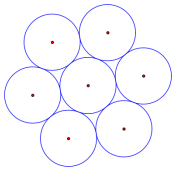



# Duality and augmented Lagrangian

## Lecture notes

TAF MCE - UE Advanced optimization for IA

thierry.chonavel@imt-atlantique.fr

2021

## Contents



## 1 Introduction

In these notes we introduce duality principle that is often useful in optimization. In constrained optimization problems, it can be noted that if Lagrange multipliers were known, finding solution often would be much simpler. Thus, focussing on the search for Lagrange multipliers can be of particular interest. The problem of finding them is known as the **dual problem**. In the case of convex problems, we will see that that there is some symmetry between the initial (also called **primal**) problem and the dual problem, while in the general case, the difference between the solutions of these two problems is measured by the **duality gap**.
Let us consider the standard constrained optimization problem

$$\begin{cases} \min_{\mathbf{x}} f(\mathbf{x}) \\ \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \\ \mathbf{h}(\mathbf{x}) = \mathbf{0} \end{cases} \tag{1}$$

with $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^p$. Problem (1) is often referred to as the **primal problem**. Alternatively, the **dual problem** looks for optimization of Lagrange multipliers. Before going into details, let us give a few reasons why duality is worth being considered.

1. **the dual problem of (1) is a concave optimization problem**, regardless problem (1) is convex or not. Thus the dual problem is often simpler than the initial problem.

2. If problem (1) has a global solution given by $f(\mathbf{x}^*) = p^*$ and $d^*$ denotes the optimum of the dual problem, we have $p^* \geq d^*$. In other words **solving the dual problem yields a lower bound for the solution of problem (1)**, what can be helpful to decide wether an approximate solution $\mathbf{x} \in \mathcal{D} = dom(\mathbf{g}) \cap dom(\mathbf{h})$ of (1) could be accepted by comparing $f(\mathbf{x})$ to $d^*$.

3. **The initial problem involves $n$ variables and $m+p$ constraints while its dual problem involves $m+p$ variables and $p$ simple constraints**. What makes a problem difficult is often handling a large number of complex inequality constraints rather than a large number of variables, making thus the dual problem often appealing when $p$ is large.

4. Interesting algorithms can be derived from the study of duality principle.

In the following we are going to define the dual problem and justify these arguments.

## 2 Dual problem and duality gap

### 2.1 Definition

The Lagrangian of problem (1) writes

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}).$$

Clearly, for $\boldsymbol{\mu} \geq \mathbf{0}$, we have

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq f(\mathbf{x}) + \mathcal{I}_{\{0\}}(\mathbf{h}(\mathbf{x})) + \mathcal{I}_{\mathbb{R}^m_-}(\mathbf{g}(\mathbf{x})), \tag{2}$$

where $\mathcal{I}_A(\mathbf{u}) = 0$ if all entries of $\mathbf{u}$ are in $A$ and $\mathcal{I}_A(\mathbf{x}) = +\infty$ otherwise. Thus, $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is a continuous underestimate of

the right hand term which is an objective equivalent to the initial constrained problem. The **Dual function** is defined by

$$\phi(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}). \tag{3}$$

## 2.2 Weak duality

Letting

$$d^* \doteq \sup_{\boldsymbol{\lambda}; \boldsymbol{\mu} \geq \mathbf{0}} \phi(\boldsymbol{\lambda}, \boldsymbol{\mu}),$$

the quantity $p^* - d^*$ is called the **duality gap**:

**Theorem 1 (weak duality)** *The duality gap $p^* - d^*$ is positive:*

$$p^* \geq d^*.$$

**Proof** Relations (2) and (3) yield

$$\phi(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \min_{\{\mathbf{x}; \mathbf{h}(\mathbf{x})=0, \mathbf{g}(\mathbf{x}) \leq 0\}} f(\mathbf{x}) = p^*.$$

Then,

$$\mathbf{d}^* = \max_{\boldsymbol{\lambda}; \boldsymbol{\mu} \geq \mathbf{0}} \phi(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq p^*$$

□

In fact , this result is a particular case of the **max-min inequality** [4] which states that for any function $k(\mathbf{w}, \mathbf{z})$ the **max-min inequality** [4] is satisfied:

$$\sup_{\mathbf{z}} \inf_{\mathbf{w}} k(\mathbf{w}, \mathbf{z}) \leq \inf_{\mathbf{w}} \sup_{\mathbf{z}} k(\mathbf{w}, \mathbf{z}). \tag{4}$$

Indeed, since $k(\mathbf{w}, \mathbf{z}) \leq \sup_{\mathbf{z}} k(\mathbf{w}, \mathbf{z})$, we have $\inf_{\mathbf{w}} k(\mathbf{w}, \mathbf{z}) \leq \inf_{\mathbf{w}} \sup_{\mathbf{z}} k(\mathbf{w}, \mathbf{z})$ and as the right hand side of this inequality does not depend on $\mathbf{z}$, considering the supremum of both sides with respect to $\mathbf{z}$ yields the max-min inequality. Theorem 1 is obtained letting $k(\mathbf{w}, \mathbf{z}) = L(\mathbf{w}, \mathbf{z})$, with $\mathbf{w} = \mathbf{x}$ and $\mathbf{z} = (\boldsymbol{\mu}, \boldsymbol{\lambda})$.

## 2.3 Example

As an example, let us consider the problem $\min \frac{1}{2} \parallel \mathbf{x} \parallel^2$, s.t. (subject to) $\mathbf{A}\mathbf{x} = \mathbf{b}$. Then, $\nabla L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{x} + \mathbf{A}^T\boldsymbol{\lambda}$ and for fixed $\boldsymbol{\lambda}$ the minimum of $L$ is reached for $\mathbf{x} = -\mathbf{A}^T\boldsymbol{\lambda}$. This leads to

$$\begin{aligned} \phi(\boldsymbol{\lambda}) &= L(-\mathbf{A}^T\boldsymbol{\lambda}, \boldsymbol{\lambda}) \\ &= (1/2)\boldsymbol{\lambda}^T\mathbf{A}\mathbf{A}^T\boldsymbol{\lambda} + \boldsymbol{\lambda}^T(-\mathbf{A}\mathbf{A}^T\boldsymbol{\lambda} - \mathbf{b}) \\ &= -(1/2)\boldsymbol{\lambda}^T\mathbf{A}\mathbf{A}^T\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\mathbf{b}. \end{aligned}$$

## 2.4 Concavity of the dual function

Now, for the sake of simplicity, we can assume inequality constraints only since $\mathbf{h}(\mathbf{x}) = 0 \Leftrightarrow [\mathbf{h}(\mathbf{x}) \leq 0$ and $\mathbf{h}(\mathbf{x}) \geq 0]$. In the previous example $\phi$ is clearly a concave function. In fact, this is a general property:

**Theorem 2** *The set of $\mathbb{R}^p_+$ where $\phi(\boldsymbol{\mu})$ is lower bounded is convex and in this set $\phi$ is concave.*

**Proof** For $0 \leq \alpha \leq 1$, and $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^p_+$ where $\phi(\boldsymbol{\mu}_1), \phi(\boldsymbol{\mu}_2) > -\infty$

$$\begin{aligned} &\phi(\alpha\boldsymbol{\mu}_1 + (1-\alpha)\boldsymbol{\mu}_2) \\ &\quad = \inf_{\mathbf{x}}\{f(\mathbf{x}) + (\alpha\boldsymbol{\mu}_1 + (1-\alpha)\boldsymbol{\mu}_2)^T\mathbf{g}(\mathbf{x})\} \\ &\quad \geq \alpha \inf_{\mathbf{x}}\{f(\mathbf{x}) + \boldsymbol{\mu}_1^T\mathbf{g}(\mathbf{x})\} \\ &\qquad + (1-\alpha)\inf_{\mathbf{x}}\{f(\mathbf{x}) + \boldsymbol{\mu}_2^T\mathbf{g}(\mathbf{x})\} \\ &\quad \geq \alpha\phi(\boldsymbol{\mu}_1) + (1-\alpha)\phi(\boldsymbol{\mu}_2). \end{aligned}$$

□

## 2.5 Strong duality and saddle point interpretation

Let us consider again the example in subsection 2.3 for which first order necessary optimality conditions (NC1) clearly write

$$(\mathbf{x}^*, \boldsymbol{\lambda}^*) = (\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b},\ -(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}).$$

We observe that

$$\begin{aligned} d^* &= \max_{\boldsymbol{\lambda}} \phi(\boldsymbol{\lambda}) \\ &= \phi(\boldsymbol{\lambda}^*) \\ &= (1/2)\ \mathbf{b}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b} \\ &= (1/2) \parallel \mathbf{x}^* \parallel^2 \\ &= p^* \end{aligned}$$

Thus, it appears that there are cases where maximizing $\phi$ yields the multipliers that solve NC1 optimality conditions and the duality gap is 0. When it is satisfied, this property is called **strong duality**.

