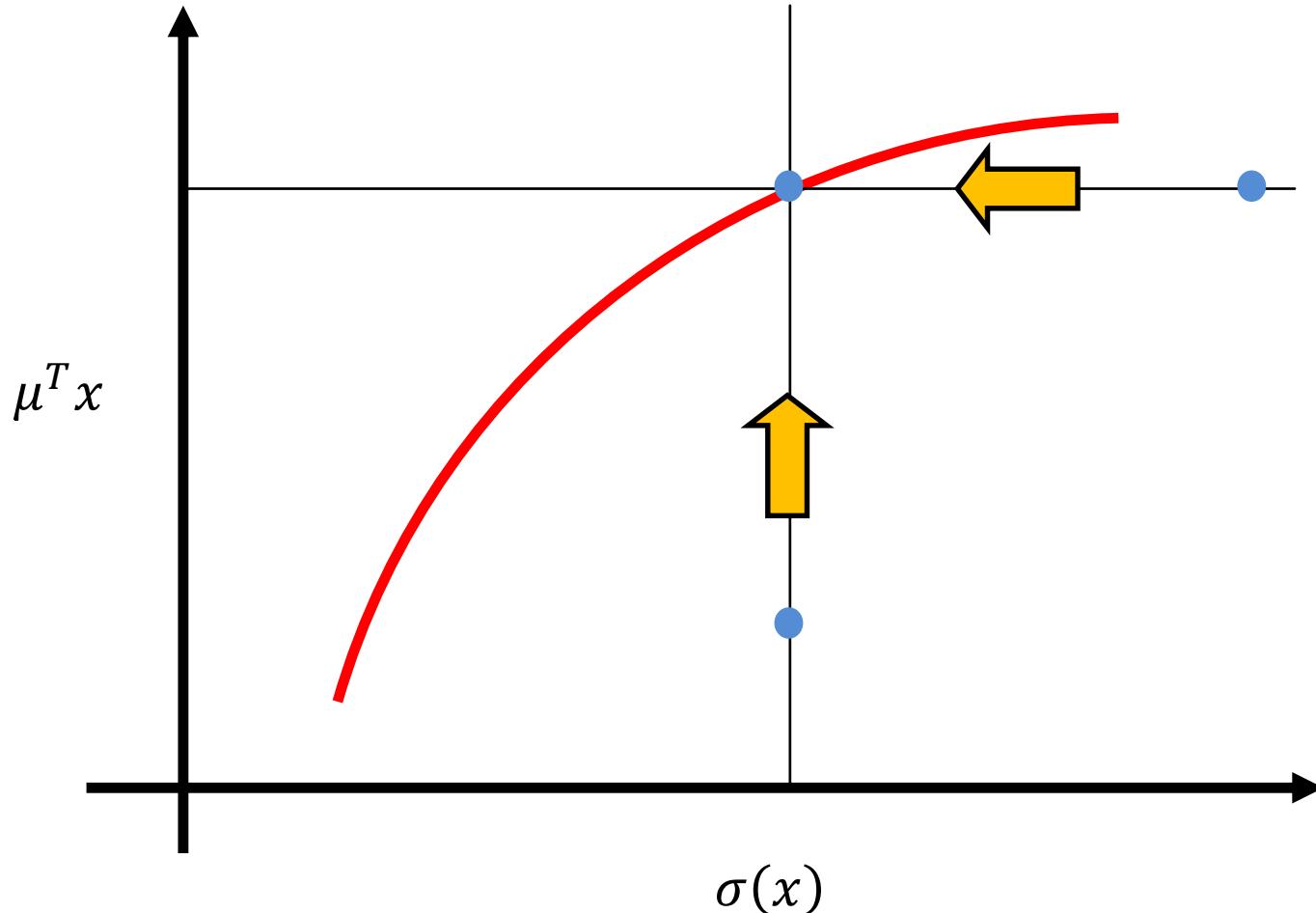
- (3) its **risk-adjusted return is maximized** among all admissible portfolios:

# Efficient Portfolios (Cont.)

$$\max_{x \in \mathbb{R}^n} \underbrace{\mu^\top x - \delta x^\top \Sigma x}_{\text{risk-adjusted return}} \quad (EP3) \quad \text{subject to} \quad x \in \mathcal{X}$$

- ▶ The parameter  $\delta > 0$  sets the relative importance of the **expected return**  $\mu^\top x$  and the **variance**  $x^\top \Sigma x$ 
  - ▶ Choose a small value of  $\delta$  if you have small sensitivity to risk
  - ▶ Choose a large value of  $\delta$  if you have large sensitivity to risk
- ▶ Problem: neither **expected return** nor **variance** are controlled directly

# Efficient Portfolios: Visualization via Efficient Frontier



Risk of portfolio  $x$ :  $\sigma(x) \stackrel{\text{def}}{=} (x^\top \Sigma x)^{1/2}$  (standard deviation of the returns)

# Minimum Risk as a Function of Required Expected Return

Consider this family of optimization problems parameterized by  $R$ :

$$\begin{aligned}\sigma(R) &\stackrel{\text{def}}{=} \min_{x \in \mathbb{R}^n} && (x^\top \Sigma x)^{1/2} \\ (\dagger) \quad &\text{subject to} && \mu^\top x \geq R \\ &&& x \in \mathcal{X}\end{aligned}$$

- ▶  $\sigma(R)$  is the **smallest risk** of an admissible portfolio whose **expected return is at least  $R$**
- ▶ Define

$$\mathcal{X}(R) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : \mu^\top x \geq R, x \in \mathcal{X}\}$$

and let  $R_{max} > 0$  be a constant for which the set  $\mathcal{X}(R_{max})$  is nonempty.

## Exercise 255

Show that  $\mathcal{X}(R) \neq \emptyset$  for all  $R \leq R_{max}$ .

