Continuous Optimization

Chapter 1: Basics of Optimization

1 Optimization problems

In this course we will focus on finite-dimensional continuous optimization problems

$$\begin{aligned} \min \ & f(x) \\ \text{s.t.} \ & g_i(x) \leq 0 \quad i = 1, \dots, m \\ & h_j(x) = 0 \quad j = 1, \dots, p \end{aligned} \tag{1}$$

- $x \in \mathbb{R}^n$: vector of variables to be chosen (n scalar variables $x_1, \dots, x_n$)
- f: objective function to be minimized
- $g_1, \dots, g_m$: inequality constraint functions to be satisfied
- $h_1, \dots, h_p$: equality constraint functions to be satisfied

The feasible set can also be denoted by $S = \{x \in \mathbb{R}^n : g_i(x) \leq 0 \ i = 1, \dots, m, \ \ h_j(x) = 0 \ j = 1, \dots, p\}$.

1.1 Classification

Concerning the dimensionality, optimization problems can be

- infinite-dimensional: when the space of variables is infinite dimensional, e.g., elements can be vectors of functions of one ore more variables;
- finite-dimensional: when the space of variables is finite-dimensional, e.g., $\mathbb{R}^n$.

Concerning the granularity of the space, optimization problems can be

- discrete: when at least one variable is an integer;
- continuous: when all variables involved are continuous.

Concerning the hypothesis of linearity, optimization problems are also called

- linear programming: when the objective function is linear and the feasible set is defined by a system of linear equalities and inequalities;
- nonlinear problems: when the objective function is nonlinear and/or at least one constraint is defined by a nonlinear function.

Concerning the hypothesis of differentiability, we can distinguish between

- smooth optimization: when the functions defining the objective and the constraints are continuously differentiable;
- nonsmooth optimization problems: when the functions defining the objective and the constraints are continuous, but typically not differentiable in all points of the feasible set. Weaker notions of derivatives are used to solve this problem.
- derivative-free optimization: this branch of optimization consider cases in which evaluating the objective function/constraints is very costly, in particular the functions defining the objective and the constraints might not be continuous. A similar field is that of black-box optimization, in which the optimizer assumes to not have any information on the objective function or on the constraints.

Concerning the hypothesis of convexity, we can distinguish between

- convex optimization, when the functions defining the objective and the constraints are all convex;

- nonconvex optimization, when the functions defining the objective and the constraints might be nonconvex.

Example 1 (The transportation problem (Linear Programming)) *Consider a company producing and selling a certain commodity. The company has a set of N sources, where the commodity is produced and a set of M demand centers where the commodity is sold. At each source $i \in \{1, \dots, N\}$, a given quantity q_i of commodity is available. Each demand center $j \in \{1, \dots, M\}$ requires a given quantity d_j. We indicate by c_{ij} the cost per unit of transporting the commodity from source i to demand center j.*

The problem is to determine the quantity to be transported from each source to each demand center in such a way that

- *the availability constraints at the sources are satisfied;*
- *the requirements of the demand centers are satisfied;*
- *the total transportation cost is minimized.*

To formulate the problem, we indicate by x_{ij} the quantity of the commodity transported from source i to demand center j. Then the optimization problem is

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} \sum_{j=1}^{M} c_{ij} x_{ij} \\ s.t. \quad & \sum_{j=1}^{M} x_{ij} \leq q_i \quad i = 1, \dots, m \\ & h_j(x) = 0 \quad j = 1, \dots, p \end{aligned}$$

Example 2 (Training of Neural Networks (Unconstrained Optimization)) *Let us start from describing a single (artificial) neuron. In particular, this unit takes as an input a vector and performs a scalar product with the weights of the neuron. In a second step, this value passes through a thresholding function σ called activation function. Inspired by real neurons, also this unit propagates its output only if the weighted sum of its inputs (in the natural neuron, the potential difference) is over a certain threshold. Neural networks stack various neurons in parallel to build a layer and, in turn, concatenate various layer (deep networks). This model is very expressive, in fact already a wide-enough 2-layer neural network is able to approximate any possible function of the input.*

Neural networks are the most widely used machine learning model and they can be formally defined as follows.

