**Soft Computing manuscript No.**
(will be inserted by the editor)

# The use of Grossone in Elastic Net Regularization and Sparse Support Vector Machines

**Renato De Leone · Nadaniela Egidi · Lorella Fatone**

Received: date / Accepted: date

**Abstract** New algorithms for the numerical solution of optimization problems involving the $l_0$ pseudo-norm are proposed. They are designed to use a recently proposed computational methodology that is able to deal numerically with finite, infinite and infinitesimal numbers. This new methodology introduces an infinite unit of measure expressed by the numeral ① (*grossone*) and indicating the number of elements of the set $\mathbb{N}$, of natural numbers. We show how the numerical system built upon ① and the proposed approximation of the $l_0$ pseudo-norm in terms of ① can be successfully used in the solution of elastic net regularization problems and sparse Support Vector Machines classification problems.

**Keywords** Elastic Net Regularization · Grossone · Sparse Support Vector Machines

**Mathematics Subject Classification (2010)** MSC 90C30 · MSC 90C90 · MSC 65K05

R. De Leone
School of Science and Technology, University of Camerino, Camerino, Italy
Tel.: +39-0737-402532
Fax: +39-0737-632525
E-mail: renato.deleone@unicam.it

N. Egidi
School of Science and Technology, University of Camerino, Camerino, Italy
Tel.: +39-0737-402507
Fax: +39-0737-632525
E-mail: nadaniela.egidi@unicam.it

L. Fatone
School of Science and Technology, University of Camerino, Camerino, Italy
Tel.: +39-0737-402558
Fax: +39-0737-632525
E-mail: lorella.fatone@unicam.it

## 1 Introduction

Given a vector $x$ of $n$ components, the $l_0$ pseudo-norm

$$\|x\|_0 := \text{number of nonzero components in } x,$$

has often been used in optimization problems arising in various fields. However, the introduction of $\|x\|_0$ makes these problems extremely complicated to solve, so that various approximations and iterative schemes have been proposed in the scientific literature to efficiently solve them.

The use of this pseudo-norm arises in many different fields, such as machine learning, signal processing, pattern recognition, portfolio optimization, subset selection problem in regression, and elastic-net regularization. Cardinality constrained optimization problems are difficult to solve and a common approach is to apply global discrete optimization techniques.

Quite recently Sergeyev, in a book and in a series of papers, proposed a novel approach to infinite and infinitesimal numbers. By introducing the new numeral *grossone* (indicated by ①), defined as the number of elements of the set of the natural numbers, Sergeyev demonstrated how it is possible to operate with finite, infinite and infinitesimal quantities using the same arithmetics. This new numerical system allows to treat infinite and infinitesimal numbers as particular cases of a single structure, and offers a new view and an alternative approach for many fundamental aspects of mathematics such as limits, derivatives, sums of series and so on.

The aim of this paper is to show how this new numeral system and in particular ① can be used in different optimization problems, by replacing the $l_0$ pseudo-norm $\|x\|_0$ with

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

Indeed in literature there are many contributes for approximating the $l_0$ pseudo-norm. For example in [23] two new smooth approximations of the $l_0$ pseudo-norm are presented and other approximations are recalled in the following.

The paper is organized as follows. In Section 2 the new numeral system is presented by describing its main properties: Infinity, Identity, and Divisibility. Moreover, the new numeral positional system and the concept of *gross-number* are discussed. In Section 3 the properties of $\|x\|_0$ are introduced and some approximations proposed in literature are presented. The definition of $\|x\|_{0,①^{-1}}$ is then introduced and it is shown that $\|x\|_0$ and $\|x\|_{0,①^{-1}}$ coincide for the finite term and may differ only for infinitesimal terms. Two different applications of $\|x\|_{0,①^{-1}}$ are presented in the Sections 4 and 5. The first, studied in Section 4, concerns elastic net regularization and an algorithm for solving the optimization problem

$$\min_x \frac{1}{2} \|Ax - b\|_2^2 + \lambda_0 \|x\|_{0,①^{-1}} + \frac{\lambda_1}{2} \|x\|_2^2 .$$

In Section 5 the newly proposed definition of $\|x\|_{0,①^{-1}}$ is used in classification problems using sparse Support Vector Machines (SVMs). In particular, we suggest an interpretation of an updating rule already proposed in literature based on the KKT (Karush-Kuhn-Tucker) conditions and the expansion of gross-numbers.

We briefly describe some notations used throughout the paper. With $\mathbb{N}$ we indicate the set of natural numbers. Given $n, m \in \mathbb{N}$, let $\mathbb{R}^n$ be the space of the $n$-dimensional vectors with real components and let $\mathbb{R}^{m\times n}$ be the space of matrices with real elements, $m$ rows and $n$ columns. All vectors are column vectors and are indicated with lower case Latin letter (e.g. $x$, $y$, $z \in \mathbb{R}^n$). Subscripts indicate components of a vector, while superscripts are used to identify different vectors. Matrices are indicated with upper case Roman letter (e.g. $A$, $B \in \mathbb{R}^{m\times n}$). If $A \in \mathbb{R}^{m\times n}$, $A_{i\cdot}^T$ is the $i$-th row of $A$. The symbol $\|x\|$ indicates the norm of a vector $x$. Specific norms or parameters of the norm are indicated with subscripts. The scalar product of two vectors $x$, $y$ in $\mathbb{R}^n$ is denoted by $x^T y$, while in a generic Hilbert space we use the notation $\langle x, y\rangle$. The symbol $:=$ denotes definition of the term. The gradient of a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ at a point $x \in \mathbb{R}^n$ is indicated by $\nabla f(x)$.

