Seminar: Overview and Applications of the KKT Conditions And Duality

Luca Solbiati

October 2022

Abstract

In the field of constrained nonlinear optimization the KKT conditions and the theory of duality have played a central role in the analysis and development of algorithms and modeling techniques.

While even a naive understanding of these concepts is very useful and practical, a deep knowledge of these results can aid significantly in the proper application of algorithms and models, delivering meaningful insights into the field of optimization and operation reasearch in general.

The purpose of this seminar is to give an extensive and rigorous presentation of these results, together with examples, intuitions, and applications that motivate their study.

1 Introduction

1.1 The nonlinear constrained optimization problem

In this seminar we shall consider the constrained optimization problem:

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{subject to} \quad \begin{cases} c_i(x) = 0 & i \in \mathcal{E} \\ c_i(x) \geq 0 & i \in \mathcal{I} \end{cases} \tag{1}$$

This kind of problem is found in many practical applications, such as finance, engineering and machine learning. In these fields there is often a need to minimize (or maximize) a quantity, described by the objective function f, that depends on certain characteristics of the system, called variables (x).
Often these variables are subject to constraints (c_i), that limit their possible configurations. These constraints arise naturally in many problems. Consider for example manufacturing, where one cannot have a negative or nearly infinite production of something, hence we must impose $0 \leq x_i \leq b \in \mathbb{R}^+$.

We may distinguish between two kinds of constraints:

1. *Equality constraints*: that are indexed in problem (1) by $i \in \mathcal{E}$
2. *Inequality constraints*: that are indexed in problem (1) by $i \in \mathcal{I}$

By defining the feasible set Ω as the set of points that satisfy all the constraints:

$$\Omega := \{x \in \mathbb{R}^n \mid c_i(x) = 0 \; \forall \; i \in \mathcal{E} \quad c_i(x) \geq 0 \; \forall \; i \in \mathcal{I}\} \tag{2}$$

We can rewrite problem (1) as:

$$\min_{x\in\Omega} f(x) \tag{3}$$

which is known as the geometric formulation of the problem. We may observe that the geometric formulation is unique, while the algebraic formulation (1) in general is not.
The goal of this seminar is to discuss the necessary conditions that must be satisfied by solutions of problem (1). These necessary conditions are not only used to check if a given point is a (local) optimizer, but are also used to aid in the development of optimization algorithms and in the modelling of optimization problems, through the study of sensitivity and dual formulations.

1.2 Local solutions and smoothness

For many nonlinear optimization problems there is little hope to find the actual minimum of (1), also known as global solution. What one could reasonably hope for is instead to find a local solution of the problem, that we define as follows:

Definition 1.1 *A vector $x^* \in \Omega$ is a local solution of problem (1) if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for all $x^* \neq x \in \mathcal{N} \cap \Omega$. If such inequality is strict we shall say that x^* is a strict local solution.*

In order to characterize local solutions we shall find a set of necessary conditions that they need to satisfy. These conditions, known as KKT conditions, will rely on the smoothness of the objective function and of the constraints, in particular we assume throughout this seminar that they are $\mathcal{C}^1$ functions. Smoothness is also required by many optimization algorithms, as it ensures that the objective function and the constraints all behave in a reasonably predictable way, therefore allowing for algorithms to make good choices for search directions.
We may observe that smoothness in the constraints does not necessarily produce a smooth feasible region (in the sense of manifolds). For example, even simple linear constraints can produce a diamond-shaped feasible region with sharp edges.
This also suggests that we can transform apparently nonsmooth optimization problems into smooth optimization problems by rewriting their algebraic formulations. For example consider the nonsmooth constraint:

$$|x_1| + |x_2| \leq 1$$

This can be rewritten into a smooth constraint as:

$$x_1 + x_2 \leq 1, \quad x_1 - x_2 \leq 1, \quad -x_1 + x_2 \leq 1, \quad -x_1 - x_2 \leq 1$$

Similarly, nonsmooth optimization problems can sometimes be reformulated as smooth constrained problems by introducing slack variables. Consider for example:

$$\min_{x \in \mathbb{R}} f(x) = \max(x^2, x)$$

By adding an artificial variable t we can rewrite this problem as:

$$\min_{t \in \mathbb{R}} t \quad \text{s.t} \quad t \geq x, \quad t \geq x^2$$

These reformulation techniques are often used in cases where f is the maximum of a collection of functions or f is a 1-norm or ∞-norm of a vector function.

2 Active Set And The Lagrangian Function

2.1 Active set definition

We introduce in this section two very important concepts for the characterization of optimal points.
We may notice that, by smoothness of the constraint functions, if an inequality constraint c_i is such that $c_i(x^*) > 0$ then that constraint is actually not playing any part in the characterization of the local minimum. Indeed, by continuity we can always find a neighbourhood of x^* such that any x in that neighbourhood has $c_i(x) > 0$. In this sense the constraint is *inactive*, because whether or not it was present x^* would still be a local minimum.
For this reason we define the following set:

Definition 2.1 *The active set $\mathcal{A}(x)$ at any feasible point $x \in \Omega$ is the set of indices of active constraints at x, i.e. :*

$$\mathcal{A}(x) := \mathcal{E} \cup \{i \in \mathcal{I} \mid c_i(x) = 0\} \tag{4}$$

2.2 Two simple examples

2.2.1 Single equality constraint

Consider the simple optimization problem:

$$\min x_1 + x_2 \quad \text{s.t} \quad x_1^2 + x_2^2 - 2 = 0$$

For this problem we have $f(x) = x_1 + x_2$, $\mathcal{I} = \emptyset$ and $\mathcal{E} = \{1\}$ with $c_1(x) = x_1^2 + x_2^2 - 2$. We can easily see that the feasible set for this problem is the circle of radius $\sqrt{2}$ and that the optimal solution is $x^* = (-1, -1)$. Indeed, from any other point it's easy to decrease f while staying feasible.
This is clearly not possible for x^*, and, while we could prove it rigorously with some simple arguments, it is worth to look at this in an informal way by using first-order Taylor approximations.
When moving from a feasible point x with a small step s, to retain feasibility we require:

$$0 = c_1(x + s) \approx c_1(x) + \nabla c_1(x)^T s = \nabla c_1(x)^T s$$

Hence, in a first order sense, we must impose:

$$\nabla c_1(x)^T s = 0 \tag{5}$$

Similarly, if we want s to produce a decrease in f we ask that:

$$0 > f(x + s) - f(x) \approx \nabla f(x)^T s$$

That in a first order sense translates into:

$$\nabla f(x)^T s < 0 \tag{6}$$

But for x^* it is easy to see that:

$$\nabla f(x^*) = \lambda_1^* \nabla c_1(x^*) \quad \text{for} \quad \lambda_1^* = -1/2 \tag{7}$$

and so satisfying both conditions (5) and (6) is impossible for any s ! Viceversa, we can see that the only way for those conditions to not be satisfied by any s is for $\nabla f(x)$ and $\nabla c_1(x)$ to be parallel. In fact, if they are not parallel we can set:

$$\bar{s} := -const \cdot \left(Id - \frac{\nabla c_1(x) \nabla c_1(x)^T}{||\nabla c_1(x)||^2} \right) \nabla f(x)$$

to obtain a first order feasible decrease (easy check: use Cauchy-Schwarz for (6)).