- *$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{M}$ be a training set of M instances*
  - *$x_i \in \mathbb{R}^d$ d is the amount of features of the dataset (e.g., pixels)*
  - *$y_i \in \mathbb{R}$*
- *the j-th layer applied on a generic input n_{j-1}-dimensional input x:*
  $g_j(x) := \sigma(W_j x + b_j)$, $W_j \in \mathbb{R}^{n_{j-1} \times n_j}$, $b_j \in \mathbb{R}^{n_j}$, σ is applied component-wise,
  e.g., ReLU $\sigma(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{otherwise} \end{cases}$
- *the weights: $w = \begin{pmatrix} vec(W_1) \\ \vdots \\ vec(W_L) \end{pmatrix} \in \mathbb{R}^n$*
- *the network applied on the i-th instance: $h_i(w) := g_L \circ \cdots \circ g_1(x_i)$*
- *the losses: $f_i(w) := \mathcal{L}(h_i(w), y_i)$, e.g., $\mathcal{L}(h_i(w), y_i) = (h_i(w) - y_i)^2$*

$$\min_{w \in \mathbb{R}^n} f(w) = \frac{1}{M} \sum_{i=1}^{M} f_i(w)$$

Example 3 (Portfolio Optimization (Convex Quadratic Programming)) *Portfolio selection theory studies how to allocate an investor's available capital into a prefixed set of assets with the aims of maximizing the expected return and minimizing the investment risk. Let n be the number of available assets, let $\mu \in \mathbb{R}^n$ be the vector of expected returns of the assets, and let $Q \in \mathbb{R}^{n\times n}$ be a symmetric positive semidefinite matrix whose generic element q_{ij} is the covariance of returns of assets i and j. We assume that the available (normalized) capital is fully invested. Then, let $x \in \mathbb{R}^n$ be the vector of decision variables, where x_i is the fraction of the available capital to be invested into asset i, with $i = 1, \dots, n$.*

By this notation, $\mu^T x$ is the expected return of the portfolio and $x^T Q x$ is the covariance of the portfolio which can be used as a measure of the risk connected with the investment (diversification of the portfolio). In the traditional Markowitz portfolio selection model [1], the optimization problem is stated as the following convex quadratic programming problem

$$\begin{aligned} \min\ & x^T Q x \\ s.t.\ & \mu^T x \geq \beta \\ & e^T x = 1 \\ & x \geq 0, \end{aligned}$$

where β is the desired expected return of the portfolio and $e \in \mathbb{R}^n$ denotes the column vector of all ones.

Example 4 (Shortest Path (Discrete Optimization, not covered by this course)) *Let $G = (V, E)$ be a graph with V the set of vertices and E the set of edges (directed, $(i, j) \neq (j, i)$, with i, j being two vertices) with $n = |E|$. Let $c_{ij} \geq 0$ be the cost of choosing the edge (i, j), e.g., travel time, travel fee, difference in hight, and let $x_{ij} \in \{0, 1\}^n$ be the variable telling if we select or not the edge (i, j). Also, let s be the starting node and d the destination, then the optimization problem can be stated as*

$$\begin{aligned} \min\ & c^T x \\ s.t.\ & x \geq 0 \\ & \sum_j x_{ij} - \sum_j x_{ji} = \begin{cases} 1 & i = s \\ 0 & i \neq s, d \\ -1 & i = d \end{cases} \qquad \forall i \in V \end{aligned}$$

2 Mathematical Preliminaries

2.1 The space $\mathbb{R}^n$

The vector space $\mathbb{R}^n$ is the set of n-dimensional column vectors with real components endowed with the component-wise addition operator

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}$$

and the scalar product

$$\lambda \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{pmatrix}$$

The standard basis of $\mathbb{R}^n$ is $e_1, e_2, \dots, e_n$, where e_i is the n-length column vector whose ith component is 1 while all the others are 0.
The nonnegative orthant $\mathbb{R}^n_+ := \{(x_1, x_2, \dots, x_n)^T : x_1, x_2, \dots, x_n \geq 0\}$.
The positive orthant $\mathbb{R}^n_{++} := \{(x_1, x_2, \dots, x_n)^T : x_1, x_2, \dots, x_n > 0\}$.
The set of all real-valued matrices of order $m \times n$ is denoted by $\mathbb{R}^{m\times n}$.
The identity matrix will be denoted by Id, where its dimension will be clear from the context.