## 2 The algebra of ①

The numeral system, originally proposed by Sergeyev in [25,31,26], is based on the numeral ① (called *grossone*) defined as the number of elements of the set $\mathbb{N}$. This new definition of infinite unit consents to work numerically with infinities and infinitesimals. In particular, the numerical system built upon ① makes possible to treat infinite and infinitesimal numbers in a unique framework, and to work with all of them numerically.

For instance, using ①-based numerals, it is possible to execute arithmetic operations with floating-point numbers and to assign concrete infinite and infinitesimal values to variables. Moreover, ① allows to compute more precisely the number of elements of infinite sets extending the traditional set theory operating with Cantor's cardinals. For example, the set of even numbers and the set of integers that the traditional cardinalities identify both as countable, have in this new numeral system, respectively, $\frac{①}{2}$ and $2① + 1$ elements.

The new computational methodology has been successfully applied in several fields of pure and applied mathematics offering new and alternative approaches. Here we only mention (numerical) differentiation [28], ODE [34,2], optimization [7,11–14,33], hyperbolic geometry [20], infinite series and the Riemann zeta function [26,29], biological processes, cellular automata [10]. For a survey of the various aspects and applications of ①, we refer the interested reader to [25,31,5,6,17,19,24,27,30,32] and to the references therein.

Following the procedure used in the past when the numeral 0 (zero) has been introduced to extend the natural numbers to integers, Sergeyev has introduced the new numeral ①.

In particular ① is introduced by adding the Infinite Unit Axiom postulate (IUA) to the axioms of real numbers. The IUA postulate is composed of three parts: Infinity, Identity and Divisibility:

1. *Infinity*: any finite natural number $n$ is less than *grossone*, i.e., $n < ①$, $\forall n \in \mathbb{N}$.
2. *Identity*: the following relations link ① to the identity elements 0 and 1:

$$0 \cdot ① = ① \cdot 0 = 0, \quad ① - ① = 0, \quad \frac{①}{①} = 1, \quad ①^0 = 1, \quad 1^{①} = 1, \quad 0^{①} = 0. \tag{1}$$

3. *Divisibility*: for any finite natural number $n$ the sets $\mathbb{N}_{k,n}$, $1 \leq k \leq n$, called the $n$th parts of the set $\mathbb{N}$ of natural numbers and defined as

$$\mathbb{N}_{k,n} = \{k, k+n, k+2n, k+3n, \ldots\}, \quad 1 \leq k \leq n, \quad \bigcup_{k=1}^{n} \mathbb{N}_{k,n} = \mathbb{N}, \tag{2}$$

have the same number of elements indicated by the numeral $\frac{①}{n}$. Note that $\frac{①}{n}$ is larger than any finite number.

Since this postulate is added to the standard axioms of real numbers, all standard properties (i.e. commutative and associative properties, distributive property of multiplication over addition, existence of inverse elements with respect to addition and multiplication, ....) also apply to ① and to *grossone*-based numerals. On the other hand, since in this framework it is possible to execute arithmetical operations with a variety of different infinities and infinitesimals, indeterminate forms as well as various kinds of divergences are not present when working with any (finite, infinite and infinitesimal) numbers of the new numerical system.

In this new numeral positional system a *gross-number* (or *gross-scalar*) $C$ can be represented similarly to traditional positional numeral system, but with base number ①, that is:

$$C = c_{p_m} ①^{p_m} + \cdots + c_{p_1} ①^{p_1} + c_{p_0} ①^{p_0} + c_{p_{-1}} ①^{p_{-1}} + \cdots + c_{p_{-k}} ①^{p_{-k}}, \tag{3}$$

where $m, k \in \mathbb{N}$, for $i = -k, -(k-1), \ldots, -1, 0, 1, \ldots, m-1, m$, numerals $c_{p_i}$ are floating-point numbers and exponents $p_i$ are gross-numbers such that

$$p_m > p_{m-1} \cdots > p_1 > p_0 = 0 > p_{-1} > \cdots > p_{-(k-1)} > p_{-k}. \tag{4}$$

A gross-number is called finite if $m = k = 0$, it is called infinite if $m > 0$, and it is called infinitesimal if $m = 0$, $c_{p_0} = 0$ and $k > 0$. The exponents $p_i$, $i = -k, -(k-1), \ldots, -1, 0, 1, \ldots, m-1, m$, are called *gross-powers* and can be finite, infinite, and infinitesimal. In (3) all numerals $c_{p_i} \neq 0$, $i = -k, -(k-1), \ldots, -1, 0, 1, \ldots, m-1, m$, are called *gross-digits* and belong to a traditional numeral system (for example floating-point numbers).

We note that in this new numeral system the record

$$C = c_{p_m} ①^{p_m} \cdots c_{p_1} ①^{p_1} c_{p_0} ①^{p_0} c_{p_{-1}} ①^{p_{-1}} \cdots c_{p_{-k}} ①^{p_{-k}}, \tag{5}$$

represents the number $C$. Infinitesimal numbers are represented by numerals $C$ having only negative (finite or infinite) gross-powers. The infinitesimal number $①^{-1}$ verifies $①^{-1}① = ①①^{-1} = 1$. Note that all infinitesimals are not equal to zero and in particular $①^{-1} = \frac{1}{①} > 0$ because it is the result of a division between two positive gross-numbers. In the following we consider only gross-numbers having representation (3) with $p_i \in \mathbb{Z}$, $i = -k, \ldots, m-1, m$.

We conclude this section by observing that the *Infinity Computer* is a new kind of a supercomputer able to execute numerical computations with finite, infinite, and infinitesimal numbers numerically (not symbolically) using ① and the new numeral system. For more details see [25] and the references therein.

## 3 The $l_0$ pseudo-norm and ①

In many problems in optimization and numerical analysis, it is extremely important to obtain a vector with the smallest possible number of components different from zero.