We can rewrite the parallel condition (7) by introducing the Lagrangian:

$$\mathcal{L}(x, \lambda_1) = f(x) - \lambda_1 c_1(x)$$

and stating:

$$\nabla_x \mathcal{L}(x^*, \lambda_1^*) = 0$$

This suggests that we can search for solutions of the equality constrained problem by seeking stationary points of the Lagrangian function. The scalar quantity λ_1 is called a Lagrange multiplier for c_1.

2.2.2 Single inequality constraint

We consider a simple variant of the first example, by transforming the equality constraint into an inequality constraint:

$$\min x_1 + x_2 \quad \text{s.t} \quad 2 - x_1^2 - x_2^2 \geq 0$$

For this example the feasible region is given by the centered ball of radius $\sqrt{2}$. As before, it's easy to check that the optimal solution is given by $x^* = (-1, -1)$. However, the first order analysis is slightly different. By computations similar to the first example we get that a small step s remains feasible if:

$$c_1(x) + \nabla c_1(x)^T s \geq 0 \tag{8}$$

For inequality constraints we may distinguish between two cases:

1. c_1 **is inactive**: in this case any step small enough satisfies (8), and so for $\alpha > 0$ small enough we can obtain a decrease by setting $s := -\alpha \nabla f(x)$ whenever $\nabla f(x) \neq 0$

2. c_1 **is active**: in this case the step condition can be rewritten as:

$$\nabla f(x)^T s < 0, \quad \nabla c_1(x)^T s \geq 0 \tag{9}$$

   These two condition define two half-planes. It can be easily proven that the intersection of these two regions is empty only when $\nabla f(x)$ and $\nabla c_1(x)$ point in the same direction, i.e.:

$$\nabla f(x) = \lambda_1 \nabla c_1(x), \quad \text{for some} \;\; \lambda_1 \geq 0$$

   Unlike the case of equality constraints, the sign of the multiplier is relevant for the first order optimality condition.

By defining the Lagrangian:

$$\mathcal{L}(x, \lambda_1) := f(x) - \lambda_1 c_1(x)$$

we can reformulate the first order necessary conditions as:

$$\nabla_x \mathcal{L}(x^*, \lambda_1^*) = 0 \quad \text{for some} \;\; \lambda_1^* \geq 0$$

$$\lambda_1^* c_1(x^*) = 0$$

The second condition, used to include both the active and inactive case, is known as *complementarity condition*.

2.3 The Lagrangian

As seen in the previous examples, the Lagrangian plays a fundamental role in the definition of first order necessary conditions for optimality.

Definition 2.2 *The Lagrangian function associated to the optimization problem* (1) *is defined as:*

$$\mathcal{L}(x, \lambda) := f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x) \tag{10}$$

The Lagrangian can be also used to restate the optimization problem (1) as:

$$\inf_{x \in \mathbb{R}^n} \sup_{\lambda_i \geq 0 \; \forall \imath \in \mathcal{I}} \mathcal{L}(x, \lambda) \tag{11}$$

Indeed, when constraints are not respected we get an infinite penalization of the function: $\sup_{\lambda \geq 0} \mathcal{L}(x, \lambda) = +\infty$, and when they are respected we get $\sup_{\lambda \geq 0} \mathcal{L}(x, \lambda) = f(x)$.

3 Tangent Cone And Constraint Qualifications

3.1 Definitions

For the previous informal first order analysis to hold, smoothness is not sufficient. We also require that the first order description of the constraints accurately reflects the geometry of the feasible set. This will translate into a set of conditions known as constraint qualifications.

Definition 3.1 *Given a feasible point $x \in \Omega$ we call $\{z_k\}_k$ a feasible sequence approaching x if $z_k \in \Omega$ and $z_k \to x$ as $k \to \infty$.*

Definition 3.2 *We say that $d \in \mathbb{R}^n$ is tangent to Ω at x if there exists a feasible sequence $\{z_k\}_k$ approaching x and a sequence of scalars $t_k \to 0$ such that:*

$$d = \lim_{k \to \infty} \frac{z_k - x}{t_k} \tag{12}$$

The tangent cone $T_\Omega(x^*)$ is the set of all tangents to Ω at x. It's easy to verify that the tangent cone is indeed a cone.

The tangent cone is a geometric way to determine the directions tangent to Ω at x, and it will play an important role in the construction of the necessary optimality conditions.

Definition 3.3 *Given a feasible point $x \in \Omega$ the set of linearized feasible directions $\mathcal{F}(x)$ is:*

$$\mathcal{F}(x) := \left\{ d \in \mathbb{R}^n \mid \begin{array}{ll} d^T \nabla c_i(x) = 0 & \forall i \in \mathcal{E} \\ d^T \nabla c_i(x) \geq 0 & \forall i \in \mathcal{I} \cap \mathcal{A}(x) \end{array} \right\} \tag{13}$$

It's easy to verify that $\mathcal{F}(x)$ is also a cone. Observe that inactive constraints don't play any role in determining the first order tangent directions. We also note that changing the algebraic formulation of the constraints may change $\mathcal{F}(x)$. For example $c_1(x) := x_1^2 + x_2^2 - 2 = 0$ is equivalent to $\bar{c}_1(x) := (x_1^2 + x_2^2 - 2)^2 = 0$ but produces different linearized feasible directions.