This leads us to a geometric interpretation of strong duality. As discussed at the end of section 2.2, for any multivariate function $k(\mathbf{w}, \mathbf{z})$ the max-min inequality is always satisfied: $\sup_{\mathbf{z}} \inf_{\mathbf{w}} k(\mathbf{w}, \mathbf{z}) \leq \inf_{\mathbf{w}} \sup_{\mathbf{z}} k(\mathbf{w}, \mathbf{z})$. On another hand, a **saddle point** for $k$ is defined as a point $(\mathbf{w}^*, \mathbf{z}^*)$, such that

$$\forall \mathbf{w}, \mathbf{z},\ k(\mathbf{w}^*, \mathbf{z}) \leq k(\mathbf{w}^*, \mathbf{z}^*) \leq k(\mathbf{w}, \mathbf{z}^*),$$

that is, at $(\mathbf{w}^*, \mathbf{z}^*)$ $k$ has a minimum w.r.t. (with respect to) $\mathbf{w}$ and a maximum w.r.t. $\mathbf{z}$. The next theorem shows that if there is a saddle point, at this point the max-min inequality for $k$ can be replaced by an equality.

**Theorem 3** *If $(\mathbf{w}^*, \mathbf{z}^*)$ is a saddle point of $k(\mathbf{w}, \mathbf{z})$, then*

$$\sup_{\mathbf{z}} \inf_{\mathbf{w}} k(\mathbf{w}, \mathbf{z}) = k(\mathbf{w}^*, \mathbf{z}^*) = \inf_{\mathbf{w}} \sup_{\mathbf{z}} k(\mathbf{w}, \mathbf{z})$$

**Proof** As $(\mathbf{w}^*, \mathbf{z}^*)$ is a saddle point,

$$\inf_{\mathbf{w}} \sup_{\mathbf{z}} k(\mathbf{w}, \mathbf{z}) \leq \sup_{\mathbf{z}} k(\mathbf{w}^*, \mathbf{z}) \leq k(\mathbf{w}^*, \mathbf{z}^*)$$

$$k(\mathbf{w}^*, \mathbf{z}^*) \leq \inf_{\mathbf{w}} k(\mathbf{w}, \mathbf{z}^*) \leq \sup_{\mathbf{z}} \inf_{\mathbf{w}} k(\mathbf{w}, \mathbf{z}).$$

Thus,

$$\inf_{\mathbf{w}} \sup_{\mathbf{z}} k(\mathbf{w}, \mathbf{z}) \leq k(\mathbf{w}^*, \mathbf{z}^*) \leq \sup_{\mathbf{z}} \inf_{\mathbf{w}} k(\mathbf{w}, \mathbf{z}), \tag{5}$$

and from the max-min inequality we also have $\sup_{\mathbf{z}} \inf_{\mathbf{w}} k(\mathbf{w}, \mathbf{z}) \leq \inf_{\mathbf{w}} \sup_{\mathbf{z}} k(\mathbf{w}, \mathbf{z})$ so that all inequalities in Eq. (5) become equalities. □

Then, it appears that strong duality holds when $(\mathbf{x}^*, \boldsymbol{\mu}^*)$ represents a saddle point of the Lagrangian:

$$d^* = \sup_{\boldsymbol{\mu} \geq \mathbf{0}} \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu}) = \inf_{\mathbf{x}} \sup_{\boldsymbol{\mu} \geq \mathbf{0}} L(\mathbf{x}, \boldsymbol{\mu}) = p^*, \tag{6}$$

The connections between local optima and saddle points are summarized in the following theorem.

**Theorem 4** *If $(\mathbf{x}^*, \boldsymbol{\mu}^*)$, with $\boldsymbol{\mu}^* \in \mathbb{R}^p_+$ is a saddle point of $L(\mathbf{x}, \boldsymbol{\mu})$, then $\mathbf{x}^*$ is a solution of the primal problem. Conversely, if the primal problem is convex with solution $\mathbf{x}^*$, there exists $\boldsymbol{\mu}^* \in \mathbb{R}^p_+$ such that $(\mathbf{x}^*, \boldsymbol{\mu}^*)$, is a saddle point of $L(\mathbf{x}, \boldsymbol{\mu})$.*

**Proof** If $(\mathbf{x}^*, \boldsymbol{\mu}^*)$ is a saddle point of $L(\mathbf{x}, \boldsymbol{\mu})$, the condition $L(\mathbf{x}^*, \boldsymbol{\mu}) \leq L(\mathbf{x}^*, \boldsymbol{\mu}^*)$ writes $(\boldsymbol{\mu} - \boldsymbol{\mu}^*)^T \mathbf{g}(\mathbf{x}^*) \leq \mathbf{0}$, what can occur for any $\boldsymbol{\mu} \in \mathbb{R}^p_+$ only if $\mathbf{g}(\mathbf{x}^*) \leq 0$, showing thus that $\mathbf{x}^*$ is feasible. Then $\boldsymbol{\mu}^{*T} \mathbf{g}(\mathbf{x}^*) \leq \mathbf{0}$, since $\boldsymbol{\mu}^* \in \mathbb{R}^p_+$. But letting $\boldsymbol{\mu} = \mathbf{0}$, we also get $(\mathbf{0} - \boldsymbol{\mu}^*)^T \mathbf{g}(\mathbf{x}^*) \leq \mathbf{0}$, that is, $\boldsymbol{\mu}^{*T} \mathbf{g}(\mathbf{x}^*) \geq \mathbf{0}$. Finally, $\boldsymbol{\mu}^{*T} \mathbf{g}(\mathbf{x}^*) = \sum_i \mu_i^{*T} \mathbf{g}_i(\mathbf{x}^*) = \mathbf{0}$. Then, for any $\mathbf{x} \in \mathcal{D}$, inequality $L(x^*, \boldsymbol{\mu}^*) \leq L(\mathbf{x}, \boldsymbol{\mu}^*)$ yields

$$L(\mathbf{x}^*, \boldsymbol{\mu}^*) = f(\mathbf{x}^*) \leq L(\mathbf{x}, \boldsymbol{\mu}^*) = f(\mathbf{x}) + \boldsymbol{\mu}^{*T} \mathbf{g}(\mathbf{x}) \leq f(\mathbf{x}),$$

which completes the proof of the first part of the theorem. Conversely, from KKT conditions, if $\mathbf{x}^*$ is a solution of the primal problem, there exists $\boldsymbol{\mu}^* \in \mathbb{R}^p_+$ with $\boldsymbol{\mu}^{*T} \mathbf{g}(\mathbf{x}^*) = \mathbf{0}$ and $\nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\mu}^*) = \mathbf{0}$. Consider any $\boldsymbol{\mu} \in \mathbb{R}^p_+$ and $\mathbf{x} \in \mathcal{D}$. first note that

$$L(\mathbf{x}^*, \boldsymbol{\mu}) \leq f(\mathbf{x}^*) = L(\mathbf{x}^*, \boldsymbol{\mu}^*).$$

In addition, for a convex problem, $\mathbf{x} \to L(\mathbf{x}, \boldsymbol{\mu}^*)$ is convex and the necessary condition $\nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\mu}^*) = \mathbf{0}$ is also sufficient to guarantee that $L(\mathbf{x}^*, \boldsymbol{\mu}^*) \leq L(\mathbf{x}, \boldsymbol{\mu}^*)$. Finally, we get $L(\mathbf{x}^*, \boldsymbol{\mu}) \leq L(\mathbf{x}^*, \boldsymbol{\mu}^*) \leq L(\mathbf{x}, \boldsymbol{\mu}^*)$ which proves that $(\mathbf{x}^*, \boldsymbol{\mu}^*)$ is a saddle point of $L$. □

This result suggests solving the primal problem by looking for a saddle point of the Lagrangian. A natural idea that will be developed in the next section consists in performing iteratively moves of $\mathbf{x}$ that decrease the Lagrangian and moves of $\boldsymbol{\mu}$ (and $\boldsymbol{\lambda}$ when there are equality constraints), that increase the Lagrangian. This can be done via gradient steps.
Recall the primal and dual problems:

$$\begin{array}{l} (P): \text{ solve } \inf_{\mathbf{x}} \sup_{\boldsymbol{\mu}} L(\mathbf{x}, \boldsymbol{\mu}) = p^* \\ (Q): \text{ solve } \sup_{\boldsymbol{\mu} \geq \mathbf{0}} \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu}) = d^* \end{array} \tag{7}$$

Consider also the problems $(P_{\boldsymbol{\mu}})$ : solve $\inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu})$. From theorem 4, for a saddle point $(\mathbf{x}^*, \boldsymbol{\mu}^*)$, $\mathbf{x}^*$ is a solution of $(P)$. In addition, since $\inf_{\mathbf{x}} \sup_{\boldsymbol{\mu}} L(\mathbf{x}, \boldsymbol{\mu}) = L(\mathbf{x}^*, \boldsymbol{\mu}^*) = \sup_{\boldsymbol{\mu}} \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu})$, $\boldsymbol{\mu}^*$ is a solution of the dual problem $(Q)$. Conversely, assuming that we can define uniquely $\mathbf{x}_{\boldsymbol{\mu}} = \arg\min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu})$ (then $\phi(\boldsymbol{\mu}) = L(\mathbf{x}_{\boldsymbol{\mu}}, \boldsymbol{\mu})$), we would like the solution $\mathbf{x}_{\boldsymbol{\mu}^*}$ of $(P_{\boldsymbol{\mu}^*})$ where $\boldsymbol{\mu}^*$ denotes a solution of $(Q)$ to be a solution of $(P)$. This is guaranteed by the following result:

**Theorem 5** *If $\mathbf{g}$ is continuous and for any $\boldsymbol{\mu} \in \mathbb{R}^p_+$ there is a unique solution $\mathbf{x}_{\boldsymbol{\mu}}$ of $(P_{\boldsymbol{\mu}})$, where $\boldsymbol{\mu} \to \mathbf{x}_{\boldsymbol{\mu}}$ is continuous, then $\phi(\boldsymbol{\mu})$ is derivable with $\nabla_{\boldsymbol{\mu}} \phi(\boldsymbol{\mu}) = \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}})$. In addition, if $\boldsymbol{\mu}^*$ denotes any solution of $(Q)$, $\mathbf{x}_{\boldsymbol{\mu}^*}$ is a solution of $(P)$.*

**Proof** First, let us show that $\phi(\boldsymbol{\mu})$ is derivable. Let $\boldsymbol{\mu}, \boldsymbol{\mu} + \boldsymbol{\delta} \in \mathbb{R}^p_+$. Then, since $\phi(\boldsymbol{\mu}) = L(\mathbf{x}_{\boldsymbol{\mu}}, \boldsymbol{\mu}) \leq L(\mathbf{x}_{\boldsymbol{\mu}+\boldsymbol{\delta}}, \boldsymbol{\mu})$ and $\phi(\boldsymbol{\mu} + \boldsymbol{\delta}) = L(\mathbf{x}_{\boldsymbol{\mu}+\boldsymbol{\delta}}, \boldsymbol{\mu} + \boldsymbol{\delta}) \leq L(\mathbf{x}_{\boldsymbol{\mu}}, \boldsymbol{\mu} + \boldsymbol{\delta})$, we have

$$\begin{array}{ll} \phi(\boldsymbol{\mu} + \boldsymbol{\delta}) - \phi(\boldsymbol{\mu}) & \leq L(\mathbf{x}_{\boldsymbol{\mu}}, \boldsymbol{\mu} + \boldsymbol{\delta}) - L(\mathbf{x}_{\boldsymbol{\mu}}, \boldsymbol{\mu}) \\ & \leq \boldsymbol{\delta}^T \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}}) \end{array}$$

$$\begin{array}{ll} \phi(\boldsymbol{\mu} + \boldsymbol{\delta}) - \phi(\boldsymbol{\mu}) & \geq L(\mathbf{x}_{\boldsymbol{\mu}+\boldsymbol{\delta}}, \boldsymbol{\mu} + \boldsymbol{\delta}) - L(\mathbf{x}_{\boldsymbol{\mu}+\boldsymbol{\delta}}, \boldsymbol{\mu}) \\ & \geq \boldsymbol{\delta}^T \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}+\boldsymbol{\delta}}). \end{array}$$

Then, $\phi(\boldsymbol{\mu} + \boldsymbol{\delta}) - \phi(\boldsymbol{\mu}) \in [\boldsymbol{\delta}^T \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}+\boldsymbol{\delta}}), \boldsymbol{\delta}^T \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}}]$ and there exists $\alpha \in [0, 1]$ such that

$$\begin{array}{ll} \phi(\boldsymbol{\mu} + \boldsymbol{\delta}) - \phi(\boldsymbol{\mu}) & = \boldsymbol{\delta}^T \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}}) + \alpha \boldsymbol{\delta}^T [\mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}+\boldsymbol{\delta}}) - \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}})] \\ & = \boldsymbol{\delta}^T \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}}) + \| \boldsymbol{\delta} \| \epsilon(\boldsymbol{\delta}) \end{array}$$

with $\lim_{\boldsymbol{\delta} \to \mathbf{0}} \epsilon(\boldsymbol{\delta}) = 0$, from the continuity of $\mathbf{g}$. This shows that $\phi(\boldsymbol{\mu})$ is derivable, with

$$\nabla_{\boldsymbol{\mu}} \phi(\boldsymbol{\mu}) = \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}}). \tag{8}$$

Now, as $\boldsymbol{\mu}^*$ maximizes $\phi$ over $\mathbb{R}^p_+$, Euler inequality writes $(\boldsymbol{\mu} - \boldsymbol{\mu}^*)^T \nabla_{\boldsymbol{\mu}} \phi(\boldsymbol{\mu}^*) \leq 0$ for any $\boldsymbol{\mu} \in \mathbb{R}^p_+$, that is, $\boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}^*}) \leq \boldsymbol{\mu}^{*T} \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}^*})$. This leads to

$$\begin{array}{ll} L(\mathbf{x}_{\boldsymbol{\mu}^*}, \boldsymbol{\mu}) & = f(\mathbf{x}_{\boldsymbol{\mu}^*}) + \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}^*}) \\ & \leq f(\mathbf{x}_{\boldsymbol{\mu}^*}) + \boldsymbol{\mu}^{*T} \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}^*}) \\ & \leq L(\mathbf{x}_{\boldsymbol{\mu}^*}, \boldsymbol{\mu}^*) \end{array} \tag{9}$$

Then we clearly have $L(\mathbf{x}_{\boldsymbol{\mu}^*}, \boldsymbol{\mu}) \leq L(\mathbf{x}_{\boldsymbol{\mu}^*}, \boldsymbol{\mu}^*) \leq L(\mathbf{x}, \boldsymbol{\mu}^*)$ which shows that $(\mathbf{x}_{\boldsymbol{\mu}^*}, \boldsymbol{\mu}^*)$ is a saddle point. Thus, from the first part of theorem 4, $\mathbf{x}_{\boldsymbol{\mu}^*}$ is a solution of $(P)$. □

**Remarks**

1. It is interesting to note that theorem 5 does not even require the constraints $\mathbf{g}$ to be derivable (one can check that this applies to the first part of theorem 4 it resorts to).

2. The property $\nabla_{\boldsymbol{\mu}} \phi(\boldsymbol{\mu}) = \mathbf{g}(\mathbf{x}_{\boldsymbol{\mu}})$ is very interesting for using gradient ascent to increase $L$ w.r.t. $\boldsymbol{\mu}$ from point $(\mathbf{x}_{\boldsymbol{\mu}}, \boldsymbol{\mu})$ as will be done in the next section.

3. Consider $\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu}) : \mathbb{R}^{n+p} \to \mathbb{R}^n$. Then, using arguments similar to those used in the sensitivity analysis ([5] section 2.1) that involves implicit function theorem, it can be shown that a sufficient condition for the function $\mu \to \mathbf{x}_{\boldsymbol{\mu}}$ to be uniquely defined about $(\mathbf{x}_{\boldsymbol{\mu}^*}, \boldsymbol{\mu}^*)$ is that $\mathbf{x}_{\boldsymbol{\mu}^*}$ is a regular point and $\nabla^2 L(\mathbf{x}_{\boldsymbol{\mu}^*}, \boldsymbol{\mu}^*) \succ \mathbf{0}$.

# 3 Uzawa and Arrow-Hurwicz algorithms

Considering saddle point interpretation of strong duality suggests the **Uzawa algorithm** or the **Arrow-Hurwicz algorithm** (algorithm 1) for a problem in the form $\min_{\mathbf{x}} f(\mathbf{x})$ s.t. $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$

**Algorithm 1** Uzawa-Arrow-Hurwicz algorithms

**init** $\mathbf{x}_0$ with $\mathbf{g}(\mathbf{x}_0) < 0$, $\boldsymbol{\mu}_0 > 0$
**while** stopping condition $\neq$ true **do**
  $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu}_k)$ ▷ Uzawa
  or
  $\begin{cases} \alpha_k & \approx \arg\min_r L\left(\mathbf{x}_k - r \nabla_{\mathbf{x}} L(\mathbf{x}_k, \boldsymbol{\mu}_k)\right) \\ \mathbf{x}_{k+1} & = \mathbf{x}_k - \alpha_k \nabla_{\mathbf{x}} L(\mathbf{x}_k, \boldsymbol{\mu}_k) \end{cases}$
  ▷ Arrow-Hurwicz
  $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \beta_k \mathbf{h}(\mathbf{x}_{k+1})$
  $\boldsymbol{\mu}_{k+1} = [\boldsymbol{\mu}_k + \gamma_k \mathbf{g}(\mathbf{x}_{k+1})]_+$
**end while**

Arrow-Hurwicz algorithm replaces the minimization of the Lagrangian w.r.t $\mathbf{x}$ in Uzawa algorithm by a gradient step. In algorithm 1, $[\mathbf{z}]_+$ denotes the vector with $i$-th entry $[\mathbf{z}]_{+,i} = \max(\mathbf{z}_i, 0)$. Note also that a gradient step projected on $\mathbb{R}^p_+$ is considered for the update of $\boldsymbol{\mu}$.
Convergence guarantees can be obtained for Uzawa algorithm. In particular, restricting our interest to the case of linear inequality constraints, we have

