Chapter I: Preliminaries

Organization

- Contact:
  - Email: `sagnol -at- math.tu-berlin.de`
  - Office: MA 518
  - Assistant: Daniel Schmidt g.W., MA 524
- Timing:
  - Monday 10-12, H1029
  - Thursday 14-16, H3006
  - Lecture every Monday
  - Lecture or Exercises every second week on Thursday
  - Ocasionally: Tutorials with python notebooks
- Resources:
  - There will be a handout, put online as and when the corresponding chapters are finished
  - The course is mainly based on the book "Convex Optimization", S. Boyd & L. Vandenberghe, 2004, freely available online at `http://stanford.edu/ boyd/cvxbook/`.
  - Selected chapters are also based on the following references:
    - "Lecture on Modern Convex Optimization", A. Ben-Tal & A. Nemirovski, 2001.
    - "Topics in Convex Optimisation", Lecture notes of H. Fawzi at Cambridge.
    - "Approximation Algorithms and Semidefinite Programming", Lecture notes of B. Gärtner & J. Matoušek at ETH Zurich.
    - "Semidefinite Optimization", Lecture notes of M. Laurent & F. Vallentin at Utrecht.
- Evaluation:
  - Exercises will be given on week in advance. At the beginning of exercise sessions, check the exercises you've prepared. One student will be asked to explain his solution. You need 50% of all exercises *checked* to pass the exam.
  - Oral examination. You should not learn everything *by heart*, but rather know which result exists and be able to find it in your handout if needed. You will not be asked to prove results from the course, but rather to solve some new exercises & show your modelling skills.
  - Mathematical rigor: the handout contains precise mathematical statements. On the blackboard however, I will handwavy sometimes. IMHO it's OK to handwavy as long as:
    - You are aware that you're not handling all small details of the problem;
    - You have an idea of how to solve the problem is a rigorous manner.

- Objectives:
  - Understand the concept of duality in convex optimization, for both the "standard" nonlinear programming approach and the "modern" conic programming approach.
  - Learn to recognize convexity in a mathematical optimization problem, and modelling tricks to represent or reformulate a problem as a linear program (LP), second-order cone program (SOCP), semidefinite program (SDP), or geometric program (GP).
  - Review many applications of convex optimization, in particular in the fields of engineering, data analysis, combinatorial optimization, and robust optimization.
  - Learn to use modern interface such as PICOS or CVXPY, to formulate conic optimization problems in python and solve them with state-of-the-art solvers.
  - Understand the algorithms (interior point methods).

Notation for scalars, vectors, and matrices

In this script, we use the following standard notation:

- $[n] := \{1, \dots, n\}$
- Scalar numbers are denoted by plain lower case letters, e.g. $c \in \mathbb{R}$.
- Vectors are denoted by boldface lower case letters, e.g. $\boldsymbol{v} \in \mathbb{R}^n$, with elements $v_1, \dots, v_n$. Unless stated otherwise, the symbol $\boldsymbol{v}$ always denote a *column vector*. The associated row vector is $\boldsymbol{v}^T$. On the blackboard, we'll make no distinction between plain and boldface letters, but it should be clear from the context whether the symbol x denotes a scalar or a vector.
- Matrices are denoted by upper case letters, e.g. $A \in \mathbb{R}^{m \times n}$, with elements A_{ij} ($i \in [m],\ j \in [n]$).
- Random variables are denoted by boldface upper case letters, e.g. $\boldsymbol{X}$. (Note that we do not distinguish between random scalars and random vectors, this should be clear from the context. The coordinates of a random vector $\boldsymbol{X} \in \mathbb{R}^n$ are the $\boldsymbol{X}_i$'s, also written in boldface since they are random variables).
- Sometimes, we will write $A = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_n] \in \mathbb{R}^{m \times n}$, which means that A is a matrix with columns $\boldsymbol{a}_j \in \mathbb{R}^m$ ($\forall j \in [n]$).
- Similarly, $A = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_m]^T \in \mathbb{R}^{m \times n}$, means that A is a matrix with rows $\boldsymbol{a}_i \in \mathbb{R}^n$ ($\forall i \in [m]$).
- $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$.
- $\mathbb{R}_{++} = \{x \in \mathbb{R} : x > 0\}$.
- $\mathbb{S}^n = \{X \in \mathbb{R}^{n \times n} : X = X^T\}$.
- $\mathbb{S}^n_+ = \{X \in \mathbb{S}^n : X \text{ is positive semidefinite }\}$.
- $\mathbb{S}^n_{++} = \{X \in \mathbb{S}^n : X \text{ is positive definite }\}$.