Constraint qualifications will ensure that the first order description $\mathcal{F}(x)$ of the feasible set is equal to the tangent cone $T_\Omega(x)$, so that we can translate the tangent cone optimality conditions into first order (algebraic) conditions. The most used constraint qualification is the following:

Definition 3.4 *Given a feasible point $x \in \Omega$ we say that the linear independence constraint qualification (LICQ) holds if the set of active constraint gradients $\{\nabla c_i(x) \mid i \in \mathcal{A}(x)\}$ is linearly independent.*

While LICQ is not the only set of constraint qualifications, it is the most widely used in practice, both for analysis and algorithm design.

3.2 Relationship between $\mathcal{F}(x)$ and $T_\Omega(x)$

The following result states that the set of linearized feasible directions $\mathcal{F}(x)$ is in general bigger than $T_\Omega(x)$, however, thanks to the LICQ, we can guarantee that they are equal.
For the proof we will need the matrix $A(x^*)$, which is the matrix whose rows are the active constraint gradients, i.e. :

$$A(x^*) = \begin{bmatrix} \nabla c_{i_1}(x^*)^T \\ \vdots \\ \nabla c_{i_{|\mathcal{A}(x^*)|}}(x^*)^T \end{bmatrix}_{i_j \in \mathcal{A}(x^*)} \tag{14}$$

Lemma 3.1 *Let $x^* \in \Omega$ be a feasible point, then:*

1. $T_\Omega(x^*) \subseteq \mathcal{F}(x^*)$;

2. *If the LICQ holds at* x^* *then* $\mathcal{F}(x^*) = T_\Omega(x^*)$

Proof. W.L.O.G. we can assume that all the constraints are active and are indexed by $i = 1, .., m$ (we can achieve this by dropping all the inactive constraints).
1) For any $d \in T_\Omega(x^*)$ let $\{z_k\}_k$ and $\{t_k\}_k$ be the sequences that define d, i.e. :

$$d = \lim_{k\to\infty} \frac{z_k - x^*}{t_k}$$

From this definition we have that:

$$z_k = x^* + t_k d + o(t_k) \tag{15}$$

For equality constraints $i \in \mathcal{E}$ this implies:

$$\begin{aligned} 0 = \frac{1}{t_k} c_i(z_k) &= \frac{1}{t_k}[c_i(x^*) + t_k \nabla c_i(x^*)^T d + o(t_k)] \\ &= \nabla c_i(x^*)^T d + \frac{o(t_k)}{t_k} \end{aligned}$$

hence for $k \to +\infty$ we get $\nabla c_i(x^*)^T d = 0$, as required. We conclude in a similar way for inequality constraints that $\nabla c_i(x^*)^T d \geq 0$ and so $d \in \mathcal{F}(x^*)$.
2) Viceversa, suppose that LICQ holds at x^* and that $d \in \mathcal{F}(x^*)$. By hypothesis, $A(x^*)$ is a $m \times n$ matrix with full row rank m. Let Z be a matrix whose columns are a basis for the null space of $A(x^*)$, i.e. :

$$Z \in \mathbb{R}^{n\times(n-m)}, \quad Z \text{ has full column rank}, \quad A(x^*)Z = 0 \tag{16}$$

We define the parametrized system of equations $R : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ by:

$$R(z,t) = \begin{bmatrix} c(z) - tA(x^*)d \\ Z^T(z - x^* - td) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{17}$$

where $c = (c_1, ...c_m)$ is the vector of constraint functions. Given any sequence of scalars $t_k \to 0^+$ we claim that for small t_k the solutions z_k associated to $R(z, t_k) = 0$ give a feasible sequence that approaches x^* and that defines a tangent equal to d. The main idea behind the proof is to use the implicit function theorem. We observe that z^* is a solution of the system for $t = 0$ and that:

$$\nabla_z R(x^*, 0) = \begin{bmatrix} A(x^*) \\ Z^T \end{bmatrix}$$

which, by construction of Z, is a full rank matrix. Hence, according to the implicit function theorem, the system (17) has a unique solution z_k for all values of t_k sufficiently small. Moreover, being z_k solutions of (17) and $d \in \mathcal{F}(x^*)$ it holds that:

$$\begin{aligned} c_i(z_k) &= t_k \nabla c_i(x^*)^T d = 0 \quad \forall i \in \mathcal{E} \\ c_i(z_k) &= t_k \nabla c_i(x^*)^T d \geq 0 \quad \forall i \in \mathcal{I} \cap \mathcal{A}(x^*) \end{aligned}$$

so z_k is feasible. We prove now that the tangent produced by this approaching sequence is indeed d. By a Taylor's theorem we can write:

$$\begin{aligned} 0 = R(z_k, t_k) &= \begin{bmatrix} c(z) - tA(x^*)d \\ Z^T(z - x^* - td) \end{bmatrix} = \begin{bmatrix} A(x^*)(z_k - x^*) + o(||z_k - x^*||) - t_k A(x^*)d \\ Z^T(z_k - x^* - t_k d) \end{bmatrix} = \\ &= \begin{bmatrix} A(x^*) \\ Z^T \end{bmatrix} (z_k - x^* - t_k d) + o(||z_k - x^*||) \end{aligned}$$

By dividing the above by t_k and by multiplying the inverse of the coefficient matrix we conclude, since we get:

$$d = \frac{z_k - x^*}{t_k} + o\left(\frac{||z_k - x^*||}{t_k}\right)$$

□

3.3 A Necessary Optimality Condition For The Tangent Cone

As we have previously discussed, the relationship between the linearized feasible directions $\mathcal{F}(x)$ and the tangent cone $T_\Omega(x^*)$ is important because through the tangent cone we are able to get a simple necessary condition of optimality. This condition is stated in the following theorem:

Theorem 3.1 *If x^* is a local solution of (1), then we have:*

$$\nabla f(x^*)^T d \geq 0 \quad \forall\, d \in T_\Omega(x^*) \tag{18}$$

Proof. We proceed by contradiction. Suppose that there exists $d \in T_\Omega(x^*)$ such that $\nabla f(x^*)^T d < 0$ that is defined by the sequences $\{z_k\}$ and $\{t_k\}$. Using Taylor and definition (12) we get:

$$\begin{aligned} f(z_k) &= f(x^*) + (z_k - x^*)^T \nabla f(x^*) + o(||z_k - x^*||) \\ &= f(x^*) + t_k d^T \nabla f(x^*) + o(t_k) \end{aligned}$$

For t_k small enough $t_k d^T \nabla f(x^*)$ dominates $o(t_k)$ and, since $d^T \nabla f(x^*) < 0$ we can write:

$$f(z_k) < f(x^*) + \frac{1}{2} t_k d^T \nabla f(x^*) \quad \forall\, k \text{ big enough}$$

which leads to a contradiction, since x^* is a local minimum. □

4 First-Order Optimality Conditions

4.1 Farkas' Lemma

We are almost ready to state and prove the first-order optimality conditions for constrained optimization. In order to prove the main result we will need a classical theorem of the alternative known as Farkas' lemma.