2.2 Inner Products

Definition 1 (Inner product) *An inner product on $\mathbb{R}^n$ is a map $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the following properties*

1. *Symmetry:* $\langle x, y\rangle = \langle y, x\rangle \quad \forall x, y \in \mathbb{R}^n$;
2. *Additivity:* $\langle x, y + z\rangle = \langle x, y\rangle + \langle x, z\rangle \quad \forall x, y, z \in \mathbb{R}^n$;
3. *Homogeneity:* $\langle \lambda x, y\rangle = \lambda\langle x, y\rangle \quad \forall x, y \in \mathbb{R}^n$;
4. *Positive definiteness:* $\langle x, x\rangle \geq 0 \quad \forall x \in \mathbb{R}^n$ *and* $\langle x, x\rangle = 0 \Leftrightarrow x = 0$.

In this course the inner product we will use is the standard dot product:

$$\langle x, y\rangle = x^T y = \sum_{i=1}^{n} x_i y_i \quad \forall x, y \in \mathbb{R}^n$$

Notice that the dot product is not the only inner product in $\mathbb{R}^n$, for instance let $w \in \mathbb{R}^n_{++}$, it is easy to show that the following weighted dot product is also an inner product

$$\langle x, y\rangle_w = \sum_{i=1}^{n} w_i x_i y_i.$$

2.3 Vector norms

Definition 2 *(Vector Norm) A norm* $||.||$ *on* $\mathbb{R}^n$ *is a function* $||.|| : \mathbb{R}^n \to \mathbb{R}$ *satisfying the following:*

1. *Nonnegativity:* $||x|| \geq 0 \quad \forall x \in \mathbb{R}^n$ *and* $||x|| = 0 \Leftrightarrow x = 0$;
2. *Positive Homogeneity:* $||\lambda x|| = |\lambda|||x|| \quad \forall x \in \mathbb{R}^n$ *and* $\lambda \in \mathbb{R}$;
3. *Triangle Inequality:* $||x + y|| \leq ||x|| + ||y|| \quad \forall x, y \in \mathbb{R}^n$.

One natural way to generate a norm on $\mathbb{R}^n$ is to take any inner product $\langle \cdot, \cdot\rangle$ on $\mathbb{R}^n$ and define the associated norm

$$||x|| := \sqrt{\langle x, x\rangle} \quad \forall x \in \mathbb{R}^n$$

which can be proved to be a norm. If the inner product is the dot product, then the associated norm is the Euclidean norm or l_2 norm, i.e., $|| \cdot ||_2$. In the rest of the course, the default norm will be the Euclidean norm and the subscript 2 will be omitted. This norm belongs to the class of l_p norms (for $p \geq 1$, with $0 \leq p < 1$ these are quasi-norms) defined by

$$||x||_p := \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}.$$

By computing the limit of the l_p-norm with $p \to \infty$ we achieve the l_∞ norm, i.e.,

$$||x||_\infty := \max_{i \in [n]} |x_i|.$$

where $[n] := \{1, \ldots, n\}$ for any $n \in \mathbb{N}$.

An important inequality connecting the dot product of two vectors and their norms is the Cauchy–Schwarz inequality, which will be used frequently throughout the book.

Lemma 1 (Cauchy–Schwarz inequality) *For any* $x, y \in \mathbb{R}^n$,

$$|x^T y| \leq ||x|| \cdot ||y||.$$

Equality is satisfied if and only if x *and* y *are linearly dependent.*

2.4 Matrix norms

Similarly to vector norms, we can define the concept of a matrix norm.