The inequalities between vectors are *elementwise*, that is, if $A = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_m]^T \in \mathbb{R}^{m \times n}$ is a matrix with rows $\boldsymbol{a}_i^T$'s, then $A\boldsymbol{x} \leq \boldsymbol{b}$ means

$$\boldsymbol{a}_i^T \boldsymbol{x} \leq b_i \quad (\forall i \in [m]).$$

The ith vector of the (canonical) basis of $\mathbb{R}^n$ is $\boldsymbol{e}_i = [0, \dots, 0, 1, 0, \dots, 0]^T$, with the 1 in ith position.

Example:
For a vector $\boldsymbol{v} \in \mathbb{R}^n$, we have $\boldsymbol{e}_i^T \boldsymbol{v} = v_i$.
For a vector $A \in \mathbb{R}^{m\times n}$, we have $\boldsymbol{e}_i^T A \boldsymbol{e}_j = A_{ij}$. (here, it is implicit that the vectors $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ are of appropriate size, that is, $\boldsymbol{e}_i \in \mathbb{R}^m$ and $\boldsymbol{e}_j \in \mathbb{R}^n$.
The matrix $E_{ij} = \boldsymbol{e}_i \boldsymbol{e}_j^T$ is a matrix (whose size depend on the dimensions of $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$, and should be clear from the context) with zeroes everywere, except for a one in position (i, j).

#1

The vector of all ones is $\mathbf{1} = \sum_i \boldsymbol{e}_i$. We sometimes write $\mathbf{1}_n$ to make it clear that $\mathbf{1}_n \in \mathbb{R}^n$.

The matrix of all ones is $J = \mathbf{1}\mathbf{1}^T$. When the dimension is not clear from the context, we can write $J_n = \mathbf{1}_n \mathbf{1}_n^T$ $J_{m,n} = \mathbf{1}_m \mathbf{1}_n^T$.

The identity matrix is I (or $I_n \in \mathbb{S}^n$). The zero vector is $\mathbf{0}$ (or $\mathbf{0}_n \in \mathbb{R}^n$), and the zero matrix is O (or $O_n \in \mathbb{S}^n$).

Let $\boldsymbol{u} \in \mathbb{R}^n$. The diagonal matrix with elements $u_1, \ldots, u_n$ is denoted by $\mathrm{Diag}(\boldsymbol{u})$.
The vector of diagonal elements of a matrix $X \in \mathbb{R}^{n\times n}$ is denoted by $\mathbf{diag}\, X$.

The image space of $A \in \mathbb{R}^{m\times n}$, i.e., the space spanned by the columns of A, is denoted by

$$\mathbf{Im}\, A := \{A\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^n\} \subseteq \mathbb{R}^m.$$

The nullspace of A is denoted by

$$\mathbf{Ker}\, A := \{\boldsymbol{y} \in \mathbb{R}^n : A\boldsymbol{y} = \mathbf{0}\}.$$

We also recall that the rank of $A \in \mathbb{R}^{m\times n}$ is the dimension of $\mathbf{Im}\, A$, and we say that A has *full column rank* whenever $\mathbf{rank}\, A = n$, which means that the n columns of A are linearly independent. Similarly, we say that A has *full row rank* if A^T has full column rank, i.e., if the m rows of A are linearly independent.

Scalar products and norms

The scalar product of two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ is

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle := \boldsymbol{u}^T \boldsymbol{v} = \sum_{i=1}^n u_i v_i$$

The scalar product of two matrices of the same size is

$$\langle A, B \rangle := \mathrm{trace}\, A^T B = \sum_j (A^T B)_{jj} = \sum_{i,j} A_{ij} B_{ij}.$$

In particular, note that when A and B are symmetric, we simply have $\langle A, B \rangle = \mathrm{trace}\, AB$.