Consider two matrices $B \in \mathbb{R}^{n \times m}$ and $C \in \mathbb{R}^{n \times p}$. We define the cone K as:

$$K := \{By + Cw \mid w \in \mathbb{R}^p, \quad y \in \mathbb{R}^m, \quad y \geq 0\} \tag{19}$$

Given a vector $g \in \mathbb{R}^n$ Farkas' lemma states that either $g \in K$ or else there is a vector $d \in \mathbb{R}^n$ such that:

$$g^t d < 0, \quad B^T d \geq 0, \quad C^T d = 0 \tag{20}$$

In other terms, either g is in the cone K or there exists a "particular" separating hyperplane between K and g.

Theorem 4.1 (Farkas' lemma) *Let K be defined as in (19). Given $g \in \mathbb{R}^n$ we have either that $g \in K$ or that there exists $d \in \mathbb{R}^n$ satisfying (20), but not both.*

Proof. We show first that both cannot hold simultaneously. If such were the case then $g = By + Cw$ and there would be a vector $d \in \mathbb{R}^n$ such that:

$$0 > d^T g = d^T By + d^t Cw \overset{(20)}{\geq} 0$$

which is absurd. We show now that one of the alternatives must hold, in particular we show that if $g \notin K$ then we can construct d that satisfies (20). it can be easily proven that K is a closed convex set, in particular the euclidean projection on K is well defined. Let $\hat{s}$ be the euclidean projection of

g on K. Since K is a cone it holds that $\alpha\hat{s} \in K$ for every $\alpha \geq 0$, in particular by definition of $\hat{s}$ the quantity $||\alpha\hat{s} - g||^2$ is minimized by $\alpha = 1$, so:

$$\frac{d}{d\alpha}||\alpha\hat{s} - g||^2\Big|_{\alpha=1} = 0 \implies (2\hat{s}^T\hat{s} - 2\alpha\hat{s}^T\hat{s})\Big|_{\alpha=1} = 0$$

concluding that:

$$\hat{s}^T(\hat{s} - g) = 0 \tag{21}$$

For any $s \in K$, by basic properties of projection, it also holds that:

$$(s - \hat{s})^T(\hat{s} - g) \geq 0$$

coupled with (21) this implies:

$$s^T(\hat{s} - g) \geq 0 \tag{22}$$

We claim that the vector $d := \hat{s} - g$ satisfies (20). Indeed we have that:

$$d^T g = d^T(\hat{s} - d) = (\hat{s} - g)^T\hat{s} - d^T d \overset{(21)}{=} -||d||^2 < 0$$

and for each $y \geq 0$, w we have $s = By + Cw \in K$ so by (22):

$$d^T(By + Cw) \geq 0$$

For $y = 0$ this implies $(C^T d)^T w \geq 0$ for every w, hence $C^T d = 0$. Similarly, for $w = 0$ we have $(B^T d)^T y \geq 0$ for all $y \geq 0$, which implies $B^T d \geq 0$, therefore concluding the proof. □

Farkas' lemma can be applied to limit the possibilities on the objective function and constraint gradients configurations:

Corollary 4.1 *Given a feasible point $x^* \in \Omega$, one and only one of the following holds:*

1. *There exists λ^* such that:*

$$\nabla f(x^*) = \sum_{i \in \mathcal{A}(x^*)} \lambda_i \nabla c_i(x^*) \quad \textit{with} \;\; \lambda_i \geq 0 \textit{ for } i \in \mathcal{I} \cap \mathcal{A}(x^*) \tag{23}$$

2. *There exists $d \in \mathcal{F}(x^*)$ such that $d^T \nabla f(x^*) < 0$.*

Proof. It follow directly from Farkas' lemma by setting $g := \nabla f(x^*)$ and:

$$K := \left\{ \sum_{i \in \mathcal{A}(x^*)} \lambda_i \nabla c_i(x^*) \mid \lambda_i \geq 0 \; \forall i \in \mathcal{I} \cap \mathcal{A}(x^*) \right\}$$

□

4.2 KKT Optimality Conditions

We are now ready to state and prove the first-order conditions necessary for a point to be a local minimum. These conditions, also known as Karush-Kuhn-Tucker (KKT) conditions, will follow almost trivially from the examples and lemmas that we have presented thus far. Nevertheless, this theory, that can be considered an extension to (the quite old theory of) Lagrange multipliers, was first presented only in 1951, and has since then been fundamental in the development of nonlinear constrained optimization.

Theorem 4.2 (KKT Conditions) *Suppose that x^* is a local minimum of problem (1) and that the functions f and c_i are $\mathcal{C}^1$. If the LICQ holds at x^*, then there exists a Lagrange multiplier vector λ^* such that the following conditions are satisfied:*

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0 \qquad \textit{(stationarity)} \tag{24a}$$
$$c_i(x^*) = 0 \qquad \forall i \in \mathcal{E} \qquad \textit{(primal feasibility)} \tag{24b}$$
$$c_i(x^*) \geq 0 \qquad \forall i \in \mathcal{I} \qquad \textit{(primal feasibility)} \tag{24c}$$
$$\lambda_i^* \geq 0 \qquad \forall i \in \mathcal{I} \qquad \textit{(dual feasibility)} \tag{24d}$$
$$\lambda_i^* c_i(x^*) = 0 \qquad \forall i \in \mathcal{E} \cup \mathcal{I} \qquad \textit{(complementary slackness)} \tag{24e}$$

Proof.
Since x^* is a feasible point, primal feasibility must obviously be satisfied. By corollary 4.1 we either have that:

1. There exists λ^* such that:
$$\nabla f(x^*) = \sum_{i \in \mathcal{A}(x^*)} \lambda_i \nabla c_i(x^*) \quad \text{with} \;\; \lambda_i \geq 0 \text{ for } i \in \mathcal{I} \cap \mathcal{A}(x^*) \tag{25}$$
2. There exists $d \in \mathcal{F}(x^*)$ such that $d^T \nabla f(x^*) < 0$.