Definition 3 *(Operator/Matrix Norm) A norm* $||.||$ *on* $\mathbb{R}^{m \times n}$ *is a function* $||.|| : \mathbb{R}^{m \times n} \to \mathbb{R}$ *satisfying the following:*

1. *Nonnegativity:* $||A|| \geq 0 \quad \forall A \in \mathbb{R}^{m \times n}$ *and* $||A|| = 0 \Leftrightarrow A = 0$;
2. *Positive Homogeneity:* $||\lambda A|| = |\lambda|||A|| \quad \forall A \in \mathbb{R}^{m \times n}$ *and* $\lambda \in \mathbb{R}$;

3. Triangle Inequality: $||A+B|| \leq ||A|| + ||B|| \quad \forall A, B \in \mathbb{R}^{m\times n}$.

Given a matrix $A \in \mathbb{R}^{m\times n}$ and two norms $||.||_p$ and $||.||_q$ on $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, the induced operator/matrix norm $||A||_{a\to b}$ is defined by

$$||A||_{p\to q} = \max_{||x||_p=1} ||Ax||_q.$$

It can be shown that the above definition implies that for any $x \in \mathbb{R}^n$ it holds the inequality

$$||Ax||_q \leq ||A||_{p\to q}||x||_p.$$

An induced matrix norm is indeed a norm in the sense that it satisfies the three properties required above. We refer to the matrix norm $||\cdot||_{p,q}$s the (p,q)-norm. When $p = q$, we will simply refer to it as an p-norm and omit one of the subscripts in its notation.

An important operator norm is the spectral norm, where $p = q = 2$. In this case, it can be proved that the operator norm coincide with the maximum singular value of $A \in \mathbb{R}^{m\times n}$ (see below for more details on eigenvalues and singular values)

$$||A||_2 = ||A||_{2\to 2} = \sqrt{\lambda_1(A^TA)} =: \sigma_1(A).$$

The default operator norm for this course will be the spectral norm, so when the subscript is omitted we will implicitly refer to this norm.

When $p = q = 1$, the operator norm reduces to

$$||A||_1 = ||A||_{1\to 1} = \max_{j\in[n]} \sum_{i=1}^{m} A_{ij}.$$

This norm is also called maximum absolute column sum norm.

When $p = q = \infty$, the operator norm reduces to

$$||A||_\infty = ||A||_{\infty\to\infty} = \max_{j\in[m]} \sum_{i=1}^{n} A_{ij}.$$

This norm is also called the maximum absolute row sum norm.

An example of a matrix norm that is not induced by any norm is the Frobenius norm defined by

$$||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} = ||\text{vec}(A)||,$$

in fact, you can think of this norm as the Euclidean norm of the vector obtained by flattening the matrix A into a vector.

2.5 Eigenvalues and Eigenvectors

Let $A \in \mathbb{R}^{n\times n}$. Then a nonzero vector $v \in \mathbb{R}^n$ is called an eigenvector of A if there exists a $\lambda \in \mathbb{C}$ (the complex field) for which

$$Av = \lambda v$$

The scalar λ is the eigenvalue corresponding to the eigenvector v. In general, real-valued matrices can have complex eigenvalues, but it is well known that all the eigenvalues of symmetric matrices are real. The eigenvalues of a symmetric $n \times n$ matrix A are denoted by

$$\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A),$$

where $\lambda_1(A)$ is the maximum eigenvalue and $\lambda_n(A)$ is the minimum.

Definition 4 *A matrix U is said to be orthogonal when $U^TU = UU^T = Id$.*

One of the most useful results related to eigenvalues is the spectral decomposition theorem, which states that any symmetric matrix A has an orthonormal basis of eigenvectors.