**Theorem 6** *For a problem in the form $\min_{\mathbf{x}} f(\mathbf{x})$ s.t. $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, where $f$ is $\alpha$-strongly convex [6], variables $\mathbf{x}_k$ generated by Uzawa algorithm converge to a solution $\mathbf{x}^*$ of the primal problem for fixed step size $\gamma_k = \gamma$ provided*

$$0 < \gamma < \frac{2\alpha}{\| \mathbf{A} \|^2},$$

*with $\| \ \|$ the operator norm [7]. In addition, if $\mathbf{A}$ is full rank, $\boldsymbol{\mu}_k$ converges to the corresponding Lagrange multiplier $\boldsymbol{\mu}^*$ solution of KKT conditions.*

**Proof** Uzawa iterations write

$$\left\{ \begin{array}{ll} \mathbf{x}_{k+1} & = \arg\min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\mu}_k) \\ \boldsymbol{\mu}_{k+1} & = [\boldsymbol{\mu}_k + \gamma \mathbf{g}(\mathbf{x}_k)]_+ \end{array} \right. \tag{10}$$

Then, $\nabla_{\mathbf{x}} f(\mathbf{x}_k) + \boldsymbol{\mu}_k^T \nabla_{\mathbf{x}} \mathbf{g}(\mathbf{x}_k) = \mathbf{0}$. On another hand, a fixed point solution $(\mathbf{x}^*, \boldsymbol{\mu}^*)$ should satisfy $\mathbf{x}^* = \arg\min_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\mu}^*)$ and $\boldsymbol{\mu}^* = [\boldsymbol{\mu}^* + \gamma \mathbf{g}(\mathbf{x}^*)]_+$. Indeed, the necessary Euler maximality condition for $\phi$ at $\boldsymbol{\mu}^*$ writes

$$\nabla_{\boldsymbol{\mu}} \phi(\mathbf{x}^*)^T (\boldsymbol{\mu} - \boldsymbol{\mu}^*) \leq 0, \quad \forall \boldsymbol{\mu} \in \mathbb{R}^p_+.$$

But since $\nabla_{\boldsymbol{\mu}} \phi(\boldsymbol{\mu}^*) = \mathbf{g}(\mathbf{x}^*)$ (Eq. 8), for $\gamma > 0$, we also get

$$(\boldsymbol{\mu}^* + \gamma \nabla \phi(\mathbf{x}^*) - \boldsymbol{\mu}^*)^T (\boldsymbol{\mu} - \boldsymbol{\mu}^*) \leq 0, \quad \forall \boldsymbol{\mu} \in \mathbb{R}^p_+.$$

Then, from the characterization of the projection on a closed convex set (here $\mathbb{R}^p_+$) [8], it appears that $\mu^* = [\boldsymbol{\mu}^* + \mathbf{g}(\mathbf{x}^*)]_+$. Then, since $\boldsymbol{\mu}_{k+1} = [\boldsymbol{\mu}_k + \gamma \mathbf{g}(\mathbf{x}_{k+1})]_+$, we get

$$\begin{array}{l} \| \ \boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}^* \ \|^2 \\ \\ \quad = \| \ [\boldsymbol{\mu}_k + \gamma \mathbf{g}(\mathbf{x}_{k+1})]_+ - [\boldsymbol{\mu}^* + \gamma \mathbf{g}(\mathbf{x}^*)]_+ \ \|^2 \\ \quad \leq \| \ \boldsymbol{\mu}_k + \gamma \mathbf{g}(\mathbf{x}_{k+1}) - \boldsymbol{\mu}^* + \gamma \mathbf{g}(\mathbf{x}^*) \ \|^2 \\ \quad \leq \| \ \boldsymbol{\mu}_k - \boldsymbol{\mu}^* + \gamma \mathbf{A}(\mathbf{x}_{k+1} - \mathbf{x}^*) \ \|^2 \\ \\ \quad \leq \| \ \boldsymbol{\mu}_k - \boldsymbol{\mu}^* \ \|^2 + \gamma^2 \ \| \ \mathbf{A} \ \|^2 \| \ \mathbf{x}_{k+1} - \mathbf{x}^* \ \|^2 \\ \qquad + 2\gamma (\boldsymbol{\mu}_k - \boldsymbol{\mu}^*)^T \mathbf{A} (\mathbf{x}_{k+1} - \mathbf{x}^*) \\ \\ \quad \leq \| \ \boldsymbol{\mu}_k - \boldsymbol{\mu}^* \ \|^2 + \gamma^2 \ \| \ \mathbf{A} \ \|^2 \| \ \mathbf{x}_{k+1} - \mathbf{x}^* \ \|^2 \\ \qquad - 2\gamma \left( \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}^*) \right)^T (\mathbf{x}_{k+1} - \mathbf{x}^*) \\ \\ \quad \leq \| \ \boldsymbol{\mu}_k - \boldsymbol{\mu}^* \ \|^2 - \gamma \left( 2\alpha - \gamma \ \| \ \mathbf{A} \ \|^2 \right) \| \ \mathbf{x}_{k+1} - \mathbf{x}^* \ \|^2 \end{array}$$

where the first inequality holds because projection onto a convex set is contractive [9], the second to last from necessary optimality conditions $\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}) + \mathbf{A}^T \boldsymbol{\mu}_k = \mathbf{0}$ and $\nabla_{\mathbf{x}} f(\mathbf{x}^*) + \mathbf{A}^T \boldsymbol{\mu}^* = \mathbf{0}$. and the last one from strong convexity characterization:

$$(\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}^*))^T (\mathbf{x}_{k+1} - \mathbf{x}^*) \geq \alpha \ \| \ \mathbf{x}_{k+1} - \mathbf{x}^* \ \|^2 .$$

Then, from hypothesis $2\alpha - \gamma \ \| \ \mathbf{A} \ \|^2 > 0$, $\| \ \boldsymbol{\mu}_k - \boldsymbol{\mu}^* \ \|$ decreases to a limit and we must have $\| \ \mathbf{x}_{k+1} - \mathbf{x}^* \ \| \to 0$. In addition, if $\mathbf{A}$ is full rank, relations $\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}) + \mathbf{A}^T \boldsymbol{\mu}_k = \mathbf{0}$ yield $\boldsymbol{\mu}_k = (\mathbf{A}\mathbf{A}^T)^{-1} \mathbf{A} \nabla_{\mathbf{x}} f(\mathbf{x}_k)$ and $\boldsymbol{\mu}_k$ converges to $\boldsymbol{\mu}^* = (\mathbf{A}\mathbf{A}^T)^{-1} \mathbf{A} \nabla_{\mathbf{x}} f(\mathbf{x}^*)$. □

# 4 Augmented Lagrangian

If we look at the Lagrangian, for a linear program $\min \mathbf{f}^T \mathbf{x}$ s.t. $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, for fixed $\boldsymbol{\mu}$ the Lagrangian

$$L(\mathbf{x}, \boldsymbol{\mu}) = (\mathbf{f}^T + \boldsymbol{\mu}^T \mathbf{A})\mathbf{x} - \boldsymbol{\mu}^T \mathbf{b}$$

has no lower bound. Thus, theorem 5 does not apply and we cannot recover $\mathbf{x}^*$ by solving the dual problem. **Augmented Lagrangian** techniques permit to combat this kind of difficulty by adding a penalty term to the objective.
In the case of equality constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ the penalty can be $\frac{c}{2} \ \| \ \mathbf{h}(\mathbf{x}) \ \|^2$ while for inequality constraints $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$ it can be taken as $\frac{c}{2} \sum_j \max(0, \mathbf{g}_j(\mathbf{x}))^2$ since condition $\mathbf{g}_j(\mathbf{x}) \leq 0$ also writes $\max(0, \mathbf{g}_j(\mathbf{x})) = 0$. Thus in what follows we simply assume equality constraints. We will come back to the case of inequality constraints at the end of the section.

## 4.1 Equality constraints

For the problem

$$(P) \left\{ \begin{array}{l} \min f(\mathbf{x}) \\ \mathbf{h}(\mathbf{x}) = 0, \end{array} \right.$$

the augmented Lagrangian method combines ideas from dual algorithms described in the previous section and penalty techniques, simply replacing the initial problem $(P)$ by problem

$$(P_c) \left\{ \begin{array}{l} \min f(\mathbf{x}) + \frac{c}{2} \ \| \ \mathbf{h}(\mathbf{x}) \ \|^2 \\ \mathbf{h}(\mathbf{x}) = 0. \end{array} \right.$$

It is clear that problems $(P)$ and $(P_c)$ share the same minima. But we are going to see that introducing the penalty term $\frac{c}{2} \ \| \ \mathbf{h}(\mathbf{x}) \ \|^2$ can significantly improve performance.
For the problem $(P)$ the augmented Lagrangian writes

$$L_c(\mathbf{x}, \lambda) = f(\mathbf{x}) + \frac{c}{2} \ \| \ \mathbf{h}(\mathbf{x}) \ \|^2 + \lambda^T \mathbf{h}(\mathbf{x})$$

with $c > 0$. In other words, $L_c$ is the Lagrangian of the problem $(P_c)$.
Note that

$$\nabla_{\mathbf{x}} (f(\mathbf{x}) + \frac{c}{2} \ \| \ \mathbf{h}(\mathbf{x}) \ \|^2) = \nabla_{\mathbf{x}} f(\mathbf{x}) + \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}) [c\mathbf{h}(\mathbf{x})].$$