Since by theorem 3.1 it cannot happen that there exists a $d \in T_\Omega(x^*) \overset{lemma 3.1}{=} \mathcal{F}(x^*)$ with $d^T \nabla f(x^*) < 0$, then the first option must hold, hence there exists a λ^* that respects stationarity. In particular, by corollary 4.1, we can pick $\lambda_i = 0$ for the inactive constraints i, and $\lambda_i \geq 0$ for the active inequality constraints, hence dual feasibility and complementary slackness are satisfied. □

4.3 Other Constraint Qualifications

We have stated the KKT Conditions under the assumption that the LICQ held at x^*. It is important to note that while the LICQ is the most widely used assumption in constrained optimization, there are other constraint qualifications that can be considered.
An important example is the one of linear constraints:

$$c_i(x) = a_i^T x + b_i \tag{26}$$

for these kind of constraints it's not difficult to prove a version of lemma 3.1:

Lemma 4.1 *Suppose that at some $x^* \in \Omega$ all active constraints are linear. Then $\mathcal{F}(x^*) = T_\Omega(x^*)$*

Proof. As before we can assume W.L.O.G. that all constraints are active. We need to prove that for any $w \in \mathcal{F}(x^*)$ we have $w \in T_\Omega(x^*)$. By definition of $\mathcal{F}(x^*)$ it holds that:

$$a_i^T w = 0 \quad \forall i \in \mathcal{E}, \qquad a_i^T w \geq 0 \quad \forall i \in \mathcal{I}$$

In particular it's easy to check by linearity that $z_k := x^* + (1/k)w$ are all feasible points. By picking $t_k = 1/k$ we conclude that $w \in T_\Omega(x^*)$ since:

$$\lim_k \frac{z_k - x^*}{t_k} = \lim_k \frac{(1/k)w}{1/k} = w$$

□

Since the LICQ is only used in the KKT proof to have $\mathcal{F}(x^*) = T_\Omega(x^*)$ we can immediately conclude that for linear constraints the KKT-theorem holds.
Another useful constraint qualification is the Mangasarian-Fromovitz constraint qualification (MFCQ):

Definition 4.1 *We say that the Mangasarian-Fromovitz constraint qualification (MFCQ) holds at x^* if there exists a vector $w \in \mathbb{R}^n$ such that:*

$$\nabla c_i(x^*)^T w > 0 \qquad \forall i \in \mathcal{I} \cap \mathcal{A}(x^*) \tag{27a}$$

$$\nabla c_i(x^*)^T w = 0 \qquad \forall i \in \mathcal{E} \tag{27b}$$

This set of conditions is weaker than LICQ, since by linear independence of the constraint gradients we can easily find such a w as a solution to a linear system. A version of the KKT conditions can be proven under MFCQ, but these results are out of the scope of this seminar.

4.4 Intuitive Explanation Of Lagrange Multipliers And Sensitivity

While the importance of the Lagrangian and of its multipliers in optimality theory is clear, their intuitive significance may be lost when studying them only through a rigorous treatment. We will argue that Lagrange multipliers say something about the sensitivity of the optimal objective value $f(x^*)$ to the presence of the constraint c_i, i.e. how much perturbations (or noise) to the constraint may influence the solution.
This kind of information can be useful when modeling a problem, as it is usually undesirable for a model to be too sensitive to variations of the data.

Let us assume that the constraint $i \in \mathcal{I}$ is active and let us relax the constraint by perturbing its r.h.s., asking that:

$$c_i(x) \geq -\epsilon ||\nabla c_i(x^*)||$$

Let $x^*(\epsilon)$ be the new solution to this perturbed problem, and assume that for ϵ small enough the Lagrange multipliers don't change much (in particular i is still active). Then:

$$-\epsilon ||\nabla c_i(x^*)|| = c_i(x^*(\epsilon)) - c_i(x^*) \approx \nabla c_i(x^*)^T (x^*(\epsilon) - x^*)$$

$$0 = c_j(x^*(\epsilon)) - c_j(x^*) \approx \nabla c_j(x^*)^T (x^*(\epsilon) - x^*) \quad \forall i \neq j \in \mathcal{A}(x^*)$$

By the KKT Condition (24a) and previous computations we get:

$$f(x^*(\epsilon)) - f(x^*) \approx \nabla f(x^*)^T (x^*(\epsilon) - x^*) = \sum_{j \in \mathcal{A}(x^*)} \lambda_j^* \nabla c_j(x^*)^T (x^*(\epsilon) - x^*) \approx$$

$$\approx -\epsilon ||\nabla c_i(x^*)|| \lambda_i^*$$

So by taking limits of the above we can conclude that:

$$\frac{df(x^*(\epsilon))}{d\epsilon} = -\lambda_i^* ||\nabla c_i(x^*)|| \tag{28}$$

So the larger $\lambda_i^* ||\nabla c_i(x^*)||$ is, the more the optimal value is sensitive to the placement of the i-th constraint.

Another interesting way to look at the Lagrange multipliers is through the lens of physics.
We can imagine f as the potential energy of a force (gravity, for example). The force associated to this potential energy is $-\nabla f(x)$, and it tends to push the "particle" x towards lower values of f (in a sense, this is a dynamical description of the classical gradient descent method). When we have reached a minimum x^*, we have either that the force is null, or that it is pushing against the constraint. For the system to stay still, the reaction from the constraints must be equal to the force. The reaction of each constraint can only be orthogonal to the constraint itself. Under regularity assumptions it can be seen that the orthogonal directions of each constraint are indeed $\nabla c_i(x^*)$. It's easy to understand that the higher λ_i is, the harder the force is pushing against the constraint, which means that in that direction there is a rapid decrease of potential energy (i.e. the optimal value is sensitive to the constraint). In particular, we may observe that inactive constraints cannot react to the force, because the particle it's not pushing against them, which explains the complementary slackness condition $\lambda_i^* = 0$.

5 Duality

Another important concept in constrained optimization is *duality*. Duality theory builds an alternative *dual problem* related to the original optimization problem (1), that in this context is usually called *primal*.
The dual formulation is often times considered because it can deliver important insights of the primal problem, either by simplifying it, or by adding descriptive variables that aid in system modelization or in the development of optimization algorithms.

We shall consider only the special case where there are no equality constraints and $f, -c_i$ are all convex functions.