In [23] the problem of determining a vector belonging to a polyhedral set and having the minimum number of nonzero components is studied, and two smooth approximations of the $l_0$ pseudo-norm are proposed. The general optimization problem with cardinality constraints is considered in [4], where a reformulation as a smooth optimization problem is proposed. In [22] and [15] the cardinality-constrained optimization problem is reformulated using a DC (difference of convex functions) approach. We refer the interested reader to [15] for additional references to optimization problems where sparsity of the solution is required.

Determining a vector having the minimum number of nonzero components can be generally obtained by adding to the original problem a further term penalizing the number of components different from zero or a term that can approximately achieve the same goal.

Let $x \in \mathbb{R}^n$. The $l_0$ pseudo-norm is defined as

$$\|x\|_0 := \text{number of nonzero components in } x = \sum_{i=1}^{n} 1_{x_i \neq 0}, \tag{6}$$

where $1_a$ is the characteristic (indicator) function that is equal to 1 if $a \neq 0$ and zero otherwise. To be precise $\|\cdot\|_0$ is called $l_0$ pseudo-norm since it is not a norm. In fact for $x \in \mathbb{R}^n$, $x \neq 0$ and $0 \neq \lambda \in \mathbb{R}$, we have:

$$\|\lambda x\|_0 = \|x\|_0 ,$$

and hence $\|\lambda x\|_0 = |\lambda| \, \|x\|_0$ if and only if $|\lambda| = 1$.

In successive sections we present some specific use of the $l_0$ pseudo-norm for regularization and sparse solutions problems.

In [21] it is shown that computing sparse approximate solutions to linear systems is NP-hard. Moreover, in [1] it is shown that, for system of linear relations, the problems of determining a solution violating the minimum number

of relations (when the system itself is infeasible) and determining a solution with as few nonzero variables as possible (if feasible) are both NP-hard, and various strong bounds on the approximability of different variants of these problems are discussed.

Therefore, various approximations of $\|x\|_0$ have been proposed in the literature. In the context of Feature Selection and Machine Learning, in [3] the following approximation is proposed

$$\|x\|_0 \approx \sum_{i=1}^{n} \left(1 - e^{-\alpha|x_i|}\right),$$

where $\alpha$ is a given positive number for which the authors suggest to set the value 5.

In [18], in the context of elastic net regularization (discussed in detail in Section 4), the authors proposed the following approximation:

$$\|x\|_0 \approx \|x\|_{0,\delta} := \sum_{i=1}^{n} \frac{x_i^2}{x_i^2 + \delta}, \tag{7}$$

where $\delta > 0$ and smaller positive values of $\delta$ provide a better approximation of $\|x\|_0$.

Following this suggestion and by using the new numeral system, we propose to approximate the quantity $\|x\|_0$ with

$$\|x\|_{0,①^{-1}} := \sum_{i=1}^{n} \frac{x_i^2}{x_i^2 + ①^{-1}}. \tag{8}$$

Let us study in detail the connections between $\|x\|_0$ and $\|x\|_{0,①^{-1}}$. Let $\psi(t) := \dfrac{t^2}{t^2 + ①^{-1}}$. Hence $\|x\|_{0,①^{-1}} = \sum_{i=1}^{n} \psi(x_i)$. For $i = 1, \ldots, n$, we assume

$$x_i = x_i^{(0)} + R_i ①^{-1},$$

where $R_i$ includes only finite and infinitesimal terms.
When $x_i^{(0)} = 0$

$$\psi(x_i) = \frac{R_i^2 ①^{-2}}{R_i^2 ①^{-2} + ①^{-1}} = ①^{-1} \frac{R_i^2 ①^{-1}}{R_i^2 ①^{-2} + ①^{-1}} =$$

$$= ①^{-1} \frac{R_i^2}{R_i^2 ①^{-1} + 1} = 0①^0 + R_i' ①^{-1},$$

where $R_i'$ includes only finite and infinitesimal terms. Instead, when $x_i^{(0)} \neq 0$

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

where again $S_i'$ includes only finite and infinitesimal terms.

Therefore

$$\|x\|_{0,①^{-1}} = \|x\|_0 + T①^{-1}, \tag{9}$$

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

## 4 Elastic net regularization and ①

Given a matrix $A \in \mathbb{R}^{m\times n}$ and a vector $b \in \mathbb{R}^m$, in many important applications it is essential to determine a solution $x \in \mathbb{R}^n$ of the system of linear equations $Ax = b$ with the smallest number of nonzero components:

$$\begin{array}{ll} \min\limits_{x} & \|x\|_0\,, \\ \text{subject to} & Ax = b. \end{array}$$

The associated generalized elastic net regularization problem [18] is

$$\min_x \frac{1}{2}\|Ax-b\|_2^2 + \lambda_0\|x\|_0 + \frac{\lambda_2}{2}\|x\|_2^2\,, \tag{10}$$

where $\lambda_0 > 0$ and $\lambda_2 > 0$ are regularization parameters. In [18] the authors suggest to substitute $\|x\|_0$ with $\|x\|_{0,\delta}$, as defined in (7), for fixed positive $\delta$, and a convergent algorithm for the solution of the corresponding optimization problem is proposed. Clearly the obtained solution only approximates the optimal solution of (10).

We propose to use in (10) the approximation (8) of $\|x\|_0$. Replicating some of the proofs in [18] a convergent algorithm can be constructed. Moreover, due to the position (9), apart from terms of order $①^{-1}$ or below, the obtained solution solves also problem (10).