Asume that $\mathbf{x}_c$ minimizes the unconstrained penalized objective $f_c(\mathbf{x}) = f(\mathbf{x}) + \frac{c}{2} \ \| \ \mathbf{h}(\mathbf{x}) \ \|^2$ and $(\mathbf{x}^*, \lambda^*)$ is the solution of $(P)$. Then,

$$\begin{array}{ll} \nabla_{\mathbf{x}} f_c(\mathbf{x}_c) & = \nabla_{\mathbf{x}} f(\mathbf{x}_c) + \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}_c) [c\mathbf{h}(\mathbf{x}_c)] \\ & = \nabla_{\mathbf{x}} f(\mathbf{x}_c) + \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}_c) \boldsymbol{\lambda}_c \\ & = 0. \end{array}$$

where $\boldsymbol{\lambda}_c = c\mathbf{h}(\mathbf{x}_c)$. As $c$ becomes large, the minimum of $f_c$ tends to come closer to set $\{\mathbf{x};\ \mathbf{h}(\mathbf{x}) = \mathbf{0}\}$. Assuming that $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is the solution of $(P)$, $\mathbf{x}_c$ approaches $\mathbf{x}^*$ while $\boldsymbol{\lambda}_c$ approaches $\boldsymbol{\lambda}^*$. An interesting aspect of problem $(P_c)$ compared to the unconstrained minimization of $f_c$ with increasing $c$ is that it can benefit from improvement of the behavior of the objective brought by the penalty for sufficiently large $c$ without having to let it tend to infinity since problems $(P)$ and $(P_c)$ are equivalent. In addition, adapting the dual algorithms considered in the previous section to augmented Lagrangian problems $(P_c)$ yields several benefits that are discussed in the next section.

## 4.2 Dual approach for augmented Lagrangian

**Convergence speed of gradient algorithms** With a view to study the benefits of dual approaches for augmented Lagrangian let us first recall the following result related to the convergence of gradient algorithms

**Theorem 7** *Let $f$ a convex function with minimum $f^*$ such that $m\mathbf{I} \leq \nabla^2 f(x) \leq M\mathbf{I}$. Then, the gradient algorithm with optimal step-size yields a sequence $(\mathbf{x}_k)_k$ such that*

$$f(\mathbf{x}_{k+1}) - f^* \leq (1 - m/M)(f(\mathbf{x}_k) - f^*).$$

**Proof** Note first that Taylor-Lagrange formula yields

$$\begin{array}{ll} f(\mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)) \leq & f(\mathbf{x}_k) - \alpha \ \| \ \nabla f(\mathbf{x}_k) \ \|^2 \\ & + \frac{\alpha^2 M}{2} \ \| \ \nabla f(\mathbf{x}_k) \ \|^2 . \end{array}$$

The minimum of the right hand side is reached for $\alpha = 1/M$, leading to $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{1}{2M} \ \| \ \nabla f(\mathbf{x}_k) \ \|^2$. Then,

$$f(\mathbf{x}_{k+1}) - f^* \leq f(\mathbf{x}_k) - f^* - \frac{1}{2M} \ \| \ \nabla f(\mathbf{x}_k) \ \|^2 . \tag{11}$$

We also have

$$f(\mathbf{x}) \geq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x}-\mathbf{x}_k) + \frac{m}{2} \parallel \mathbf{x}-\mathbf{x}_k \parallel^2 .$$

and the minimum of the right hand side is obtained for $\mathbf{x} = \mathbf{x}_k - \frac{1}{m}\nabla f(\mathbf{x}_k)$. thus,

$$f^* \geq f(\mathbf{x}_k) - \frac{1}{2m} \parallel \nabla f(\mathbf{x}_k) \parallel^2,$$

leading to $- \parallel \nabla f(\mathbf{x}_k) \parallel^2 \leq 2m(f^* - f(\mathbf{x}_k))$. Inserting this inequality in Eq. (11) yields

$$f(\mathbf{x}_{k+1}) - f^* \leq (1 - \frac{m}{M})(f(\mathbf{x}_k) - f^*). \tag{12}$$

□

This result shows that the convergence speed of gradient descent algorithms is related to the condition number $K = M/m$ ($K \geq 1$) of the Hessian matrix of the objective $f$: the lower $K$ the better for convergence speed.

**Hessian matrices of Lagrangians and dual functions** Now, we are going to calculate the Hessian matrices of Lagrangians $L$ and $L_c$ and of the dual functions $\phi(\boldsymbol{\lambda})$ and $\phi_c(\boldsymbol{\lambda})$ related to problems $(P)$ and $(P_c)$ respectively.
Letting $\mathbf{x}_{\boldsymbol{\lambda}} = \arg\min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda})$, we have the following equations:

$$\begin{array}{rcl} \nabla_{\mathbf{x}} L(\mathbf{x}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) & = & \nabla_{\mathbf{x}} f(\mathbf{x}_{\boldsymbol{\lambda}}) + \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}})\boldsymbol{\lambda} \\ & = & 0 \\ \nabla_{\boldsymbol{\lambda}}[\nabla_{\mathbf{x}} L(\mathbf{x}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda})] & = & \nabla_{\boldsymbol{\lambda}}\mathbf{x}_{\boldsymbol{\lambda}} \left(\nabla^2_{\mathbf{x}} f(\mathbf{x}_{\boldsymbol{\lambda}}) + \nabla^2_{\mathbf{x}} \mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}})\boldsymbol{\lambda}\right) \\ & & +\nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}})^T \\ & = & \nabla_{\boldsymbol{\lambda}}\mathbf{x}_{\boldsymbol{\lambda}} \nabla^2_{\mathbf{x}} L(\mathbf{x}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}) + \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}})^T \\ & = & 0. \end{array} \tag{13}$$

On another hand, as discussed at the end of section 2.5, $\nabla_{\boldsymbol{\lambda}}\phi(\boldsymbol{\lambda}) = \mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}})$ and

$$\nabla^2_{\boldsymbol{\lambda}}\phi(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}}\mathbf{x}_{\boldsymbol{\lambda}} \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}}) \tag{14}$$

The last equation of (13) yields

$$\nabla_{\boldsymbol{\lambda}}\mathbf{x}_{\boldsymbol{\lambda}} = -\nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}})^T [\nabla_{\mathbf{x}} L(\mathbf{x}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda})]^{-1}, \tag{15}$$

and replacing (15) inside (14), we get

$$\nabla^2_{\boldsymbol{\lambda}}\phi(\boldsymbol{\lambda}) = -\nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}})^T [\nabla_{\mathbf{x}} L(\mathbf{x}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda})]^{-1} \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}}). \tag{16}$$

Now, let us consider the same calculations for $\phi_c$: omitting variables to simplify notations, at $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ we get

$$\begin{array}{rcl} \nabla_{\mathbf{x}} L_c & = & \nabla_{\mathbf{x}} L + c\nabla_{\mathbf{x}} \mathbf{h} \times \mathbf{h} \\ & = & 0 \\ \nabla_{\boldsymbol{\lambda}}\nabla_{\mathbf{x}} L_c & = & \nabla_{\boldsymbol{\lambda}}\mathbf{x} \nabla^2_{\mathbf{x}} L + \nabla_{\mathbf{x}} \mathbf{h}^T \\ & & +c\nabla_{\boldsymbol{\lambda}}\mathbf{x} \left(\nabla^2_{\mathbf{x}} \mathbf{h} \times \mathbf{h} + \nabla_{\mathbf{x}} \mathbf{h} \nabla_{\mathbf{x}} \mathbf{h}^T\right) \\ & = & \nabla_{\boldsymbol{\lambda}}\mathbf{x} \nabla^2_{\mathbf{x}} L_c + \nabla_{\mathbf{x}} \mathbf{h}^T \\ & = & 0, \end{array} \tag{17}$$

where

$$\nabla^2_{\mathbf{x}} L_c = \nabla^2_{\mathbf{x}} L + c\nabla_{\mathbf{x}} \mathbf{h} \nabla_{\mathbf{x}} \mathbf{h}^T + c\nabla^2_{\mathbf{x}} \mathbf{h} \times \mathbf{h}. \tag{18}$$

Then putting all pieces together just like for the study of $\nabla^2_{\boldsymbol{\lambda}}\phi$, we get

$$\nabla^2_{\boldsymbol{\lambda}}\phi_c(\boldsymbol{\lambda}) = -\nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}})^T [\nabla_{\mathbf{x}} L_c(\mathbf{x}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda})]^{-1} \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}}). \tag{19}$$

Note that for $\mathbf{x}_{\boldsymbol{\lambda}}$ feasible, $\mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}}) = \mathbf{0}$ and the last term of (18) cancels: $\nabla^2_{\mathbf{x}} L_c = \nabla^2_{\mathbf{x}} L + c\nabla_{\mathbf{x}} \mathbf{h} \nabla_{\mathbf{x}} \mathbf{h}^T$. Close to the optimum, $\mathbf{x}_{\boldsymbol{\lambda}} \approx \mathbf{x}_{\boldsymbol{\lambda}^*} = \mathbf{x}^*$ and $\mathbf{h}(\mathbf{x}_{\lambda}) \approx 0$, leading to

$$\nabla^2_{\boldsymbol{\lambda}}\phi_c(\boldsymbol{\lambda}) \approx -\nabla_{\mathbf{x}} \mathbf{h}^T [\nabla^2_{\mathbf{x}} L + c\nabla_{\mathbf{x}} \mathbf{h} \nabla_{\mathbf{x}} \mathbf{h}^T]^{-1} \nabla_{\mathbf{x}} \mathbf{h}, \tag{20}$$

with equality when $\mathbf{h}(\mathbf{x}_{\boldsymbol{\lambda}}) = 0$, and in particular at $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$.