5.1 The Dual Problem

Consider the convex optimization problem:

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to } c_i(x) \geq 0 \quad \forall i = 1, ..., m \tag{29}$$

where $f, -c_i$ are all convex functions. The associated Lagrangian is given by:

$$\mathcal{L}(x, \lambda) := f(x) - \lambda^T c(x)$$

We define the *dual objective function* $q : \mathbb{R}^n \to \mathbb{R}$ as follows:

$$q(\lambda) := \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda) \tag{30}$$

For some values of λ it may happen that $q(\lambda) = -\infty$. To avoid this problem we restrict q to the domain $\mathcal{D}$ where it is finite.

Theorem 5.1 *The dual objective function q is concave and its domain $\mathcal{D}$ is convex.*

Proof. For any λ^1, λ^2 and $\alpha \in [0, 1]$ it trivially holds that:

$$q(\alpha\lambda^1 + (1-\alpha)\lambda^2) = \inf_x [\alpha\mathcal{L}(x, \lambda^1) + (1-\alpha)\mathcal{L}(x, \lambda^2)] \geq \alpha \inf_x [\mathcal{L}(x, \lambda^1)] + (1-\alpha) \inf_x [\mathcal{L}(x, \lambda^2)] =$$
$$\geq \alpha q(\lambda^1) + (1-\alpha) q(\lambda^2)$$

hence q is concave. If both $\lambda^1, \lambda^2 \in \mathcal{D}$ it follows immediately that $q(\alpha\lambda^1 + (1-\alpha)\lambda^2) > -\infty$, hence $\alpha\lambda^1 + (1-\alpha)\lambda^2 \in \mathcal{D}$, which implies that $\mathcal{D}$ is convex. □

The dual problem is defined as follows:

$$\max_{\lambda \in \mathcal{D}} q(\lambda) \quad \text{subject to } \lambda \geq 0 \tag{31}$$

We recall that the original problem (29) can be stated in terms of its Lagrangian as:

$$\inf_{x \in \mathbb{R}^n} \sup_{\lambda \geq 0} \mathcal{L}(x, \lambda)$$

The dual problem is obtained by simply swapping the infimum with the supremum:

$$\sup_{\lambda \geq 0} \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda) = \sup_{\lambda \geq 0} q(\lambda)$$

In general the dual has only weak connections with the primal, but under proper assumptions there are strong links between the two.
One may ask what is the advantage of the dual formulation, since it requires to compute the global minimum of a function. We will see that we can exploit the special structure of some problems in order to simplify the dual, usually leading to an easier formulation, or at least a formulation that can give meaningful information for algorithm and model design.

5.2 From Primal To Dual

In this subsection we will see how the primal problem gives information on the dual.

Theorem 5.2 (Weak duality) *For any feasible $\bar{x} \in \Omega$ and $\bar{\lambda} \geq 0$, we have:*

$$q(\bar{\lambda}) \leq f(\bar{x}) \tag{32}$$

Proof. The proof is quite straightforward:

$$q(\bar{\lambda}) = \inf_x f(x) - \bar{\lambda}^T c(x) \leq f(\bar{x}) - \bar{\lambda}^T c(\bar{x}) \leq f(\bar{x})$$

□

For the following results we will use the KKT conditions related to the primal problem (29):

$$\nabla f(\bar{x}) - \nabla c(\bar{x})\bar{\lambda} = 0 \tag{33a}$$
$$c(\bar{x}) \geq 0 \tag{33b}$$
$$\bar{\lambda} \geq 0 \tag{33c}$$
$$\bar{\lambda}_i c_i(\bar{x}) = 0, \qquad \forall i = 1, ..., m \tag{33d}$$

The next results shows how solutions of the primal problem lead to solutions of the dual:

Theorem 5.3 (from primal to dual) *Suppose that $\bar{x}$ is a solution of the primal problem* (29) *and that $f, -c_i$ are all convex functions differentiable at $\bar{x}$. Then any $\bar{\lambda}$ for which $(\bar{x}, \bar{\lambda})$ satisfies the KKT conditions is a solution of the dual problem* (31)*.*

Proof. By hypothesis we have that $\bar{\lambda} \geq 0$ and that $\mathcal{L}(\cdot, \bar{\lambda})$ is a convex differentiable function, hence it holds that:

$$\mathcal{L}(x, \bar{\lambda}) \geq \mathcal{L}(\bar{x}, \bar{\lambda}) + \nabla_x \mathcal{L}(\bar{x}, \bar{\lambda})^T (x - \bar{x}) \overset{(33a)}{=} \mathcal{L}(\bar{x}, \bar{\lambda})$$

Therefore, we have:

$$q(\bar{\lambda}) = \inf_x \mathcal{L}(x, \bar{\lambda}) = \mathcal{L}(\bar{x}, \bar{\lambda}) = f(\bar{x}) - \bar{\lambda}^T c(\bar{x}) \overset{(33d)}{=} f(\bar{x}) \tag{34}$$

By weak duality $q(\lambda) \leq f(\bar{x})$ for all $\lambda \geq 0$, hence $\bar{\lambda}$ is the maximum of q and a solution of the dual problem. □

Note that if the functions are $\mathcal{C}^1$ and LICQ holds, then by the KKT conditions an optimal Lagrange multiplier does exist, and so any solution to the primal implies a solution of the dual.

5.3 From Dual To Primal

Conversely, in this subsection we will see how the dual problem can give information on the primal under certain conditions. In particular, we will ask for strict convexity of the Lagrangian $\mathcal{L}(\cdot, \bar{\lambda})$, which can be achieved by asking any of the functions $f, -c_i$ to be strictly convex.