In detail, consider the problem

$$\min_x f(x),$$

where

$$\begin{aligned} f(x) &:= \frac{1}{2}\|Ax-b\|_2^2 + \lambda_0\|x\|_{0,①^{-1}} + \frac{\lambda_2}{2}\|x\|_2^2 \\ &= \frac{1}{2}(Ax-b)^T(Ax-b) + \lambda_0\sum_{i=1}^{n}\frac{x_i^2}{x_i^2+①^{-1}} + \frac{\lambda_2}{2}x^Tx. \end{aligned} \tag{11}$$

Let $D(x) \in \mathbb{R}^{n\times n}$ be given by

$$D_{ii}(x) = \frac{2①^{-1}}{\left((x_i)^2 + ①^{-1}\right)^2}, \quad D_{ij}(x) = 0,\ i \neq j. \tag{12}$$

Then

$$\nabla f(x) = \left( A^T A + \lambda_2 I + \lambda_0 D(x) \right) x - A^T b.$$

Following [18, Algorithm 1] the iterative scheme we propose is the following.

**Algorithm 4.1:**

**Choose** Choose $x^0 \in \mathbb{R}^n$;
**For** $k = 0, 1, \ldots$
  **Compute** $x^{k+1}$ by solving

$$\left( A^T A + \lambda_2 I + \lambda_0 D(x^k) \right) x^{k+1} = A^T b. \tag{13}$$

**End**

The following Lemma is the basis for establishing the convergence result.

**Lemma 1** *Let $f$ be given by (11). Let $x^{k+1}$ be obtained from $x^k$ by solving the system of equations (13). Then*

$$f\left(x^k\right) - f\left(x^{k+1}\right) \geq \frac{1}{2} \left\| Ax^k - Ax^{k+1} \right\|_2^2 + \frac{\lambda_2}{2} \left\| x^k - x^{k+1} \right\|_2^2, \tag{14}$$

*and hence*

$$\left\| Ax^k - Ax^{k+1} \right\|_2^2 \leq 2 \left( f\left(x^k\right) - f\left(x^{k+1}\right) \right), \tag{15}$$

$$\left\| x^k - x^{k+1} \right\|_2^2 \leq \frac{2}{\lambda_2} \left( f\left(x^k\right) - f\left(x^{k+1}\right) \right). \tag{16}$$

*Proof* See Appendix A.

The lemma above shows that the sequence $\{f(x^k)\}$ is a non-increasing sequence. We are now ready to state the following convergence theorem.

**Theorem 1** *Let $\mathcal{L}_0 := \{x : f(x) \leq f(x^0)\}$ be a compact set, and let $\{x^k\}$ be the sequence produced by the iterative scheme (13). Then*

1. *the sequence $\{x^k\}$ is all contained in $\mathcal{L}_0$;*
2. *the sequence $\{x^k\}$ has at least one accumulation point;*
3. *each accumulation point of $\{x^k\}$ belongs to $\mathcal{L}_0$;*
4. *each accumulation point $x^*$ satisfies the condition*

$$\left( A^T A + \lambda_2 I + \lambda_0 D(x^*) \right) x^* = A^T b,$$

*and hence is a stationary point of $f$.*

*Proof* First, from condition (14) in Lemma 1 we have $f\left(x^{k+1}\right) \leq f\left(x^k\right)$ and hence the entire sequence $\{x^k\}$ is contained in $\mathcal{L}_0$. Moreover, the sequence $\left\{f\left(x^k\right)\right\}$ is a bounded non–increasing sequence and hence it is a convergent seguence. The existence of accumulation points for the subsequence follows from the compactness of $\mathcal{L}_0$. Let now $x^*$ be an accumulation point of $\{x^k\}$ and $\left\{x^{k_l}\right\}$ be a subsequence indexed by $l$ converging to $x^*$. From (16), it follows that

$$\left\|x^{k_l} - x^{k_{l+1}}\right\|_2^2 \leq \frac{2}{\lambda_2}\Big(f\left(x^{k_l}\right) - f\left(x^{k_{l+1}}\right)\Big).$$

The right term converges to 0, and also the subsequence $\left\{x^{k_{l+1}}\right\}$ converges to the accumulation point $x^*$. Moreover

$$\left(A^T A + \lambda_2 I + \lambda_0 D(x^{k_l})\right) x^{k_{l+1}} = A^T b,$$

and hence

$$\left(A^T A + \lambda_2 I + \lambda_0 D(x^*)\right) x^* = A^T b.$$

Therefore, $x^*$ is a stationary point of $f$.

## 5 Sparse Support Vector Machine

In this section we show how ① and the results of Section 3 can be used in the context of Sparse Support Vector Machines (SSVMs).

Assume that empirical data (training set) $(x^i, y_i)$, $i = 1, \ldots, l$, are given, where $x^i \in \mathbb{R}^n$, and $y_i \in \{-1, 1\}$, $i = 1, \ldots, l$. Note that when the index $i$ is used as a superscript the corresponding object is an input, while when it is used as a subscript the corresponding object is an output. The aim is to determine an hyperplane (and hence a vector $w \in \mathbb{R}^n$ and a scalar $\theta$) such that:

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

The classification function is

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

Classification in the feature space (instead of the original space) requires to introduce a map

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

where $\mathcal{E}$ is an Hilbert Space with scalar product $\langle \cdot, \cdot \rangle$. In classical SVM (see [8,9,35] and references therein) the construction of the optimal hyperplane requires to solve the following (primal) optimization problem

$$\begin{array}{ll} \min\limits_{w,\theta,\xi} & \frac{1}{2}\langle w, w\rangle + Ce^T\xi, \\ \text{subject to} & y_i\left(\left\langle w, \phi(x^i)\right\rangle + \theta\right) \geq 1 - \xi_i, \quad i = 1, \ldots, l, \\ & \xi_i \geq 0, \quad i = 1, \ldots, l, \end{array} \tag{17}$$

where $e$ is a vector with all elements equal to 1 and $C$ is a positive scalar.