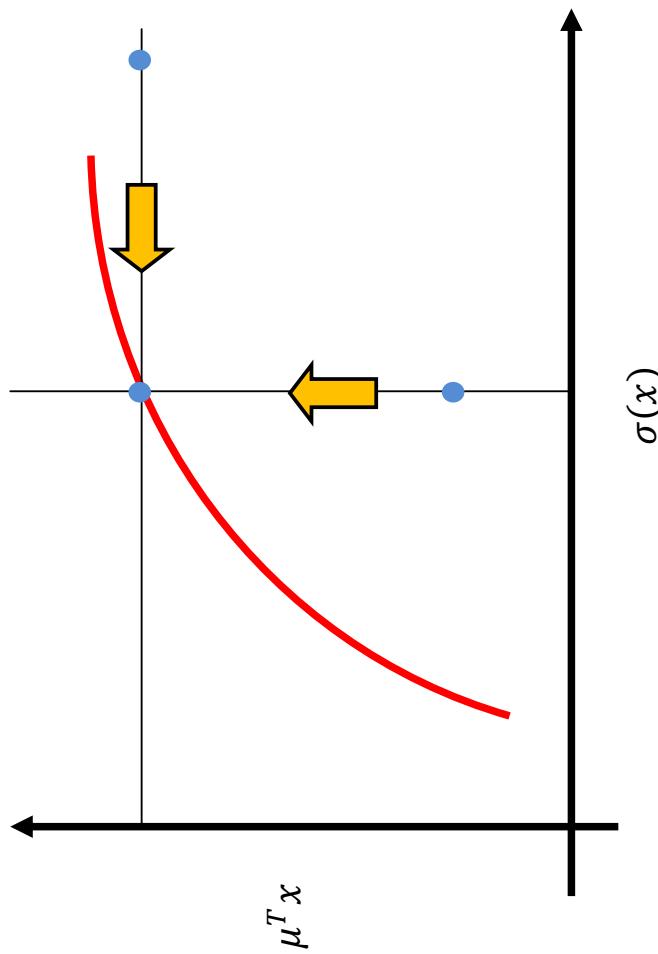
## Exercise 256

Show that  $f(x) = (x^\top \Sigma x)^{1/2}$  is a convex function.

# Convexity of $\sigma(R)$

## Theorem 257

The function  $\sigma : (\infty, R_{max}] \rightarrow \mathbb{R}$ , defined by (†), is convex.



## Convexity of $\sigma(R)$ : Proof (1/2)

We proceed directly from definition of convexity. Fix arbitrary  $0 < \alpha < 1$  and  $R_1, R_2 \leq R_{max}$ . We need to show that

$$\underbrace{\sigma(\alpha R_1 + (1 - \alpha) R_2)}_{R_3} \leq \alpha \sigma(R_1) + (1 - \alpha) \sigma(R_2). \quad (143)$$

Let

$$f(x) = (x^\top \Sigma x)^{1/2}$$

and let  $x_i$  for  $i = 1, 2, 3$  be the optimal solution in  $(\dagger)$  for  $R = R_i$ . Clearly,

$$f(x_i) = \sigma(R_i),$$

and hence (143) can be equivalently written as

$$f(x_3) \leq \alpha f(x_1) + (1 - \alpha) f(x_2). \quad (144)$$

## Convexity of $\sigma(R)$ : Proof (2/2)

Since  $x_i$  is feasible for  $(\dagger)$  with right hand side  $R = R_i$ , we have

$$\mu^\top x_i \geq R_i, \quad i = 1, 2. \quad (145)$$

$$x_i \in \mathcal{X}, \quad i = 1, 2. \quad (146)$$

By adding an  $\alpha$  multiple of the first inequality in (145) to the  $(1 - \alpha)$  multiple of the second inequality, we get

$$\mu^\top (\alpha \textcolor{blue}{x}_1 + (1 - \alpha) \textcolor{green}{x}_2) \geq \alpha \textcolor{blue}{R}_1 + (1 - \alpha) \textcolor{green}{R}_2 = \textcolor{red}{R}_3. \quad (147)$$

# Example: Stocks, Bonds and Money Market

## Example 258

Consider investing into Stocks (S&P 500), Bonds (10y US Treasury Bond) and Money Market (1-day Federal Fund Rate).

### Step 1. Get Historical Data

$$l_{i,t} = \text{return on asset } i = 1, \dots, n \text{ at time } t = 0, 1, \dots, T$$

|                  | S (asset 1) | B (asset 2) | MM (asset 3) |
|------------------|-------------|-------------|--------------|
| 1960 ( $t = 0$ ) | 20.26       | 262.94      | 100.00       |
| 1961 ( $t = 1$ ) | 25.69       | 268.73      | 102.33       |
| 1962 ( $t = 2$ ) | 23.43       | 284.09      | 105.33       |
| 1963 ( $t = 3$ ) | 28.75       | 289.16      | 108.89       |
| :                | :           | :           | :            |

# Example: Stocks, Bonds and Money Market (Cont.)

## Step 2. Transform into Historical Return Rates

$$r_{i,t} = \frac{I_{i,t} - I_{i,t-1}}{I_{i,t-1}}$$

|                  | S (asset 1) | B (asset 2) | MM (asset 3) |
|------------------|-------------|-------------|--------------|
| 1960 ( $t = 0$ ) | -           | -           | -            |
| 1961 ( $t = 1$ ) | 26.81%      | 2.20%       | 2.33%        |
| 1962 ( $t = 2$ ) | -8.78%      | 5.72%       | 2.93%        |
| 1963 ( $t = 3$ ) | 22.69%      | 1.79%       | 3.38%        |
| :                | :           | :           | :            |

# Example: Stocks, Bonds and Money Market (Cont.)

## Step 3. Estimate Mean Return Rates

$$\bar{r}_i = \underbrace{\frac{1}{T} \sum_{t=1}^T r_{i,t}}_{\text{arithmetic mean}}$$
$$\mu_i = \underbrace{\left( \prod_{t=1}^T (1 + r_{i,t}) \right)^{1/T} - 1}_{\text{geometric mean}}$$

|             | S (asset 1) | B (asset 2) | MM (asset 3) |
|-------------|-------------|-------------|--------------|
| $\bar{r}_i$ | 12.06%      | 7.85%       | 6.32%        |
| $\mu_i$     | 10.73%      | 7.37%       | 6.27%        |

$$\mu = (0.1073, 0.0737, 0.0627)^\top$$

# Example: Stocks, Bonds and Money Market (Cont.)

## Step 4. Estimate Covariance Matrix

$$\Sigma_{ij} = \frac{1}{T} \sum_{i=1}^T (r_{i,t} - \bar{r}_i)(r_{j,t} - \bar{r}_j), \quad i, j \in \{1, 2, \dots, n\}$$

