DIPARTIMENTO DI INFORMATICA
E SISTEMISTICA ANTONIO RUBERTI

SAPIENZA
UNIVERSITÀ DI ROMA

Finding sparse solutions to problems with convex constraints via concave programming

Francesco Rinaldi

Technical Report n. 8, 2009

Finding sparse solutions to problems with convex constraints via concave programming

F. Rinaldi

Dipartimento di Informatica e Sistemistica
Sapienza Università di Roma
Via Ariosto, 25 - 00185 Roma - Italy

e-mail: rinaldi@dis.uniroma1.it

Abstract

In this work, we consider a class of nonlinear optimization problems with convex constraints with the aim of computing sparse solutions. This is an important task arising in various fields such as machine learning, signal processing, data analysis. We adopt a concave optimization-based approach, we define an effective version of the Frank-Wolfe algorithm, and we prove the global convergence of the method. Finally, we report numerical results on test problems showing both the effectiveness of the concave approach and the efficiency of the implemented algorithm.

Keywords Zero-norm, concave programming, Frank-Wolfe method.

1 Introduction

We consider a class of constrained nonsmooth optimization problems of the form:

$$\begin{aligned} &\min_{x \in R^n} \; g(x) + \lambda \|x\|_0 \\ &x \in C \end{aligned} \tag{1}$$

where $\lambda > 0$, C is a compact convex set, g is a continuously differentiable function, and $\|x\|_0$ is the **zero-norm** of x defined as

$$\|x\|_0 = card\{x_i : x_i \neq 0\}.$$

The problem (1) is quite general and includes as special cases a wide variety of problems arising from different fields (e.g. machine learning, signal processing, data analysis).
In machine learning, for instance, an interesting problem that can be formulated as in (1) is the Sparse Linear Discriminant Analysis (SLDA) (see, e.g., [10]). Given a pair of symmetric matrices:

(i) *between-class* covariance matrix: A positive semi-definite;

(ii) *within-class* covariance matrix: B positive definite;

in SLDA, we want to find a sparse vector x which maximizes a class-separability criterion defined by the *generalized* Rayleigh quotient:

$$R(x; A, B) = \frac{x^T A x}{x^T B x} \; .$$

Namely, we want to solve the following optimization problem:

$$\begin{aligned} &\min_{x \in R^n} \; -\frac{1}{2} x^T A x + \lambda \|x\|_0 \\ &x^T B x \leq 1. \end{aligned} \tag{2}$$

Sparse Principal Component Analysis (SPCA) is a well-known problem in data analysis (see, e.g., [4, 6, 14]). In SPCA, given a (symmetric positive semi-definite) covariance matrix C, the goal is finding a sparse vector x which explains the maximum amount of variance. The zero-norm formulation related to this problem is:

$$\begin{aligned} &\min_{x \in R^n} \; -\frac{1}{2} x^T C x + \lambda \|x\|_0 \\ &x^T x \leq 1. \end{aligned} \tag{3}$$

In signal analysis, a widely-studied problem is the sparse representation of noisy signals (see, e.g., [3, 5]). Given a dictionary $A \in R^{m \times n}$ of elementary signals and a real noisy signal b the goal is finding a sparse representation x of signal b in terms of the dictionary A. This problem can be formulated as follows:

$$\begin{aligned} &\min_{x \in R^n} \; \|x\|_0 \\ &\|Ax - b\|^2 \leq \delta \end{aligned} \tag{4}$$

where δ is a fixed error tolerance.
In order to make problem (1) tractable, a simple approach can be that of replacing the zero-norm, which is a nonconvex discontinuous function, with the ℓ_1 norm (see, e.g., [3, 13, 14]) thus obtaining the problem:

$$\begin{array}{l} \min\limits_{x \in R^n} \; g(x) + \lambda \|x\|_1 \\ \\ x \in C \end{array} \tag{5}$$

which can be efficiently solved even when the dimension of the problem is large. However, some experiments reported in [2, 11] show that a concave optimization-based approach, for the special case of a polyhedral feasible set, performs better than the ℓ_1 norm-based one. In this paper, inspired by the idea developed in [9, 11, 12], we propose a concave programming approach for solving problem (1). We replace the zero-norm with a separable concave function thus obtaining the following formulation:

$$\begin{array}{l} \min\limits_{x \in R^n} \; g(x) + \lambda \sum\limits_{j=1}^{n} h_j(x_j, u) \\ \\ x \in C \end{array} \tag{6}$$

where $h_j : R \to R$, for $j = 1, \ldots, n$ are concave, continuously differentiable functions depending on a vector $u \in R^m$ of parameters. Then, in order to solve problem (6), we use a new suitably developed version of the Frank-Wolfe algorithm.
The paper is organized as follows. In Section 2, we describe various smooth concave functions that can be used in place of the zero-norm when searching for sparse solutions to problems with convex constraints. In Section 3, we state some well-known optimality conditions for constrained problems based on Lagrange multipliers. In Section 4, we report a result related to a Big-M method for convex programming problems. In Section 5, after a brief review of the well-known Frank-Wolfe method, we derive some new theoretical results which have an important impact on the computational efficiency of the method. These results suggest the definition of a version of the method that eliminates the variables set to zero, thus allowing for a dimensionality reduction which greatly increments the speed of the procedure. We formally prove, by means of the results reported in Section 4, the global convergence of this modified version of the Frank-Wolfe method. In section 6, we describe a version of the reduced Frank-Wolfe algorithm with unitary stepsize that can be used when the problem we want to solve has a concave objective function. Finally, in section 7, we report the numerical results on test problems showing both the usefulness of the new concave formulations and the efficiency in terms of computational time of the implemented minimization algorithm.