The corresponding dual problem is

$$\begin{array}{ll} \min\limits_{\alpha} & \frac{1}{2}\alpha^T Q\alpha - e^T\alpha, \\ \text{subject to} & y^T\alpha = 0, \\ & 0 \leq \alpha \leq Ce, \end{array} \tag{18}$$

where

$$Q_{ij} = y_i y_j K_{ij}, \text{ and } K_{ij} = K(x^i, x^j) := \left\langle \phi(x^i), \phi(x^j)\right\rangle.$$

The function

$$K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R},$$

is called the kernel function. It is well known that for the construction of the dual problem and the classification function the complete knowledge of the function $\phi(\cdot)$ is not necessary: only the quantities $K_{ij} = \left\langle \phi(x^i), \phi(x^j)\right\rangle$ are needed. In fact, from KKT conditions

$$w = \sum_{i=1}^{l} \alpha_i y_i \phi(x^i) \tag{19}$$

and the classification function is

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

For a sparse representation of SVM, the vector $w$ is substituted by its expansion in terms of the vector $\alpha$. Moreover, let $K_{i.}$ be the column vector that corresponds to the $i$th row of matrix $(K_{ij})$, note that

$$\begin{aligned} {K_{i.}}^T\alpha + \theta &= \sum_{j=1}^{l} K_{ij}\alpha_j + \theta \\ &= \sum_{j=1}^{l} \left\langle \phi(x^i), \phi(x^j)\right\rangle \alpha_j + \theta \\ &= \left\langle \phi(x^i), \sum_{j=1}^{l} \phi(x^j)\alpha_j \right\rangle + \theta. \end{aligned}$$

In [16] the authors consider the following optimization problem instead of (17), obtained by replacing $\frac{1}{2}\langle w, w\rangle$ with $\|\alpha\|_0$ and by using the expansion (19) of $w$ in terms of $\alpha$:

$$\begin{aligned} \min_{\alpha,\theta,\xi} \quad & \|\alpha\|_0 + Ce^T\xi, \\ \text{subject to } & y_i\left[K_{i.}{}^T\alpha + \theta\right] \geq 1 - \xi_i, \quad i = 1, \ldots, l, \\ & \xi \geq 0. \end{aligned} \tag{20}$$

In problem (20) the term $\|\alpha\|_0$ is then replaced by $\frac{1}{2}\alpha^T\Lambda\alpha$, where $\Lambda$ is the diagonal matrix with $\Lambda_{ii} = \lambda_i$, $i = 1, \ldots, l$.

The following iterative scheme was proposed in [16] to solve the above problem. In particular, given a very small positive value $\epsilon$, in the Algorithm 5.1 the new value $\lambda_r^{k+1}$ for $\lambda_r$ is set to $1/\left(\alpha_r^k\right)^2$ if $\left|\alpha_r^k\right|$ is "significantly" different from zero, otherwise, $\lambda_r^{k+1} = 1/\epsilon^2$.

**Algorithm 5.1:**

**Set** $\lambda_r^0 = 1$, $r = 1, \ldots, l$;
**For** $k = 0, 1, \ldots$
  **Solve**

$$\begin{aligned} \min_{\alpha,\theta,\xi} \quad & \frac{1}{2}\sum_{r=1}^{l}\lambda_r^k\alpha_r^2 + Ce^T\xi, \\ \text{subject to } & y_i\left[K_{i.}{}^T\alpha + \theta\right] \geq 1 - \xi_i, \quad i = 1, \ldots, l, \\ & \xi \geq 0. \end{aligned} \tag{21}$$

  and let $\alpha^k$ be the optimal solution;
  **Update** $\lambda^{k+1}$ according to the formula

$$\lambda_r^{k+1} = \begin{cases} \dfrac{1}{\left(\alpha_r^k\right)^2} & \text{if } \left|\alpha_r^k\right| \geq \epsilon, \\ \dfrac{1}{\epsilon^2} & \text{otherwise}, \end{cases} \qquad r = 1, \ldots, l,$$

**End**

The KKT conditions for Problem (21) are

$$\Lambda^k\alpha - \sum_{j=1}^{l}\beta_j y_j K_{j.} = 0, \tag{22a}$$

$$\beta^T y = 0, \tag{22b}$$

$$Ce - \beta \geq 0, \tag{22c}$$

$$\beta \geq 0, \tag{22d}$$

$$\beta^T(Ce - \beta) = 0, \tag{22e}$$

where $\beta$ is the vector of multipliers associated to the constraints $y_i\Big[K_{i.}{}^T\alpha + \theta\Big] \geq 1 - \xi_i, \quad i = 1, \ldots, l.$

From (22a) it follows that

$$\lambda_r^k \alpha_r = \sum_{j=1}^{l} \beta_j y_j K_{jr} = \bar{K}_{r.}^T \beta, \quad r = 1, \ldots, l,$$

where $\bar{K}_{rj} = y_j K_{jr}$, with $r, j = 1, \ldots, l$.