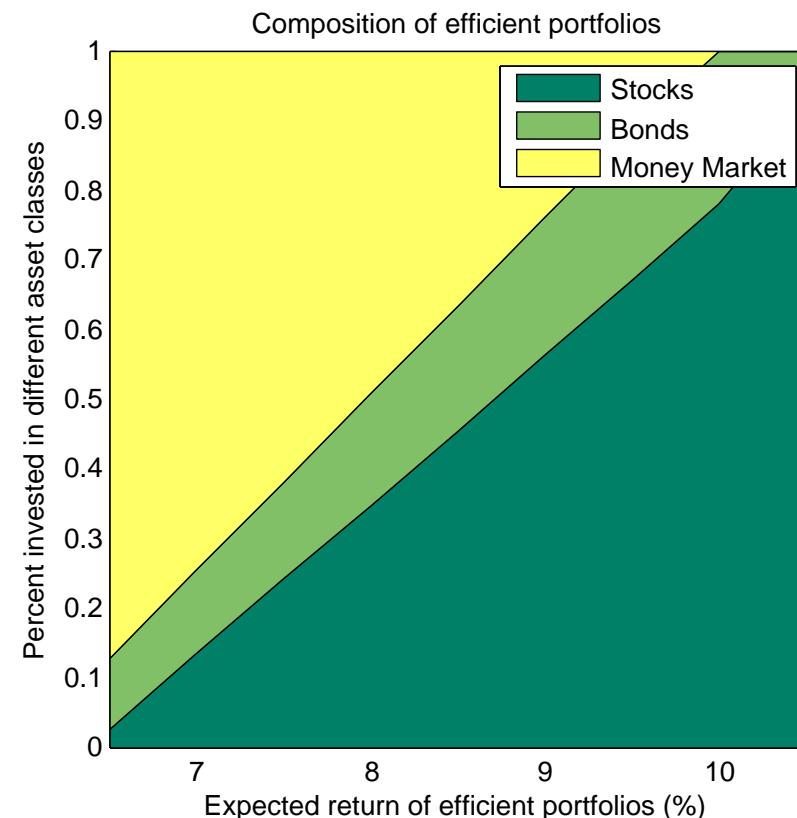
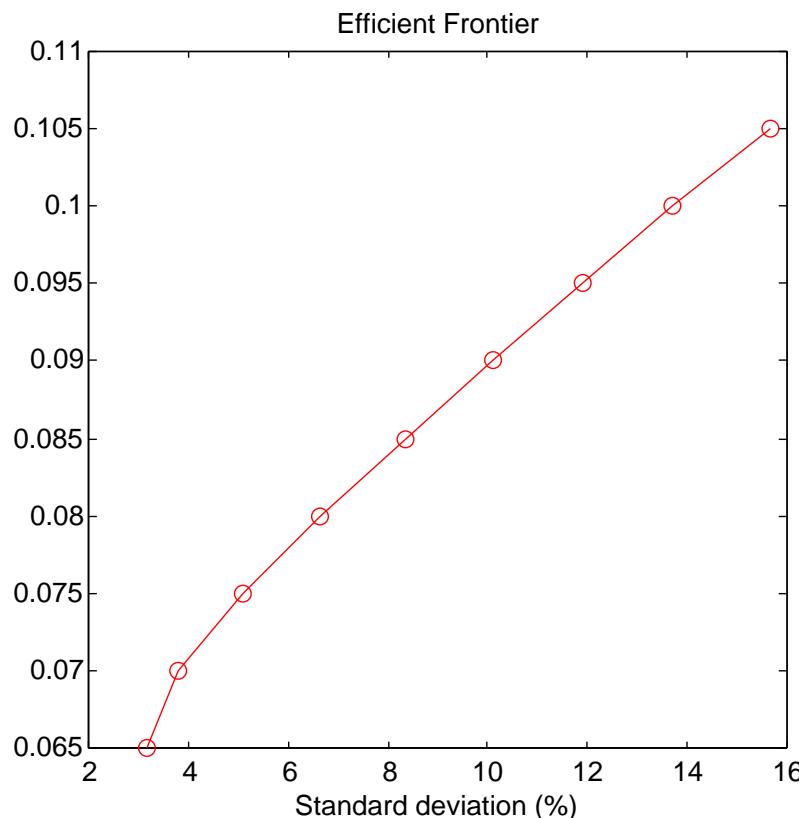
$$\Sigma = \begin{pmatrix} 0.02778 & 0.00387 & 0.00021 \\ 0.00387 & 0.01112 & -0.00020 \\ 0.00021 & -0.00020 & 0.00115 \end{pmatrix}$$

## Step 5. Find Efficient Portfolio using CVXPY

$$(EP2) \quad \begin{array}{ll} \min_{x \in \mathbb{R}^3} & x^\top \Sigma x \\ \text{subject to} & \mu^\top x \geq R \\ & x \in \mathcal{X} \end{array}$$

for  $\mathcal{X} = \{x \in \mathbb{R}^3 : x \geq 0, x_1 + x_2 + x_3 = 1\}$  and  $R \in [6.5\%, 10.5\%]$ .

# Example: Stocks, Bonds and Money Market (CODE OUTPUT)





# Introduction to Optimization

Peter Richtárik



## Lecture 25: Sharpe Ratio - Part 1

# Lecture Outline

- ▶ Sharpe Ratio
- ▶ Capital Allocation Line

# All Assets in Markowitz' Model are Risky

- ▶ Indeed, the variance of the return rate of asset  $S_i$  is

$$\begin{aligned}\text{Var}[r_i] &= \mathbb{E}[(r_i - \mu_i)^2] \\ &= \mathbb{E}[((r - \mu)(r - \mu)^\top)_{ii}] \\ &= (\mathbb{E}[(r - \mu)(r - \mu)^\top])_{ii} \\ &= \Sigma_{ii}.\end{aligned}$$

- ▶ Now, since  $\Sigma$  is positive definite, we have

$$\Sigma_{ii} = e_i^\top \Sigma e_i > 0$$

( $e_i$  is the  $i$ th unit coordinate vector). So the variance of the rates of return of all assets is positive, i.e., the assets are **risky**.