2 Concave formulations for finding a sparse vector over a convex set

Consider the general problem of finding a vector belonging to a compact convex set C and having the minimum number of nonzero components, that is

$$\begin{array}{l} \min\limits_{x \in R^n} \; \|x\|_0 \\ \\ x \in C. \end{array} \tag{7}$$

Since the objective function in (7) is discontinuous, we can use a continuously differentiable, concave function that somehow approximates the behaviour of the zero-norm function. A similar approach has already been proposed in [9, 11, 12] for finding sparse solutions to linear systems. In order to illustrate the idea underlying the concave approach, we observe that the objective function of problem (7) can be written as follows

$$\|x\|_0 = \sum_{i=1}^{n} s(|x_i|)$$

where $s : R \to R^+$ is the *step function* such that $s(t) = 1$ for $t > 0$ and $s(t) = 0$ for $t \leq 0$. Following the approach described in [9], we replace the discontinuous step function by a continuously differentiable concave function $v(t) = 1 - e^{-\alpha t}$, with $\alpha > 0$, thus obtaining a problem of the form

$$\begin{aligned} &\min_{x,y \in R^n} \sum_{i=1}^{n} (1 - e^{-\alpha y_i}) \\ &x \in C \\ &-y_i \leq x_i \leq y_i \quad i = 1, \ldots, n. \end{aligned} \tag{8}$$

The approach is well-motivated from a theoretical point of view. In fact, it is easy to see that

$$\lim_{\alpha \to \infty} \sum_{i=1}^{n} (1 - e^{-\alpha y_i}) = \|y\|_0,$$

and the objective function is a smooth approximation of the zero-norm. Another way to solve problem (7) can be that of using the logarithm function instead of the step function [12], and this leads to a concave smooth problem of the form

$$\begin{aligned} &\min_{x,y \in R^n} \sum_{i=1}^{n} \ln(\epsilon + y_i) \\ &x \in P \\ &-y_i \leq x_i \leq y_i \quad i = 1, \ldots, n, \end{aligned} \tag{9}$$

with $0 < \epsilon \ll 1$. Formulation (9) is practically motivated by the fact that, due to the form of the logarithm function, it is better to increase one variable y_i while setting to zero another one rather than doing some compromise between both, and this should facilitate the computation of a sparse solution. The following two concave formulations, related to the ideas underlying (8) and (9) respectively, have been proposed in [11] for finding a sparse solution to a linear system:

$$\begin{aligned} &\min_{x \in R^n, y \in R^n} \sum_{i=1}^{n} (y_i + \epsilon)^p \\ &x \in C \\ &-y_i \leq x_i \leq y_i \quad i = 1, \ldots, n \end{aligned} \tag{10}$$

with $0 < p < 1$, and $0 < \epsilon$;

$$\begin{aligned} &\min_{x \in R^n, y \in R^n} -\sum_{i=1}^{n} (y_i + \epsilon)^{-p} \\ &x \in C \\ &-y_i \leq x_i \leq y_i \quad i = 1, \ldots, n \end{aligned} \tag{11}$$

with $1 \leq p$, and $0 < \epsilon$.

3 Optimality Conditions for Constrained Problems based on Lagrange Multipliers

In this section we state some well-known optimality conditions for constrained problems based on Lagrange multipliers, namely Karush-Kuhn-Tuker conditions (see [1] for further details). We consider the problem:

$$\begin{aligned} &\min f(x) \\ &g(x) \leq 0 \\ &h(x) = 0 \end{aligned} \tag{12}$$

where $f : R^n \to R$, $g : R^n \to R^m$, and $h : R^n \to R^p$ are continuously differentiable function.

Definition 1 *a feasible vector x is said to be* regular *if the equality constraints gradients $\nabla h_i(x)$, $i = 1, \ldots, m$, and the active inequality constraint gradients $\nabla g_i(x)$, $i \in A(x) = \{i \ : \ g_i(x) = 0\}$, are linearly independent.*

We now state necessary conditions for optimality:

Proposition 1 *Let $x^\star$ be a local minimum of the problem* 12. *Assume that $x^\star$ is regular. Then there exists Lagrange multipliers $\lambda^\star \in R^m$ and $\mu^\star \in R^p$ satisfying the following conditions:*

$$\begin{aligned} &\nabla f(x^\star) + \nabla g(x^\star)\lambda^\star - \nabla h(x^\star)\mu^\star = 0 \\ &\lambda^{\star T} g(x^\star) = 0 \\ &\lambda^\star \geq 0. \end{aligned} \tag{13}$$

There are are a number of conditions (so called constraint qualifications) that guarantee the existence of Lagrange multipliers. The following proposition is due to Slater:

Proposition 2 *Let $x^\star$ be a local minimum of the problem* 12. *Assume that g_i are convex functions and that there exists a feasible vector $\bar{x}$ satisfying the following condition:*

$$g_j(\bar{x}) < 0 \quad \forall \ j \in A(x^\star). \tag{14}$$

Then $x^\star$ satisfies the necessary conditions of proposition 1.

