Journal of Global Optimization
https://doi.org/10.1007/s10898-021-01089-w

Sparse regularization via bidualization

Amir Beck[1] · **Yehonathan Refael**[1]

Received: 10 June 2020 / Accepted: 18 August 2021

Abstract
The paper considers the sparse envelope function, defined as the biconjugate of the sum of a squared ℓ_2-norm function and the indicator of the set of k-sparse vectors. It is shown that both function and proximal values of the sparse envelope function can be reduced into a one-dimensional search that can be efficiently performed in linear time complexity in expectation. The sparse envelope function naturally serves as a regularizer that can handle both sparsity and grouping information in inverse problems, and can also be utilized in sparse support vector machine problems.

Keywords Biconjugate · Sparsity · Convex envelope · Randomized root search

1 Introduction

In this paper we investigate the extended real-valued function $s_k : \mathbb{R}^n \to (-\infty, \infty]$ given by

$$s_k(\mathbf{x}) = \begin{cases} \frac{1}{2}\|\mathbf{x}\|_2^2, & \|\mathbf{x}\|_0 \leq k, \\ \infty, & \text{else}, \end{cases}$$

where k is a positive integer and $\|\cdot\|_0$ is the so-called ℓ_0-norm[1] function that counts the number of nonzero elements in the input vector: $\|\mathbf{x}\|_0 \equiv \#\{i : x_i \neq 0\}$. Two motivating examples for considering s_k are given in the next subsection.

1.1 Prototype examples

(I) Sparse and grouping inducing regularizer Many inverse problems in science have the form

[1] Obviously, the ℓ_0-"norm" is not actually a norm.

✉ Amir Beck
becka@tauex.tau.ac.il

Yehonathan Refael
yonatanrefael100@gmail.com

[1] School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel

Published online: 25 September 2021

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + s(\mathbf{x}), \tag{1.1}$$

where f is a data fidelity term (e.g., least squares) and s is a regularizer that models some kind of an a priori knowledge on the vector that needs to be recovered. In many applications, it is reasonable to assume that the sought vector should have a small number of nonzero components, meaning that it is sparse. Perhaps the most natural regularizer in this case is the ℓ_0-norm. Unfortunately, the ℓ_0-norm is a difficult function to handle being nonconvex and even non-continuous. One way to circumvent this difficulty is by replacing the ℓ_0-norm by the ℓ_1-norm, which is also a sparsity-inducing regularizer. It was actually shown that in some important settings, the usage of the ℓ_1-norm leads to the same solution as the one that would be obtained by using the ℓ_0-norm, see for example the review paper [5] and references therein. One extremely popular model is to take f as a least squares fidelity term and s as an ℓ_1 expression, leading to one of the formulations of the so-called LASSO problem [11].

In [13] it was observed that in some statistical applications that posses the "grouping effect", the ℓ_1 regularizer does not yield satisfactory results. This is why the elastic net regularizer was proposed in [13]; the elastic net regularizer function is a weighted sum of the squared ℓ_2-norm and the ℓ_1-norm.

The function s_k, much like the elastic net function, also takes into account sparsity and grouping properties. The grouping property is handled as in the elastic net regularizer, by a squared ℓ_2-norm, but the sparsity property is treated in a straightforward manner by taking into account the ℓ_0-norm constraint.

(II) Support Vector Machines In the linear separation problem, we are given n vectors in a p-dimensional space $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ that belong to two classes described by the labels vector $\mathbf{y} \in \{-1, 1\}^n$ ($y_i = 1$ means that $\mathbf{x}_i$ belongs to the first class; otherwise, it belongs to the second class). The support vector machine (SVM) problem finds a hyperplane $H_{\mathbf{w},\beta} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{w}^T\mathbf{x} + \beta = 0\}$ that aims to separate the two classes with a small classification error.

A well-known formulation of the SVM problem is given by (see e.g., [7])

$$\begin{array}{ll} \min & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\mathbf{e}^T\boldsymbol{\xi} \\ \text{s.t.} & \mathbf{Y}(\mathbf{X}\mathbf{w} + \beta\mathbf{e}) \geq \mathbf{e} - \boldsymbol{\xi}, \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{1.2}$$

where the decision variables are $\mathbf{w} \in \mathbb{R}^p$, $\boldsymbol{\xi} \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$. The parameters are $\mathbf{Y} = \text{diag}(\mathbf{y})$ (the diagonal matrix whose diagonal elements are the components of $\mathbf{y}$), the data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ whose rows are $\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T$ and the constraint violation parameter $C > 0$.