# Introduction of a Risk-Free Asset to the Model

We now amend Markowitz' model with a **risk-free asset  $S^0$** :

- ▶ We assume that
  - ▶ The variance of (the rate of return of)  $S^0$  is zero:

$$\text{Var}[r_0] = 0.$$

(We assume this since this makes the asset risk-free)

- ▶ The rate of return of  $S^0$  satisfies:

$$0 < r_0 < \mu^\top x \quad \text{for some} \quad x \in \mathcal{X}$$

(It is natural to assume that there should be a portfolio yielding higher expected return than the risk-free asset

Remarks:

- ▶ We implicitly assumed (in Markowitz' model) that  $0 \in \mathcal{X}$  (i.e., that we are allowed to not form any portfolio)

# Portfolio Including the Risk-Free Asset

- ▶ We now consider a portfolio of assets  $S^0, S^1, \dots, S^n$ ; that is, we include the **risk-free asset  $S^0$** , too.
- ▶ We consider the following **admissible set**:

$$\mathcal{Y} \stackrel{\text{def}}{=} \{y = (1 - \alpha, \alpha x) \in \mathbb{R}^{n+1} : x \in \mathcal{X}, 0 \leq \alpha \leq 1\}.$$

- ▶  $\mathcal{Y}$  is a subset of  $\mathbb{R}^{n+1}$  because we now have  $n + 1$  assets
- ▶  $\mathcal{Y}$  contains only portfolios which arise by picking some  $x \in \mathcal{X}$ , scaling it by the factor of  $\alpha \in [0, 1]$ , and adding to the portfolio  $1 - \alpha$  units of the risk-free asset  $S^0$ . In other words, we combine  $\alpha$  units of  $x \in \mathcal{X}$  with  $1 - \alpha$  units of  $S^0$ .

# New Model: Budget Preservation Property

## Example 259

Assume that there exists  $B \geq 1$  such that the following **budget constraint** holds

$$\sum_{i=1}^n x_i \leq B \quad \text{for all } x \in \mathcal{X}.$$

Show that the budget constraint also holds for any admissible portfolio in  $\mathcal{Y}$ . That is, show that

$$\sum_{i=0}^n y_i \leq B \quad \text{for all } y \in \mathcal{Y}.$$

**Solution:** Pick any  $y \in \mathcal{Y}$ . Then  $y = (1 - \alpha, \alpha x)$  for some  $x \in \mathcal{X}$  and  $0 \leq \alpha \leq 1$ . We now have

$$\sum_{i=0}^n y_i = 1 - \alpha + \alpha \sum_{i=1}^n x_i \leq 1 - \alpha + \alpha B \leq B. \quad (148)$$

# New Model: Return and Risk of a Portfolio

Pick  $y \in \mathcal{Y}$  such that  $y = (1 - \alpha, \alpha x)$ .

- The **expected return rate of portfolio  $y$**  is given by

$$\begin{aligned}\mathbb{E}[r(y)] &= \mathbb{E}\left[\sum_{i=0}^n y_i r_i\right] = \mathbb{E}\left[(1 - \alpha)r_0 + \sum_{i=1}^n (\alpha x_i)r_i\right] \\ &= (1 - \alpha)\underbrace{\mathbb{E}[r_0]}_{r_0} + \alpha \sum_{i=1}^n x_i \mathbb{E}[r_i] = (1 - \alpha)r_0 + \alpha \mathbb{E}[r(x)].\end{aligned}$$

- Since

$$\begin{aligned}\text{Var}[r(y)] &= \text{Var}\left[\sum_{i=0}^n y_i r_i\right] = \text{Var}[y_0 r_0] + \text{Var}\left[\sum_{i=1}^n y_i r_i\right] \\ &= y_0^2 \underbrace{\text{Var}[r_0]}_{=0} + \text{Var}[\alpha r(x)] = \alpha^2 \text{Var}[r(x)],\end{aligned}$$

the **risk (standard deviation) of portfolio  $y$**  is given by

$$\sigma(y) = \sqrt{\text{Var}[r(y)]} = \alpha \sqrt{\text{Var}[r(x)]} = \alpha \sigma(x). \quad (149)$$

# Capital Allocation Line (CAL)

Now fix portfolio  $x \in \mathcal{X}$  and consider writing the expected return rate of portfolio  $y = (1 - \alpha, \alpha x)$  as a function of  $\sigma(y)$ :

$$\begin{aligned}\mathbb{E}[r(y)] &= \alpha \mathbb{E}[r(x)] + (1 - \alpha)r_0 \\ &= \alpha \frac{\sigma(x)}{\sigma(x)} \mathbb{E}[r(x)] + \left(1 - \alpha \frac{\sigma(x)}{\sigma(x)}\right) r_0 \\ &\stackrel{(149)}{=} \frac{\sigma(y)}{\sigma(x)} \mathbb{E}[r(x)] + \left(1 - \frac{\sigma(y)}{\sigma(x)}\right) r_0 \\ &= \underbrace{r_0 + \left(\frac{\mathbb{E}[r(x)] - r_0}{\sigma(x)}\right) \sigma(y)}_{\text{linear function of } \sigma(y)}. \tag{150}\end{aligned}$$

## Definition 260 (Capital Allocation Line)

From (150) we see that the expected rate of return of portfolio  $y$  is a **linear function** of its risk  $\sigma(y)$ . This linear function is called the **Capital Allocation Line (CAL)**.

Remarks:

1.  $\sigma(x) > 0$  for all  $0 \neq x \in \mathcal{X}$  (since  $\Sigma$  is positive definite)
2.  $\mathbb{E}[r(x)] > r_0$  for some  $x \in \mathcal{X}$

# Sharpe Ratio

## Definition 261 (Sharpe Ratio)

Sharpe ratio of portfolio  $y \in \mathcal{Y}$  is defined as

$$SR(y) \stackrel{\text{def}}{=} \frac{\mathbb{E}[r(y)] - r_0}{\sigma(y)}.$$

Remarks:

- ▶ Sharpe ratio is a **reward-to-volatility** ratio of portfolio  $y$
- ▶ In some sense **it is reasonable to prefer portfolios with a higher Sharpe ratio: one gets more reward per unit of volatility.**
- ▶ Hence, we will seek to find / compute portfolios with maximal Sharpe ratio.

