DOI: 10.1478/AAPP.1021A1

AAPP | Atti della Accademia Peloritana dei Pericolanti
*Classe di Scienze Fisiche, Matematiche e Naturali*
ISSN 1825-1242

Vol. 102, No. 1, A1 (2024)

# KARUSH-KUHN-TUCKER CONDITIONS AND LAGRANGIAN APPROACH FOR IMPROVING MACHINE LEARNING TECHNIQUES: A SURVEY AND NEW DEVELOPMENTS

TIZIANA CIANO [a] AND MASSIMILIANO FERRARA [b]*

ABSTRACT. In this work we propose new proofs of some classical results of nonlinear programming milestones, in particular for the Kuhn-Tucker conditions and Lagrangian methods and functions. This study is concerned with some interesting features found in the well-known tools and methods, connecting them with a technical analysis of the "Maximal Margin Classifier" designed specifically for linearly separable data, while referring to the condition in which data can be separated linearly by using a hyperplane. In this context of analysis, we technically point out the centrality played by these mathematical tools when obtaining robustness in Machine Learning procedures analyzing some support vector machine (SVM) models, as they are used in various contexts and applications (*e.g.*, Soft Margin SVM and Maximum Margin SVM). This paper represents the first study reinforcing the ongoing Machine Learning Modeling and the research project we will launch in the near future on this fascinating frame of analysis. In this work we examine the problem of estimating the bias into a decision-making process. A new decision function algorithm is introduced as well.

## 1. Introduction

Optimization modeling is nowadays well-established as a fundamental framework driving the examination of numerous challenging allocations or decision-making problems following the Artificial Intelligence frontier of knowledge. It offers a certain amount of philosophical elegance that is difficult to argue against, and it frequently provides a necessary amount of operational simplicity. It is a complicated choice problem including the selection of values for a number of linked variables using this optimization methodology, focusing on one or more objective(s) intended to quantify performance and gauge the quality of the decision. Consequently, a specific optimization formulation should only be thought of as an approximation analysis, as is the case with all quantitative analysis methodologies. It takes modeling expertise to capture the key components of an issue and common sense to analyze the outcomes to arrive at useful conclusions. Therefore, optimization should be viewed as a conceptualization and analysis tool rather than as a theory that offers a logically sound solution. Through practical experience in the field and a solid comprehension of pertinent theory, one can develop abilities and common sense in framing problems and interpreting outcomes. Every issue formulation process includes a trade-off between the competing objectives of creating a mathematical model that is also tractable and sufficiently complex to contain the description of the situation. Set optimization is the use of set-valued maps for optimization. Many set optimization publications make use of the idea of a minimizer and its variations, which can easily be researched via vector optimization.

Today, however, we use an ordering relation for set comparison that is more oriented toward applications; this ordering relation was initially presented to optimization by Kuroiwa (1998); a first publication was presented by Kuroiwa, Tanaka, and Ha (1997). Young (1931) utilized this idea in algebra, Nishnianidze (1984) used it in fixed point theory, and Chiriaev and Walster (1998) used it in interval analysis and computer science. A discussion of even more plausible order relations can be found in the paper by Jahn and Ha (2011). These order interactions have significant socioeconomic applications, as demonstrated by Neukel (2013). For both scalar and multiobjective optimization, Karush-Kuhn-Tucker (KKT) optimality requirements are crucial in the field of optimization theory. If an appropriate constraint qualification holds, the KKT criteria are met at a weak efficient point. Jahn (2017) studied set optimization problems in finite-dimensional spaces with the property that the images of the set-valued objective map are described by inequalities and equalities and that the sets are compared with the lower-order relation of the sets. For these problems, the new Karush–Kuhn–Tucker conditions are shown as necessary and sufficient optimality conditions. Furthermore, the conditions of optimality without multiplier of the target map are presented. The utility of these results is demonstrated with a standard example.

The so-called “approximate optimality conditions”, also known as “asymptotic optimality conditions” or “sequential optimality conditions”, in which suitable sequences of points and multipliers are taken into account, are another type of necessary optimality conditions in scalar problems that do not require a constraint qualification. These conditions have been studied for many years. For pseudoconcave programming, we can cite the works of Fiacco and McCormick (1967), Kortanek and Evans (1968), and Zlobec (1971), and the book by Hestenes (1975). This last author generalizes the traditional optimality requirements stated by Guignard in an asymptotic manner. The Karush-Kuhn-Tucker conditions are also taken into consideration by Craven (1984) and Trudzik (1982).

For example, the papers of Andreani, Martínez, and Svaiter (2010), Andreani, Haeser, and Martínez (2011), Haeser and Schuverdt (2011), Dutta *et al.* (2013), and Haeser and de Melo (2013) are a few recent works that have focused on the study of such conditions due to their interest in the design and analysis of algorithms used to check this type of approximate optimality conditions. Ye and Zhang (2013) studied optimality conditions for non-fluid optimization problems with equality, inequality, and constraints of abstract sets. They derived the enhanced Fritz John condition, the Karush-Kuhn-Tucker condition, and the pseudonormal and quasinormal conditions. They also provided a tighter upper estimate for the Fréchet subdifferential and the limiting subdifferential of the value function.

Giorgi, Jiménez, and Novo (2016) dealt with the study of the approximate KKT condition for a continuously differentiable multi-objective problem in finite-dimensional spaces, whose feasible set is defined by inequality and equality constraints. They extended to this context the necessary optimality conditions obtained for scalar problems through KKT approximate conditions. Furthermore, we also proved that these conditions are sufficient conditions under the assumption of convexity as well. Ghosh *et al.* (2019) presented an extended Karush-Kuhn-Tucker condition to characterize efficient solutions to constrained interval optimization problems. It is derived from Gordan's theorem and applied to binary classification with range-valued data using Support Vector Machines (SVM). In fact, SVM classifiers prioritize increasing the class separation over exploiting class-specific internal models in the training set. However, it has recently been discovered that structure information, as an implicit prior knowledge, is essential for the development of an effective classifier in various real-world applications.

Therefore, over the last decades the Artificial Intelligence techniques, models and tools were developed. In this context of analysis the Machine Learning modeling has been pushed towards other strictly correlated fields of experimental applications in different scientific fields. From a mathematical point of view it should be interesting to analyze some technical aspects related to the reinforcement of Support Vector Machine modeling, and in particular, in relation to increasing the robustness, efficiency and accuracy of classical models normally applied in studying forecasting and predictive phenomena. In this vein, we intend to study the Karush-Kuhn-Tucker conditions and the connected Lagrangian approach, developing a new point of view opening arising research scenarios.

Xue, Chen, and Yang (2011) proposed a new wide-margin classifier, the structural regularized support vector machine (SRSVM). It sits at the intersection of cluster granularity and quadratic programming and follows the same optimization formulation as LapSVM (Laplacian Support Vector Machine), simultaneously integrating compactness within classes with separability between classes. Experimental results have demonstrated that SRSVM is often superior in classification and generalization performance compared to state-of-the-art algorithms.

