The Use of Infinities and Infinitesimals for Sparse Classification Problems

Renato De Leone, Nadaniela Egidi, and Lorella Fatone

Abstract In this chapter we discuss the use of *grossone* and the new approach to infinitesimal and infinite proposed by Sergeyev in determining sparse solutions for special classes of optimization problems. In fact, in various optimization and regression problems, and in solving overdetermined systems of linear equations it is often necessary to determine a sparse solution, that is a solution with as many as possible zero components. Expanding on the results in [16], we show how continuously differentiable concave approximations of the l_0 pseudo–norm can be constructed using *grossone*, and discuss the properties of some new approximations. Finally, we will conclude discussing some applications in elastic net regularization and Sparse Support Vector Machines.

1 Introduction

In many optimization problems, in regression methods and when solving over-determined systems of equations, it is often necessary to determine a

R. De Leone (✉) · N. Egidi · L. Fatone
School of Science and Technology, University of Camerino, Camerino, MC, Italy
e-mail: renato.deleone@unicam.it

N. Egidi
e-mail: nadaniela.egidi@unicam.it

L. Fatone
e-mail: lorella.fatone@unicam.it

Y. D. Sergeyev and R. De Leone (eds.), *Numerical Infinities and Infinitesimals in Optimization*, Emergence, Complexity and Computation 43,
https://doi.org/10.1007/978-3-030-93642-6_6

sparse solution, that is, a solution with the minimum number of nonzero components. This kind of problems are known as sparse approximation problems and arise in different fields. In Machine Leaning, the Feature Extraction problem requires, for a given problem, to eliminate as many features as possible, while still maintaining a good accuracy in solving the assigned task (for example, a classification task). Sparse solutions are also required in signal/image processing problem, for example in sparse approximation of signals, image denoising, etc. [4, 12, 36].

In these cases the l_0 pseudo-norm is utilized. This pseudo-norm counts the number of nonzero elements of a vector. Problems utilizing the l_0 pseudo-norm have been considered by many researchers, but they seem "to pose many conceptual challenges that have inhibited its widespread study and application" [4]. Moreover, the resulting problem is NP–hard and, in order to construct a more tractable problem, various continuously differentiable concave approximations of the l_0 pseudo-norm are used, or the l_0 pseudo-norm is replaced by the simpler to handle 1–norm. In [27] two smooth approximations of the l_0 pseudo-norm are proposed in order to determine a vector that has the minimum l_0 pseudo-norm.

Recently, Sergeyev proposed a new approach to infinitesimals and infinities[1] based on the numeral ①, the number of elements of $\mathbb{N}$, the set of natural numbers. It is crucial to note that ① is not a symbol and is not used to perform symbolic calculations. In fact, the ① is a natural number, and it has both cardinal and ordinal properties, exactly as the "standard", finite natural numbers. Moreover, the new proposed approach is different from non–Standard Analysis, as demonstrated in [33]. A comprehensive description of the grossone–based methodology can also be found in [32].

The use of ① and the new approach to infinite and infinitesimals has been beneficial in several fields of pure and applied mathematics including optimization [6–9, 14, 15, 17–19, 23], numerical differentiation [29], ODE [1, 22, 34], hyperbolic geometry [25], infinite series and the Riemann zeta function [28, 30], biology [31], and cellular automata [13].

Moreover, this new computational methodology has been also utilized in the field of Machine Learning allowing to construct new spherical separations for classification problems [2], and novel sparse Support Vector Machines (SSVMs) [16].

In this chapter we discuss the use of ① to obtain new approximations for the l_0 pseudo-norm, and two applications are considered in detail. More specifically, the chapter is organized as follows. In Sect. 2 some of the most utilized smooth approximations of the l_0 pseudo–norm proposed in the literature are

[1] See Chap. 1 for an in–depth description of the properties of the new system and its advantages

discussed. Then, in the successive Sect. 3 it is shown how to utilize ① for constructing approximations of the l_0 pseudo-norm. Finally, in Sect. 4, mostly based on [16], two relevant applications of the newly proposed approximation scheme for the l_0 pseudo norm are discussed in detail: the elastic net regulation problem and sparse Support Vector Machines.