# Sharpe Ratio: More Insight

## Theorem 262

Fix  $x \in \mathcal{X}$  and consider the set of portfolios

$$\mathcal{Y}(x) \stackrel{\text{def}}{=} \{y = (1 - \alpha, \alpha x) : 0 < \alpha \leq 1\}.$$

Then all portfolios  $y \in \mathcal{Y}(x)$  have the same Sharpe ratio.

### Proof.

Choose any  $y \in \mathcal{Y}(x)$ . By rewriting identity (150), we immediately get

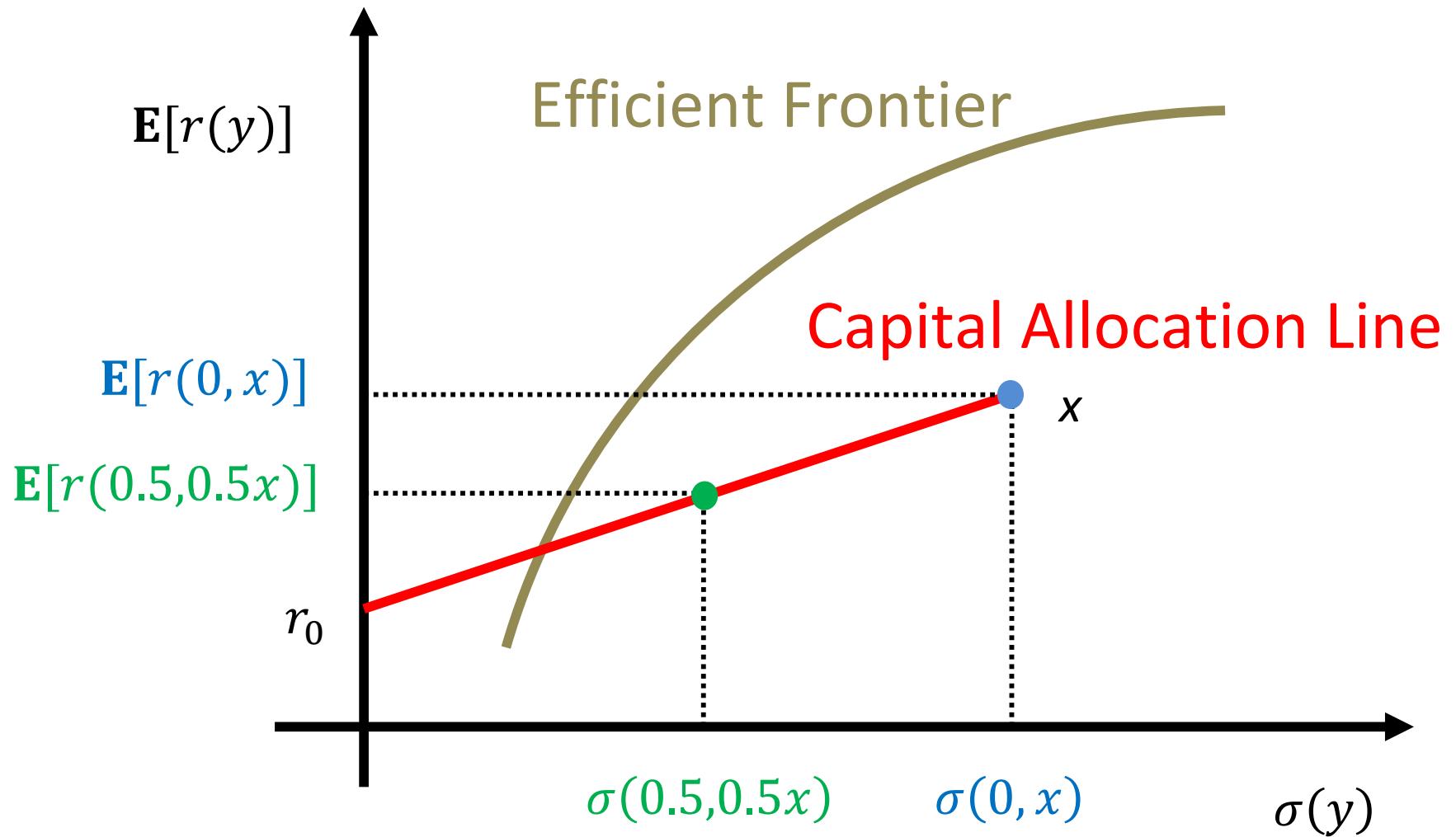
$$\frac{\mathbb{E}[r(y)] - r_0}{\sigma(y)} = \frac{\mathbb{E}[r(x)] - r_0}{\sigma(x)}. \quad (151)$$

□

Remarks:

- ▶ This result says that **all portfolios belonging to  $\mathcal{Y}(x)$  have the same Sharpe ratio**
- ▶ So, for investors interested *only* in Sharpe ratio, all portfolios  $y \in \mathcal{Y}(x)$  are equally good (or bad!)

# Capital Allocation Line: Visualization



# Capital Allocation Line: Explanations

1. The figure (on the previous slide) is an attempt to draw portfolios from  $\mathcal{Y}$  into the standard 2D picture we used to draw portfolios from  $\mathcal{X}$
2. This is done as follows:
  - ▶ point  $x \in \mathcal{X}$  corresponds to portfolio  $y = (0, x) \in \mathcal{Y}$  (that is, no risk-free asset; unit of  $x$ )
  - ▶ point  $r_0$  on the vertical axis corresponds to portfolio  $y = (1, 0)$  (that is, a unit of risk-free asset, no  $x$ )
  - ▶ the midpoint between  $r_0$  and  $x$  corresponds to the portfolio  $y = (0.5, 0.5x)$  (that is, 50% risk-free asset, 50% portfolio  $x$ )
  - ▶ In this plot, **CAL corresponds to portfolios in the set  $\mathcal{Y}(x)$**
3. Notice that coordinates of all  $y$ -portfolios on the capital allocation line make sense: they correspond to their risk (horizontal axis) and expected rate of return (vertical axis)
4. For each  $x \in \mathcal{X}$  (remember, there are no  $x$ -portfolios above the efficient frontier), the line joining  $r_0$  and  $x$  is a capital allocation line each point of which corresponds to some combination of  $x$  and the risk-free asset.