The research of maximal margin classifiers has received a considerable amount of attention lately. Their extraordinary generalization skills are partly to blame for this attraction. In a nutshell, the maximal margin hyperplane accurately classifies all of the data given a collection of linearly separable data and maximizes the least distance between the data and the hyperplane. Computing the maximal margin hyperplane relates to the now-classic Support Vector Machines training problem if the Euclidean norm is used to quantify the distance. The formulation of this assignment as a quadratic programming problem is natural. If an arbitrary norm $p$ is employed, this task transforms into a more general mathematical programming issue that must be handled using general-purpose (and computationally demanding) optimization techniques (see, for example, Nachbar, Nossek, and Strobl 1993; Mangasarian 1997). When the target to be learned is sparse, a more general task called feature selection difficulties occurs.

Gentile (2000) discussed the study of maximum margin classifiers, and in particular hyperplanes of maximum margin, which classify linearly separable data and maximize the minimum distance. This problem is related to SVM training and can be solved using generic optimization methods. A decision method that separates the training data with the maximal margin is discovered by maximum margin classifiers. This approach minimizes the structural risk and improves the classifier's generalization skills, which is supported by the theoretical studies of both Vapnik and Chervonenkis (1974) and Cortes and Vapnik (1995), and experimental data. The Support Vector Machines illustrated by Cortes and Vapnik (1995) convert the quadratic programming problem into the supervised learning of the maximal margin classifier. Although quadratic programming has been extensively studied, there is still a need to process enormous amounts of data using complex optimization strategies.

Xu *et al.* (2005) proposed a new method for clustering based on the search for maximum margin hyperplanes through the data, which can be solved with a semi-defined program and leads to semi-supervised training for the support vector machines. Hein, Bousquet, and Schölkopf (2005) presented a framework for generating maximum margin algorithms for metric spaces. The authors consider two general cases: trusting the metric globally and believing only in the local structure of the metric. The algorithm optimization problem in this Banach space cannot be solved exactly, but an approximation is given that is exact when considering the training data plus a test point as a finite metric space. Isometric embeddings in a Hilbert space are restricted to the subclass of Hilbert metrics and the results suggest that the SVM has a better generalization performance. Nonlinear and non-smooth scalar and multi-objective optimizations were used in the past by different authors to introduce extended generalized support vector machine formulations (for instance, by means of the LC1 functions). One can refer to the papers of La Torre and Vercellis (2002), Cusano and La Torre (2003), La Torre (2003), and Orsenigo and Vercellis (2004), and references therein and, for general aspects in this fascinating direction, to Mangasarian (1998). Further work is needed in this direction.

The paper is structured as follows. In Section **2** a classic nonlinear programming problem is described by deepening; in Sections **2.1** and **2.2**, we discuss the theorems related to the Kuhn-Tucker-Uzawa conditions. In Sections **3** and **3.1**, we introduce the Support Vector Machine and relate the Lagrangian approach to it. Finally, in Section **4** we present the conclusions and future developments.

## 2. Nonlinear programming: a brief and useful sketch of classical milestones by new proofs and new points

A stimulating diversity of pure, widely applicable mathematics, numerical analysis, and computers can be explored using nonlinear programming. This strength is also clear for what concerns Artificial Intelligence – in particular for what concerns Machine Learning developments. Usually, a typical mathematical programming problem is formally introduced in the following way:

$$\begin{gathered} \max f(x) \\ s.t.\ g(x) \leq b, \text{ with } x \geq 0 \\ \text{where } f:\ \mathbb{R}^n \to \mathbb{R} \text{ and } g: \mathbb{R}^n \to \mathbb{R}^m. \end{gathered} \tag{1}$$

The “limitations” – the constraints and non-negativity of variables – that are visible in this suggested generic scheme are only apparent for $x$. The nonlinear programming problem (NLP) is the name of the mathematical programming issue that was introduced at the beginning of this essay (see (1)). One can fit the majority of the static optimization point issues using the same formal scheme. Strong inequalities and situations where one or more variables can take values from a set that does not comprise an interval ($>$) are two situations in which it is not possible to move on with this formal conceptual unification. According to the generalizations that can be drawn about the functions that affect problem (1), there are two extremely significant subtypes that have been deeply studied in time:

(1) A concave programming problem when the objective function $f$ is concave, and the admissible region is also concave (*i.e.*, the components of the vector function are all convex $g$: $g_1, g_2, \dots, g_m$).
(2) An NLP problem with differentiable functions when they are all differentiable, namely $f$, $g_1, g_2, \dots, g_m$.

Important findings such as separation theorems that enable research into the overall behavior of the objective function are made completely exploitable by the concavity hypothesis (or convexity). The derivative, which, on the other hand, lends itself to a local study, serves as the analysis and investigation instrument in the case of the differentiability hypothesis. In fact, it is stated in principle that the requirements of the global maximum/minimum are given in Hypothesis 1, and the local maximum/minimum are assumed in Hypothesis 2.

**2.1. Concave/convex programming: some useful aspects strictly connected to Machine learning applications.** We present an essential proposition that is required in order to prove a key NLP theorem.

**Proposition 1.** *Let* $\varphi_1, \varphi_2, \dots, \varphi_m : R^n \to R$ *be concave defined on the same convex set* $X \subseteq R^n$ *and consider a function* $\bar{\varphi} : R^n \to R^m$ *that has* $\varphi_1, \varphi_2, \dots, \varphi_m\}$ *as its components. If the system* $\varphi(x) > 0, \forall x \in X$, *it is impossible, and there will exist in* $R^m$ *a row vector* $\hat{\lambda} \geq 0$ *such that:*

$$\widehat{\lambda}\varphi(x) \leq 0 \quad \forall x \in X \tag{2}$$

*Proof.* $\forall x \in X$ we define the set

$$H(x) = \{h : h \in \mathbb{R}^m, h < \varphi(x)\}. \tag{3}$$

$H$ is the union of all $H(x)$ when changing $(x)$ in $X$, that is:

$$H = \bigcup_{x \in X} H(x) \tag{4}$$

To this end, we say that:

- $H$ is convex. Consider $h_1$ and $h_2$ as two generic elements of $H$: for the (4) each of them belongs to at least one among the sets (3) and therefore $h_1 \in H(x^1)$ and $h_2 \in H(x^2)$. Starting from the hypothesis of concavity of the components we will have $\varphi$:

  $$\alpha h_1 + (1-\alpha) h_2 < \alpha\varphi\left(x^1\right) + (1-\alpha)\varphi\left(x^2\right) \leq \varphi\left(\alpha x^1 + (1-\alpha)x^2\right),$$

  with $0 < \alpha < 1$. Therefore:

  $$\alpha h_1 + (1-\alpha) h_2 \in H\left(\alpha x^1 + (1-\alpha)x^2\right) \leq H, \text{given that } \alpha x^1 + (1-\alpha)x^2 \in X.$$

- $H$ does not contain the origin. In fact, if that were the case, the origin would belong to at least one, and we will have $H(x)\ \varphi(x) > 0$ for some $x$. As $H$ convex and $0 \notin H$, we know that there is a vector $p \neq 0$ :
$$p^t \cdot h \geq p^t \cdot 0 = 0 \quad \forall h \in H.$$
Since the components of $h$ take arbitrarily large negative values in the absolute value, it must be $p \leq 0$. Indicating the vector $-p^t$ with $\widehat{\lambda}$ , we have:
$$\widehat{\lambda} \cdot h \leq 0 \qquad \forall h \in H.$$
At this point it is observed that we can write $h = \varphi(x) - \varepsilon$ with $x \in X$ and $\varepsilon > 0$. Varying $x$ in $X$ and $\varepsilon > 0$ we could generate all of the $h \in H$. However, then:
$$\widehat{\lambda} \cdot (\varphi(x) - \varepsilon) \leq 0 \qquad \forall x \in X \quad \text{and} \quad \forall \varepsilon > 0$$
by placing $a = \widehat{\lambda} \cdot \varepsilon$:
$$\widehat{\lambda} \cdot \varphi(x) \leq a \quad \forall x \in X \quad \text{and} \quad \forall a > 0$$
Having to hold this last inequality $\forall a > 0$ it will coincide with (2).