Under some suitable convexity assumptions we can state sufficient conditions for optimality:

Proposition 3 *Let f and g_i $i = 1, \ldots, m$ be convex continuously differentiable functions, and let equality constraints $h_i(x)$ $i = 1, \ldots, p$ be affine functions. If there exists Lagrange multipliers $\lambda^\star \in R^m$ and $\mu^\star \in R^p$ satisfying the following conditions:*

$$\begin{aligned} &\nabla f(x^\star) + \nabla g(x^\star)\lambda^\star - \nabla h(x^\star)\mu^\star = 0 \\ &g(x^\star) \leq 0, \;\; h(x^\star) = 0 \\ &\lambda^{\star T} g(x^\star) = 0 \\ &\lambda^\star \geq 0, \end{aligned} \tag{15}$$

then $x^\star$ is a global minimum of the problem (12).

4 Big-M for convex programming problems

We report a result (and its proof) that we will use to derive some new convergence results of a modified Frank-Wolfe method we will present.

Proposition 4 *Consider the convex programming problems*

$$\begin{aligned} &\min f(x) \\ &g(x) \leq 0 \\ &Ax = b \end{aligned} \tag{16}$$

$$\begin{aligned} &\min f(x) + Me^T z \\ &g(x) + h(z) \leq 0 \\ &Ax + Qz = b \\ &z \geq 0 \end{aligned} \tag{17}$$

where

1) $e \in R^{n_z}$ is a vector of ones, $b \in R^p$, $A \in R^{p \times n}$, $Q \in R^{p \times n_z}$;

2) $f : R^n \to R$ is a convex, continuously differentiable function;

3) $g : R^n \to R^m$ is a convex, continuously differentiable function;

4) $h : R^{n_z} \to R^m$ is a convex, continuously differentiable function such that $h(0) = 0$.

Assume that problem (16) *admits a solution $x^\star$, and that there exists a feasible vector $\bar{x}$ satisfying the following condition:*

$$g(\bar{x}) < 0. \tag{18}$$

Then there exists a value M_0 such that for all $M \geq M_0$ we have that

(i) *the vector $(x^\star, 0)^T$ is a solution of* (17)*;*

(ii) *if $(\bar{x}, \bar{z})^T$ is a solution of* (17)*, then $\bar{z} = 0$ and $\bar{x}$ is a solution of* (16).

Proof. (i) Since $x^\star$ is a solution of problem (16) we have from Proposition 2 that there exist Lagrange multipliers $\lambda^\star \in R^m$ and $\mu^\star \in R^p$ satisfying conditions (15).
Now consider problem (17) and the KKT system related to it

$$\begin{array}{l} \nabla f(x) + \nabla g(x)\lambda - A^T\mu = 0 \\ Me + \nabla h(z)\lambda - Q^T\mu - \tau = 0 \\ \lambda^T[g(x) + h(z)] = 0 \\ \tau^T z = 0 \\ z, \lambda, \tau \geq 0. \end{array} \tag{19}$$

Since $h(0) = 0$, the vector $(x^\star, 0)^T$ is a feasible point for (17), and, as for M sufficiently large we have

$$-\nabla h(0)\lambda^\star + Q^T\mu^\star \leq Me,$$

it is possible to find a value $\tau^\star \geq 0$ such that the vector $(x^\star, 0, \lambda^\star, \mu^\star, \tau^\star)^T$ is a solution of (19). Thus, from Proposition 3 we have that $(x^\star, 0)$ is a global optimum of problem 17 and the assertion is proved.

(ii) By contradiction let us assume that there exist a sequence of positive scalars $\{M^k\}$, with $M^k \to \infty$ for $k \to \infty$, and a corresponding sequence of vectors $\{(x^k, z^k)^T\}$ such that $z^k \neq 0$, and $(x^k, z^k)^T$ is solution of (17) when $M = M^k$. We can then define an infinite subset K such that, for all $k \in K$ we have $z_i^k > 0$ for some index $i \in \{1, \ldots, n_z\}$. Using (18) and the fact that $h(0) = 0$, we have

$$g(\bar{x}) + h(0) < 0.$$

By Proposition 2, there exist Lagrange multipliers $\lambda^\star, \mu^\star, \tau^\star$ such that $(x^\star, z^\star, \lambda^\star, \mu^\star, \tau^\star)^T$ is a solution of (19) when $M = M^k$. Then, using the complementarity condition

$${\tau^k}^T z^k = 0,$$

we obtain $\tau_i^k = 0$. Hence, we can write

$$\left(M + e_i^T \nabla h(z^\star)\lambda^\star - e_i^T Q^T \mu^\star\right) = 0 \qquad \forall k \in K,$$