# Introduction to Optimization

Peter Richtárik



## Lecture 26: Sharpe Ratio - Part 2

# Lecture Outline

- ▶ Computing the optimal risky portfolio by maximizing the Sharpe ratio
- ▶ Solution 1
- ▶ Solution 2

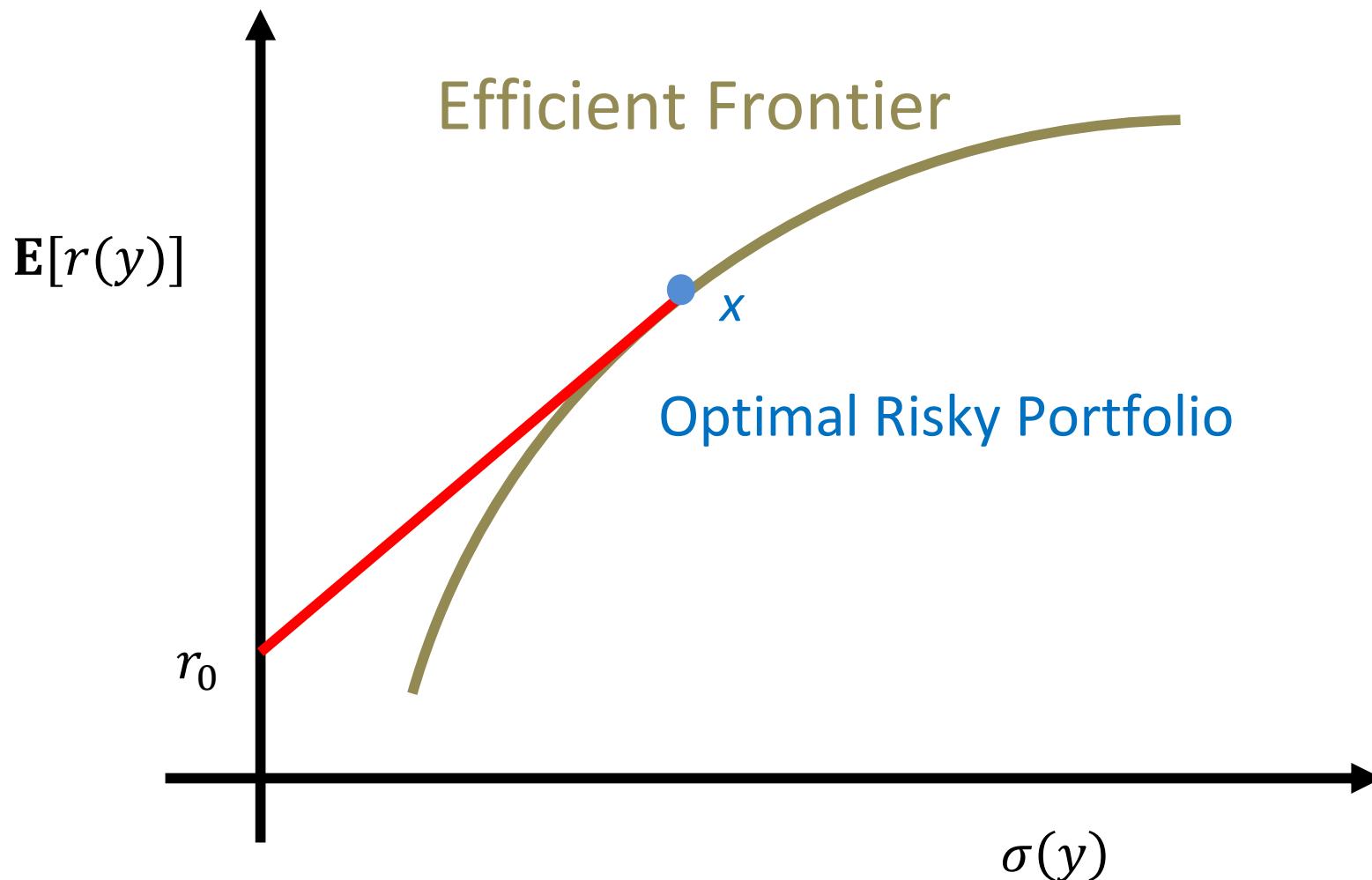
# Optimal Risky Portfolio: Maximizing Sharpe Ratio

Definition 263 (Optimal Risky Portfolio)

**Optimal risky portfolio** is the  $x^* \in \mathcal{X}$  whose Sharpe ratio is maximal among all portfolios  $x \in \mathcal{X}$ . That is, it is the solution of the optimization problem

$$\begin{aligned} & \text{maximize} && \frac{\mathbb{E}[r(x)] - r_0}{\sigma(x)} \quad \left\{ = \frac{\mu^\top x - r_0}{\sqrt{x^\top \Sigma x}} \right\} \\ & \text{subject to} && x \in \mathcal{X} \end{aligned}$$

# Optimal Risky Portfolio: Visualization



This is the portfolio  $x \in \mathcal{X}$  with the highest Sharpe ratio, i.e., whose CAL has the largest slope.

# Computing the Optimal Risky Portfolio

- ▶ It is not easy to compute the optimal risky portfolio by maximizing the Sharpe ratio since this is a **nonconvex** function of  $x$ .
- ▶ We will now rewrite the problem into an equivalent **convex** (and hence computationally tractable) form.

# Conic Optimization: Computing the Optimal Risky Portfolio

## Theorem 264

Let  $e \in \mathbb{R}^n$  be the vector of all ones and assume that

- a)  $\mathcal{X} \subseteq \mathbb{R}^n$  is closed and convex
- b)  $\mathcal{X} \subset \{x : e^\top x = 1\}$  (unit budget)
- c) there exists  $\hat{x} \in \mathcal{X}$  such that  $\mu^\top \hat{x} > r_0$

Let  $(z^*, \kappa^*)$  be the optimal solution of the following **convex conic optimization problem**:

$$\begin{aligned} \min \quad & z^\top \Sigma z \\ (z, \kappa) \in & \mathcal{X}^+ \\ (\mu - r_0 e)^\top z = & 1, \end{aligned} \tag{152}$$

$$\mathcal{X}^+ \stackrel{\text{def}}{=} \{(z, \kappa) \in \mathbb{R}^n \times \mathbb{R} : \kappa > 0, \frac{z}{\kappa} \in \mathcal{X}\} \cup \{(0, 0)\}$$

is a **closed convex cone**. Then the optimal risky portfolio is given by

$$x^* = z^*/\kappa^*.$$

# Cones

## Definition 265 (Cone)

A set  $\mathcal{S} \subset \mathbb{R}^d$  is a **cone** if it is closed under multiplication by a nonnegative scalar. That is, if for all  $s \in \mathcal{S}$  and  $t \geq 0$ , we have  $ts \in \mathcal{S}$ .