We briefly describe our notation now. All vectors are column vectors and will be indicated with lower case Latin letter (i.e. $x, y, \ldots$). Subscripts indicate components of a vector, while superscripts are used to identify different vectors. Matrices will be indicated with upper case Roman letter (i.e. A, B, $\ldots$). The set of natural and real numbers will be denoted, respectively, by $\mathbb{N}$ and $\mathbb{R}$. The space of the n–dimensional vectors with real components will be indicated by $\mathbb{R}^n$. Superscript ^{T} indicates transpose. The scalar product of two vectors x and y in $\mathbb{R}^n$ will be denoted by $x^T y$. Instead, for a generic Hilbert space, the scalar product of two elements x and y will be indicated by $\langle x, y\rangle$. The Euclidean norm of a vector x will be denoted by $\|x\|$. The space of the $m \times n$ matrices with real components will be indicated by $\mathbb{R}^{m\times n}$. For a $m \times n$ matrix A, A_{ij} is the element in the ith row, jth column.

In the new positional numeral system with base ①, a *gross-scalar* (or *gross–number*) C has the following representation:

$$C = C^{(p_m)}①^{p_m} + \cdots + C^{(p_1)}①^{p_1} + C^{(p_0)}①^{p_0} + C^{(p_{-1})}①^{p_{-1}} + \cdots + C^{(p_{-k})}①^{p_{-k}}, \tag{1}$$

where $m, k \in \mathbb{N}$, for $i = -k, -k+1, \ldots, -1, 0, 1, \ldots, m-1, m$, the quantities $C^{(p_i)}$ are floating-point numbers and p_i are gross-numbers such that

$$p_m > p_{m-1} > \cdots > p_1 > p_0 = 0 > p_{-1} > \cdots > p_{-k+1} > p_{-k}. \tag{2}$$

If $m = k = 0$ the gross-number C is called finite; if $m > 0$ it is called infinite; if $m = 0$, $C^{(p_0)} = 0$ and $k > 0$ it is called infinitesimal; the exponents p_i, $i = -k, -k+1, \ldots, -1, 0, 1, \ldots, m-1, m$, are called *gross-powers*.

2 The l_0 Pseudo-norm in Optimization Problems

Given a vector $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$, the l_0 pseudo-norm of x is defined as the number of its components different from zero, that is:

$$\|x\|_0 = \text{number of nonzero components of } x = \sum_{i=1}^{n} 1_{x_i}, \tag{3}$$

where 1_a is the characteristic (indicator) function, that is, the function which is equal to 1 if $a \neq 0$ and zero otherwise.

Note that $\|\cdot\|_0$ is not a norm and hence is called, more properly, pseudo-norm. In fact, for a non zero vector $x \in \mathbb{R}^n$, and a not null constant $\lambda \in \mathbb{R}$, we have:

$$\|\lambda x\|_0 = \|x\|_0 .$$

Consequently $\|\lambda x\|_0 = |\lambda| \, \|x\|_0$, $\lambda \in \mathbb{R}$, if and only if $|\lambda| = 1$.

The l_0 pseudo-norm plays an important role in several numerical analysis and optimization problems, where it is important to get a vector with as few non-zero components as possible. For example, this pseudo-norm has important applications in elastic-net regularization, pattern recognition, machine learning, signal processing, subset selection problem in regression and portfolio optimization. For example, in signal and image processing many media types can be sparsely represented using transform-domain methods, and sparsity of the representation is fundamental in many highly used techniques of compression (see [4] and references therein). In [20, 26] the cardinality-constrained optimization problem is studied and opportunely reformulated. In [5] the general optimization problem with cardinality constraints has been reformulated as a smooth optimization problem.

The l_0 pseudo-norm is strongly related to the l_p norms. Given a vector $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$, the l_p norm of x is defined as

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} .$$

It is not too difficult to show that

$$\|x\|_0 = \lim_{p \to 0} \|x\|_p^p = \lim_{p \to 0} \sum_{i=1}^{n} |x_i|^p .$$

In Fig. 1 the behavior of $|\sigma|^p$ ($\sigma \in \mathbb{R}$) for different values of p is shown (see also [4]). Note that, as the figure suggests, for $0 < p < 1$, the function $\|x\|_p$ is a concave function.

It must be noted that the use of $\|x\|_0$ makes the problems extremely complicated to solve, and various approximations of the l_0 pseudo-norm have been proposed in the scientific literature, For example, in [27] two smooth approximations of the l_0 pseudo-norm are proposed in order to determine a particular vector that has the minimum l_0 pseudo-norm.