Once again, instead of $\|\alpha\|_0$, we use $\|\alpha\|_{0,①^{-1}}$ and we propose to solve the following ①–Sparse SVM problem in place of (20):

$$\begin{aligned} \min_{\alpha,\theta,\xi} \quad & \frac{①}{2}\|\alpha\|_{0,①^{-1}} + Ce^T\xi, \\ \text{subject to } & y_i\Big[K_{i.}{}^T\alpha + \theta] \geq 1 - \xi_i, \quad i = 1, \ldots, l, \\ & \xi \geq 0. \end{aligned} \tag{23}$$

Let by (8)

$$\begin{aligned} h(\alpha) &:= \frac{①}{2}\|\alpha\|_{0,①^{-1}} \\ &= \frac{①}{2}\sum_{j=1}^{l}\frac{\alpha_j^2}{\alpha_j^2 + ①^{-1}} = \frac{①}{2}\sum_{j=1}^{l}\frac{\alpha_j^2 + ①^{-1} - ①^{-1}}{\alpha_j^2 + ①^{-1}} \\ &= \frac{l}{2}① - \sum_{j=1}^{l}\frac{①}{2}\frac{①^{-1}}{\alpha_j^2 + ①^{-1}} \\ &= \frac{l}{2}① - \frac{1}{2}\sum_{j=1}^{l}\frac{1}{\alpha_j^2 + ①^{-1}}. \end{aligned}$$

Then

$$\Big[\nabla h(\alpha)\Big]_r = \frac{\alpha_r}{\left(\alpha_r^2 + ①^{-1}\right)^2}.$$

The KKT conditions for the above problem (23) are

$$\nabla h(\alpha) - \sum_{j=1}^{l}\beta_j y_j K_{j.} = 0, \tag{24a}$$

$$\beta^T y = 0, \tag{24b}$$

$$Ce - \beta \geq 0, \tag{24c}$$

$$\beta \geq 0, \tag{24d}$$

$$\beta^T(Ce-\beta)=0. \tag{24e}$$

Note that conditions (24a) can be rewritten as

$$\frac{1}{\left(\alpha_r^2+①^{-1}\right)^2}\alpha_r=\bar{K}_{r.}^T\beta, \quad r=1,\ldots,l. \tag{25}$$

From the above formula it is "natural" to set the new value for $\lambda_r$

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

Now, for $r=1,\ldots,l$, let

$$\alpha_r=\alpha_r^{(0)}+\alpha_r^{(1)}①^{-1}+\ldots=\alpha_r^{(0)}+A①^{-1},$$

with $A\in\mathbb{R}$ finite or infinitesimal. When $\alpha_r^{(0)}=0$

$$\alpha_r^2+①^{-1}=①^{-1}+A^2①^{-2},$$

and

$$\frac{1}{\left(\alpha_r^2+①^{-1}\right)^2}=\frac{1}{\left(①^{-1}+A^2①^{-2}\right)^2}=\frac{1}{①^{-2}}+A'\frac{1}{①^{-3}}, \tag{26}$$

with $A'$ finite or infinitesimal. In place, when $\alpha_r^{(0)}\neq 0$

$$\alpha_r^2+①^{-1}=\left(\alpha_r^{(0)}\right)^2+A''①^{-1},$$

and

$$\frac{1}{\left(\alpha_r^2+①^{-1}\right)^2}=\frac{1}{\left(\alpha_r^{(0)}\right)^4+A'''①^{-1}}, \tag{27}$$

with $A''$, $A'''$ finite or infinitesimal.

Formulas (26) and (27) mimic almost perfectly the updating formulas for $\lambda^{k+1}$ proposed in [16] in order to solve (20) by using problem (21). The main difference is that, for the nonzero case there is a power 4 instead of 2. However, using ① and considering problem (23) we can shed some light on the updating formulas for $\beta$, otherwise quite arbitrary. The term $\frac{1}{①^{-2}}$ perfectly corresponds to the updating formula $\lambda_r^{k+1}=\frac{1}{\epsilon^2}$. In the other case the updating formula $\lambda_r^{k+1}=\frac{1}{\bar{\alpha}_r^2}$ is replaced by the term $\frac{1}{\left(\alpha_r^{(0)}\right)^4}$.

The use of ① allows to easily obtain both formulas from the KKT condition (24a) and observations on the expansion of the gross-number $\alpha$.

## 6 Conclusions

In this paper we have presented some possible uses in optimization problems of the novel approach to infinite and infinitesimal numbers proposed by Sergeyev [25,31,26]. In particular, optimization problems including smoothed $l_0$ penalty and classification problems involving sparse Support Vector Machines are studied. In order to avoid the difficulties due to the use of the $l_0$ pseudo-norm of a vector, we propose to approximate the $l_0$ pseudo-norm by using a smooth function defined in terms of ①. The results obtained using this new approximation in the two optimization problems perfectly match with those presented in the literature. Actually the new ①-based methodology may represent a fruitful and promising tool to be exploited within other classification and regression problems offering new views and different perspectives.

## A Proof of Lemma 1

We prove Lemma 1 of Section 4 by adapting to this context the proof proposed by the authors in [18, Lemma 2].

*Proof* of Lemma 1. Let $f$ be defined in (11). We have:

$$\begin{aligned} f\left(x^k\right) - f\left(x^{k+1}\right) &= \frac{1}{2}\left\|Ax^k - b\right\|_2^2 + \lambda_0 \left\|x^k\right\|_{0,①^{-1}} + \frac{\lambda_2}{2}\left\|x^k\right\|_2^2 \\ &- \frac{1}{2}\left\|Ax^{k+1} - b\right\|_2^2 - \lambda_0 \left\|x^{k+1}\right\|_{0,①^{-1}} - \frac{\lambda_2}{2}\left\|x^{k+1}\right\|_2^2 \\ &= \frac{1}{2}\left\|Ax^k - b\right\|_2^2 - \frac{1}{2}\left\|Ax^{k+1} - b\right\|_2^2 + \lambda_2\left(\frac{1}{2}\left\|x^k\right\|_2^2 - \frac{1}{2}\left\|x^{k+1}\right\|_2^2\right) \\ &+ \lambda_0\left(\left\|x^k\right\|_{0,①^{-1}} - \left\|x^{k+1}\right\|_{0,①^{-1}}\right). \end{aligned} \tag{28}$$