□

**Remark 1.** *If there is a need, one can take in* (2)
$$\sum_{r=1}^{m} \widehat{\lambda}_r = 1, \textit{ and divide each } \widehat{\lambda}_r \textit{ for } \sum_{r=1}^{m} \widehat{\lambda}_r > 0$$
*The Kuhn-Tucker-Uzawa theorem, which offers a required condition of a global maximum for concave programming problems, can be introduced thanks to the proposition 1 that has just been stated and shown.*

**Theorem 2** (Kuhn-Tucker-Uzawa). *If* $\hat{\lambda}\ (b - g(\hat{x})) = 0$ *is the global maximum point for problem* (1)*, there exist* $m+1$ *non-negative quantities* $\hat{\lambda}_0, \hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_m$*, of which not all are simultaneously zero:*
$$\hat{\lambda}_0 f(x) + \hat{\lambda}\ (b - g(x)) \leq \hat{\lambda}_0 f(x) \qquad \forall x \in X \tag{5}$$
*where* $\widehat{\lambda} = \left[\hat{\lambda}_0,\ \hat{\lambda}_1, \hat{\lambda}_2, \ldots,\ \hat{\lambda}_m\right]$. *In particular, we have:*
$$\hat{\lambda}\ (b - g(\hat{x})) = 0 \tag{6}$$

*Proof.* Having $g_1,\ g_2,\ \ldots,\ g_m$ all convex functions defined by $b_1 - g_1(x),\ b_2 - g_2(x),\ \ldots,$ $b_m - g_m(x)$ are all concave. Starting from this hypothesis we will have the following system:
$$f(x) - f(\hat{x}) > 0$$
$$b - g(x) \geq 0$$
which has no solutions for $x \geq 0$. Finally, the following system will have no solution for $x \geq 0$,
$$f(x) - f(\hat{x}) > 0$$
$$b - g(x) > 0$$

The set of $x \geq 0$ is convex and therefore the assumptions of proposition 1 for $m+1$ functions defined by $f(x) - f(\hat{x})$, $b_1 - g_1(x)$, $b_2 - g_2(x)$, $\ldots$, $b_m - g_m(x)$ are that there exist $m+1$ non-negative numbers $\hat{\lambda}_0,\ \hat{\lambda}_1,\ \hat{\lambda}_2,\ \ldots,\ \hat{\lambda}_m$ that are not all null such that:

$$\hat{\lambda}_0 (f(x) - f(\hat{x})) + \hat{\lambda}\ \ (b - g(x)) \leq 0 \quad \forall x \geq 0$$

where from these it follows that $\hat{\lambda} = \left[\hat{\lambda}_0,\ \hat{\lambda}_1,\ \hat{\lambda}_2,\ \ldots,\ \hat{\lambda}_m\right]$. Let us show (6).

Let $\hat{\lambda}(b - g(\hat{x})) \geq 0$, being $\hat{\lambda} \geq 0$ and $b - g(\hat{x}) \geq 0$. Replacing $\hat{x}$ in (5) one also has $\hat{\lambda}(b - g(\hat{x})) \leq 0$, and therefore it is equal to (6). We introduce the *Slater condition (S)*, namely, that there is at least one $\hat{x} \geq 0$ such that $g(\hat{x}) < b$. □

We will then have the following theorem:

**Theorem 3.** *If the assumptions of Theorem 2 are satisfied (K.T.U.) and according to Slater's condition we have that* $\hat{\lambda}_0 > 0$ *and one can make sure that* $\hat{\lambda}_0 = 1$*, we have the following proof.*

*Proof.* We will have to prove that $\lambda_0$ cannot be zero. If it were equal to zero equation 5 would be:

$$\hat{\lambda}\ \ (b - g(\hat{x})) \leq 0 \quad \forall x \geq 0$$

which is obviously absurd, since $\hat{\lambda}(b - g(\hat{x})) \leq 0, \hat{\lambda} \geq 0$ and $(b - g(\hat{x})) > 0$. From this we have that $\hat{\lambda}_0 > 0$. In compliance with Theorem 3, one can act on the structure of (5), obtaining:

$$f(x) + \hat{\lambda}\ \ (b - g(x)) \leq f(\hat{x}) \quad \forall x \geq 0. \tag{7}$$

Called $L$ the function defined by

$$L\left(x, \hat{\lambda}\right) = f(x) + \lambda\ \ (b - g(x)) \quad x \geq 0 \quad \text{and} \quad \lambda\ \geq 0. \tag{8}$$

it can be said that $\left(\hat{x},\ \hat{\lambda}\right)$ is a saddle point, namely:

$$L\left(x, \hat{\lambda}\right) \leq L\left(\hat{x},\ \hat{\lambda}\right) \leq L\ (\hat{x},\ \lambda) \quad \forall x \geq 0 \quad \forall \lambda \geq 0.$$

Noting that $L\left(\hat{x},\ \hat{\lambda}\right) = f(\hat{x})$, it turns out that the first inequality is (7), while the second follows immediately from the fact that $\lambda \geq 0$ and $b - g(\hat{x}) \geq 0$.

The function $L$ defined by $(f)$ is called the Lagrangian. The variables $\lambda_r$, are the Lagrange multipliers. The denoted function with $\widetilde{L}$ is defined by:

$$\widetilde{L}(x, \lambda_0,\ \lambda) = \lambda_0\ f(x) + \lambda . (b - g(x))$$

which is defined as an "augmented" Lagrangian. The fundamental result to which we have arrived so far is that it has been shown that a necessary condition for $\hat{x}$ is a global maximum point for problem (1) and that $L$ has a saddle point $\left(\hat{x},\ \hat{\lambda}\right)$. The link between a Mathematical Programming Problem (MPP) with $n$ variables and $m$ constraints and the behavior of a function with $n+m$ variables is a fundamental property of MPP: it is not in fact a characteristic exclusively proper only to concave programming but, on the contrary, it is found in all other types of programming. Slater's condition is essential to ensure the presence of a saddle point for $L$. Let us present, however, a counter example in which ($S$) is violated: for this purpose we consider a problem of Linear Programming with only one variable and one constraint

$$\max_x x$$
$$\text{with constraint} \quad x^2 \leq 0$$

where the only point allowed by the constraint is 0 which represents the maximum point. Hence, the function $L$ is defined by:

$$L(x,\lambda) = x - \lambda x^2$$

and it does not have any saddle. Let us introduce another result, which can be considered the reverse of Theorem 2. □