Theorem 5.4 (From dual to primal) *Suppose that f and $-c_i$ are convex and $\mathcal{C}^1$. Let $\bar{x}$ be a solution of the primal problem* (29) *at which LICQ holds. Suppose that $\hat{\lambda}$ solves the dual problem* (31) *with $q(\hat{\lambda}) = \mathcal{L}(\hat{x}, \hat{\lambda})$. If $\mathcal{L}(\cdot, \hat{\lambda})$ is a strictly convex function then $\hat{x} = \bar{x}$. That is $\hat{x}$ is the unique solution of the primal problem* (29) *and $f(\hat{x}) = \mathcal{L}(\hat{x}, \hat{\lambda})$.*

Proof. We proceed by contradiction. Assume that $\bar{x} \neq \hat{x}$, then, by LICQ and continuous differentiability, there exists a $\bar{\lambda}$ that satisfies the KKT conditions. By equation (34) and theorem 5.3 then:

$$\mathcal{L}(\bar{x}, \bar{\lambda}) \overset{(34)}{=} q(\bar{\lambda}) \overset{\text{thm } 5.3}{=} q(\hat{\lambda}) = \mathcal{L}(\hat{x}, \hat{\lambda})$$

Since by definition $\hat{x}$ is the minimum of $\mathcal{L}(\cdot, \hat{\lambda})$ we have $\nabla_x \mathcal{L}(\hat{x}, \hat{\lambda}) = 0$. Moreover, by strict convexity of $\mathcal{L}(\cdot, \hat{\lambda})$ it follows that:

$$\mathcal{L}(\bar{x}, \hat{\lambda}) > \mathcal{L}(\hat{x}, \hat{\lambda}) + \nabla_x \mathcal{L}(\hat{x}, \hat{\lambda})^T (\bar{x} - \hat{x}) = \mathcal{L}(\hat{x}, \hat{\lambda})$$

which implies:

$$-\hat{\lambda}^T c(\bar{x}) > -\bar{\lambda}^T c(\bar{x}) \overset{(33\text{d})}{=} 0$$

But $\hat{\lambda} \geq 0$ and $c(\bar{x}) \geq 0$, so $-\hat{\lambda}^T c(\bar{x}) > 0$ is absurd! □

5.4 Wolfe Dual

A slightly different formulation of the dual problem, that is more convenient for computations, is given by the *Wolfe dual* problem, that can be stated as follows:

$$\max_{x,\lambda} \mathcal{L}(x, \lambda) \tag{35a}$$

$$\text{subject to:} \quad \nabla_x \mathcal{L}(x, \lambda) = 0, \qquad \lambda \geq 0 \tag{35b}$$

In this formulation we don't ask to find a global minimum to define the dual objective q, but we restrict ourselves to stationary points of the Lagrangian so to satisfy the stationarity KKT condition. By using convexity, this usually leads to a comparable dual problem:

Theorem 5.5 (From primal to Wolfe dual) *Suppose that $f, -c_i$ are convex $\mathcal{C}^1$ functions. Let $(\bar{x}, \bar{\lambda})$ be a solution pair of the primal problem at which LICQ holds. Then $(\bar{x}, \bar{\lambda})$ solves the Wolfe dual problem* (35)

Proof. By the KKT conditions of the primal, $(\bar{x}, \bar{\lambda})$ satisfy the constraints of the Wolfe dual. We just need to prove that they are the maximum of the Lagrangian. By primal feasibility we have $\mathcal{L}(\bar{x}, \bar{\lambda}) = f(\bar{x})$ and $c(\bar{x}) \geq 0$, therefore for any other (x, λ) satisfying the Wolfe dual constraints we have:

$$\begin{aligned} \mathcal{L}(\bar{x}, \bar{\lambda}) = f(\bar{x}) \geq f(\bar{x}) - \lambda^T c(\bar{x}) = \mathcal{L}(\bar{x}, \lambda) \overset{(\text{convexity})}{\geq} \\ \geq \mathcal{L}(x, \lambda) + \nabla_x \mathcal{L}(x, \lambda)^T (\bar{x} - x) = \mathcal{L}(x, \lambda) \end{aligned}$$

hence concluding the proof. □

5.5 Notable Examples

5.5.1 Linear Programming

Consider the linear programming problem:

$$\min_{x \in \mathbb{R}^n} c^T x \quad \text{subject to} \quad Ax - b \geq 0 \tag{36}$$

The dual objective of this problem is given by:

$$q(\lambda) = \inf_x [c^T x - \lambda^T (Ax - b)] = \inf_x [(c - A^T \lambda)^T x + b^T \lambda]$$

If $c - A^T \lambda \neq 0$ we have $q(\lambda) = -\infty$, otherwise $q(\lambda) = b^T \lambda$. Hence the dual problem is given by:

$$\max_{\lambda} b^T \lambda \quad \text{subject to } A^T \lambda = c, \ \lambda \geq 0 \tag{37}$$

We obtain the same problem by considering the Wolfe Dual. For some matrices, the dual problem may be computationally easier to solve than the original problem.

5.5.2 Convex Quadratic Programming

Consider the quadratic programming problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^TGx + c^Tx \quad \text{subject to } Ax - b \geq 0 \tag{38}$$

Where G is a symmetric positive definite matrix. The dual objective for this problem is:

$$q(\lambda) = \inf_{x\in\mathbb{R}^n} \frac{1}{2}x^TGx + c^Tx - \lambda^T(Ax - b) \tag{39}$$

Since G is positive definite, the Lagrangian is strictly convex in x, so the minimum is achieved when $\nabla_x\mathcal{L}(x,\lambda) = 0$, that is:

$$Gx + c - A^T\lambda = 0$$

and by substituting in the expression we get:

$$q(\lambda) = -\frac{1}{2}(A^T\lambda - c)^TG^{-1}(A^T\lambda - c) + b^T\lambda$$

so the dual problem is:

$$\max_{\lambda\geq 0} -\frac{1}{2}(A^T\lambda - c)^TG^{-1}(A^T\lambda - c) + b^T\lambda \tag{40}$$

we get a similar result using the Wolfe formulation, only that we just need G to be positive semidefinite.

6 Applications

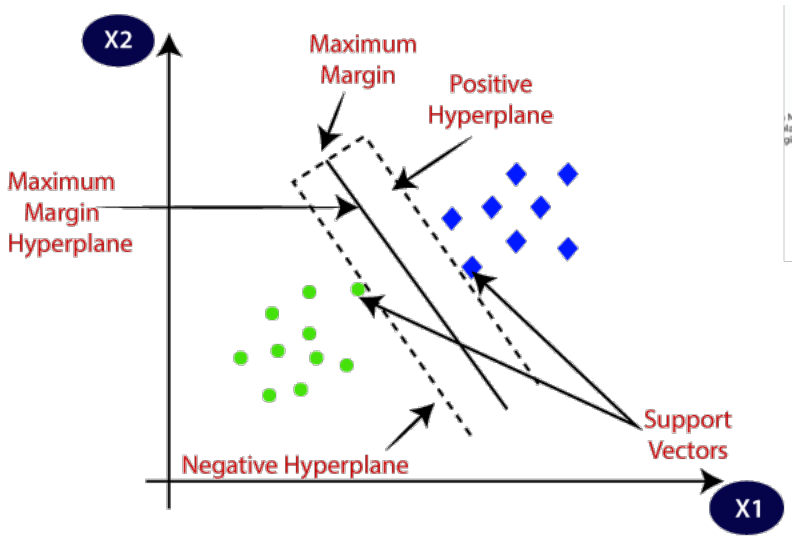
6.1 Revised And Dual Simplex Method

The nonlinear programming theory that we have developed thus far can be used to derive quite trivially many classical results of linear programming. These results are otherwise proved with a significant amount of effort, and can be considered for the most part subcases of the optimality conditions we have found.