We analyse separately the different terms of (28). We can write:

$$\begin{aligned} &\frac{1}{2}\left\|Ax^k - b\right\|_2^2 - \frac{1}{2}\left\|Ax^{k+1} - b\right\|_2^2 \\ &= \frac{1}{2}\left(Ax^k\right)^T\left(Ax^k\right) - \frac{1}{2}\left(Ax^{k+1}\right)^T\left(Ax^{k+1}\right) + b^T\left(Ax^{k+1} - Ax^k\right) \\ &= \frac{1}{2}\left\|Ax^k - Ax^{k+1}\right\|_2^2 + \left(Ax^k - Ax^{k+1}\right)^T\left(Ax^{k+1}\right) + b^T\left(Ax^{k+1} - Ax^k\right) \\ &= \frac{1}{2}\left\|Ax^k - Ax^{k+1}\right\|_2^2 + \left(Ax^{k+1} - b\right)^T\left(Ax^k - Ax^{k+1}\right) \\ &= \frac{1}{2}\left\|Ax^k - Ax^{k+1}\right\|_2^2 + \left(x^k - x^{k+1}\right)^T\left(A^TAx^{k+1} - A^Tb\right) \\ &= \frac{1}{2}\left\|Ax^k - Ax^{k+1}\right\|_2^2 - \lambda_2\left(x^{k+1}\right)^T\left(x^k - x^{k+1}\right) - \lambda_0\sum_{i=1}^{n}\frac{\left(x_i^k - x_i^{k+1}\right)x_i^{k+1}2①^{-1}}{\left(\left(x_i^k\right)^2 + ①^{-1}\right)^2}, \end{aligned} \tag{29}$$

where in the last step the definitions of the proposed iterative scheme (13) and the matrix $D$ in (12) have been used.
Moreover,

$$\frac{1}{2}\left\|x^k\right\|_2^2 - \frac{1}{2}\left\|x^{k+1}\right\|_2^2 = \tag{30}$$

$$= \frac{1}{2}(x^k)^T x^k + \frac{1}{2}(x^{k+1})^T x^{k+1} - (x^k)^T x^{k+1} - (x^{k+1})^T x^{k+1} + (x^k)^T x^{k+1}$$
$$= \frac{1}{2}\left\|x^k - x^{k+1}\right\|_2^2 + (x^{k+1})^T\left(x^k - x^{k+1}\right).$$

Substituting (29) and (30) into (28) we have:

$$\begin{aligned} f\left(x^k\right) - f\left(x^{k+1}\right) = & \frac{1}{2}\left\|Ax^k - Ax^{k+1}\right\|_2^2 + \frac{\lambda_2}{2}\left\|x^k - x^{k+1}\right\|_2^2 \\ & - \lambda_0 \sum_{i=1}^{n} \frac{\left(x_i^k - x_i^{k+1}\right) x_i^{k+1} 2①^{-1}}{\left(\left(x_i^k\right)^2 + ①^{-1}\right)^2} \\ & + \lambda_0\left(\left\|x^k\right\|_{0,①^{-1}} - \left\|x^{k+1}\right\|_{0,①^{-1}}\right). \end{aligned} \tag{31}$$

From the definition (8)

$$\left\|x^k\right\|_{0,①^{-1}} - \left\|x^{k+1}\right\|_{0,①^{-1}} = \sum_{i=1}^{n}\left(\frac{\left(x_i^k\right)^2}{\left(x_i^k\right)^2 + ①^{-1}} - \frac{\left(x_i^{k+1}\right)^2}{\left(x_i^{k+1}\right)^2 + ①^{-1}}\right). \tag{32}$$

It is easy to prove [18, Lemma 1] that given $\delta > 0$, for any $a, b \in \mathbb{R}$ the following inequality holds:

$$\frac{a^2}{a^2+\delta} - \frac{b^2}{b^2+\delta} - \frac{2\delta b(a-b)}{(a^2+\delta)^2} \geq \frac{\delta(a-b)^2}{(a^2+\delta)^2}. \tag{33}$$

As a result, we can write

$$\begin{aligned} & \sum_{i=1}^{n}\left(\frac{\left(x_i^k\right)^2}{\left(x_i^k\right)^2 + ①^{-1}} - \frac{\left(x_i^{k+1}\right)^2}{\left(x_i^{k+1}\right)^2 + ①^{-1}} - \frac{2①^{-1}\left(x_i^k - x_i^{k+1}\right) x_i^{k+1}}{\left(\left(x_i^k\right)^2 + ①^{-1}\right)^2}\right) \\ & \geq \sum_{i=1}^{n} \frac{①^{-1}\left(x_i^k - x_i^{k+1}\right)^2}{\left(\left(x_i^k\right)^2 + ①^{-1}\right)^2}. \end{aligned} \tag{34}$$

By using (32)-(34) into (31) we can conclude:

$$\begin{aligned} f\left(x^k\right) - f\left(x^{k+1}\right) \geq & \frac{1}{2}\left\|Ax^k - Ax^{k+1}\right\|_2^2 + \frac{\lambda_2}{2}\left\|x^k - x^{k+1}\right\|_2^2 \\ & + \lambda_0 \sum_{i=1}^{n} \frac{①^{-1}\left(x_i^k - x_i^{k+1}\right)^2}{\left(\left(x_i^k\right)^2 + ①^{-1}\right)^2} \\ \geq & \frac{1}{2}\left\|Ax^k - Ax^{k+1}\right\|_2^2 + \frac{\lambda_2}{2}\left\|x^k - x^{k+1}\right\|_2^2 \end{aligned} \tag{35}$$

since $\sum_{i=1}^{n} \frac{①^{-1}\left(x_i^k - x_i^{k+1}\right)^2}{\left(\left(x_i^k\right)^2 + ①^{-1}\right)^2} \geq 0$ for any $x_i^k$ and $x_i^{k+1}$.