In [24], in the framework of elastic net regularization, the following approximation of $\|x\|_0$ is studied:

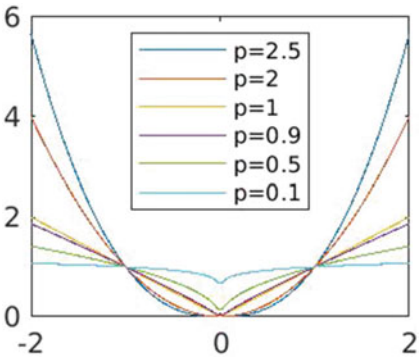

Fig. 1 The value of $|\sigma|^p$ for different values of p

$$\|x\|_{0,\delta} := \sum_{i=1}^{n} \frac{x_i^2}{x_i^2 + \delta}, \tag{4}$$

where $\delta \in \mathbb{R},\ \delta > 0$, and a small δ is suggested in order to provide a better approximation of $\|x\|_0$.

Instead, in the context of Machine Learning and Feature Selection [3], the following approximation of $\|x\|_0$:

$$\|x\|_{0,\alpha} := \sum_{i=1}^{n} \left(1 - e^{-\alpha|x_i|}\right), \tag{5}$$

where $\alpha \in \mathbb{R}$, $\alpha > 0$, is proposed, and the value $\alpha = 5$ is recommended.

By using ①, in [16] a new approximation of $\|x\|_0$ has been suggested. In the next section we discuss in detail this approximation and we also propose other approximations that use the new numeral system based on ①. Moreover, we provide the connections between $\|x\|_0$ and the new approximations.

Note that the approximation introduced in [16] has been used in connection to two different applications. The first application is an elastic net regularization. The second application concerns classification problems using sparse Support Vector Machines. These two applications are extensively reviewed in Sect. 4.

3 Some Approximations of the l_0 Pseudo-norm Using ①

The first approximation of the l_0 pseudo-norm in terms of ① was proposed in [16], where, following the idea suggested in [24] of approximating the l_0 pseudo-norm
by (4), the following approximation has been suggested:

$$\|x\|_{0,①,1} := \sum_{i=1}^{n} \frac{x_i^2}{x_i^2 + ①^{-1}}. \tag{6}$$

In this case, we have that

$$\|x\|_{0,①,1} = \|x\|_0 + C①^{-1}, \tag{7}$$

for some gross-number C which includes only finite and infinitesimal terms. Therefore, the finite parts of $\|x\|_0$ and $\|x\|_{0,①,1}$ coincide.

To this scope, let

$$\psi_1(t) = \frac{t^2}{t^2 + ①^{-1}}, \quad t \in \mathbb{R}. \tag{8}$$

We have that $\psi_1(0) = 0$ and $\psi_1(t) = 1 - ①^{-1}S$, when $t \neq 0$, where S is a gross-number such that

$$0 < S = \frac{1}{t^2 + ①^{-1}} < \frac{1}{t^2}.$$

Therefore, S has only finite and infinitesimal terms. Moreover,

$$\sum_{i=1}^{n} \frac{x_i^2}{x_i^2 + ①^{-1}} = \sum_{i=1}^{n} \psi_1(x_i) = \sum_{i=1, x_i \neq 0}^{n} \psi_1(x_i) = \|x\|_0 + C①^{-1}, \tag{9}$$

where

$$C = \begin{cases} -\displaystyle\sum_{i=1, x_i \neq 0}^{n} S_i, & \text{when } \|x\|_0 \neq 0, \\ 0, & \text{otherwise,} \end{cases}$$

and S_i is a gross-number such that

$$0 < S_i = \frac{1}{x_i^2 + ①^{-1}} < \frac{1}{x_i^2}, \quad x_i \neq 0, \quad i = 1, \ldots, n.$$

Hence C is a gross-number with only finite and infinitesimal terms and the finite part of of $\|x\|_0$ and $\|x\|_{0,①,1}$ are the same.

A different proof of this result is provided in [16]. For $i = 1, \ldots, n$, let assume that

$$x_i = x_i^{(0)} + R_i ①^{-1},$$

where R_i includes only finite and infinitesimal terms.

When $x_i^{(0)} = 0$:

$$\psi_1(x_i) = \frac{R_i^2 ①^{-2}}{R_i^2 ①^{-2} + ①^{-1}} = ①^{-1} \frac{R_i^2}{R_i^2 ①^{-1} + 1} = 0\, ①^0 + R_i' ①^{-1},$$

where R_i' includes only finite and infinitesimal terms.