Example:
The sum of all entries of a vector $\boldsymbol{v} \in \mathbb{R}^n$ is $\sum_{i=1}^n v_i = \mathbf{1}^T \boldsymbol{v}$.
The sum of all entries of a matrix $Ain\mathbb{R}^{m\times n}$ is $\sum_{i=1}^m \sum_{j=1}^n A_{ij} = \langle J, A \rangle$.

#2

Unless stated otherwise, the symbol $\|\boldsymbol{v}\|$ denotes the Euclidean norm of the vector $\boldsymbol{v} \in \mathbb{R}^n$:

$$\|\boldsymbol{v}\| := \sqrt{\boldsymbol{v}^T \boldsymbol{v}} = (\sum_{i=1}^n v_i^2)^{1/2}$$

When there might be an ambiguity, we use $\|\boldsymbol{v}\|_2$ to denote the Euclidean norm of $\boldsymbol{v}$. For all $p \geq 1$, the L_p-norm of $\boldsymbol{v}$ is $\|\boldsymbol{v}\|_p := (\sum_i v_i^p)^{1/p}$.

The *Frobenius norm* of a matrix is

$$\|A\|_F := \sqrt{\langle A, A\rangle} = (\sum_{i,j} A_{ij}^2)^{1/2}.$$

The *vectorization* of a matrix $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \in \mathbb{R}^{m\times n}$ with columns $\boldsymbol{a}_j's$ is

$$\operatorname{vec}(A) := \begin{bmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_n \end{bmatrix} \in \mathbb{R}^{mn}.$$

In particular, note that $\langle A, B\rangle = \operatorname{vec}(A)^T \operatorname{vec}(B)$ and $\|A\|_F = \|\operatorname{vec}(A)\|$.

Linear and Quadratic functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *affine* if it is of the form $\boldsymbol{x} \mapsto \sum_{i=1}^n a_i x_i + b$. In vector notation, an affine function can always be written as

$$\boldsymbol{x} \mapsto \boldsymbol{a}^T\boldsymbol{x} + b.$$

where $\boldsymbol{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. More generally, an affine function mapping $\mathbb{R}^m$ to $\mathbb{R}^n$ has the form

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

for some matrix $A \in \mathbb{R}^{m\times n}$ and a vector $\boldsymbol{b} \in \mathbb{R}^m$. We sometimes abuse the language and say that f is *linear*, although this term is normally reserved for functions of the form $\boldsymbol{x} \mapsto A\boldsymbol{x}$ (without the constant $\boldsymbol{b}$). To emphasize that a linear function has no constant part ($\boldsymbol{b} = \boldsymbol{0}$), we will speak of a *linear form*.

Quadratic functions of $\mathbb{R}^n$ to $\mathbb{R}$ are of the form

$$\boldsymbol{x} \mapsto \sum_{ij} Q_{ij}x_i x_j + \sum_i a_i x_i + b = \boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{a}^T\boldsymbol{x} + b.$$

Note that we can always assume without loss of generality (w.l.o.g.) that $Q \in \mathbb{S}^n$, because $\boldsymbol{x}^T Q\boldsymbol{x} = (\boldsymbol{x}^T Q\boldsymbol{x})^T = \boldsymbol{x}^T Q^T\boldsymbol{x}$, so:

$$\boldsymbol{x}^T Q\boldsymbol{x} = \frac{1}{2}\boldsymbol{x}^T(Q + Q^T)\boldsymbol{x}.$$

When a quadratic function has no affine part, i.e., $f(\boldsymbol{x}) = \boldsymbol{x}^T Q\boldsymbol{x}$, we speak of a *quadratic form*. An alternative formulation shows that any quadratic function can be assimilated with a quadratic form over $\mathbb{R}^n \times \{1\}$:

$$\boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{a}^T\boldsymbol{x} + b = \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix}^T \begin{bmatrix} Q & \frac{1}{2}\boldsymbol{a} \\ \frac{1}{2}\boldsymbol{a}^T & b \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix}.$$