**Theorem 4.** *Let $f, g_1, g_2, \ldots, g_m$ be real functions defined by values of $x \geq 0$. If there exists a point $(\widehat{x}, \widehat{\lambda})$ with $\widehat{x} \geq 0$ and a $\widehat{\lambda} \geq 0$ saddle for the Lagrangian function $L(x,\lambda) = f(x) + \lambda(b - g(x))$, then $\widehat{x}$ is a maximum point for $f(x)$ under the constraints $g(x) \leq b$ and $x \geq 0$. In addition, we will have that:*

$$\hat{\lambda}\,(b - g\,(x)) = 0 \tag{9}$$

**Remark 2.** *The statement of Theorem 2 did not request neither the concavity of f nor the convexity of g (or vice versa). Thus , the cited Theorem has general validity going beyond the limits of investigation of the following survey (the Concave Programming).*

*Proof.* $L(\hat{x}, \hat{\lambda}) \leq\ L\ (\hat{x}, \lambda)$ for each $\lambda \geq 0$ ensures that:

$$\hat{\lambda}(b - g(x)) \leq \lambda(b - g(x)) \tag{10}$$

Thus from (10) it can be deduced that $\lambda(b - g(x))$ is a lower bounder as $\lambda \geq 0$. As $\lambda \geq 0$ is a convex cone we will have that:

$$\lambda(b - g(\hat{x})) \geq 0 \quad \forall \lambda \geq 0$$

so that $b - g(\hat{x}) \geq 0$ in such a way that $\hat{x}$ satisfies the constraints. Placing $\lambda = 0$ in (10) we will have:

$$\hat{\lambda}(b - g(\hat{x})) \leq 0$$

However, if $\hat{\lambda} \geq 0$ and $b - g(\hat{x})$, we have:

$$\hat{\lambda}\,(b - g\,(\hat{x})) = 0$$

*i.e.*, (9). $(x, \hat{\lambda}) \leq\ L(\hat{x}, \hat{\lambda})$ means that:

$$f(x) + \hat{\lambda}(b - g(x)) \leq\ f(\hat{x}) + \hat{\lambda}(b - g(\hat{x})) \quad \forall\, x \geq 0$$

and, taking into account (9):

$$f(\hat{x}) - f(x) \geq \hat{\lambda}(b - g(x)) \quad \forall x \geq 0$$

It follows that $\forall x :\ b - g(x) \geq 0$ and we have:

$$f\,(\hat{x}) - f\,(x) \geq 0$$

*i.e.*, that $\hat{x} \geq 0$ is a maximum point under the condition $b - g(x) \geq 0$. □

**2.2. Programming with differentiable functions.** In this case we modify the starting problem (1), replacing the limitation $x \geq 0$ with the following:

$$x \in A \subseteq \mathbb{R}^n$$

where $A$ is an open.

Hence, the new problem will be as follows:

$$\max_x f(x) \quad \text{with constraints} \quad g(x) \leq b \quad x \in A, \tag{11}$$

with $A$ being an open $\mathbb{R}^n, f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ and $f, g_1, g_2, \ldots, g_m$ are all differentiable. We introduce a Constraints Regularity Condition (CRC). We indicate with $V \subseteq A$ the field of choice of the problem of NLP (11) such that:

$$V = x : x \in A, \ldots g(x) \leq b$$

and let $x^0$ be a point of $V$. Let $I(x^0)$ be the set that collects the indices $r$ such that $g_r(x^0) = b_r$ and $x$ denote by any point of $A$:

$$\bar{x}\, g'_r(x^0) \cdot (\bar{x} - x^0) \leq 0 \quad \forall r \in I(x^0)$$

We say that $x^0$ verifies the Constraints Regularity Conditions (CRC) if, however $a\bar{x}$ is taken, there exists a function $\Phi$ that maps for each value $t$, $0 \leq t \leq 1$, a point of $V$ with the following two properties:

(1) $I\Phi(0) = x^0$
(2) $\Phi(t)$ is differentiable in 0 and $\Phi'(0) = \alpha\ (\bar{x}–x^0)$ with $\alpha > 0$

To fully understand the meaning of CRC, let us consider the following set:

$$Z(x^0) = \{\bar{x} : \bar{x} \in A, \quad g'_r(x^0) \cdot\ (\bar{x} - x^0) \leq 0 \quad \text{for } r \in I(x^0)\}$$

This is the set of points obtained by substituting, for constraints satisfied with the sign of equality, with $g_r(x) = b_r$ the hyperplanes tangent to them in $x^0$ with $g'_r(x^0) \cdot (\bar{x} - x^0) = 0$ (not considering the constraints satisfied with the sign of strong inequality). It is said that the CRC requires that for each outgoing network segment from $x_0$ and all contained in $Z(x^0)$, there exists a curve all contained in $V$ and tangent to the segment of lines in $x_0$. As can be seen in Figs. 1 and 2 (case A: $I(x^0) = \{1\}$; $V$= double hatched set $Z(x^0)$= simple hatch set; case B: $I(x^0) = \{1, 2\}$; $V$= double hatched set $Z(x^0)$= line through $x^0$), the CRC can be violated when in the neighborhood of $x_0$, and the border of the field of choice has irregularities (*e.g.*, a cusp). In case B just presented, $Z(x_0)$ is reduced to a straight line and then $\bar{x}$ must be taken on it. This line $Z(x_0)$, in the proposed case, enters $V$ on the left side, and on the right hand side, it does not. It follows to construct a curve tangent to the segment that starts from $x_0$ and $\bar{x}$ and that is contained in $V$. On the basis of the considerations just expressed, it can be said that the CRC is quite analogous to Slater's condition, but it holds locally while S acts globally $(S)$.

We introduce the Kuhn-Tucker Theorem which gives us the necessary condition for a local maximum:

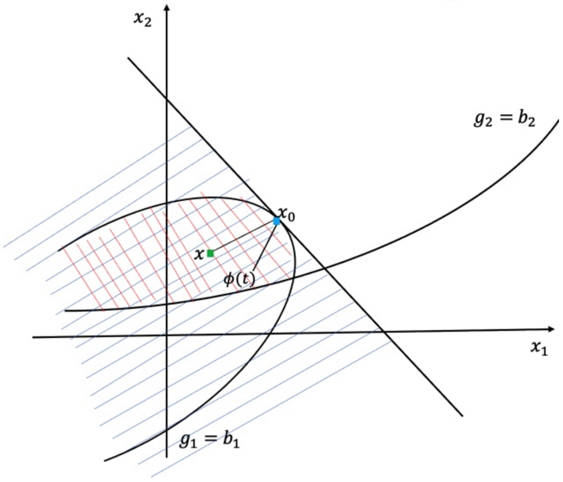

FIGURE 1. Case A: CRC is satisfied

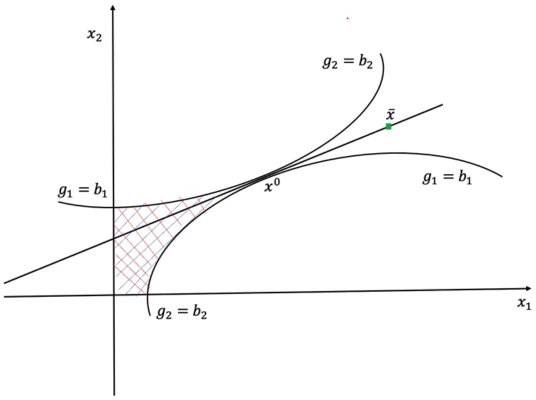

FIGURE 2. Case B: CRC is violated

**Theorem 5** (K-T). *Let there be a local maximum point for problem* (11) *and let the CRC hold in $\hat{x}$. Then there will be a row vector $\hat{\lambda} \in \mathbb{R}^m$ such that:*