When, instead, $x_i^{(0)} \neq 0$:

$$\psi_1(x_i) = \frac{\left(x_i^{(0)} + R_i ①^{-1}\right)^2}{\left(x_i^{(0)} + R_i ①^{-1}\right)^2 + ①^{-1}} = 1 - \frac{①^{-1}}{\left(x_i^{(0)} + R_i ①^{-1}\right)^2 + ①^{-1}} = 1 + R_i' ①^{-1},$$

where, again, R_i' includes only finite and infinitesimal terms. Therefore,

$$\|x\|_{0,①,1} = \sum_{i=1}^{n} \psi_1(x_i) = \|x\|_0 + S ①^{-1}$$

where S includes only finite and infinitesimal terms and hence $\|x\|_{0,①,1}$ and $\|x\|_0$ coincide in their finite part.

Following the idea suggested in [3], we now propose three novel approximation schemes of the l_0 pseudo-norm all based on the use of ①. In [3] the authors proposed to approximate the l_0 pseudo-norm using (5) and suggest to take a fixed value for α, i.e. $\alpha = 5$, or an increasing sequence of values of α.

Based on this idea, we propose the following approximation formula for $\|x\|_0$:

$$\|x\|_{0,①,2} := \sum_{i=1}^{n} \left(1 - ①^{-\alpha |x_i|}\right), \quad \alpha > 0. \tag{10}$$

Also in this case, the finite parts of $\|x\|_0$ and $\|x\|_{0,①,2}$ coincide. More precisely, let us show that

$$\|x\|_{0,①,2} = \|x\|_0 - ①^{-\alpha m_x} C, \tag{11}$$

where

$$m_x = \begin{cases} \min\left\{|x_i| \ : \ x_i \neq 0\right\}, & \text{when } x \neq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

and C is a gross-number which is null when $\|x\|_0 = 0$ and, otherwise, includes only finite and infinitesimal terms.

Let us define

$$\psi_2(t) = 1 - ①^{-\alpha|t|}, \quad t \in \mathbb{R}. \tag{13}$$

Since $\psi_2(0) = 0$, we have:

$$\begin{aligned} \|x\|_{0,①,2} &= \sum_{i=1}^{n} \left(1 - ①^{-\alpha|x_i|}\right) = \\ &= \sum_{i=1}^{n} \psi_2(x_i) = \sum_{i=1, x_i \neq 0}^{n} \psi_2(x_i) = \\ &= ||x||_0 - \sum_{i=1, x_i \neq 0}^{n} ①^{-\alpha|x_i|} = ||x||_0 - ①^{-\alpha m_x} C. \end{aligned} \tag{14}$$

It is easy to see that $C = 0$ when $\|x\|_0 = 0$. Instead, if $\|x\|_0 \neq 0$ then C has only finite and infinitesimal terms. This shows that $\|x\|_0$ and $\|x\|_{0,①,2}$ coincide in their finite part.

Another approximation of the l_0 pseudo–norm is given by:

$$\|x\|_0 \approx \|x\|_{0,①,3} := \sum_{i=1}^{n} \left(1 - e^{-①|x_i|}\right). \tag{15}$$

In this case, it is possible to show that

$$\|x\|_{0,①,3} = \|x\|_0 - e^{-①m_x} C, \tag{16}$$

where, as in the previous cases, C is a gross-number which includes only finite and infinitesimal terms and is null when $\|x\|_0 = 0$. Hence, also in this case we have that the finite parts of $\|x\|_0$ and $\|x\|_{0,①,3}$ coincide.

To prove the above result, let

$$\psi_3(t) = 1 - e^{-①|t|}, \quad t \in \mathbb{R}. \tag{17}$$

Since $\psi_3(0) = 0$, we have:

$$\begin{aligned} \|x\|_{0,①,3} &= \sum_{i=1}^{n} \left(1 - e^{-①|x_i|}\right) = \\ &= \sum_{i=1}^{n} \psi_3(x_i) = \sum_{i=1,x_i\neq 0}^{n} \psi_3(x_i) = \\ &= ||x||_0 - \sum_{i=1,x_i\neq 0}^{n} e^{-①|x_i|} = ||x||_0 - e^{-①m_x} C, \end{aligned} \tag{18}$$

where m_x is defined in (12) and C is a gross-number with only finite and infinitesimal terms. Moreover, C is null when $\|x\|_0 = 0$.