We may observe that the KKT conditions associated to a linear program are not only necessary, but also sufficient, due to the convexity of the problem. It does not come up as a surprise then, that when designing the revised and dual simplex method to solve a linear programming problem, the KKT conditions are extensively used to relate the primal and dual formulation, and the information given by the multipliers is exploited to determine each step of the algorithm.
For a complete treatment on this refer to [5]

6.2 Support Vector Machines And The Kernel Trick

Support Vector Machines (SVM) are a machine learning model used for supervised and unsupervised learning.
In the case of a supervised binary classification problem, we want to classify two classes of points $x^i \in \mathbb{R}^n$ labeled by $y^i \in \{+1, -1\}$. The main idea of SVMs is to find a hyperplane that separates almost all the points of the two classes, finding the hyperplane with maximum margin between the two.



Finding such a hyperplane leads to a quadratic optimization problem:

$$\min_{w,b,\xi_i} ||w||^2 + c\sum_i \xi_i \tag{41a}$$

$$\text{s.t. :} \quad y^i[w^T x^i + b] \geq 1 - \xi_i \tag{41b}$$

$$\xi_i \geq 0 \tag{41c}$$

where $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ define the separating hyperplane, which classifies points as $+1$ or -1 based on the sign of $w^T x + b$. ξ_i are slack variables introduced to allow some misclassification errors.
Due to the strict convexity of the problem, we can equivalently consider the dual formulation of the problem (simplified through the use of KKT conditions), given by:

$$\max_{\alpha \in \mathbb{R}^n} -\frac{1}{4}||\sum_{i=1}^n \alpha_i y^i x^i||^2 + \sum_{i=1}^n \alpha_i \tag{42a}$$

$$\text{s.t} \quad 0 \leq \alpha_i \leq c \tag{42b}$$

$$\sum_{i=1}^n \alpha_i y^i = 0 \tag{42c}$$

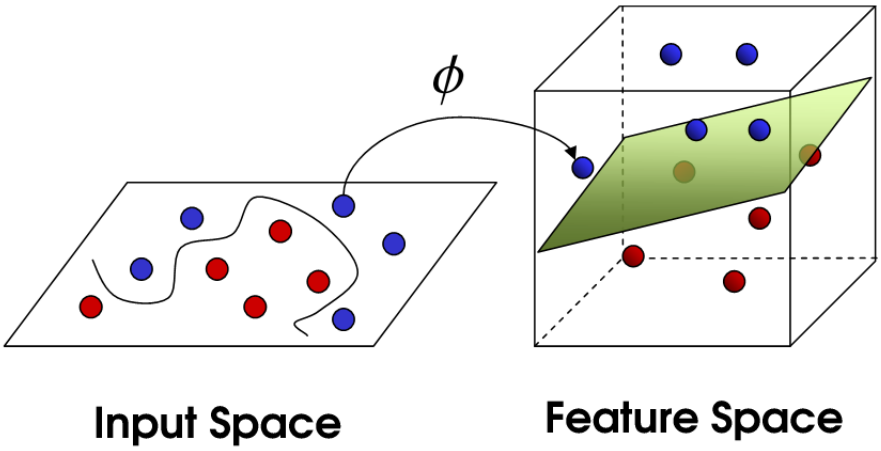
The solution of the dual problem α defines a separating hyperplane with the following formulas:

$$w = \frac{1}{2}\sum_{i=1}^n \alpha_i y^i x^i, \quad b = y^i - w^T x^i \quad \text{for any } i \text{ s.t. } \alpha_i \neq 0, c \tag{43}$$

While the primal formulation is somewhat easier to tackle from an optimization point of view, the dual formulation offers the advantage of interpretability. Indeed, the resulting separating hyperplane w is given by the linear combination of the x^i vectors for which $\alpha_i \neq 0$. These kind of vectors are known as support vectors (hence the name Support Vector Machine). Knowing how many support vectors there are is important to qualitatively establish the robustness of the SVM, since having too many of them usually indicates that we are relying too much on the data, therefore overfitting it.

The real power of the dual formulation though comes from the application of the *kernel trick*.

Most complex data isn't usually linearly separable, so a simple SVM would be a bad model for classification. However, by mapping the vectors x^i into a space of higher dimension, the data may become separable.



If we were to use the primal formulation, we would need to compute for each data point the transformation into this higher dimension, and also the resulting hyperplane w would be described by a higher dimensional object, thus slowing down computations.

Mercer's theorem relates the computation of a kernel function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ to the computation of the scalar product of transformed data $x \to \phi(x)$, where $\phi : \mathbb{R}^n \to \ell^2(\mathbb{R})$ is a transformation function associated to K. In particular :

$$K(x, y) =< \phi(x), \phi(y) >_{\ell^2}$$

So, if we were to use data transformed by ϕ, the resulting SVM dual formulation would be:

$$\max_{\alpha \in \mathbb{R}^n} -\frac{1}{4} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^i y^j K(x^i, x^j) + \sum_{i=1}^{n} \alpha_i \tag{44a}$$

$$\text{s.t} \quad 0 \leq \alpha_i \leq c \tag{44b}$$

$$\sum_{i=1}^{n} \alpha_i y^i = 0 \tag{44c}$$

which is an optimization problem with the same dimensions as before. Note that this cannot be done with the primal formulation. Considering that some transformations map the data into a space of infinite dimension, it's easy to understand how the primal formulation is not fit to tackle most feature mappings. The dual problem allows us to transform the data without explicitly computing the function ϕ, thus leading to a more tractable optimization problem. This computational trick is known as the kernel trick.

SVMs are a classical example of how the theory of duality can be used not only as an optimization tool, but also as a way to significantly improve and describe models.

References

[1] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[2] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[3] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[4] Philip Wolfe. A duality theorem for non-linear programming. *Quarterly of applied mathematics*, 19(3):239–244, 1961.

[5] Stephen Wright, Jorge Nocedal, et al. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.