$$\begin{cases} A. \quad \nabla^{\,t} f(\hat{x}) - \hat{\lambda} \cdot \nabla^t g(\hat{x}) = 0 \\ B. \quad g(\hat{x}) \le b \\ C. \quad \hat{\lambda}\ (b - g(\hat{x})) = 0 \\ D. \quad \hat{\lambda} \ge 0 \end{cases}$$

*Proof.* $g_r(\hat{x}) \le b_r$ $r = 1,2,...,m,$, if the $I(\hat{x}) = \emptyset$ assumption follows from the following theorem. □

**Theorem 6.** *$f : \mathbb{R}^n \to \mathbb{R}$ is defined on $X \subseteq \mathbb{R}^n$. If $\hat{x}$ is a local maximum point and $f$ is differentiable in x, then:*

$$f'(\hat{x}) \cdot v \leq 0 \tag{12}$$

*where v represents a non zero vector $v \in \mathbb{R}^n$ and is eligible (i.e., $x \in X \subseteq \mathbb{R}^n$, and v is admissible, with respect to $x \in X$, if $x + \varepsilon v \in X$ at least for all values of $\varepsilon > 0$ and less than a certain $\varepsilon_0 > 0$). In particular, if $\hat{x}$ is a neighborhood of X then:*

$$f'(\hat{x}) = 0 \tag{13}$$

*Proof.* Through the Taylor formula stopped at the first order, we have

$$f(\hat{x} + \varepsilon v) = f(\hat{x}) + \varepsilon\, f'(\hat{x}) \cdot v + o(\varepsilon)$$

with admissible $v$ and $\varepsilon > 0$, and $\hat{x} + \varepsilon v \in X$ being the maximum local, we have

$$f(\hat{x} + \varepsilon v) \leq f(\hat{x})$$

which follows from (12).

If $\hat{x}$ and a neighborhood of $X$ $\forall v$ admissible, and (12) is worth $\forall v$, it reduces to (13). Returning to the KT Theorem, we will have, by virtue of what was seen above $f'(\hat{x}) = 0$, and the condition are trivially verified by placing $\widehat{\lambda}, = 0$.

Suppose:

$$\begin{aligned} g_r(\hat{x}) &= b_r \quad \text{with } r \in I(\hat{x}) \\ g_r(\hat{x}) &< b_r \quad \text{with } r \notin I(\hat{x}) \end{aligned}$$

and $I(\hat{x}) \neq \emptyset$.

Arbitrarily choose an $\bar{x} \in A : g'_r(\hat{x})(\bar{x} - \hat{x}) \leq 0$ for $r \in I(\hat{x})$. Since it is a differentiable choice, we can write $f$:

$$f(x) - f(\hat{x}) = f'(\hat{x})(x - \hat{x}) + o\ (||x - \hat{x}||)$$

Considering the function $\Phi$ of CRC taking $\Phi(t) = x$, we have

$$x - \hat{x} = \Phi(t) - \Phi(0) = \phi'(0)t + o(t)$$

where $o(t)$ is nothing more than $(||x - \hat{x}||)$; therefore:

$$f(x) - f(\hat{x}) = f'(\hat{x})[\phi'(0)t + o(t)] + o(t) = f'(\hat{x}) \cdot \phi'(0)t + o(t)$$

and finally leveraging the CRC:

$$f(x) - f(\hat{x}) = \alpha f'(\hat{x})(x - \hat{x})t + o(t) \quad \text{with } \alpha > 0$$

Since $\Phi(t) \in v, \forall\ 0 \leq t \leq 1$ and $\hat{x}$ is a local maximum, we have that $f(x) - f(\hat{x}) \leq 0$ at least for sufficiently small values of $t$. It is therefore obtained that $f'(\hat{x})(x - \hat{x}) \leq 0$. Let us say that $y = x - \hat{x}$, and it can be concluded that we have $f'(\hat{x}) \cdot y \geq 0, \forall y :\ -g'_r(\hat{x}) \cdot y \geq 0$ and $r \in I(\hat{x})$. By virtue of the Farkas-Minkowski Theorem then there exist numbers $\lambda_r \geq 0\ \forall r \in I(\hat{x})$:

$$-f'(\hat{x}) = \sum_{r \in I(\hat{x})} \widehat{\lambda}_r(-g'_r(\hat{x}))$$

or

$$f'(\hat{x}) - \sum_{r \in I(\hat{x})} \widehat{\lambda_r}\, g'_r(\hat{x}) = 0$$

Let us say now that $\widehat{\lambda_r} = 0$ for $r \notin I(\hat{x})$ and we will obtain

$$f'(\hat{x}) - \sum_{r=1}^{m} \widehat{\lambda_r}\, g'_r(\hat{x}) = f'(\hat{x}) - \hat{\lambda} g'(\hat{x}) = 0$$

which results in condition A of the KT Theorem (see Theorem 2). Condition B is obvious: if $\hat{x}$ is a local maximum for problem (11), it respects the constraints. As regards condition C, since $\widehat{\lambda_r} = 0$ with $r \notin I(\hat{x})$ and $b_r - g_r(\hat{x}) = 0$ for $r \in I(\hat{x})$ we have that $\hat{\lambda}(b - g(\hat{x})) = 0$. □

## 3. Support Vector Machine: some remarks and new developments

A Support Vector Machine is a tool that data scientists can use to tackle classification and regression issues. A non-probabilistic binary classifier is an SVM. The non-probabilistic component contrasts with probabilistic classifiers, such as naive Bayes, which determine the likelihood of belonging to the class based on training instances (Aggarwal 2015). A decision boundary, which is a plane in the space of multidimensional qualities, is used by an SVM to divide data. Support vectors was the name given by the creators since only a small portion of the data contacts or supports the decision boundary. Data can only be divided into two classes by an SVM. Research on multiclass data management using SVM's is still ongoing. There are some workarounds, though. According to Bhavsar and Ganatra (2012), the strategies entail building numerous SVMs that compare feature vectors against one another using different techniques such us one-versus-rest (OVR) or one-versus-one (OVO). The OVR technique trains $k$ classifiers for $k$ classes so that each class is biased towards the other $k-1$ classes. OVO needs $k(k-1)/2$ classifiers, since it generates a binary classification issue for every conceivable class coupling. After building the necessary number of binary classifiers for the OVR or OVO techniques, the algorithm sorts new objects into categories based on which classifiers received the most votes. Each point in the supervised machine learning technique known as SVM is made up of a collection of attributes $\{x_1, \ldots, x_n\}$ and a class label $(y_i)$. Each data object is treated by an SVM as a feature space point that belongs to one of only two classes. According to the SVM, the class labels are either $y_i = 1$ or $y_i = -1$ The dataset archive is therefore mathematically represented by the sets $\bar{X}$:

$$\bar{X} = \{(x_i, y_i)\, | x_i \in \Re^p, y_i \in (-1, +1)\}_{i=1}^{n} \tag{14}$$

where $n$ is the number of vectors and $p$ is the characteristic's vector dimension.