**Local strict convexity of the augmented Lagrangian** When second order minimality conditions are satisfied for problem $(P)$ at $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$, it can be shown that for sufficiently large $c$ the augmented Lagrangian $L_c$ is strictly convex locally, w.r.t. $\mathbf{x}$:

**Theorem 8** *Assume $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ satisfies second order sufficient optimality conditions for problem $(P)$. Then, there exists $c^*$ such that for $c > c^*$, the function $\mathbf{x} \rightarrow L_c(\mathbf{x}, \boldsymbol{\lambda}^*)$ has a strict local minimum at $\mathbf{x}^*$.*

**Proof** First, let us show that there exists $c^* > 0$ such that $c > c^* \Rightarrow \nabla^2_{\mathbf{x}} L_c(\mathbf{x}^*, \boldsymbol{\lambda}^*) \succ \mathbf{0}$. Indeed, if this was untrue, there would exist a sequence $(\mathbf{x}_n)_{n\geq 0}$ with $\parallel \mathbf{x}_n \parallel = 1$ such that $\mathbf{x}_n^T \nabla^2_{\mathbf{x}} L_c(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{x}_n \leq 0$, that is,

$$\mathbf{x}_n^T \nabla^2_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{x}_n + n \parallel \nabla \mathbf{h}(\mathbf{x}^*)\mathbf{x}_n \parallel^2 \leq 0.$$

$\mathbf{x}_n$ admits a converging subsequence with limit $\mathbf{x}$ such that $\parallel \nabla \mathbf{h}(\mathbf{x}^*)\mathbf{x}_n \parallel^2 = 0$ since $\mathbf{x}_n^T \nabla^2_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}^*)\mathbf{x}_n$ is bounded. Thus, $\mathbf{x}$ belongs to the tangent space at $\mathbf{x}^*$. But then we should also have $\mathbf{x}^T \nabla^2_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{x} \leq 0$ what would be in contradiction with optimality second order sufficient conditions.
Thus $\nabla^2_{\mathbf{x}} L_c(\mathbf{x}^*, \boldsymbol{\lambda}^*) \succ \mathbf{0}$ for $c$ large enough. In addition, $\nabla_{\mathbf{x}} L_c(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) + c\nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}^*) \times \mathbf{h}(\mathbf{x}^*) = 0$ since $\mathbf{h}(\mathbf{x}^*) = 0$. Thus second order sufficient conditions are also satisfied for $(P_c)$ and $\mathbf{x} \rightarrow L_c(\mathbf{x}, \boldsymbol{\lambda}^*)$ has a strict local minimum at $\mathbf{x}^*$. □

The proof of theorem 8 shows that second order necessary conditions yield strict convexity property $\nabla^2 L_c \succ \mathbf{0}$ at $(\mathbf{x}^*, \boldsymbol{\mu}^*)$ when $c$ is large enough. Slightly moving about $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ maintains the strict convexity of $L_c$ what guarantees that the Lagrangian can be minimized, hence the existence of the dual function.
**Example** For the problem $(P)$ : $\min f(x, y) = x^2 + y$ s.t. $y = 0$, the dual cannot be calculated while it can be checked that for any $c > 0$, $\nabla^2 L_c \succ \mathbf{0}$ and the dual of $(P_c)$ is well defined and dual optimization techniques can be considered.

**Behavior of the dual Hessian** After considering the Hessian of the Lagrangian, we now study the eigenvalues of the Hessian of the dual function for the augmented problem $(P_c)$. Recall that at $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ we have

$$\nabla^2_{\boldsymbol{\lambda}}\phi_c = -\nabla_{\mathbf{x}} \mathbf{h}^T [\nabla^2_{\mathbf{x}} L + c\nabla_{\mathbf{x}} \mathbf{h} \nabla_{\mathbf{x}} \mathbf{h}^T]^{-1} \nabla_{\mathbf{x}} \mathbf{h}.$$

Letting $\mathbf{A} = \nabla^2_{\mathbf{x}} L$, $\mathbf{B} = \nabla_{\mathbf{x}} \mathbf{h}$ and $\mathbf{M} = -\nabla^2_{\boldsymbol{\lambda}}\phi = \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B}$, Woodburry formula (a.k.a. the matrix inversion lemma) [10] yields

$$\begin{array}{l} -\nabla^2_{\boldsymbol{\lambda}}\phi_c \\ \quad = \mathbf{B}^T(\mathbf{A} + c\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B} \\ \quad = \mathbf{B}^T \left[\mathbf{A}^{-1} - c\mathbf{A}^{-1}\mathbf{B}(\mathbf{I} + c\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}\right] \mathbf{B} \\ \quad = \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B} \\ \qquad -c\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}(\mathbf{I} + c\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B} \\ \quad = \mathbf{M} - c\mathbf{M}(\mathbf{I} + c\mathbf{M})^{-1}\mathbf{M} \\ \quad = \mathbf{M}(\mathbf{I} + c\mathbf{M})^{-1}(\mathbf{I} + c\mathbf{M} - c\mathbf{M}) \\ \quad = \mathbf{M}(\mathbf{I} + c\mathbf{M})^{-1} \\ \quad = \mathbf{U}\mathbf{D}(\mathbf{I} + c\mathbf{D})^{-1}\mathbf{U}^T \end{array}$$

where $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ denotes the eigen-decomposition of $\mathbf{M}$. Then, if $\mathbf{u}$ is an eigenvector of $\mathbf{M} = -\nabla^2_{\boldsymbol{\lambda}}\phi$ with corresponding eigenvalue $\alpha$, it is also an eigenvector of $-\nabla^2_{\boldsymbol{\lambda}}\phi_c$ with corresponding eigenvalue

$$\alpha_c = \frac{\alpha}{1 + c\alpha} = (\frac{1}{\alpha} + c)^{-1}$$

Letting $\alpha_m$ and $\alpha_M$ denote the smallest and largest eigenvalues of $-\nabla^2_{\boldsymbol{\lambda}}\phi$, the condition number of $-\nabla^2_{\boldsymbol{\lambda}}\phi_c$ is given by

$$K_c = \frac{\frac{1}{\alpha_m} + c}{\frac{1}{\alpha_M} + c}.$$

$K_c$ decreases from $\alpha_M/\alpha_m$ to 1 as $c$ grows. This makes the steepest ascent for the maximization of $\phi_c$ interesting due to the faster convergence when the condition number is close to 1, as shown in theorem 7 about convergence speed of gradient algorithms.

Now, note that for large $c$ the eigenvalues of $-\nabla^2_{\boldsymbol{\lambda}}\phi_c$ are close to $c^{-1}$ so that $\nabla^2_{\boldsymbol{\lambda}}\phi_c \approx -c^{-1}\mathbf{I}$. Thus, updating $\lambda$ via Newton algorithm for $\phi_c$ maximization simply writes

$$\lambda_{k+1} = \lambda_k + c\mathbf{h}(\mathbf{x}_{\lambda_k}). \tag{21}$$

### 4.3 Augmented Lagrangian with equality and inequality constraints

In order to complete the discussion about the augmented Lagrangian approach, we now briefly discuss one popular formulation for the standard problem (1) with equality and inequality constraints. Clearly, such a problem can be written in the form

$$\begin{cases} \min f(\mathbf{x}) \\ \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ \mathbf{g}(\mathbf{x}) + \mathbf{z}^2 = \mathbf{0}, \end{cases} \tag{22}$$

with $\mathbf{z}^2 = [\mathbf{z}_1^2, \ldots, \mathbf{z}_p^2]^T$. Then, the augmented Lagrangian writes

$$\begin{aligned} L_c(\mathbf{x}, \mathbf{z}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \quad & f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T(\mathbf{g}(\mathbf{x}) + \mathbf{z}^2) \\ & + \tfrac{c}{2}\left(\| \mathbf{h}(\mathbf{x}) \|^2 + \| \mathbf{g}(\mathbf{x}) + \mathbf{z}^2 \|^2\right). \end{aligned} \tag{23}$$

Then, letting $\mathbf{z}_i^2 = \mathbf{s}_i$, the dual writes

$$\begin{aligned} \phi(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \quad & \min_{\mathbf{x}, \mathbf{s} \geq 0} f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T(\mathbf{g}(\mathbf{x}) + \mathbf{s}) \\ & + \tfrac{c}{2}\left(\| \mathbf{h}(\mathbf{x}) \|^2 + \| \mathbf{g}(\mathbf{x}) + \mathbf{s} \|^2\right). \end{aligned} \tag{24}$$