## Conflict of interest

The authors declare that they have no conflict of interest. This article does not contain any studies with human participants or animals performed by any of the authors.

## References


1. Amaldi, E., Kann, V.: On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theoretical Computer Science **209**(1-2), 237–260 (1998)
2. 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. Mathematics and Computers in Simulation **141**, 24–39 (2017)
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, CA, USA (1998)
4. Burdakov, O., Kanzow, C., Schwartz, A.: Mathematical programs with cardinality constraints: reformulation by complementarity-type conditions and a regularization method. SIAM Journal on Optimization **26**(1), 397–425 (2016)
5. Caldarola, F.: The Sierpinski curve viewed by numerical computations with infinities and infinitesimals. Applied Mathematics and Computation **318**, 321–328 (2018)
6. Calude, C., Dumitrescu, M.: Infinitesimal probabilities based on grossone. SN Computer Science (2020)
7. Cococcioni, M., Pappalardo, M., Sergeyev, Y.D.: Lexicographic multi-objective linear programming using grossone methodology: Theory and algorithm. Applied Mathematics and Computation **318**, 298–311 (2018)
8. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)
9. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press (2000)
10. D'Alotto, L.: A classification of two-dimensional cellular automata using infinite computations. Indian Journal of Mathematics **55**, 143–158 (2013)
11. De Cosmis, S., De Leone, R.: The use of grossone in mathematical programming and operations research. Applied Mathematics and Computation **218**(16), 8029–8038 (2012)
12. De Leone, R.: Nonlinear programming and grossone. Applied Mathematics and Computation **318**(C), 290–297 (2018)
13. De Leone, R., Fasano, G., Sergeyev, Y.D.: Planar methods and grossone for the conjugate gradient breakdown in nonlinear programming. Computational Optimization and Applications **71**(1), 73–93 (2018)
14. Gaudioso, M., Giallombardo, G., Mukhametzhanov, M.: Numerical infinitesimals in a variable metric method for convex nonsmooth optimization. Applied Mathematics and Computation **318**, 312–320 (2018)
15. Gotoh, J., Takeda, A., Tono, K.: DC formulations and algorithms for sparse optimization problems. Mathematical Programming **169**, 141–176 (2018)
16. Huang, K., Zheng, D., Sun, J., Hotta, Y., Fujimoto, K., Naoi, S.: Sparse learning for support vector classification. Pattern Recognition Letters **31**(13), 1944–1951 (2010)
17. Iudin, D., Sergeyev, Y.D., Hayakawa, M.: Interpretation of percolation in terms of infinity computations. Applied Mathematics and Computation **218**(16), 8099–8111 (2012)
18. Li, S., Ye, W.: A generalized elastic net regularization with smoothed $l_0$ penalty. Advances in Pure Mathematics **7**, 66–74 (2017)
19. Lolli, G.: Metamathematical investigations on the theory of grossone. Applied Mathematics and Computation **255**, 3–14 (2015)
20. Margenstern, M.: An application of grossone to the study of a family of tilings of the hyperbolic plane. Applied Mathematics and Computation **218**(16), 8005–8018 (2012)
21. Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM Journal on Computing **24**(2), 227–234 (1995)
22. Pham Dinh, T., Le Thi, H.A.: Recent Advances in DC Programming and DCA, pp. 1–37. Springer Berlin Heidelberg, Berlin, Heidelberg (2014)
23. Rinaldi, F., Schoen, F., Sciandrone, M.: Concave programming for minimizing the zero-norm over polyhedral sets. Computational Optimization and Applications **46**, 467–486 (2010)

24. Rizza, D.: Numerical methods for infinite decision-making processes. International Journal of Unconventional Computing **14**(2), 139–158 (2019)
25. Sergeyev, Y.D.: Arithmetic of Infinity. Edizioni Orizzonti Meridionali, CS (2001)
26. Sergeyev, Y.D.: Numerical point of view on Calculus for functions assuming finite, infinite, and infinitesimal values over finite, infinite, and infinitesimal domains. Nonlinear Analysis Series A: Theory, Methods & Applications **71**(12), e1688–e1707 (2009)
27. Sergeyev, Y.D.: Lagrange Lecture: methodology of numerical computations with infinities and infinitesimals. Rendiconti del Seminario Matematico dell'Università e del Politecnico di Torino **68**(2), 95–113 (2010)
28. Sergeyev, Y.D.: Higher order numerical differentiation on the infinity computer. Optimization Letters **5**(4), 575–585 (2011)
29. 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 Analysis and Applications **3**(2), 129–148 (2011)
30. Sergeyev, Y.D.: The exact (up to infinitesimals) infinite perimeter of the Koch snowflake and its finite area. Communications in Nonlinear Science and Numerical Simulation **31**(1–3), 21–29 (2016)
31. Sergeyev, Y.D.: Numerical infinities and infinitesimals: Methodology, applications, and repercussions on two Hilbert problems. EMS Surveys in Mathematical Sciences **4**, 219–320 (2017)
32. Sergeyev, Y.D.: Independence of the grossone-based infinity methodology from non-standard analysis and comments upon logical fallacies in some texts asserting the opposite. Foundations of Science **24**(1), 153–170 (2019)
33. Sergeyev, Y.D., Kvasov, D.E., Mukhametzhanov, M.S.: On strong homogeneity of a class of global optimization algorithms working with infinite and infinitesimal scales. Communications in Nonlinear Science and Numerical Simulation **59**, 319–330 (2018)
34. Sergeyev, Y.D., Mukhametzhanov, M., Mazzia, F., Iavernaro, F., Amodio, P.: Numerical methods for solving initial value problems on the Infinity Computer. International Journal of Unconventional Computing **12**(1), 3–23 (2016)
35. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing **14**(3), 199–222 (2004)