The SVM classifier discovers a linear decision boundary in the feature space during training that most effectively divides the input objects into the two classes. The corresponding optimization problem identifies two parallel hyperplanes that, in the absence of any data items, produce the biggest gap. A subspace with a dimension smaller than the surrounding space is called a hyperplane. The margin is the angular separation of two parallel hyperplanes. A multidimensional decision boundary that divides the data into two halves is defined by the hyperplane that equidistantly forks the space between parallel hyperplanes.

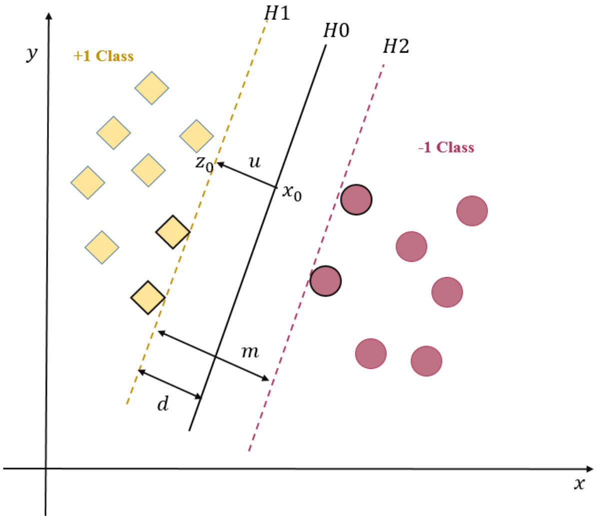

FIGURE 3. Maximizing the margin of a hyperplane

Because a nonlinear hyperplane will typically prefer to fit too closely (overfit) to a boundary that would perfectly segregate the training data but not necessarily the new data, an SVM utilizes a linear hyperplane instead. In other words, by incorrectly predicting the class of fresh data, overfitting the model on training data might result in poor generalization of the classifier (see Bhavsar and Ganatra 2012). Figure 3 's graphic demonstrates that there are no data points in the area enclosed by the two hyperplanes H1 and H2. The items provided at the borders of the two hyperplanes are referred to as "support vectors". The four support vectors (points) in this example are highlighted in black edges in Fig. 3. A linear hyperplane H0 that is equidistant between the two hyperplanes H1 and H2 is the decision boundary. The separation between the two hyperplanes, H1 and H2, is the hyperplane's margin. In other words, the decision boundary's hyperplane H0 is equivalently offset from the hyperplanes H1 and H2. In this framework, in our opinion it is really interesting to point out some mathematical aspects related to an optimization model by a Lagrangian approach normally used in SVM issues. In the sequel we will analyze each model that in our idea can present a collection of remarks related to these mathematical aspects. We consider a Lagrangian approach by introducing a generalized Lagrange functional and duality considering the following:

$$\begin{aligned} &\min_{w} f(w) \\ \text{s.t.}\quad & g_i(w) \leq 0, \quad i = 1,\ldots,n \\ & h_j(w) = 0, \quad j = 1,\ldots,n \end{aligned}$$

This is really a primal problem and technically we could consider a generalized Lagrangian as

$$L(w,\lambda,\bar{\lambda}) = f(w) + \sum_{i=1}^{k} \lambda_i g_i(w) + \sum_{j=1}^{e} \bar{\lambda}_i h_j(w)$$

Another very useful tool in this direction can be the min-max Lagrangian:

$$\max_{\lambda,\bar{\lambda}:\lambda_1\geq 0} L(w,\lambda,\bar{\lambda}) = \max_{\lambda,\lambda:\lambda_1\geq 0}\left( f(w) + \sum_{i=1}^{k} \lambda_i g_i(w) + \sum_{j=1}^{e} \bar{\lambda} h_j(w)\right) =$$

$$= \begin{cases} f(x) \; if \; g_i(w) \leq 0, \quad h_j(w) = 0 \\ \infty \qquad\qquad\qquad \text{otherwise} \end{cases}$$

and hence

$$\min_{w} \max_{\lambda,\bar{\lambda}:\lambda_i\geq 0} L\left(w,\lambda,\bar{\lambda}\right) = \min_{w} \max_{\lambda,\bar{\lambda}:\lambda_i\geq 0} L\left(w,\lambda,\bar{\lambda}\right) = \min_{w} f(w)$$

$$\text{s.t.} \qquad g_i(w) \leq 0, \quad i = 1,\ldots,n$$

$$h_j(w) = 0, \quad j = 1,\ldots,n$$

From this mathematical framework, we can obtain

(1) the primal problem (min-max of a Lagrangian function), *i.e*:

$$\min_{w} \max_{\lambda,\bar{\lambda}:\lambda_i\geq 0} L\left(w,\lambda,\bar{\lambda}\right)$$

(2) its dual form

$$\max_{\lambda,\bar{\lambda}:\lambda_i\geq 0} \min_{w} L\left(w,\lambda,\bar{\lambda}\right)$$

We can perform a comparison between 1 and 2 asking when does the following equality hold:

$$\max_{\lambda,\bar{\lambda}:\lambda_i\geq 0} \min_{w} L\left(w,\lambda,\bar{\lambda}\right) \leq \min_{w} \max_{\lambda,\bar{\lambda}:\lambda_i\geq 0} L\left(w,\lambda,\bar{\lambda}\right)$$

This equality holds when:

- $f$ and $g_i's$ are convex and $h_i's$ are affine;
- $g_i$ is strictly feasible: this means that there exists some $w$ so that $g_i(w) < 0$; under these conditions we obtain the equivalency among the primal and dual problems.

$$\min_{w} f(w)$$

$$\text{s.t.} \quad g_i(w) \leq 0 = \min_{w} \max_{\lambda,\bar{\lambda}:\lambda_i\geq 0\}} L\left(w,\alpha,\bar{\lambda}\right) = \max_{\lambda,\bar{\lambda}:\lambda_i\geq 0} \min_{w} L\left(w,\lambda,\bar{\lambda}\right)$$

$$h_i(w) = 0$$

Primal problem Dual problem

Practically the solution of the primal and dual problems satisfies the KKT conditions:

$$\frac{\partial}{\partial\, w_i}\, L\left(w^*,\, \lambda^*,\, \bar{\lambda}^*\right) = 0 \qquad i = 1, \ldots, n \tag{15}$$

$$\frac{\partial}{\partial \beta_i}\, L\left(w^*,\, \lambda,\, \bar{\lambda}^*\right) = 0 \qquad i = 1, \ldots, l \tag{16}$$

$$\alpha_i^*\; f_i\,(w^*) = 0 \qquad i = 1, \ldots, k \tag{17}$$

$$f_i(w^*) \leq 0 \qquad i = 1, \ldots, k \tag{18}$$

$$\alpha^* \geq 0 \qquad i = 1, \ldots .k \tag{19}$$

From (15) to (19) are some necessary and sufficient conditions and in particular (15) is the stationary environment, (16) is the primal feasibility, (17) represents the complementary conditions,(18) is the primal feasibility, and (19) is the dual feasibility.