Note that $L_c$ depends on $\mathbf{s}_i$ only via the term

$$\xi(\mathbf{s}_i) = (\boldsymbol{\mu}_i + c\mathbf{g}_i(\mathbf{x}))\mathbf{s}_i + (c/2)\mathbf{s}_i^2.$$

and the minimum of $\xi(s)$ is achieved for $s = -\mathbf{g}_i(\mathbf{x}) - \boldsymbol{\mu}_i/c$. Then, accounting for the positivity of $\mathbf{s}_i$, we get

$$\mathbf{s}_i = \max[0, -\mathbf{g}_i(\mathbf{x}) - \boldsymbol{\mu}_i/c].$$

So, if $\mathbf{s}_i = 0$, $\boldsymbol{\mu}_i(g_i(\mathbf{x}) + \mathbf{s}_i) + \frac{c}{2}(\mathbf{g}_i(\mathbf{x}) + \mathbf{s}_i)^2$ is equal to

$$\boldsymbol{\mu}_i\mathbf{g}_i(\mathbf{x}) + \frac{c}{2}\mathbf{g}_i(\mathbf{x})^2 = \frac{1}{2c}[(c\mathbf{g}_i(\mathbf{x}) + \boldsymbol{\mu}_i)^2 - \mu_i^2],$$

and if $\mathbf{s}_i > 0$, it is equal to $\frac{-\boldsymbol{\mu}_i^2}{2c}$. Then, letting

$$\psi_c(v, \mu) = \frac{1}{2c}[\max(0, \mu + cv)^2 - \mu^2], \tag{25}$$

we have

$$\boldsymbol{\mu}_i(g_i(\mathbf{x}) + \mathbf{s}_i) + \frac{c}{2}(\mathbf{g}_i(\mathbf{x}) + \mathbf{s}_i)^2 = \psi_c(\mathbf{g}_i(\mathbf{x})\boldsymbol{\mu}_i)$$

and after minimization w.r.t. $\mathbf{z}$, $L_c$ finally writes

$$\begin{aligned} L_c(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \quad & f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{h}(\mathbf{x}) + \tfrac{c}{2} \| \mathbf{h}(\mathbf{x}) \|^2 \\ & + \textstyle\sum_i \psi_c(\mathbf{g}_i(\mathbf{x}), \boldsymbol{\mu}_i). \end{aligned} \tag{26}$$

Considering again the definition (23) of $L_c$, we get

$$\begin{aligned} \nabla_{\mathbf{x}} L_c(\mathbf{x}, \mathbf{z}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \nabla_{\mathbf{x}} f(\mathbf{x}) + \nabla_{\mathbf{x}}\mathbf{h}(\mathbf{x})\boldsymbol{\lambda} + \mathbf{g}(\mathbf{x})\boldsymbol{\mu} \\ + c\left([\nabla_{\mathbf{x}}\mathbf{h}(\mathbf{x})]\mathbf{h}(\mathbf{x}) + [\nabla_{\mathbf{x}}\mathbf{g}(\mathbf{x})](\mathbf{g}(\mathbf{x}) + \mathbf{z}^2)\right). \end{aligned} \tag{27}$$

Then, solutions of necessary optimality conditions for problem (22) must satisfy $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, $\mathbf{g}(\mathbf{x}) + \mathbf{z}^2 = \mathbf{0}$, and

$$\nabla_{\mathbf{x}} f(\mathbf{x}) + \nabla_{\mathbf{x}}\mathbf{h}(\mathbf{x})\boldsymbol{\lambda} + \mathbf{g}(\mathbf{x})\boldsymbol{\mu} = \mathbf{0}.$$

For such a solution, $(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is also solution of problem (1) if we add the constraint $\boldsymbol{\mu} \in \mathbb{R}^p_+$. This constraint is also necessary to be able to apply second order sufficient conditions in the presence of inequality constraints and extend theorem (8) to this situation.
To express the gradient ascent iterations for $\boldsymbol{\mu}$, let us remark that

$$\begin{aligned} \frac{\partial L_c}{\partial \boldsymbol{\mu}_i} &= \frac{\partial}{\partial \boldsymbol{\mu}_i}\psi_c(\mathbf{g}_i(\mathbf{x}), \boldsymbol{\mu}_i) \\ &= \tfrac{1}{2c}[\mathbb{1}_{\{\boldsymbol{\mu}_i + c\mathbf{g}_i \geq 0\}} \frac{\partial}{\partial \boldsymbol{\mu}_i}((\boldsymbol{\mu}_i + c\mathbf{g}_i(\mathbf{x}))^2 - \boldsymbol{\mu}_i^2) \\ &\qquad + \mathbb{1}_{\{\boldsymbol{\mu}_i + c\mathbf{g}_i < 0\}} \frac{\partial}{\partial \boldsymbol{\mu}_i}(-\boldsymbol{\mu}_i^2)] \\ &= \mathbb{1}_{\{\boldsymbol{\mu}_i + c\mathbf{g}_i \geq 0\}}\mathbf{g}_i(\mathbf{x}) - \mathbb{1}_{\{\boldsymbol{\mu}_i + c\mathbf{g}_i < 0\}}(\boldsymbol{\mu}_i/c) \end{aligned} \tag{28}$$

Then, from the discussion about the behavior of the Hessian of the dual function as $c$ grows, it is clear that at a minimum of (1), for active constraints the curvature of the dual function is close to $-c$. Then the following Newton step $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + c\nabla_{\boldsymbol{\mu}} L_c$writes

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + c[\mathbb{1}_{\{\boldsymbol{\mu}_i + c\mathbf{g}_i \geq 0\}}\mathbf{g}_i(\mathbf{x}) - \mathbb{1}_{\{\boldsymbol{\mu}_i + c\mathbf{g}_i < 0\}}(\boldsymbol{\mu}_i/c)],$$

that is, $\boldsymbol{\mu} \leftarrow [\boldsymbol{\mu} + c\mathbf{g}]_+$, a form that clearly accounts for positivity constraints $\boldsymbol{\mu} \geq \mathbf{0}$. In particular, it is clear that in the neighborhood of a solution, $\boldsymbol{\mu}_i$ remains equal to 0 for an inactive constraint $\mathbf{g}_i(\mathbf{x}) < 0$.
Finally, the corresponding complete augmented Lagrangian method is given in algorithm 2

**Algorithm 2** Augmented Lagrangian algorithm

**define**

$$\begin{aligned} &\psi_c(v, \mu) \quad = \tfrac{1}{2c}[\max(0, \mu + cv)^2 - \mu^2], \\ &L_c(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \\ &f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{h}(\mathbf{x}) + \tfrac{c}{2} \| \mathbf{h}(\mathbf{x}) \|^2 + \textstyle\sum_i \psi_c(\mathbf{g}_i(\mathbf{x}), \boldsymbol{\mu}_i). \end{aligned}$$

**init** $\mathbf{x}_0$, $\boldsymbol{\lambda}_0$, $\boldsymbol{\mu}_0$, $c > 0$
**while** $c < c_{\max}$ **do**
  **while** stopping condition $\neq$ true **do**
    $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} L_c(\mathbf{x}, \boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$
    or

$$\begin{cases} \alpha_k & \approx \arg\min_r L_c\left(\mathbf{x}_k - r\nabla_{\mathbf{x}} L(\mathbf{x}_k, \boldsymbol{\mu}_k)\right) \\ \mathbf{x}_{k+1} & = \mathbf{x}_k - \alpha_k \nabla_{\mathbf{x}} L(\mathbf{x}_k, \boldsymbol{\mu}_k) \end{cases}$$

    $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + c\mathbf{h}(\mathbf{x}_{k+1})$
    $\boldsymbol{\mu}_{k+1} = [\boldsymbol{\mu}_k + c\mathbf{g}(\mathbf{x}_{k+1})]_+$
  **end while**
  $c = \rho c$
**end while**

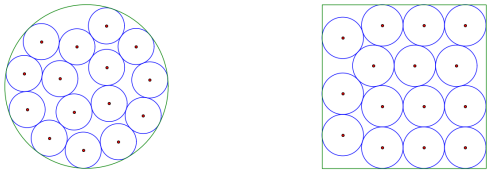
## 5 Example: sphere packing

Let us consider the disk packing problem that consists in packing $n$ disks of radius $r = 1$ within a minimum area. Here we assume in addition some shape constraint for the

enclosing surface. For instance, assuming a circular boundary with radius $z$ for the enclosing surface, we want to solve

$$(P)\begin{cases} \min_{\mathbf{x},\mathbf{y},z,\mathbf{p}} z \\ (\mathbf{x}_i-\mathbf{x}_j)^2+(\mathbf{y}_i-\mathbf{y}_j)^2>(2r)^2 \quad 1\le i<j\le n \\ (\mathbf{x}_i-\mathbf{p}_x)^2+(\mathbf{y}_i-\mathbf{p}_y)^2\le(z-r)^2 \quad 1\le i\le n \end{cases}$$

The objective $z$ is the radius that must be minimized, $\mathbf{x}$ and $\mathbf{y}$ are the vectors of $x$ and $y$ coordinates for the centers of the disks, The first set of constraints, $(\mathbf{x}_i-\mathbf{x}_j)^2+(\mathbf{y}_i-\mathbf{y}_j)^2>(2r)^2$, are the non overlap constraints for disks. Constraints $(\mathbf{x}_i-\mathbf{p}_x)^2+(\mathbf{y}_i-\mathbf{p}_y)^2\le(z-r)^2$ represent circular boundary constraints, where $\mathbf{p}$ denotes the center of the enclosing circle. Note that the problem is not convex and using convex optimization approaches can lead to local minima. Letting the center $\mathbf{p}$ vary instead of setting for instance $\mathbf{p}=\mathbf{0}$ supplies additional degrees of freedom to the problem that can be helpful to avoid some local minima.
Figure 1 shows examples of this approach for different boundary conditions using augmented Lagrangian.