Now consider the case where we seek to find a linear separator that is sparse, meaning with only a few nonzero elements. In this way, we perform classification and feature selection simultaneously, see for example the paper [12]. Given that we know a bound on the sparsity level of the separator, a natural mathematical formulation of the problem would be to incorporate an ℓ_0-norm constraint in problem (1.2):

$$\begin{array}{ll} \min & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\mathbf{e}^T\boldsymbol{\xi} \\ \text{s.t.} & \|\mathbf{w}\|_0 \leq k, \\ & \mathbf{Y}(\mathbf{X}\mathbf{w} + \beta\mathbf{e}) \geq \mathbf{e} - \boldsymbol{\xi}, \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array}$$

Clearly, as highlighted in blue in the above formulation, the problem can be written explicitly in terms of s_k as

$$\begin{aligned} \min\ & s_k(\mathbf{w}) + C\mathbf{e}^T\boldsymbol{\xi} \\ \text{s.t.}\ & \mathbf{Y}(\mathbf{X}\mathbf{w} + \beta\mathbf{e}) \geq \mathbf{e} - \boldsymbol{\xi}, \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{aligned}$$

1.2 Convexifying s_k

Unfortunately, s_k is a nonconvex and noncontinuous function, and therefore, it is in general a difficult task to solve optimization problems incorporating it. Therefore, the path taken in this paper is to consider the best convex estimator of the function, namely the biconjugate function $\mathcal{S}_k = s_k^{**}$, which we call *the sparse envelope function*. In Sect. 2 we show that $\mathcal{S}_k$ is equal to half times the square of the so-called k-support norm, introduced and studied in [1]. We show that the evaluation of the sparse envelope function reduces to a monotone one-dimensional root search problem that can be solved in linear time complexity in expectation by a randomized root search method.

In Sect. 3 we consider the proximal mapping of the sparse envelope function and show that it can also be computed in linear time using a reduction to a one-dimensional monotone root search. The obtained $O(n)$ linear complexity result is an improvement of the[2] $O(n \log n)$ and $O(n(k + \log n))$ complexities known [1] for evaluating the function value and proximal operator of the k-support norm. The ability to efficiently compute the proximal operator of the sparse envelope function implies that it is possible to employ fast proximal gradient methods such as FISTA [3] to solve the composite problem (1.1) in the case where f is convex and smooth. Section 4 describes how to construct a dual problem of a relaxation of the sparse SVM problem which is based on the composite envelope function. The dual formulation has a simple constraint set, and a smooth objective function, and can thus be tackled through accelerated first-order methods.

Notation The underlying space in the paper is $\mathbb{R}^n$ - the space of all real-valued n-length column vectors endowed with the dot product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{y}$. For $p \geq 1$, The ℓ_p-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is given by $\|\mathbf{x}\|_p \equiv \sqrt[p]{\sum_{i=1}^n |x_i|^p}$. The ℓ_∞-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is $\|\mathbf{x}\|_\infty = \max_{i=1,2,\ldots,n} |x_i|$. $\mathbf{e}$ is the vector of all ones and $\mathbf{0}$ is the vector of all zeros. For a positive integer m, we denote $[m] \equiv \{1, 2, \ldots, m\}$. For a vector $\mathbf{x} \in \mathbb{R}^n$ and $k \in [n]$, the n-dimensional vector $H_k(\mathbf{x})$ is a vector generated by keeping the k largest absolute value components of $\mathbf{x}$ and setting all the others to zeros; the set of indices of the k largest absolute value components is not unique, and we assume that in those situations an arbitrary vector $H_k(\mathbf{x})$ is chosen. It is well-known that $H_k(\mathbf{x})$ is a vector which is closest to $\mathbf{x}$ among all the k-sparse vectors, meaning that (see e.g., [2, Section 6.8.3])

$$H_k(\mathbf{x}) \in \operatorname*{argmin}_{\mathbf{y} \in C_k} \|\mathbf{y} - \mathbf{x}\|_2, \tag{1.3}$$

where C_k is the set of all k-sparse vectors: $C_k = \{\mathbf{x} : \|\mathbf{x}\|_0 \leq k\}$. Given an extended real-valued function $g : \mathbb{R}^n \to (-\infty, \infty]$, its conjugate is given by

$$g^*(\mathbf{y}) = \max_{\mathbf{x} \in \mathbb{R}^n} \left\{ \mathbf{x}^T\mathbf{y} - g(\mathbf{x}) \right\}.$$

[2] n being the underlying dimension and k being the sparsity level.

2 Sparse envelope evaluation

2.1 Reduction to one-dimensional search

As mentioned in the introduction, the paper is concerned with the function

$$s_k(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2 + \delta_{C_k}(\mathbf{x}). \tag{2.1}$$

The main purpose of the section is to find the biconjugate of s_k, which we call *the sparse envelope function*.