Theorem 1 (Spectral Decomposition) *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix and let $v_1, \ldots, v_n$ be the eigenvectors corresponding to the eigenvalues $\lambda_1(A), \ldots, \lambda_n(A)$. Then $U = (v_1|\ldots|v_n)$ is an orthonormal matrix and A can be decomposed as follows*

$$A = UDU^T, \quad \textit{with } D := \textit{diag}(\lambda_1(A), \ldots, \lambda_n(A)).$$

A direct result of the spectral decomposition theorem is that the trace and the determinant of a matrix A can be expressed via its eigenvalues:

$$\mathrm{Tr}(A) = \sum_{i=1}^{n} \lambda_i(A)$$
$$\det(A) = \prod_{i=1}^{n} \lambda_i(A)$$

Another important consequence of the spectral decomposition theorem is the bounding of the so-called Rayleigh quotient. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, the Rayleigh quotient is defined as

$$R_A(x) := \frac{x^T A x}{||x||^2} \quad \forall x \neq 0,$$

and its upper and lower bounds are given by the following lemma.

Lemma 2 *Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Then*

$$\lambda_n(A) \leq R_A(x) \leq \lambda_1(A) \quad \forall x \neq 0.$$

Moreover, we have the following corollary.

Lemma 3 *Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Then*

$$\min_{x\neq 0} R_A(x) = \lambda_n(A),$$

and the eigenvector(s) of A corresponding to the minimal eigenvalue are the minimizers of this problem. In addition,

$$\max_{x\neq 0} R_A(x) = \lambda_1(A),$$

and the eigenvectors of A corresponding to the maximal eigenvalue are maximizers of this problem.

2.6 Singular values and Singular Vectors

Similar objects exist also for non-squared matrices $A \in \mathbb{R}^{m\times n}$.

Theorem 2 (Singular Value Decomposition) *Let $A \in \mathbb{R}^{m\times n}$. Then there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ and the uniquely defined real values*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_\ell \geq 0, \quad \textit{with } \ell = \min\{m, n\},$$

such that

$$A = U\Sigma V^T \quad \textit{with } \Sigma = \textit{diag}(\sigma_1, \ldots, \sigma_\ell) \in \mathbb{R}^{m\times n}.$$

The matrices $U = (u_1, \ldots, u_m)$ and $V = (v_1, \ldots, v_m)$ are composed respectively of the left and right singular vectors, thus we can decompose A also as follows

$$A = \sum_{j=1}^{\ell} \sigma_j u_j v_j^T.$$

If $\mathrm{rank}(A) = r$, then $\sigma_1, \geq \ldots, \geq \sigma_r > 0$ and $\sigma_{r+1} = \cdots = \sigma_\ell = 0$ so that the reduced singular value decomposition is

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T.$$

Also, note that

$$A^T A = (U\Sigma V^T)^T U\Sigma V^T = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T\Sigma V^T = V\mathrm{diag}(\sigma_1^2, \ldots, \sigma_\ell^2, 0, \ldots, 0)V^T$$

and

$$AA^T = U\Sigma V^T (U\Sigma V^T)^T = U\Sigma V^T V\Sigma^T U^T = U\Sigma\Sigma^T U^T = U\mathrm{diag}(\sigma_1^2, \ldots, \sigma_\ell^2, 0, \ldots, 0)U^T,$$

which implies that $\sigma_1^2, \ldots, \sigma_\ell^2$ are eigenvalues of A^TA and AA^T, $u_1, \ldots, u_m$ are eigenvectors of AA^T and $v_1, \ldots, v_n$ are eigenvectors of A^TA. In other words the SVD can be obtained from the spectral decomposition of AA^T and A^TA. In particular,

$$\sigma_j = \sigma_j(A) = \sqrt{\lambda_j(A^TA)} = \sqrt{\lambda_j(AA^T)}$$

To conclude this section, the following function is called nuclear norm

$$||A||_* = \sum_{j=1}^{\ell} \sigma_j(A) \quad \text{with } \ell = \min\{m, n\},$$

and can actually proved to be a norm for $\mathbb{R}^{m \times n}$.

References

[1] Harry M Markowits. Portfolio selection. *Journal of finance*, 7(1):71–91, 1952.