(a) (b)

Figure 1: Packing 15 spheres (a) in a disk and (b) in a box.

## 6 Exercise: SVM and duality

In this exercise, we recall first some elements of SVM classification. For additional information about SVM principles see for instance the corresponding Wikipedia page.
SVM techniques rely on two main ideas: classify data by (i) searching for a separation boundary between two classes such that labelled data [11] exhibit largest minimum distance to the boundary and (ii) transform data into a space of larger, possibly infinite, dimension where linear separation of classes works better (but then in the original classes separation becomes nonlinear).
When possible, linear separation between 2 classes, say $c\in\{-1,1\}$, can simply be achieved by looking for an hyperplane with parameters $(\mathbf{w},b)$ such that in the learning set $\{(\mathbf{x_k},c_k)\}_{k=1:K}$ elements satisfy $c_k(\mathbf{w}^T\mathbf{x_k}-b)\ge 0$ . We can also search for two hyperplanes such that $\mathbf{w}^T\mathbf{x}-b\le -1$ for class -1 and $\mathbf{w}^T\mathbf{x}-b\ge 1$ for class 1. We would like that they represent limiting hyperplanes for both classes with maximum separation. Then, $c_k(\mathbf{w}^T\mathbf{x_k}-b)\ge 1$ with equality for at least one point in each class.
1) - Assuming the hyperplanes $\mathbf{w}^T\mathbf{x_k}-b\le -1$ and $\mathbf{w}^T\mathbf{x_k}-b\ge 1$ are limiting hyperplanes for classes $c=-1$ and $c=1$ respectively, prove that the maximum distance beween these class limiting hyperplanes is $2\parallel\mathbf{w}\parallel^{-1}$, that is,

$$\max_{c_k=1,c_l=-1}\parallel\mathbf{x_k}-\mathbf{x_l}\parallel^2=\frac{2}{\parallel\mathbf{w}\parallel}.$$

Then, check that we can get $(\mathbf{w},b)$ by solving

$$\begin{cases} \min_{\mathbf{w},b}\frac{1}{2}\parallel\mathbf{w}\parallel^2 \\ c_k(\mathbf{w}^T\mathbf{x_k}-b)\ge 1, \quad k=1:K. \end{cases}$$

This way of determining $\mathbf{w}$ enables maximum margin between class limiting hyperplanes $\mathbf{w}^T\mathbf{x_k}-b\le -1$ and $\mathbf{w}^T\mathbf{x_k}-b\ge 1$.

2) - When both classes are not exactly linearly separable, a possible linear approximation is obtained by weighting the objective $\parallel\mathbf{w}\parallel^2$ by the sum of distances to the class boundary for wrongly classified data. Explain how this can be expressed by the following problem by discussing the meaning of variables $\zeta_k$ and parameter $\lambda$:

$$\begin{cases} \min_{\mathbf{w},b}\frac{1}{2}\parallel\mathbf{w}\parallel^2+\lambda\sum_{k=1:K}\zeta_k \\ c_k(\mathbf{w}^T\mathbf{x_k}-b)\ge 1-\zeta_k, \quad \zeta_k\ge 0, \quad k=1:K. \end{cases}$$

3) Write the Lagrangian of the problem and express KKT conditions.
4) From KKT conditions, show that the dual of the problem writes

$$\begin{cases} \max_{\boldsymbol{\alpha}} f(\alpha)=\sum_{1:K}\alpha_k-\frac{1}{2}\sum_{i,j=1:K}\alpha_i\alpha_j c_i c_j(\mathbf{x_j}^T\mathbf{x_i}) \\ \sum_{k=1:K}\alpha_k c_k=0 \\ 0\le\alpha_k\le\lambda, \quad k=1:K. \end{cases}$$

What do variables $(\alpha_k)_{k=1:K}$ represent ?
Note that, just as for the primal problem, this is a standard quadratic optimization problem. However, it appears that the dual formulation presents the benefit of decoupled variables in the formulation of inequality constraints, what can help algorithms to enforce these constraints. Indeed, for instance we can combine a penalty approach to handle the equality constraint and projection on the cube satisfying inequality constraints. Then, the gradient is calculated on the penalized convex criterion $f_t(\boldsymbol{\alpha})=-f(\boldsymbol{\alpha})+t(\boldsymbol{\alpha}^T\mathbf{c})^2$ and the updated $\boldsymbol{\alpha}$ is projected:

$$\boldsymbol{\alpha}^{(n+1)}=\Pi_{[0,\lambda]^K}\left(\boldsymbol{\alpha}^{(n)}-\mu_n\nabla f_t(\boldsymbol{\alpha}^{(n)})\right),$$

where $\Pi_{[0,\lambda]^K}$ represents the projection onto the cube $[0,\lambda]^K$ of $\mathbb{R}^K$. $t$ is slightly increased at each iteration ( $t\to\beta t$, with e.g. $\beta=1.1$).
5) Check that the solution of the primal for $\mathbf{w}$ is then given by $\mathbf{w}=\sum_{\mathbf{k}}\alpha_k c_k\mathbf{x_k}$. To find $b$, check first that if $0<\alpha_k<\lambda$ then $\mathbf{x_k}$ lies on the margin. Then , $c_k(\mathbf{w}^T\mathbf{x_k}-b)=1$, that is, $b=\mathbf{w}^T\mathbf{x_k}-c_k$.
Finally, the decision variable writes $sign((\mathbf{w}^T\mathbf{x}-b))=sign(\sum_k\alpha_k c_k\mathbf{x_k}^T\mathbf{x}-b)$ .

In order to improve classes separation, we now consider possibly nonlinear separation. A nice way to achieve nonlinear separation of classes is to apply a nonlinear transform of $\mathbf{x_k}$ into a space of larger (possibly infinite) dimension, say $\mathbf{x_k}\to\varphi(\mathbf{x_k})$ and consider the linear separation of transformed variables $\varphi(\mathbf{x_k})$. Note that the decision variable is now in the form $sign(\sum_k a_k c_k\varphi(\mathbf{x_k})^T\varphi(\mathbf{x})-b)=sign(\sum_k a_k c_k k(\mathbf{x_k},\mathbf{x})-b)$ and $\varphi$ needs not to be known: only the bilinear function $k(.,.)$ must be set. In fact $k(\mathbf{x},\mathbf{x}')$ represents a scalar product $\varphi(\mathbf{x}')^T\varphi(\mathbf{x})$ if and only if it is positive definite. This result is known as Mercer's theorem. The Gaussian kernel $k(\mathbf{x},\mathbf{x}')=\exp(-\parallel\mathbf{x}-\mathbf{x}'\parallel^2/h)$ $(h>0)$ is a classical exemple of such a similarity function. The dual of the optimisation problem for data $(\varphi(\mathbf{x_k}),c_k)_k$ directly formulates in terms of $k(.,.)$ and does not require $\varphi$.
With all the calculations above you should be able to implement the dual approach for SVM proposed for the laboratory session.

## References


[1] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
https://web.stanford.edu/ boyd/cvxbook/

[2] D.G. Luenberger, Y. Ye, *Linear and Nonlinear Programming*, 3rd Ed., Springer, 2008.

[3] P.G. Ciarlet, *Introduction à l'Analyse Numérique Matricielle et à l'Optimisation*, Masson, 1982.

[4] Wikipedia: Max–min inequality
https://en.wikipedia.org/wiki/Max-min_inequality

[5] T. Chonavel, *Notes on unconstrained and constrained optimization algorithms*, IMT Atlantique, TAF MCE, UE NUMMET, 2021
https://moodle.imt-atlantique.fr/mod/resource/view.php?id=43984

[6] Wikipedia: *Strong convexity*
https://en.wikipedia.org/wiki/Convex_function#Strongly_convex_functions

[7] Wikipedia: *Operator norm*
https://en.wikipedia.org/wiki/Operator_norm

[8] Stackexchange: *projection on closed convex sets*
https://math.stackexchange.com/questions/2900841/orthogonal-projection-of-a-vector-onto-convex-set

[9] Stackexchange: *projection on closed convex sets are contractive*
https://math.stackexchange.com/questions/3809431/geometric-proof-that-projections-on-convex-sets-are-contractive https://jump.dev/JuMP.jl/stable

[10] Wikipedia: *Woodbury matrix identity*
https://en.wikipedia.org/wiki/Woodbury_matrix_identity

[11] SVM, https://en.wikipedia.org/wiki/Support_vector_machine