Definition 2.1 Let $k \in [n]$. The sparse envelope function with sparsity level k, denoted by $\mathcal{S}_k$, is the biconjugate of s_k:

$$\mathcal{S}_k = s_k^{**}.$$

By its definition, $\mathcal{S}_k$ is proper closed and convex. Throughout the paper, we will assume that $k \in [n]$ is given and fixed. Our ultimate goal in this section is to construct an efficient method for computing $\mathcal{S}_k(\mathbf{x}) = s_k^{**}(\mathbf{x})$ at a given $\mathbf{x} \in \mathbb{R}^n$. We begin by proving that the conjugate function s_k^* is the squared norm of the k-hard thresholding function.

Lemma 2.1 $s_k^*(\mathbf{y}) = \frac{1}{2}\|H_k(\mathbf{y})\|_2^2$ *for any* $\mathbf{y} \in \mathbb{R}^n$.

Proof Let $\mathbf{y} \in \mathbb{R}^n$. Then

$$\begin{aligned} s_k^*(\mathbf{y}) &= \max_{\mathbf{x}\in\mathbb{R}^n}\{\mathbf{x}^T\mathbf{y} - s_k(\mathbf{x})\} = \max_{\mathbf{x}\in C_k}\left\{\mathbf{x}^T\mathbf{y} - \frac{1}{2}\|\mathbf{x}\|_2^2\right\} = \max_{\mathbf{x}\in C_k}\left\{-\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 + \frac{1}{2}\|\mathbf{y}\|_2^2\right\} \\ &\overset{(*)}{=} -\frac{1}{2}\|H_k(\mathbf{y}) - \mathbf{y}\|_2^2 + \frac{1}{2}\|\mathbf{y}\|_2^2 \\ &\overset{(**)}{=} \frac{1}{2}\|H_k(\mathbf{y})\|_2^2, \end{aligned}$$

where (*) follows from (1.3), and (**) follows by the fact that $\|H_k(\mathbf{y}) - \mathbf{y}\|_2^2$ is the sum of squares of the $n-k$ components of $\mathbf{y}$ with the smallest absolute values and $\frac{1}{2}\|\mathbf{y}\|_2^2$ is the sum of squares of all the components of $\mathbf{y}$. □

Remark 2.1 (connection to the k-support norm) The k-support norm, denoted by $\|\cdot\|_k^{sp}$ was introduced in [1], and is defined as the norm whose unit ball is given by

$$\mathrm{conv}(\{\mathbf{w} : \|\mathbf{w}\|_0 \le k, \|\mathbf{w}\|_2 \le 1\}).$$

It was shown in [1, Section 2.1] that the dual norm of the k-support norm is given by $\|\cdot\|_H \equiv \|H_k(\cdot)\|_2$, that is, $\|\cdot\|_{H*} = \|\cdot\|_k^{sp}$ ($\|\cdot\|_{H*}$ denotes the dual norm of $\|\cdot\|_H$). Therefore, by [2, Section 4.4.15],

$$\mathcal{S}_k = (s_k^*)^* = \left(\frac{1}{2}\|\cdot\|_H^2\right)^* = \frac{1}{2}(\|\cdot\|_{H*})^2 = \frac{1}{2}(\|\cdot\|_k^{sp})^2.$$

The conclusion is that the sparse envelope function equals half times the square of the k-support norm. The analysis that will follow will be aimed at showing how the sparse envelope function and its proximal operator can be efficiently computed in expected *linear* time by randomized root search methods, as opposed to the complexities of $O(n\log n)$ and $O(n(k+\log n))$ currently known [1] for the computations of the function and proximal operator of the k-support norm.

The biconjugate function of s_k is the conjugate of s_k^*, which according to Lemma 2.1 is given by

$$\mathcal{S}_k(\mathbf{x}) = s_k^{**}(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \mathbf{x}^T\mathbf{y} - \frac{1}{2}\|H_k(\mathbf{y})\|_2^2 \right\}. \tag{2.2}$$

It is easy to see that

$$\|H_k(\mathbf{y})\|_2^2 = \max_{\mathbf{u}\in D_k} \sum_{i=1}^n u_i y_i^2, \tag{2.3}$$

where $D_k = \{\mathbf{u} \in \mathbb{R}^n : \mathbf{e}^T\mathbf{u} \le k, \mathbf{0} \le \mathbf{u} \le \mathbf{e}\}$. In particular, an optimal solution of the maximization problem in (2.3) is a vector with k ones at the coordinates corresponding to the k largest absolute values in $\mathbf{y}$, and zeros elsewhere. Plugging (2.3) into (2.2), we obtain that

$$\mathcal{S}_k(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^n} \min_{\mathbf{u}\in D_k} \left\{ \mathbf{x}^T\mathbf{y} - \frac{1}{2}\sum_{i=1}^n u_i y_i^2 \right\}.$$