**3.1. Lagrangian approach for SVM: the estimation of bias.** Starting from the mathematical structure just introduced, we can consider the following optimization problem for solving an SVM issue:

$$\min_{w,b} \frac{1}{2} ||w||^2$$

$$\text{s.t.} \qquad y^i \left(w^T x^{(i)} + b\right) \geq 1 \qquad i = 1, \ldots, m$$

$$g_i\,(w) = -y^{(i)} \left(w^T x^{(i)} + b\right) + 1 \leq 0 \quad i = 1,\, \ldots m$$

we can build a Lagrangian as

$$L\,(w, b, \lambda) = \frac{1}{2} ||w||^2 - \sum_{i=1}^{m} \lambda_i \left(y^{(i)} \left(w^T x^{(i)} + b\right) - 1\right)$$

$$C(z) := \int_0^z \gamma(s)(1 + m(z - s))ds + \int_z^T \gamma(s) e^{-\delta(s-z)} ds + Re^{-\delta(T-z)}.$$

Now, we can minimize the function taking the first partial derivative:

$$\frac{\partial\; L\,(w, b, \lambda)}{\partial\; w} = w - \sum_{i=1}^{m} \lambda_i y^{(i)} x^{(i)} = 0 \;\Rightarrow\; w = \sum_{i=1}^{m} \lambda\; y^{(i)} x^{(i)}$$

$$\frac{\partial\; L\,(w, b, \lambda)}{\partial\; b} = \sum_{i=1}^{m} \lambda_i y^{(i)} = 0$$

and we obtain

$$L\,(w, b, \lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \lambda_i \lambda_j (x^{(i)})^T x^{(j)} - b\, \sum_{i=1}^{m} \lambda_i\; y^{(i)} =$$

$$= \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \lambda_i \lambda_j (x^{(i)})^T x^{(j)}$$

We can apply this Lagrangian approach to develop some support vector machines procedures. In this direction we can consider a decision function as

$$f(x) = (w^*)^T x + b$$

which mathematically represents the equation related to H1, H2 in Fig. 4. More precisely,

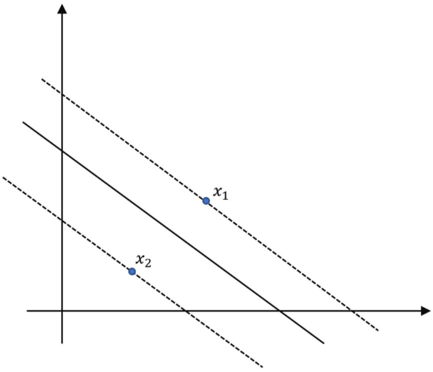

FIGURE 4. Bias solutions

we have

$$f(x) = (\sum_{i=1}^{N} \lambda_i^* y^{(i)} x^{(i)})^T \ x + b^* = \sum_{i=1}^{N} \lambda_i^* y^{(i)} (x^{(i)})^T \ x + b^*$$

The KKT conditions ask for

$$\begin{cases} \alpha_i^* f_i(w^*) = 0 \\ g_i(w^*) \leq 0 \\ \lambda_i^* \geq 0 \end{cases}$$

where $g_i(w) = \ -y^{(i)} \left( w^T x^{(i)} + b \right) + 1$.

Strictly correlated to the problem just presented, it can be useful for analyzing, from a mathematical point of view, the arising of some bias in the model. A parameter in the decision function of a support vector machine (SVM), known as the bias value, denotes the offset from the hyperplane that divides the various classes. The concept of “intercept” is another name for it. To ensure that the decision border does not always pass through the origin, the bias term is introduced into the decision function. The decision boundary’s position and slope can both be impacted by the bias, which can have a positive or negative value.

**Numerical example.** We can determine the value of the bias, where $x_1$ can be a positive support vector and $x_2$ can be a negative one (see Fig. 4).

$$\begin{cases} (w^*)^T x_1 + b = 1 \\ (w^*)^T x_2 + b = -1 \end{cases}$$

$$b^* = -\frac{(w^*)^T x_1 + (w^*)^T x_2}{2} = -\frac{\max_{i:y^{(i)}=-1} (w^*)^T x_i + \min_{i:y^{(i)}=1} (w^*)^T x_i}{2}$$

## 4. Conclusions and future developments

In this paper, new proofs of the classical results of nonlinear programming milestones, such as the Karush-Kuhn-Tucker conditions and Lagrangian methods and functions, have been proposed, opening new research scenarios by utilizing a mixture between mathematical tools and artificial intelligence environment. Artificial intelligence techniques, models, and tools have been developed to increase knowledge in related sciences in enormous fields of applications. In the second part of this work, the main scope was focused on the technical analysis of “bias”, emerging in Support Vector Machine procedures. A special decision function was introduced involving the bias, as its weight could influence the decision making. Therefore, in this paper the Karush-Kuhn-Tucker conditions and the Lagrangian approach have been studied to increase the robustness, efficiency, and accuracy of the classical models. This paper aims to start a sequence of studies on this theme involving the mixture among these fields of knowledge.

## Declarations

The authors have no conflicts of interest to declare that are relevant to the content of this article.

## References

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer. DOI: 10.1007/978-3-319-14142-8.

Andreani, R., Haeser, G., and Martínez, J. M. (2011). “On sequential optimality conditions for smooth constrained optimization”. *Optimization* **60**(5), 627–641. DOI: 10.1080/02331930903578700.

Andreani, R., Martínez, J. M., and Svaiter, B. F. (2010). “A new sequential optimality condition for constrained optimization and algorithmic consequences”. *SIAM Journal on Optimization* **20**(6), 3533–3554. DOI: 10.1137/090777189.

Bhavsar, H. and Ganatra, A. (2012). “Variations of support vector machine classification technique: A survey”. *International Journal of Advanced Computer Research* **2**(6), 223–227.

Chiriaev, D. and Walster, G. W. (1998). *Interval arithmetic specification*. Version March 13, 2000. URL: http://interval.ict.nsc.ru/Programing/IASpecfn.pdf (visited on 11/16/2023).

Cortes, C. and Vapnik, V. (1995). “Support-vector networks”. *Machine Learning* **20**, 273–297. DOI: 10.1007/BF00994018.

Craven, B. D. (1984). “Modified Kuhn-Tucker conditions when a minimum is not attained”. *Operations Research Letters* **3**(1), 47–52. DOI: 10.1016/0167-6377(84)90071-3.

Cusano, C. and La Torre, D. (2003). “On support vector machines for data classification”. In: *Recent Advances in Optimization*. Ed. by G. P. Crespi, A. Guerraggio, E. Miglierina, and M. Rocca. Milano: Guido Tommasi Editore - Datanova, pp. 101–118.