which contradicts the fact that $M^k \to \infty$. □

We notice that problem (16) includes as special case the following convex programming problem:

$$\begin{array}{l} \min c^T x \\ g(x) \leq 0 \\ Ax = b. \end{array} \tag{20}$$

5 The Frank-Wolfe - Reduced Dimension algorithm

The FrankWolfe algorithm is a well-known algorithm in operations research. It was originally proposed by Marguerite Frank and Phil Wolfe in 1956 as a procedure for solving quadratic programming problems with linear constraints [8].
In this section, we first describe the algorithm and give some results about its convergence to

a stationary point. Then we propose a new efficient version of the Frank-Wolfe algorithm for solving problems of the following form:

$$\begin{aligned} &\min\ f(x) = g(x) + h(x) = g(x) + \sum_{j=1}^{n} h_j(x_j) \\ &x \in C \\ &x_i \geq 0,\ i \in I \subseteq \{1, \ldots, n\} \end{aligned} \tag{21}$$

where:

(i) C is a compact set having the following form:

$$C = \{x \in R^n : w_l(x_{\bar{I}}) + \sum_{i \in I} s_{li}(x_i) \leq 0,\ l = 1, \ldots, m;\quad Ax = b\} \tag{22}$$

where $A \in R^{p \times n}$, $x_{\bar{I}} = \{x_i :\ i \notin I\}$, $w_l : R^{n-|I|} \to R$, and $s_{li} : R \to R$, for $l = 1, \ldots, m$ and $i \in I$, are convex, continuously differentiable functions;

(ii) $g : R^n \to R$ is a continuously differentiable function;

(ii) $h_j : R \to R$, for $j = 1, \ldots, n$ are concave, continuously differentiable functions.

We further assume that $s_{li}(0) = 0$ for $l = 1, \ldots, m$ and $i \in I$.

Herein, we report the original version of the Frank-Wolfe Algorithm:

Frank-Wolfe Algorithm

1. Let $x^0 \in C$ be the starting point;

2. For $k = 0, 1, \ldots$

   obtain solution $\overline{x}^k$ by solving the following problem:

$$\overline{x}^k = \arg\min_{x \in C} \nabla f(x^k)^T (x - x^k) \tag{23}$$

3. if $\nabla f(x^k)^T (\overline{x}^k - x^k) = 0$ then STOP

4. Otherwise, define a feasible descent direction

$$d^k = \overline{x}^k - x^k$$

   and generate a new feasible vector

$$x^{k+1} = x^k + \alpha^k d^k$$

   with $\alpha^k \in (0, 1]$ determined by means of an Armijo-like rule.

The following result, proved in [1], provides an analysis of convergence behavior of the Frank-Wolfe Algorithm.

Proposition 5 *Let $\{x^k\}$ be a sequence generated by the Frank-Wolfe Algorithm*

$$x^{k+1} = x^k + \alpha^k d^k.$$

Assume that method used for choosing stepsize α^k satisfies the following conditions:

(i) $f(x^{k+1}) < f(x^k)$, with $\nabla f(x^k) \neq 0$;

(ii) if $\nabla f(x^k) \neq 0 \ \ \forall\ k$, then we have

$$\lim_{k\to\infty} \nabla f(x^k)^T d^k = 0\ .$$

Then every limit point $\bar{x}$ of $\{x^k\}$ is a stationary point.

The next proposition shows that, under suitable conditions on the concave functions h_j, the Frank-Wolfe algorithm does not change a nonnegative variable once that it has been fixed to zero.

Proposition 6 *Let $\{x^k\}$ be any sequence generated by the Frank-Wolfe algorithm. There exists a value M such that, if $i \in I$ and*

$$h_i^{'}(0) \geq M$$

then we have that

$$x_i^k = 0 \quad \text{implies} \quad x_i^{k+1} = 0.$$

Proof. At each iteration k of the Frank-Wolfe algorithm the problem to be solved is

$$\begin{array}{l} \min \sum_{j=1}^{n} \nabla g_j(x^k) x_j + \sum_{j:x_j^k \neq 0} h_j'(x_j^k)\ x_j + \sum_{j \notin I : x_j^k = 0} h_j'(0) x_j + \sum_{j \in I : x_j^k = 0} h_j'(0) x_j \\ x \in C \\ x_i \geq 0,\ i \in I \subseteq \{1, \ldots, n\} \end{array} \tag{24}$$