## Exercise 266

Show that the following sets are cones:

- (i) any linear subspace of  $\mathbb{R}^d$
- (ii) the set  $\mathcal{S}_s = \{\alpha s : \alpha \geq 0\}$ , where  $s \in \mathbb{R}^d$
- (iii) the nonnegative orthant:  $\mathbb{R}_+^d = \{s \in \mathbb{R}^d : s \geq 0\}$

## Exercise 267

Show that the union and intersection of two cones is a cone.

# $\mathcal{X}^+$ is a closed convex cone

## Lemma 268

$\mathcal{X}^+$  is a closed convex cone.

Proof.

- ▶ **Cone.** Pick any  $(z, \kappa) \in \mathcal{X}^+$  and  $t > 0$ . Since  $t\kappa > 0$  and  $\frac{tz}{t\kappa} = \frac{z}{\kappa} \in \mathcal{X}$ , we must have  $(tz, t\kappa) \in \mathcal{X}^+$ . Hence,  $\mathcal{X}^+$  is a cone.
- ▶ **Convexity.** Pick  $(z_1, \kappa_1) \in \mathcal{X}^+$ ,  $(z_2, \kappa_2) \in \mathcal{X}^+$  and  $0 \leq \alpha \leq 1$ . We thus need to show that

$$\alpha(z_1, \kappa_1) + (1 - \alpha)(z_2, \kappa_2) = (\alpha z_1 + (1 - \alpha)z_2, \alpha \kappa_1 + (1 - \alpha)\kappa_2) \in \mathcal{X}^+.$$

Note that  $\alpha \kappa_1 + (1 - \alpha)\kappa_2 > 0$  and

$$\frac{\alpha z_1 + (1 - \alpha)z_2}{\alpha \kappa_1 + (1 - \alpha)\kappa_2} = \underbrace{\frac{\alpha \kappa_1}{\alpha \kappa_1 + (1 - \alpha)\kappa_2} \left( \frac{z_1}{\kappa_1} \right)}_{\in \mathcal{X}} + \underbrace{\frac{(1 - \alpha)\kappa_2}{\alpha \kappa_1 + (1 - \alpha)\kappa_2} \left( \frac{z_2}{\kappa_2} \right)}_{\in \mathcal{X}} \in \mathcal{X},$$

where the last step follows from convexity of  $\mathcal{X}$ .

- ▶ **Closedness.** We will skip this. Standard arguments.

# Proof of Theorem 264: First Reformulation

Let

$$h(x) \stackrel{\text{def}}{=} \frac{\mu^\top x - r_0}{\sqrt{x^\top \Sigma x}}.$$

- ▶ Since we assume that there is  $\hat{x} \in \mathcal{X}$  for which  $\mu^\top \hat{x} > r_0$ , the optimal Sharpe ratio is positive.
- ▶ Since we assume that  $e^\top x = 1$  for all  $x \in \mathcal{X}$ , we have

$$\mu^\top x - r_0 = (\mu - r_0 e)^\top x.$$

So, the **problem of maximizing the Sharpe ratio**,

$$\max\{h(x) : x \in \mathcal{X}\},$$

**can be equivalently written in the form:**

$$\begin{aligned} & \max && h(x) \\ & \text{subject to} && x \in \mathcal{A}, \end{aligned} \tag{153}$$

where

$$\mathcal{A} \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : x \in \mathcal{X}, (\mu - r_0 e)^\top x > 0\}.$$

# Proof of Theorem 264: Magical Change of Variables

We now introduce a magical **change of variables**:

$$z = \kappa x, \quad \kappa = \frac{1}{(\mu - r_0 e)^\top x}. \quad (154)$$

That is, instead of variable  $x$ , we consider the variables  $(z, \kappa)$ . Define

$$\mathcal{B} = \{(z, \kappa) \in \mathbb{R}^n \times \mathbb{R} : \kappa > 0, \frac{z}{\kappa} \in \mathcal{X}, (\mu - r_0 e)^\top z = 1\}$$

## Claim 269

$$x \in \mathcal{A} \Leftrightarrow (z, \kappa) \in \mathcal{B}.$$

## Proof.

- ▶  $\Rightarrow$  Pick  $x \in \mathcal{A}$ . Then clearly  $\kappa$  is well defined and positive. Moreover,  $\frac{y}{\kappa} = x \in \mathcal{X}$  and  $(\mu - r_0 e)^\top z = (\mu - r_0 e)^\top x \kappa = 1$ . Therefore,  $(z, \kappa) \in \mathcal{B}$ .
- ▶  $\Leftarrow$  Pick  $(z, \kappa) \in \mathcal{B}$ . Then  $x = y/\kappa \in \mathcal{X}$  and  $(\mu - r_0 e)^\top x = (\mu - r_0 e)^\top \frac{z}{\kappa} = \frac{1}{\kappa} > 0$ , whence  $x \in \mathcal{A}$ .

# Proof of Theorem 264: Claim 270 ( $h(x)$ as a function of $z$ )

Claim 270

$$h(x) = \frac{1}{\sqrt{z^\top \Sigma z}}$$

for  $x \in \mathcal{A}$  (equiv. for  $(z, \kappa) \in \mathcal{B}$ ).

Proof.

$$h(x) = \frac{(\mu - r_0 e)^\top x}{\sqrt{x^\top \Sigma x}} \stackrel{(154)}{=} \frac{1/\kappa}{\left( \left( \frac{z}{\kappa} \right)^\top \Sigma \left( \frac{z}{\kappa} \right) \right)^{1/2}} = \frac{1}{\sqrt{z^\top \Sigma z}}.$$



# Proof of Theorem 264: THE END GAME!

In view of Claim 269 and Claim 270, problem (153) can be written in the form

$$\begin{aligned} \max \quad & \frac{1}{\sqrt{z^\top \Sigma z}} \\ \text{subject to} \quad & (z, \kappa) \in \mathcal{B} \end{aligned} \tag{155}$$

Finally, the above problem has the **same solution set** as the following one:

$$\begin{aligned} \min \quad & z^\top \Sigma z \\ \text{subject to} \quad & (z, \kappa) \in \mathcal{X}^+ \\ & (\mu - r_0 e)^\top z = 1. \end{aligned} \tag{156}$$

It is easy to see that if  $(z^*, \kappa^*)$  is the optimal solution of (156), then the original variable  $x^* = z^*/\kappa^*$ , given by a reverse change of variables, is optimal for the original problem of maximizing the Sharpe ratio.