Finally, another approximation of the l_0 pseudo–norm, always in the spirit of (5), is given by:

$$\|x\|_{0,①,4} := \sum_{i=1}^{n} \left(1 - ①^{-①|x_i|}\right). \tag{19}$$

In this last case, let

$$\psi_4(t) = 1 - ①^{-①|t|}, \quad t \in \mathbb{R}. \tag{20}$$

Since even in this circumstance $\psi_4(0) = 0$, we have:

$$\begin{aligned} \|x\|_{0,①,4} &= \sum_{i=1}^{n} \left(1 - ①^{-①|x_i|}\right) = \\ &= \sum_{i=1}^{n} \psi_4(x_i) = \sum_{i=1,x_i\neq 0}^{n} \psi_4(x_i) = \\ &= ||x||_0 - \sum_{i=1,x_i\neq 0}^{n} ①^{-①|x_i|} = ||x||_0 - ①^{-①m_x} C, \end{aligned} \tag{21}$$

where m_x is again defined in (12) and C is a gross-number with only finite and infinitesimal terms that is null when $||x||_0 = 0$. As in the previous cases we have that

$$\|x\|_{0,①,4} = \|x\|_0 - ①^{-①m_x}C, \tag{22}$$

and, therefore, the finite parts of $\|x\|_0$ and $\|x\|_{0,①,4}$ coincide.

We have presented a number of different approximation schemes for the l_0 pseudo-norm. We want to stress that in all the cases the value of $\|x\|_0$ and its approximation coincide in their finite part and may only differ for infinitesimals quantities.

In the next section we will discuss some utilization of these approximating schemes in two extremely important problems: regularization and classification.

4 Applications in Regularization and Classification Problems

In this section we review some interesting uses of the proposed l_0 pseudo–norm approximations in two classes of optimization problems: elastic net regularization problems and sparse Support Vector Machine classification problems. These two applications are deeply studied in [16].

4.1 Elastic Net Regularization

There are many important applications where we want to determine a solution $x \in \mathbb{R}^n$ of a given linear system $Ax = b$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, such that x has the smallest number of nonzero components, that is

$$\begin{aligned} &\min_x && \|x\|_0\,, \\ &\text{subject to} && Ax = b. \end{aligned}$$

To this problem it is possible to associate the following generalized elastic net regularization:

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda_0\|x\|_0 + \frac{\lambda_2}{2}\|x\|_2^2\,, \tag{23}$$

where $\lambda_0 > 0$ and $\lambda_2 > 0$ are two regularization parameters (see [24] for details).

In [24] a suitable algorithm for the solution of Problem (23) with $\|x\|_{0,\delta}$ (defined in (4)) instead of $\|x\|_0$ is proposed. The corresponding solution approximates the solution of (23) and depends on the choice of $\delta > 0$ in (4).

Following the idea suggested in [24], we look for the solution of the following minimization problem:

$$\min_x f_1(x), \tag{24}$$

where

$$f_1(x) := \frac{1}{2}\|Ax - b\|_2^2 + \lambda_0 \|x\|_{0,①,1} + \frac{\lambda_2}{2}\|x\|_2^2 . \tag{25}$$

Note that Problem (24)–(25) is obtained from Problem (23) by substituting $\|x\|_{0,①,1}$ to $\|x\|_0$.

In [16] we proved that the corresponding solution coincides with the solution of the original Problem (23) apart from infinitesimal terms. In particular, in [16], the following iterative scheme for the solution of Problem (24)–(25) has been proposed: given an initial value $x^0 \in \mathbb{R}^n$, for $k = 0, 1, \ldots$, compute x^{k+1} by solving

$$\left(A^T A + \lambda_2 I + \lambda_0 D(x^k)\right) x^{k+1} = A^T b, \tag{26}$$

where $I \in \mathbb{R}^{n\times n}$ is the identity matrix and $D \in \mathbb{R}^{n\times n}$ is the following diagonal matrix:

$$D_{ii}(x) = \frac{2①^{-1}}{\left((x_i)^2 + ①^{-1}\right)^2}, \qquad D_{ij}(x) = 0, \ i \neq j. \tag{27}$$

The convergence of the sequence $\{x^k\}$ to the solution of Problem (24)–(25) is ensured by Theorem 1 in [16]. In particular, when $\mathcal{L} := \{x : f_1(x) \leq f_1(x^0)\}$ is a compact set, the above constructed sequence $\{x^k\}_k$ has at least one accumulation point, $x_k \in \mathcal{L}$ for each $k = 1, \ldots$, and each accumulation point of $\{x^k\}_k$ belongs to $\mathcal{L}$ and is a stationary point of f_1.

In [24] a similar algorithm was proposed, where $\|x\|_0$ was substituted by (4). However, in this latter case, the quality of the final solution (in terms of being also a solution of Problem (24)–(25) strongly depends on the value of δ that is utilized. In our approach, instead, taking into account that $\|x\|_0$ and our approximation with ① only differ for infinitesimal terms, the final solution solves also Problem (24)–(25).