Let $\bar{x}^k$ be a solution of (24). As g is continuously differentiable and C is compact, there exists a value $L < \infty$ such that

$$\|\nabla g(x)\|_\infty \leq L \quad \forall\ x \in C. \tag{25}$$

For any $i \in I$ such that $x_i^k = 0$, by (ii) of Proposition 4 it follows that there exists a value S such that if $\nabla g_i(x^k) + h_i^{'}(0) \geq S$ then we have $\bar{x}_i^k = 0$. Thus, if $i \in I$, $x_i^k = 0$ and $h_i^{'}(0) \geq M = S + L$, then we obtain

$$x_i^{k+1} = x_i^k + \alpha^k(\bar{x}_i^k - x_i^k) = 0.$$

□

On the basis of Proposition 6 we can define the following version of the Frank-Wolfe algorithm, where the convex problems to be solved are of reduced dimension. We denote by Ω the feasible set of problem (21), i.e.,

$$\Omega = \{x \in R^n : x \in C, x_i \geq 0, i \in I\}.$$

Frank-Wolfe - Reduced Dimension (FW-RD) Algorithm

1. Let $x^0 \in C$ be the starting point;

2. For $k = 0, 1, \ldots$, let $I^{x^k} = \{i \in I : x_i^k = 0\}$ and $C^{x^k} = \{x \in \Omega : x_i = 0, \forall\ i \in I^{x^k}\}$

   obtain solution $\overline{x}^k$ by solving the following problem:

$$\overline{x}^k = \arg \min_{x \in C^{x^k}} \nabla f(x^k)^T (x - x^k) \tag{26}$$

3. if $\nabla f(x^k)^T (\overline{x}^k - x^k) = 0$ then STOP

4. Otherwise, define a feasible descent direction

$$d^k = \overline{x}^k - x^k$$

   and generate a new feasible vector

$$x^{k+1} = x^k + \alpha^k d^k$$

   with $\alpha^k \in (0, 1]$ determined by means of an Armijo-like rule.

Note that the convex programming problem (26) is equivalent to a convex problem of dimension $n - |I^{x^k}|$, and that $I^{x^k} \subseteq I^{x^{k+1}}$, so that the problems to be solved are of nonincreasing dimensions. This yields obvious advantages in terms of computational time.

In order to show the convergence of the algorithm, we report here some definitions about correspondences (see [7] for further details):

Definition 2 *Let Θ and S be subsets of R^l and R^n, respectively. A* correspondence Φ *from Θ to S is a map that associates with each element $\theta \in \Theta$ a (nonempty) subset $\Phi(\theta) \subset S$.*

We denote a correspondence as follows

$$\Phi : \Theta \to P(S)$$

where P(S) denotes the power set of S, i.e. the set of all nonempty subsets of S.

Definition 3 *A correspondence is said to be* upper-semicontinuous *at a point $\theta \in \Theta$ if for any sequence $\{\theta^k\}$ converging to θ, and for any sequence $\{s^k\}$ converging to s, with $s^k \in \Phi(\theta^k)$, we have $s \in \Phi(\theta)$. Φ is upper-semicontinuous on Θ if is upper-semicontinuous at each $\theta \in \Theta$.*

Definition 4 *A correspondence is said to be* lower-semicontinuous *at a point $\theta \in \Theta$ if for any sequence $\{\theta^k\}$ converging to θ, and for any $s \in \Phi(\theta)$ there exists a sequence $\{s^k\}$ converging to s, with $s^k \in \Phi(\theta^k)$. Φ is lower-semicontinuous on Θ if is lower-semicontinuous at each $\theta \in \Theta$.*

Definition 5 *A correspondence $\Phi : \Theta \to P(S)$ is said to be* continuous *at a point $\theta \in \Theta$ if is lower-semicontinuous and upper-semicontinuous at θ. Φ is continuous on Θ if is lower and upper-semicontiuous at each $\theta \in \Theta$.*

It is easy to see that the correspondence C^x is lower-semicontinuous. Now we can formally prove the convergence of the proposed algorithm to a stationary point.

Proposition 7 *Let $\{x^k\}$ be a sequence generated by the FW-RD Algorithm*

$$x^{k+1} = x^k + \alpha^k d^k.$$

Assume that method used for choosing stepsize α^k satisfies the following conditions:

(i) $f(x^{k+1}) < f(x^k)$, with $\nabla f(x^k) \neq 0$;

(ii) if $\nabla f(x^k) \neq 0 \ \ \forall\ k$, then we have

$$\lim_{k\to\infty} \nabla f(x^k)^T d^k = 0\ .$$

Suppose there exists a value S such that $h_i^{'}(0) \geq S\ \forall\ x_i = 0$ with $i \in I$, then every limit point $\bar{x}$ of $\{x^k\}$ is a stationary point.

Proof. As we assumed compactness of C, a limit point $\bar{x} \in P$ exists and the norm of vector d^k is bounded above

$$\|d^k\| = \|\bar{x}^k - x^k\| \leq \|\bar{x}^k\| + \|x^k\|\ .$$

We can now define a subsequence $\{x^k\}_K$ such that

$$\lim_{k\to\infty, k\in K} x^k = \bar{x}, \quad \lim_{k\to\infty, k\in K} d^k = \bar{d}\ .$$

By using hypothesis (ii), we obtain

$$\nabla f(\bar{x})^T \bar{d} = 0.$$

Let d^k be a direction generated by the Frank-Wolfe method; we have

$$\nabla f(x^k)^T d^k \leq \nabla f(x^k)^T (x - x^k),\ \forall\ x \in C^{x^k}. \tag{27}$$