Dutta, J., Deb, K., Tulshyan, R., and Arora, R. (2013). "Approximate KKT points and a proximity measure for termination". *Journal of Global Optimization* **56**(4), 1463–1499. DOI: 10.1007/s10898-012-9920-5.

Fiacco, A. V. and McCormick, G. P. (1967). "The slacked unconstrained minimization technique for convex programming". *SIAM Journal on Applied Mathematics* **15**(3), 505–515. DOI: 10.1137/0115046.

Gentile, C. (2000). "A New Approximate Maximal Margin Classification Algorithm". In: *Advances in Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/2000/file/d072677d210ac4c03ba046120f0802ec-Paper.pdf.

Ghosh, D., Singh, A., Shukla, K. K., and Manchanda, K. (2019). "Extended Karush-Kuhn-Tucker condition for constrained interval optimization problems and its application in support vector machines". *Information Sciences* **504**, 276–292. DOI: 10.1016/j.ins.2019.07.017.

Giorgi, G., Jiménez, B., and Novo, V. (2016). "Approximate Karush–Kuhn–Tucker condition in multiobjective optimization". *Journal of Optimization Theory and Applications* **171**, 70–89. DOI: 10.1007/s10957-016-0986-y.

Haeser, G. and de Melo, V. V. (2013). "Approximate-KKT stopping criterion when Lagrange multipliers are not available". *Optimization Online*. URL: https://optimization-online.org/?p=12284.

Haeser, G. and Schuverdt, M. L. (2011). "On approximate KKT condition and its extension to continuous variational inequalities". *Journal of Optimization Theory and Applications* **149**(3), 528–539. DOI: 10.1007/s10957-011-9802-x.

Hein, M., Bousquet, O., and Schölkopf, B. (2005). "Maximal margin classification for metric spaces". *Journal of Computer and System Sciences* **71**(3), 333–359. DOI: 10.1016/j.jcss.2004.10.013.

Hestenes, M. R. (1975). *Optimization Theory: The Finite Dimensional Case*. New York: Wiley.

Jahn, J. (2017). "Karush–Kuhn–Tucker conditions in set optimization". *Journal of Optimization Theory and Applications* **172**, 707–725. DOI: 10.1007/s10957-017-1066-7.

Jahn, J. and Ha, T. X. D. (2011). "New order relations in set optimization". *Journal of Optimization Theory and Applications* **148**(2), 209–236. DOI: 10.1007/s10957-010-9752-8.

Kortanek, K. O. and Evans, J. P. (1968). "Asymptotic Lagrange regularity for pseudoconcave programming with weak constraint qualification". *Operations Research* **16**(4), 849–857. DOI: 10.1287/opre.16.4.849.

Kuroiwa, D. (1998). "On natural criteria in set-valued optimization (dynamic decision systems under uncertain environments)". *Kyoto University Research Information Repository (KURENAI)* **1048**, 86–92. URL: http://hdl.handle.net/2433/62183.

Kuroiwa, D., Tanaka, T., and Ha, T. X. D. (1997). "On cone convexity of set-valued maps". *Nonlinear Analysis: Theory, Methods & Applications* **30**(3), 1487–1496. DOI: 10.1016/S0362-546X(97)00213-7.

La Torre, D. (2003). "On generalized derivatives for $C^{1,1}$ vector optimization problems". *Journal of Applied Mathematics* **2003**, 365–376. DOI: 10.1155/S1110757X03209049.

La Torre, D. and Vercellis, C. (2002). "C Wedge(1, 1) Approximations of Generalized Support Vector Machines". *UNIMI Department of Economics* (19.2002). DOI: 10.2139/ssrn.616601.

Mangasarian, O. L. (1997). "Mathematical programming in data mining". *Data Mining and Knowledge Discovery* **1**, 183–201. DOI: 10.1023/A:1009735908398.

Mangasarian, O. L. (1998). *Generalized Support Vector Machines*. Tech. rep. 98-14. MINDS@UW Madison. URL: http://digital.library.wisc.edu/1793/64390.

Nachbar, P., Nossek, J. A., and Strobl, J. (1993). "The generalized AdaTron algorithm". In: *1993 IEEE International Symposium on Circuits and Systems*. Vol. 4. Chicago, IL, USA, pp. 2152–2155. DOI: 10.1109/ISCAS.1993.394184.

Neukel, N. (2013). "Order relations of sets and its application in socio-economics". *Applied Mathematical Sciences* **7**(115), 5711–5739. DOI: 10.12988/ams.2013.37419.

Nishnianidze, Z. G. (1984). "Fixed points of monotonic multiple-valued operators". *Bulletin of the Georgian National Academy of Sciences* **114**, 489–491.

Orsenigo, C. and Vercellis, C. (2004). "Discrete support vector decision trees via tabu search". *Computational Statistics & Data Analysis* **47**(2), 311–322. DOI: 10.1016/j.csda.2003.11.005.

Trudzik, L. I. (1982). "Asymptotic Kuhn-Tucker conditions in abstract spaces". *Numerical Functional Analysis and Optimization* **4**(4), 355–369. DOI: 10.1080/01630568208816122.

Vapnik, V. and Chervonenkis, A. Y. (1974). "On the method of ordered risk minimization. I". *Avtomatika i Telemekhanika* **8**, 21–30. URL: https://www.mathnet.ru/eng/at8452.

Xu, L., Neufeld, J., Larson, B., and Schuurmans, D. (2005). "Maximum Margin Clustering". In: *Advances in Neural Information Processing Systems 17*. Ed. by L. Saul, Y. Weiss, and L. Bottou. Vol. 17. Cambridge, MA, USA: The MIT Press, pp. 1537–1544. URL: https://proceedings.neurips.cc/paper_files/paper/2004/hash/6403675579f6114559c90de0014cd3d6-Abstract.html.

Xue, H., Chen, S., and Yang, Q. (2011). "Structural regularized support vector machine: a framework for structural large margin classifier". *IEEE Transactions on Neural Networks* **22**(4), 573–587. DOI: 10.1109/TNN.2011.2108315.

Ye, J. J. and Zhang, J. (2013). "Enhanced Karush–Kuhn–Tucker condition and weaker constraint qualifications". *Mathematical Programming* **139**(1-2) (Computational and Analytical Mathematics), 353–381. DOI: 10.1007/s10107-013-0667-7.

Young, R. C. (1931). "The algebra of many-valued quantities". *Mathematische Annalen* **104**(1), 260–290. DOI: 10.1007/BF01457934.

Zlobec, S. (1971). "Extensions of asymptotic Kuhn–Tucker conditions in mathematical programming". *SIAM Journal on Applied Mathematics* **21**(3), 448–460. DOI: 10.1137/0121047.

[a] Università della Valle d'Aosta,
Dipartimento di Scienze Economiche e Politiche,
Località Grand Chemin 181, 11020 Saint Christophe, Italy

[b] Università Mediterranea di Reggio Calabria,
Dipartimento di Giurisprudenza, Economia e Scienze Umane,
Via dell'Università, 98124 Reggio Calabria, Italy

[*] To whom correspondence should be addressed | email: massimiliano.ferrara@unirc.it

Communicated 19 May 2023; manuscript received 22 August 2023; published online 2 January 2024