The results presented here are relative to the first of the four approximation schemes for $\|x\|_0$ discussed in Sect. 3.

Considering the minimization of the following functions

$$f_i(x) := \frac{1}{2}\|Ax - b\|_2^2 + \lambda_0 \|x\|_{0,①,i} + \frac{\lambda_2}{2}\|x\|_2^2, \tag{28}$$

with $i = 2$ or $i = 3$ or $i = 4$, and computing the corresponding first order optimality conditions, new iterative schemes similar to (26) can be obtained and studied.

4.2 Sparse Support Vector Machines

The grossone ① and the different approximations of l_0-pseudo norm can be also used in Sparse Support Vector Machines.

Given empirical data (training set) (x^i, y_i), $i = 1, \ldots, l$, with inputs $x^i \in \mathbb{R}^n$, and outputs $y_i \in \{-1, 1\}$, $i = 1, \ldots, l$, we want to compute a vector $w \in \mathbb{R}^n$ and a scalar θ (and hence an hyperplane) such that:

$$\begin{aligned} w^T x^i + \theta > 0 \text{ when } y_i = 1, \\ w^T x^i + \theta < 0 \text{ when } y_i = -1. \end{aligned}$$

The classification function is

$$h(x) = \text{sign}\left(w^T x + \theta\right).$$

Given

$$\phi : \mathbb{R}^n \mapsto \mathcal{E},$$

where $\mathcal{E}$ is an Hilbert space with scalar product $\langle \cdot, \cdot \rangle$, the optimal hyperplane can be constructed by solving the following (primal) optimization problem (see [10, 11, 35] and references therein for details):

$$\begin{array}{ll} \min\limits_{w,\theta,\xi} & \frac{1}{2}\langle w, w\rangle + Ce^T\xi, \\ \text{subject to} & y_i\left(\langle w, \phi(x^i)\rangle + \theta\right) \geq 1 - \xi_i, \quad i = 1, \ldots, l, \\ & \xi_i \geq 0, \quad i = 1, \ldots, l, \end{array} \tag{29}$$

where $e \in \mathbb{R}^l$ is a vector with all elements equal to 1 and C is a positive scalar.

The dual of (29) is

$$\begin{array}{ll} \min\limits_{\alpha} & \frac{1}{2}\alpha^T Q\alpha - e^T\alpha, \\ \text{subject to} & y^T\alpha = 0, \\ & 0 \leq \alpha \leq Ce, \end{array} \tag{30}$$

where

$$Q_{ij} = y_i y_j K_{ij}, \quad K_{ij} = K(x^i, x^j) := \left\langle \phi(x^i), \phi(x^j) \right\rangle, \quad i, j = 1, \ldots, l,$$

and $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is the kernel function.

We note that, this dual problem and the classification function depend only on $K_{ij} = \left\langle \phi(x^i), \phi(x^j) \right\rangle$. In fact, from the Karush–Kuhn–Tucker conditions we have

$$w = \sum_{i=1}^{l} \alpha_i y_i \phi(x^i), \tag{31}$$

and the classification function reduces to

$$h(x) = \text{sign}\Big(\langle w, \phi(x) \rangle + \theta \Big) = \text{sign}\left(\sum_{i=1}^{l} \alpha_i y_i \left\langle \phi(x^i), \phi(x) \right\rangle + \theta \right).$$

In [21], the authors consider an optimization problem based on (29) where $\frac{1}{2} \langle w, w \rangle$ is replaced with $\|\alpha\|_0$ (and, then, this term is approximated by $\frac{1}{2}\alpha \Lambda \alpha$ for opportune values of a diagonal matrix Λ) and use the expansion (31) of w in terms of α.

Furthermore, in [16] the quantity $\|\alpha\|_0$ is replaced by $\|\alpha\|_{0,①,1}$, and the following ①–Sparse SVM problem is defined:

$$\begin{aligned} \min_{\alpha,\theta,\xi} \quad & \tfrac{①}{2} \|\alpha\|_{0,①,1} + Ce^T \xi, \\ \text{subject to } \; & y_i \Big[{K_{i.}}^T \alpha + \theta \Big] \geq 1 - \xi_i, \quad i = 1, \ldots, l, \\ & \xi \geq 0, \end{aligned} \tag{32}$$

where $K_{i.}$ denotes the column vector that corresponds to the ith row of the matrix K.