We want to show that, by taking the limit as $k \in K$, $k \to \infty$, we obtain

$$0 = \nabla f(\bar{x})^T \bar{d} \leq \nabla f(\bar{x})^T (x - \bar{x}),\ \forall\ x \in C^{\bar{x}}.$$

By contradiction, let us assume there exists a point $\tilde{s} \in C^{\bar{x}}$ satisfying the following inequality

$$\nabla f(\bar{x})^T \bar{d} > \nabla f(\bar{x})^T (\tilde{s} - \bar{x}). \tag{28}$$

By lower-semicontinuity of the correspondence C^x, as $\tilde{s} \in C^{\bar{x}}$, there exists a subsequence $\{s^k\}_K$ converging to $\tilde{s}$, with $s^k \in C^{x^k}$. For k sufficiently large we have from inequality (28)

$$\nabla f(x^k)^T d^k > \nabla f(x^k)^T (s^k - x^k),$$

but this contradicts (27).

Now we prove that $\bar{x}$ is a stationary point. Indeed, $\bar{x}$ is a solution of

$$\begin{array}{l} \min \nabla f(\bar{x})^T x = \min \sum\limits_{j:\bar{x}_j \neq 0} (\nabla g_j(\bar{x}) + h_j'(\bar{x}_j))\ x_j + \sum\limits_{j \notin I^{\bar{x}}: \bar{x}_j = 0} (\nabla g_j(\bar{x}) + h_j'(0))\ x_j \\ x \in \Omega \\ x_i = 0,\ i \in I^{\bar{x}}. \end{array} \tag{29}$$

As g is continuously differentiable and C is compact, there exists a value $L < \infty$ such that

$$\|\nabla g(\bar{x})\|_\infty \leq L \tag{30}$$

and by (i) of Proposition 4 it follows that there exists a value S such that, if $h_j'(0) \geq S = M + L$ then $\bar{x}$ is a solution of

$$\begin{array}{l} \min \sum\limits_{j:\bar{x}_j \neq 0} (\nabla g_j(\bar{x}) + h_j'(\bar{x}_j))\ x_j + \sum\limits_{j \notin I^{\bar{x}}: \bar{x}_j = 0} (\nabla g_j(\bar{x}) + h_j'(0))\ x_j + \sum\limits_{j \in I^{\bar{x}}: \bar{x}_j = 0} (\nabla g_j(\bar{x}) + h_j'(0))\ x_j \\ x \in \Omega \end{array} \tag{31}$$

Therefore we have

$$\nabla f(\bar{x})^T \bar{x} \leq \nabla f(\bar{x})^T x \quad \forall x \in \Omega,$$

and this proves that $\bar{x}$ is a stationary point of problem (21). □

Concerning the separable concave functions used in problems (8), (9), (10), (11), we have for $j = 1, \ldots, n$

- $h_j(y_j; \alpha) = 1 - e^{-\alpha y_j}$ and $h_j'(0) = \alpha$;
- $h_j(y_j; \epsilon) = \ln(y_j + \epsilon)$ and $h_j'(0) = 1/\epsilon$;
- $h_j(y_j; \epsilon, p) = (y_j + \epsilon)^p$ and $h_j'(0) = p(\epsilon)^{p-1}$ with $0 < p < 1$;
- $h_j(y_j; \epsilon, p) = -(y_j + \epsilon)^{-p}$ and $h_j'(0) = p(\epsilon)^{-p-1}$ with $1 \leq p$;

Therefore, the assumption of Proposition 7 holds for suitable values of the parameters of the above concave functions, so that Algorithm FW-RD can be applied.

6 The Frank-Wolfe - Reduced Dimension algorithm with unitary stepsize

When the function f of Problem (21) is concave, we can use a constant stepsize $\alpha = 1$ and still be sure the algorithm converges to a stationary point. The following proposition shows convergence of the Frank-Wolfe algorithm with stepsize $\alpha^k = s$ and $s \in (0, 1]$ when a concave function is minimized over a compact convex set:

Proposition 8 *Let f be a continuously differentiable, concave function. Let $\{x^k\}$ be a sequence generated by the Frank-Wolfe algorithm*

$$x^{k+1} = x^k + \alpha^k d^k\ ,$$

where a constant stepsize is chosen

$$\alpha^k = s,\ k = 0, 1, \ldots$$

with $s \in (0, 1]$. Then every limit point $\bar{x}$ of $\{x^k\}$ is a stationary point.

Proof. we have from concavity of f:

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^T(x^{k+1} - x^k) < h(x^k) \ .$$

Note that since $\{f(x^k)\}$ is monotonically decreasing, $\{f(x^k)\}$ either converges to a finite value or diverges to $-\infty$.
Let $\bar{x}$ be a limit point of $\{x^k\}$; since f is continuous $f(\bar{x})$ is a limit point of $\{f(x^k)\}$, so it follows that the entire sequence converges to $f(\bar{x})$. Therefore, we obtain

$$f(x^k) - f(x^{k+1}) \to 0$$

From concavity of f:

$$f(x^k) - f(x^{k+1}) \geq -\alpha^k \nabla f(x^k)^T d^k \ .$$

Since α^k is a constant stepsize, we have that

$$\nabla f(x^k)^T d^k \to 0 \ .$$

By Proposition 5 it follows that every limit point $\bar{x}$ of $\{x^k\}$ is a stationary point. □