The gradient of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is

$$\nabla f : \boldsymbol{x} \mapsto \begin{bmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^n,$$

and if f is twice differentiable, its Hessian is the function

$$\nabla^2 f : \boldsymbol{x} \mapsto \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\boldsymbol{x}) \end{pmatrix} \in \mathbb{S}^n,$$

In particular, the gradient and Hessian of linear and quadratic forms read

$$\begin{aligned} \nabla(\boldsymbol{x} \mapsto \boldsymbol{a}^T\boldsymbol{x}) &= \boldsymbol{a} \\ \nabla^2(\boldsymbol{x} \mapsto \boldsymbol{a}^T\boldsymbol{x}) &= O \\ \nabla(\boldsymbol{x} \mapsto \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x}) &= Q\boldsymbol{x} \\ \nabla^2(\boldsymbol{x} \mapsto \frac{1}{2}\boldsymbol{x}^T Q\boldsymbol{x}) &= Q. \end{aligned}$$

Example:
Let $\boldsymbol{u} \in \mathbb{R}^m$ and $\boldsymbol{v} \in \mathbb{R}^n$.. The function $f : \mathbb{R}^{m\times n} \to \mathbb{R},\ X \mapsto \boldsymbol{u}^T X \boldsymbol{v}$ is a linear function of X. This becomes obvious if we use the following formula:

$$\boldsymbol{u}^T X \boldsymbol{v} = \langle X, \boldsymbol{u}\boldsymbol{v}^T \rangle. \tag{1}$$

The above is a simple consequence from the fact that the trace is *invariant to cyclic permutations*: trace AB = trace BA. Indeed, #3

$$\begin{aligned} \boldsymbol{u}^T X \boldsymbol{v} &= \operatorname{trace} \boldsymbol{u}^T X \boldsymbol{v} && \text{(seen as a } 1\times 1\text{-matrix)} \\ &= \operatorname{trace} X\boldsymbol{v}\boldsymbol{u}^T && \text{(from the invariance to cyclic permutations)} \\ &= \langle X, \boldsymbol{u}\boldsymbol{v}^T \rangle && \text{(note that } \boldsymbol{u}\boldsymbol{v}^T \text{ is an } m\times n\text{-matrix)} \end{aligned}$$

Random vectors

Let $\boldsymbol{X}$ be a real-valued random variable, associated with a probability measure $\mathbb{P}$. The *expected value* (or the *expectation*) of $\boldsymbol{X}$ is

$$\mathbb{E}[\boldsymbol{X}] = \int_{x\in\mathbb{R}} x \, d\mathbb{P}(x).$$

and its *variance* is

$$\mathbb{V}[\boldsymbol{X}] = \mathbb{E}[\boldsymbol{X}^2] - \mathbb{E}[\boldsymbol{X}]^2 = \mathbb{E}\big[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^2\big] \geq 0.$$

Similarly, when $\boldsymbol{X}$ is a random vector in $\mathbb{R}^n$, it expected value is the vector

$$\mathbb{E}[\boldsymbol{X}] = \int_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{x} \, d\mathbb{P}(\boldsymbol{x}) = \begin{bmatrix} \mathbb{E}[\boldsymbol{X}_1] \\ \vdots \\ \mathbb{E}[\boldsymbol{X}_n] \end{bmatrix}.$$

The *variance-covariance matrix* of $\boldsymbol{X}$ is the matrix with the *covariance* of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ in position (i,j): $\mathbf{cov}[\boldsymbol{X}_i, \boldsymbol{X}_j] = \mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_j] - \mathbb{E}[\boldsymbol{X}_i]\,\mathbb{E}[\boldsymbol{X}_j] = \mathbb{E}\big[(\boldsymbol{X}_i - \mathbb{E}[\boldsymbol{X}_i])(\boldsymbol{X}_j - \mathbb{E}[\boldsymbol{X}_j])\big]$. In matrix notation,

$$\mathbb{V}[\boldsymbol{X}] := \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^T = \mathbb{E}\big[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T\big].$$

Note that the diagonal of ith diagonal element of $\mathbb{V}[\boldsymbol{X}]$ is nothing but $\mathbb{V}[\boldsymbol{X}_i]$. The last expression also shows

that a variance-covariance matrix is always a positive semidefinite matrix (this will become clear during the lecture): $\mathbb{V}[\boldsymbol{X}] \in \mathbb{S}_+^n$.