Since the above problem is convex in $\mathbf{u}$ and concave in $\mathbf{y}$, by Sion's minimax theorem [9, Theorem 36.3], we can replace the roles of $\mathbf{u}$ and $\mathbf{y}$, and obtain the following expression:

$$\mathcal{S}_k(\mathbf{x}) = \min_{\mathbf{u}\in D_k} \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ \mathbf{x}^T\mathbf{y} - \frac{1}{2}\sum_{i=1}^n u_i y_i^2 \right\}.$$

The optimal value of the inner maximization problem is

$$\frac{1}{2}\sum_{i=1}^n \phi(x_i, u_i),$$

where ϕ is the well-known "quadratic over linear" function given by

$$\phi(x,u) \equiv \begin{cases} \frac{x^2}{u}, & u > 0, \\ 0, & x = u = 0, \\ \infty, & \text{else.} \end{cases} \tag{2.4}$$

This function is known to be proper closed and convex, and it is an important example of a closed convex function that is not continuous (see for example [9, p. 83]). Lemma 2.2 below summarizes the above discussion and presents a variational formula for $\mathcal{S}_k$ that will be the key ingredient in the development of an efficient algorithm for computing its value.

Lemma 2.2 *For any* $\mathbf{x} \in \mathbb{R}^n$,

$$\mathcal{S}_k(\mathbf{x}) = \frac{1}{2}\min_{\mathbf{u}\in D_k} \sum_{i=1}^n \phi(x_i, u_i). \tag{2.5}$$

Our next task is to construct a more explicit expression for $\mathcal{S}_k$. For that, we will construct a dual problem to the minimization problem in (2.5). Associating a Largrange multiplier only to the inequality constraint $\mathbf{e}^T\mathbf{u} \le k$ (defining D_k), and disregarding the constant $\frac{1}{2}$, we obtain the following Lagrangian function for the minimization problem in (2.5):

$$L(\mathbf{u}, \mu) = \sum_{i=1}^n (\phi(x_i, u_i) + \mu u_i) - k\mu.$$

The dual objective function is therefore given by

$$q(\mu) \equiv \min_{\mathbf{u}:0\le\mathbf{u}\le\mathbf{e}} L(\mathbf{u},\mu) = \sum_{i=1}^{n} \varphi_{x_i,0}(\mu) - k\mu, \tag{2.6}$$

where for any $b \in \mathbb{R}$ and $\alpha \ge 0$, the function $\varphi_{b,\alpha}$ is defined by

$$\varphi_{b,\alpha}(\mu) \equiv \min_{0\le u\le 1} \{\phi(b, \alpha + u) + \mu u\}, \quad \mu \ge 0. \tag{2.7}$$

Utilizing strong duality [9], we can conclude that the problem of evaluating $\mathcal{S}_k$ is equivalent to a one-dimensional concave maximization problem.

Lemma 2.3 *For any* $\mathbf{x} \in \mathbb{R}^n$,

$$2\mathcal{S}_k(\mathbf{x}) = \max_{\mu\ge 0} q(\mu) \tag{2.8}$$

with q *being the concave function defined by*

$$q(\mu) = \sum_{i=1}^{n} \varphi_{x_i,0}(\mu) - k\mu, \tag{2.9}$$

where φ *is given in (2.7). In addition, the maximal value of the problem in (2.8) is attained at some* $\mu \ge 0$.

Our next task will be to study the properties of the function q that will enable us to compute its maximal value efficiently. For that, we require the following lemma.[3]

Lemma 2.4 *Let* $\alpha \ge 0$, $b \in \mathbb{R}$ *and consider the function* $\varphi_{b,\alpha} : \mathbb{R}_+ \to \mathbb{R}$ *given in (2.7). Then*

(a) if $b = 0$, *then* $\varphi_{b,\alpha}(\mu) = 0$ *for any* $\mu \ge 0$ *and the set of minimizers in (2.7) is* $[0, 1]$ *if* $\mu = 0$ *or the singleton* $\{0\}$ *if* $\mu > 0$;