Here is a version of the modified Frank-Wolfe Algorithm, with unitary stepsize, for concave functions:

Frank-Wolfe - Reduced Dimension Algorithm with Unitary Stepsize (FW-RDUS)

1. Let $x^0 \in C$ be the starting point;

2. For $k = 0, 1, \ldots$, let $I^{x^k} = \{i \in I : x_i^k = 0\}$ and $C^{x^k} = \{x \in \Omega : x_i = 0, \forall\ i \in I^{x^k}\}$

   obtain solution $\overline{x}^k$ by solving the following problem:

$$\overline{x}^k = \arg \min_{x \in C^{x^k}} \nabla f(x^k)^T (x - x^k) \tag{32}$$

3. if $\nabla f(x^k)^T(\overline{x}^k - x^k) = 0$ then STOP

4. Otherwise

$$x^{k+1} = \bar{x}^k.$$

The following result about the convergence of the FW-RDUS algorithm is an immediate consequence of Proposition 7.

Corollary 1 *Let $\{x^k\}$ be a sequence generated by the FW-RDUS Algorithm. Suppose there exists a value S such that $h_i^{'}(0) \geq S\ \forall\ x_i = 0$ with $i \in I$, then every limit point $\bar{x}$ of $\{x^k\}$ is a stationary point.*

The assumption of Corollary 1 holds for suitable values of the parameters of the concave functions presented in Section 2, so that Algorithm FW-RDUS can be applied when using those functions. The results obtained on computational experiments will be presented in the next section.

7 Computational experiments

In our computational experiments we have considered problem (4). We remark that the aim of experiments has been that of evaluating the effectiveness of the various formulations in finding sparse vectors (possibly the sparsest vectors) belonging to a convex set.

Test problems

For several values of n and m we randomly generated the matrix A, the vector b, and a value of the tolerance δ_1. Then we obtained two more values of the tolerance as follows: $\delta_2 = 2\delta_1$; $\delta_3 = 4\delta_1$. For each problem we performed experiments using:

- formulation (8), denoted by *exp*, with $\alpha = 5$;
- formulation (9), denoted by *log*, with $\epsilon = 10^{-5}$;
- formulation (10), denoted *Formulation II*, with $\epsilon = 10^{-7}$ and $p = 0.1$;
- formulation (11), denoted by *Formulation II*, with $\epsilon = 10^{-5}$ and $p = 1$.

We also report the results obtained using the ℓ_1 norm formulation:

$$\begin{array}{l} \min\limits_{x \in R^n} \ \|x\|_1 \\ \\ \|Ax - b\|^2 \leq \delta \end{array} \tag{33}$$

denoted by ℓ_1.

Implementation details

Algorithms FW and FW-RDUS were implemented in C using CPLEX (10.0) as solver of the quadratic programming problems. The experiments were carried out on Intel Pentium 4 3.2 GHz 1.0 GB RAM.

Results

The results obtained on the randomly generated problems are shown in Table 1, where we report

- the number n of variables, the number m of constraints;
- for formulation ℓ_1, the zero-norm of the optimal solution attained;
- for each nonlinear concave formulation:
  - the average of the zero-norm value of the stationary points determined;
  - the best zero-norm value of those stationary points;
  - percentage of runs where the best zero-norm value was attained.

From Table 1 we can see that *Formulation I* gives the best results among all the formulations. We further note that the results obtained by means of the concave formulations are clearly better than those corresponding to the ℓ_1 formulation.

Summarizing, the computational experiments confirm the effectiveness of the concave-based approach for finding sparse solutions to problems with convex constraints, and show that the concave formulations here proposed represent good alternatives to the ℓ_1 formulation. We remark that a wider availability of efficient formulations is important as it can make easier the search of sparse solutions for different classes of problems.

Finally, in order to assess the differences in terms of computational time between the standard Frank-Wolfe (FW) algorithm and a new version of the algorithm presented in the preceding section and denoted by Algorithm FW-RDUS, we report in Table 2 the results obtained by the two algorithms using *log formulation*. As we might expect, the differences are noticeable and show the usefulness of Algorithm FW-RDUS. Further experiments not here reported and performed using the other concave formulations point out the same differences between the two algorithms in terms of computational time. In all the tests we detected no difference between the two algorithms in terms of computed solution.