The algorithmic scheme, originally proposed in [21] and revised in [16], starting from $\lambda_r^0 = 1$, $r = 1, \ldots, l$, requires, at each iteration, the solution of the following optimization problem:

$$\begin{aligned} \min_{\alpha,\theta,\xi} \quad & \frac{1}{2} \sum_{r=1}^{l} \lambda_r^k \alpha_r^2 + Ce^T \xi, \\ \text{subject to } \; & y_i \Big[{K_{i.}}^T \alpha + \theta \Big] \geq 1 - \xi_i, \quad i = 1, \ldots, l, \\ & \xi \geq 0, \end{aligned} \tag{33}$$

and then the update of λ^k with a suitable formula.

From the Karush–Kuhn–Tucker conditions for Problem (32), it follows that

$$\frac{1}{\left(\alpha_r^2 + ①^{-1}\right)^2}\alpha_r = \bar{K}_{r\cdot}^T \beta, \quad r = 1, \ldots, l, \tag{34}$$

where $\bar{K}_{r\cdot}$ is the r–th row of the matrix $\bar{K}$ with $\bar{K}_{rj} = y_j K_{jr}$, for $r, j = 1, \ldots, l$.

The Conditions (34) above suggest the more natural updating formula:

$$\lambda_r^{k+1} = \frac{1}{\left(\alpha_r^2 + ①^{-1}\right)^2}, \quad r = 1, \ldots, l. \tag{35}$$

Moreover, by considering the expansion of the gross-number α, it is easy to verify that formula (35) well mimics the updating formulas for λ^k proposed in [21], also providing a more sound justification for the updating scheme.

We note that the algorithm proposed in [16], and briefly described here, is based on the first of the approximations of $\|\alpha\|_0$ discussed in Sect. 3. Using the other different approximations introduced in the same section, new different updating formulas for λ^{k+1} can be obtained.

5 Conclusions

The use of the l_0 pseudo–norm is pervasive in optimization and numerical analysis, where a sparse solution is often required. Using the new approach to infinitesimal and infinite proposed by Sergeyev, four different approximations of the l_0 pseudo–norm are presented in this chapter. In all cases, we proved that the finite value of the l_0 pseudo–norm and and its approximation coincide, being different only for infinitesimal terms. The use of such approximations is beneficial in many applications, where the discontinuity due to the use of the l_0 pseudo–norm is easily eliminated, by using one of the four proposed approaches presented in this chapter.

References

1. Amodio, P., Iavernaro, F., Mazzia, F., Mukhametzhanov, M.S., Sergeyev, Y.D.: A generalized Taylor method of order three for the solution of initial value problems in standard and infinity floating-point arithmetic. Math. Comput. Simul. **141**, 24–39 (2017)
2. Astorino, A., Fuduli, A.: Spherical separation with infinitely far center. Soft. Comput. **24**(23), 17751–17759 (2020)
3. Bradley, P.S., Mangasarian, O.L.: Feature selection via concave minimization and support vector machines. In: Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98, pp. 82–90. Morgan Kaufmann Publishers Inc., San Francisco (1998)
4. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**(1), 34–81 (2009)
5. Burdakov, O., Kanzow, C., Schwartz, A.: Mathematical programs with cardinality constraints: reformulation by complementarity-type conditions and a regularization method. SIAM J. Optim. **26**(1), 397–425 (2016)
6. Cococcioni, M., Cudazzo, A., Pappalardo, M., Sergeyev, Y.D.: Solving the lexicographic multi-objective mixed-integer linear programming problem using branch-and-bound and grossone methodology. Commun. Nonlinear Sci. Numer. Simul. **84**, 105177 (2020)
7. Cococcioni, M., Fiaschi, L.: The Big-M method with the numerical infinite M. Optim. Lett. **15**, 2455–2468 (2021)
8. Cococcioni, M., Pappalardo, M., Sergeyev, Y.D.: Towards lexicographic multi-objective linear programming using grossone methodology. In: Sergeyev, Y.D., Kvasov, D.E., Dell'Accio, F., Mukhametzhanov, M.S., (eds.) Proceedings of the 2nd International Conference "Numerical Computations: Theory and Algorithms", vol. 1776, p. 090040. AIP Publishing, New York (2016)
9. Cococcioni, M., Pappalardo, M., Sergeyev, Y.D.: Lexicographic multi-objective linear programming using grossone methodology: Theory and algorithm. Appl. Math. Comput. **318**, 298–311 (2018)
10. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
11. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press (2000)
12. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Trans. Image Process. **16**(8), 2080–2095 (2007)
13. D'Alotto, L.: Cellular automata using infinite computations. Appl. Math. Comput. **218**(16), 8077–8082 (2012)
14. De Cosmis, S., De Leone, R.: The use of grossone in mathematical programming and operations research. Appl. Math. Comput. **218**(16), 8029–8038 (2012)
15. De Leone, R.: Nonlinear programming and grossone: quadratic programming and the role of constraint qualifications. Appl. Math. Comput. **318**, 290–297 (2018)
16. De Leone, R., Egidi, N., Fatone, L.: The use of grossone in elastic net regularization and sparse support vector machines. Soft. Comput. **23**(24), 17669–17677 (2020)
17. De Leone, R., Fasano, G., Roma, M., Sergeyev, Y.D.: How Grossone Can Be Helpful to Iteratively Compute Negative Curvature Directions. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11353, pp. 180–183 (2019)