The expectation is linear, so in particular, for a scalar random variable $\boldsymbol{X}$ and constants $a, b \in \mathbb{R}$, we have $\mathbb{E}[a\boldsymbol{X} + b] = a\mathbb{E}[\boldsymbol{X}] + b$. Similarly, for a random vector $\boldsymbol{X} \in \mathbb{R}^n$, a matrix $A \in \mathbb{R}^{m\times n}$ and a (constant) vector $\boldsymbol{b} \in \mathbb{R}^m$, it holds

$$\mathbb{E}[A\,\boldsymbol{X} + \boldsymbol{b}] = A\,\mathbb{E}[\boldsymbol{X}] + \boldsymbol{b}.$$

A consequence is that when $\boldsymbol{X}$ is multiplied by a constant $a \in \mathbb{R}$, its variance (or variance-covariance matrix) gets multiplied by the square of this constant: $\mathbb{V}[a\boldsymbol{X}] = a^2\mathbb{V}[\boldsymbol{X}]$. More generally, we can show that if $\boldsymbol{X}$ takes values in $\mathbb{R}^n$ and A is an $m \times n-$matrix,

$$\mathbb{V}[A\,\boldsymbol{X}] = A\,\mathbb{V}[\boldsymbol{X}]A^T.$$

Example:
Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two real-valued random variables. We can use the above formula to derive the standard formula

$$\mathbb{V}[\boldsymbol{X} + \boldsymbol{Y}] = \mathbb{V}[\boldsymbol{X}] + \mathbb{V}[\boldsymbol{Y}] + 2\,\mathbf{cov}[\boldsymbol{X}, \boldsymbol{Y}],$$

which shows, in particular, that the variance of the sum equals the sum of the variances whenever the two variables are uncorrelated. Indeed, let $\boldsymbol{Z} = [\boldsymbol{X}, \boldsymbol{Y}]^T$ be the random vector of $\mathbb{R}^2$ obtained by concatenating $\boldsymbol{X}$ and $\boldsymbol{Y}$. Then, #4

$$\mathbb{V}[\boldsymbol{X} + \boldsymbol{Y}] = \mathbb{V}\left[\mathbf{1}^T\boldsymbol{Z}\right] = \mathbf{1}^T\mathbb{V}[\boldsymbol{Z}]\mathbf{1} = \mathbf{1}^T\begin{bmatrix} \mathbb{V}[\boldsymbol{X}] & \mathbf{cov}[\boldsymbol{X}, \boldsymbol{Y}] \\ \mathbf{cov}[\boldsymbol{X}, \boldsymbol{Y}] & \mathbb{V}[\boldsymbol{Y}] \end{bmatrix}\mathbf{1},$$

which is nothing but the sum of the 4 entries of the last $2 \times 2-$matrix.

Example:
Let $\boldsymbol{X}$ be a random vector which takes values in $\mathbb{R}^n$, with expected value $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu} \in \mathbb{R}^n$ and variance-covariance matrix $\mathbb{V}[\boldsymbol{X}] = \Sigma \in \mathbb{S}_+^n$.

The function $f : \mathbb{S}^n \to \mathbb{R}$, which associates Q to the expected value of $\boldsymbol{X}^TQ\boldsymbol{X}$ is a linear form. We will show that $f(Q)$ has a simple expression, which depends only on $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}]$ and $\Sigma = \mathbb{V}[\boldsymbol{X}]$. To see this, we can write #5

$$\begin{aligned} f(Q) &= \mathbb{E}[\boldsymbol{X}^TQ\boldsymbol{X}] \\ &= \mathbb{E}\big[\langle Q, \boldsymbol{X}\boldsymbol{X}^T\rangle\big] && \text{(cf. Eq (1) in Example \#3)} \\ &= \big\langle Q, \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]\big\rangle && \text{(linearity of expectation)} \\ &= \langle Q, \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^T\rangle && \text{(because } \Sigma = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T\text{)}. \end{aligned}$$