Problem	n	m	δ	$\boldsymbol{l_1}$	**exp**	**log**	**Form. I**	**Form. II**
1	100	20	0.93	8	7.6/4/12	4.4/4/58	4.4/4/58	5.2/4/30
2	100	20	1.86	5	5.0/2/25	2.1/2/92	2.0/2/97	3.0/2/52
3	100	20	3.71	3	3.0/1/43	1.1/1/99	1.0/1/100	1.6/1/80
4	200	40	3.00	19	14.6/6/18	6.7/6/39	6.7/6/42	8.1/6/18
5	200	40	6.00	10	10.0/3/10	3.8/3/18	3.3/3/17	4.7/3/18
6	200	40	12.01	4	6.2/2/26	2.0/2/100	2.0/2/100	3.0/2/72
7	400	80	13.24	36	29.5/16/1	13.7/12/4	13.6/12/5	17.0/13/10
8	400	80	26.49	30	20.7/7/1	6.1/6/94	6.1/6/95	8.9/6/34
9	400	80	52.99	5	11.7/4/10	4.0/4/100	4.0/4/100	5.5/4/52
10	800	160	58.77	80	57.1/43/1	26.6/23/2	26.0/23/3	38.3/24/1
11	800	160	117.54	42	39.7/22/1	11.9/11/21	11.8/11/24	20.8/11/1
12	800	160	235.08	16	21.9/7/1	7.0/7/100	7.0/7/100	10.0/7/46
13	1600	320	263.96	147	109.5/75/1	48.4/45/5	48.1/44/1	92.0/48/1
14	1600	320	527.92	82	73.1/29/1	20.2/20/77	20.2/20/81	56.3/21/1
15	1600	320	1055.80	22	37.7/14/2	12.2/12/75	12.6/12/37	19.7/12/15

Table 1: Comparison on Test problems.

Problem	FW	FW-RDUS
1	0.453	0.094
2	0.141	0.047
3	0.140	0.032
4	1.000	0.188
5	0.890	0.141
6	0.625	0.109
7	8.219	1.015
8	7.515	1.563
9	6.579	1.359
10	73.015	6.391
11	81.656	7.437
12	76.391	4.141
13	767.657	93.094
14	866.719	51.89
15	812.609	46.32

Table 2: Comparison using log Formulation between the two versions of the Frank-Wolfe algorithm in terms of CPU-time (seconds).

8 Conclusions and future work

In this work, we have considered the problem of finding a sparse solution to a problem with convex constraints, which arises in different important fields, such as signal processing and data analysis. We have proposed a concave optimization-based approach for dealing with this issue. Furthermore, we described a new efficient version of the Frank-Wolfe algorithm and we proved its convergence to a stationary point.
The computational experiments evidenced that the concave formulations can be valid alternatives to the ℓ_1 formulation, as in most cases they get sparser solutions. The results we report also show a considerable speed-up when using the variable fixing variant of the Franke-Wolfe method in place of the traditional one. This speed-up might be extremely beneficial when multiple runs of the algorithm are performed, e.g. in a Multistart method.
Future work will be devoted to the development of global optimization algorithms for finding sparse solutions to problems having convex constraints and to the definition of suitable techniques for SLDA, SPCA and sparse representation of noisy signals.

References

[1] D.P. Bertsekas, *Nonlinear Programming*, 2nd edn., Athena Scientific, 1999.

[2] P. S. Bradley, O. L. Mangasarian, *Feature selection via concave minimization and support vector machines*, in Proceedings of the 15^{th} International Conference on Machine Learning (ICML '98), J. Shavlik, editor, Morgan Kaufmann, San Francisco, California, pp. 82-90, 1998.

[3] A.M. Bruckstein, D. L. Donoho, M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, Siam Review, 51(1), pp 34–81, 2009.

[4] J. Cadima, I.T. Jolliffe, *Loading and correlations in the interpretation of principal components*, Journal of Applied Statistics, 22, pp. 203–214, 1995.

[5] S.S. Chen, D.L. Donoho, M.A. Saunders, *Atomic decomposition basis pursuit*, SIAM Rev., 43, pp. 129–159, 2001.

[6] A. d'Aspremont, L. El Ghaoui, M.I. Jordan, G.R.G. Lanckriet, *A direct formulation for sparse PCA using semidefinite programming*, SIAM Rev., 49(3), pp. 434–448, 2007.

[7] F. Facchinei, J.S. Pang, *Finite-dimensional variational inequalities and complementarity problems*, Springer-Verlag, New York, 2003.

[8] M. Frank, P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly, 3, pp. 95–110, 1956.

[9] O.L. Mangasarian, *Machine learning via polyhedral concave minimization*, in "Applied Mathematics and Parallel Computing – Festschrift for Klaus Ritter", H. Fischer, B. Riedmueller, S. Schaeffler, editors, Physica-Verlag, Germany, pp. 175-188, 1996.

[10] B. Moghaddam, Y. Weiss, S. Avidan, *Generalized Spectral Bounds for Sparse LDA*, in Proceedings of the 23^{rd} International Conference on Machine Learning, Pittsburgh, PA, 2006.

[11] F. Rinaldi, F. Schoen, M. Sciandrone, *Concave programming for minimizing the zero-norm over polyhedral sets*, in Comput. Opt. Appl., to Appear.

[12] J. Weston, A. Elisseef, B. Schölkopf, *Use of the zero-norm with linear models and kernel model*, Journal of Machine Learning Research, 3, pp. 1439–1461, 2003.

[13] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Royal Statist. Soc. Ser. B, Vol. 58(1), pp. 267–288, 1996.

[14] H. Zou, T. Hastie, R. Tibshirani, *Sparse principal component analysis*, Journal of Computer Graphic and Statistics, 15, pp. 256–286, 2006.