18. De Leone, R., Fasano, G., Sergeyev, Y.D.: Planar methods and grossone for the conjugate gradient breakdown in nonlinear programming. Comput. Optim. Appl. **71**(1), 73–93 (2018)
19. Gaudioso, M., Giallombardo, G., Mukhametzhanov, M.S.: Numerical infinitesimals in a variable metric method for convex nonsmooth optimization. Appl. Math. Comput. **318**, 312–320 (2018)
20. Gotoh, J., Takeda, A., Tono, K.: DC formulations and algorithms for sparse optimization problems. Math. Program. **169**, 141–176 (2018)
21. Huang, K., Zheng, D., Sun, J., Hotta, Y., Fujimoto, K., Naoi, S.: Sparse learning for support vector classification. Pattern Recogn. Lett. **31**(13), 1944–1951 (2010)
22. Iavernaro, F., Mazzia, F., Mukhametzhanov, M.S., Sergeyev, Y.D.: Computation of higher order lie derivatives on the infinity computer. J. Computat. Appl. Math. **383** (2021)
23. Lai, L., Fiaschi, L., Cococcioni, M.: Solving mixed Pareto-Lexicographic multiobjective optimization problems: the case of priority chains. Swarm Evolut. Comput. **55**, 100687 (2020)
24. Li, S., Ye, W.: A generalized elastic net regularization with smoothed l_0 penalty. Adv. Pure Math. **7**, 66–74 (2017)
25. Margenstern, M.: An application of grossone to the study of a family of tilings of the hyperbolic plane. Appl. Math. Comput. **218**(16), 8005–8018 (2012)
26. Pham Dinh, T., Le Thi, H.A.: Recent advances in DC programming and DCA. In: Nguyen, N.T., Le Thi, H.S., (eds.) Transactions on Computational Intelligence XIII. Lecture Notes in Computer Science, vol. 8342. Springer (2014)
27. Rinaldi, F., Schoen, F., Sciandrone, M.: Concave programming for minimizing the zero-norm over polyhedral sets. Comput. Optim. Appl. **46**, 467–486 (2010)
28. Sergeyev, Y.D.: Numerical point of view on calculus for functions assuming finite, infinite, and infinitesimal values over finite, infinite, and infinitesimal domains. Nonlinear Anal. Seri. A: Theory, Methods Appl. **71**(12), e1688–e1707 (2009)
29. Sergeyev, Y.D.: Higher order numerical differentiation on the infinity computer. Optim. Lett. **5**(4), 575–585 (2011)
30. Sergeyev, Y.D.: On accuracy of mathematical languages used to deal with the Riemann zeta function and the Dirichlet eta function. p-Adic numbers. Ultrametric Anal. Appl. **3**(2), 129–148 (2011)
31. Sergeyev, Y.D.: Using blinking fractals for mathematical modelling of processes of growth in biological systems. Informatica **22**(4), 559–576 (2011)
32. Sergeyev, Y.D.: Numerical infinities and infinitesimals: methodology, applications, and repercussions on two Hilbert problems. EMS Surv. Math. Sci. **4**(2), 219–320 (2017)
33. Sergeyev, Y.D.: Independence of the grossone-based infinity methodology from nonstandard analysis and comments upon logical fallacies in some texts asserting the opposite. Found. Sci. **24**(1), 153–170 (2019)
34. Sergeyev, Y.D., Mukhametzhanov, M.S., Mazzia, F., Iavernaro, F., Amodio, P.: Numerical methods for solving initial value problems on the infinity computer. Int. J. Unconv. Comput. **12**(1), 3–23 (2016)
35. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Stat. Comput. **14**(3), 199–222 (2004)
36. Stanković, L., Sejdić, E., Stanković, S., Daković, M., Orović, I.: A tutorial on sparse signal reconstruction and its applications in signal processing. Circuits Syst. Signal Process. **38**(3), 1206–1263 (2019)