# Computing $\mathcal{X}^+$ for $\mathcal{X}$ = Polyhedral Set

## Lemma 271 (Polyhedral $\mathcal{X}$ )

Consider the admissible set of the form

$$\mathcal{X} = \{x \in \mathbb{R}^n : \mathbf{A}x \geq b, \mathbf{C}x = d\},$$

where  $\mathbf{A}, \mathbf{C}$  are matrices and  $b, d$  vectors. Then

- (i)  $\mathcal{X}$  is bounded  $\Leftrightarrow \{s : \mathbf{A}s \geq 0, \mathbf{C}s = 0\} = \{0\}$
- (ii) If  $\mathcal{X}$  is bounded, then

$$\mathcal{X}^+ = \{(z, \kappa) : \mathbf{A}z - b\kappa \geq 0, \mathbf{C}z - d\kappa = 0, \kappa \geq 0\}.$$

## Proof (Computing $\mathcal{X}^+$ for $\mathcal{X}$ = Polyhedral Set)

- (i)  $\Rightarrow$  Assume there exists  $s \neq 0$  for which  $\mathbf{A}s \geq 0$  and  $\mathbf{C}s = 0$ . We wish to show that then  $\mathcal{X}$  is unbounded. Take  $x_0 \in \mathcal{X}$  and consider

$$x(t) \stackrel{\text{def}}{=} x_0 + ts, \quad t \geq 0.$$

Then

$$\mathbf{A}x(t) = \underbrace{\mathbf{A}x_0}_{\geq b} + t \underbrace{\mathbf{A}s}_{\geq 0} \geq b, \quad \text{for all } t \geq 0,$$

$$\mathbf{C}x(t) = \underbrace{\mathbf{C}x_0}_{=d} + t \underbrace{\mathbf{C}s}_{=0} = d, \quad \text{for all } t \geq 0,$$

from which it follows that  $x(t) \in \mathcal{X}$  for all  $t \geq 0$ , implying that  $\mathcal{X}$  is unbounded.

- (i)  $\Leftarrow$  Assume  $\mathcal{X}$  is not bounded. We wish to show that then the set  $\{s : \mathbf{A}s \geq 0, \mathbf{C}s = 0\}$  is not equal to  $\{0\}$ . There must exist  $s \neq 0$  such that  $x_0 + ts \in \mathcal{X}$  for all  $t \geq 0$ . In particular,

$$\mathbf{A}(x_0 + ts) \geq b, \quad \mathbf{C}(x_0 + ts) = d, \quad \text{for all } t \geq 0$$

$$t\mathbf{A}s \geq \underbrace{b - \mathbf{A}x_0}_{\leq 0}, \quad \underbrace{\mathbf{C}x_0}_{=d} + t\mathbf{C}s = d, \quad \text{for all } t \geq 0$$

Hence,  $\mathbf{A}s \geq 0$  and  $\mathbf{C}s = 0$ .

# Proof (Computing $\mathcal{X}^+$ for $\mathcal{X}$ = Polyhedral Set)

(ii)

$$\begin{aligned}\mathcal{X}^+ &= \{(z, \kappa) : \kappa > 0, \frac{z}{\kappa} \in \mathcal{X}\} \cup \{(0, 0)\} \\ &= \{(z, \kappa) : \kappa > 0, \mathbf{A}\frac{z}{\kappa} \geq b, \mathbf{C}\frac{z}{\kappa} = d\} \cup \{(0, 0)\} \\ &= \{(z, \kappa) : \kappa > 0, \mathbf{A}z - b\kappa \geq b, \mathbf{C}z - d\kappa = 0\} \cup \{(0, 0)\}\end{aligned}$$

Note that if we allow  $\kappa = 0$  in the above expression, then the conditions on  $z$  are:  $\mathbf{A}z \geq 0$  and  $\mathbf{C}z = 0$ . But from boundedness of  $\mathcal{X}$  we know that there is only one such  $z$ :  $z = 0$ . So, by allowing  $\kappa = 0$ , the set  $\mathcal{X}^+$  does not grow, which proves the result.

# Exercises

Exercise 272 (Computing  $\mathcal{X}^+$  for  $\mathcal{X}$  = Unit Ball)

If  $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$ , then

$$\mathcal{X}^+ = \{(z, \kappa) : \|z\| \leq \kappa, \kappa \geq 0\},$$

where  $\|z\| \stackrel{\text{def}}{=} \sqrt{z^\top z}$  is the standard Euclidean norm.

# Table of Contents

# Contents I

## Part I: Theory

1. The Space  $\mathbb{R}^d$
2. Introduction
3. Convex Sets - Part 1
4. Convex Sets - Part 2
5. Convex Functions - Part 1
6. Convex Functions - Part 2
7. Convex Functions - Part 3
8. Convex Functions - Part 4
9. Fenchel Conjugation
10. Fenchel Duality - Part 1
11. Fenchel Duality - Part 2
12. Fenchel Duality - Part 3

## Part II: Applications

13. Asset Pricing & Arbitrage Detection via Linear Programming - Part 1



## Contents II

14. Asset Pricing & Arbitrage Detection via Linear Programming - Part 2
15. Asset Pricing & Arbitrage Detection via Linear Programming - Part 3
16. Asset Pricing & Arbitrage Detection via Linear Programming - Part 4
17. Image Manipulation
18. Truss Topology Design - Part 1
19. Truss Topology Design - Part 2
20. Smoothness and Strong Convexity
21. SGD - Part 1
22. SGD - Part 2
23. SGD - Part 3
24. Markowitz Mean-Variance Portfolio via Quadratic Programming
25. Sharpe Ratio - Part 1
26. Sharpe Ratio - Part 2