(b) if $b \ne 0$, *then (using the convention that* $p/0 = \infty$ *for* $p > 0$*) for any* $\mu \ge 0$

$$\varphi_{b,\alpha}(\mu) = \begin{cases} \frac{b^2}{\alpha+1} + \mu, & \sqrt{\mu} \le \frac{|b|}{\alpha+1}, \\ 2|b|\sqrt{\mu} - \alpha\mu, & \frac{|b|}{\alpha+1} < \sqrt{\mu} < \frac{|b|}{\alpha}, \\ \frac{b^2}{\alpha}, & \sqrt{\mu} \ge \frac{|b|}{\alpha}. \end{cases} \tag{2.10}$$

$\varphi_{b,\alpha}$ is differentiable at any $\mu > 0$ and its derivative is given by

$$\varphi'_{b,\alpha}(\mu) = \begin{cases} 1, & \sqrt{\mu} \le \frac{|b|}{\alpha+1}, \\ \frac{|b|}{\sqrt{\mu}} - \alpha, & \frac{|b|}{\alpha+1} < \sqrt{\mu} < \frac{|b|}{\alpha}, \\ 0, & \sqrt{\mu} \ge \frac{|b|}{\alpha}. \end{cases} \tag{2.11}$$

In addition, problem (2.7) has a unique minimizer $\tilde{u}$ given by

$$\tilde{u} = \begin{cases} 1, & \sqrt{\mu} \le \frac{|b|}{\alpha+1}, \\ \frac{|b|}{\sqrt{\mu}} - \alpha, & \frac{|b|}{\alpha+1} < \sqrt{\mu} < \frac{|b|}{\alpha}, \\ 0, & \sqrt{\mu} \ge \frac{|b|}{\alpha}. \end{cases}$$

(c) The right derivative of $\varphi_{b,\alpha}$ at 0 is given by

$$(\varphi_{b,\alpha})'_+(0) = \begin{cases} 1, & b \ne 0, \\ 0, & b = 0. \end{cases}$$

[3] The lemma is written in a general way that will allow us to compute the prox operator of $\mathcal{S}_k$ later on.

Proof (a) If $b = 0$, then for any $\mu \geq 0$,

$$\varphi(\mu) = \min_{0\leq u\leq 1} (\phi(0, \alpha + u) + \mu u) = \min_{0\leq u\leq 1} \mu u = 0,$$

and the optimal set is either $[0, 1]$ if $\mu = 0$ or $\{0\}$ if $\mu > 0$.
(b) Assume that $b \neq 0$. We make the notation $H(u) = \phi(b, u+\alpha)+\mu u$ and split the analysis into two cases.

Case I $\boldsymbol{\alpha} > 0$. Note that in this case H is differentiable over $[0, 1]$, and that for any $u \in [0, 1]$,

$$H'(u) = -\frac{b^2}{(\alpha + u)^2} + \mu.$$

The function H is strictly convex over $[0, 1]$ and we denote its unique minimizer by $\tilde{u}$. Since H is convex, it follows that $\tilde{u} = 0$ if and only if $H'(0) \geq 0$, which is the same as $-\frac{b^2}{\alpha^2} + \mu \geq 0 \Leftrightarrow \sqrt{\mu} \geq \frac{|b|}{\alpha}$. The optimal solution is $\tilde{u} = 1$ if and only if $H'(1) \leq 0$, which is the same as $-\frac{b^2}{(\alpha+1)^2} + \mu \leq 0 \Leftrightarrow \sqrt{\mu} \leq \frac{|b|}{\alpha+1}$. In all other cases, meaning if $\frac{|b|}{\alpha+1} < \sqrt{\mu} < \frac{|b|}{\alpha}$, we have that $\tilde{u}$ is the unique scalar satisfying $H'(\tilde{u}) = 0$, that is, $\tilde{u} = \frac{|b|}{\sqrt{\mu}} - \alpha$. To conclude, the optimal solution $\tilde{u}$ is given by

$$\tilde{u} = \begin{cases} 1, & \sqrt{\mu} \leq \frac{|b|}{\alpha+1}, \\ \frac{|b|}{\sqrt{\mu}} - \alpha, & \frac{|b|}{\alpha+1} < \sqrt{\mu} < \frac{|b|}{\alpha}, \\ 0, & \sqrt{\mu} \geq \frac{|b|}{\alpha}. \end{cases}$$

and thus,

$$\varphi_{b,\alpha}(\mu) = H(\tilde{u}) = \begin{cases} \frac{b^2}{\alpha+1} + \mu, & \sqrt{\mu} \leq \frac{|b|}{\alpha+1}, \\ 2|b|\sqrt{\mu} - \alpha\mu, & \frac{|b|}{\alpha+1} < \sqrt{\mu} < \frac{|b|}{\alpha}, \\ \frac{b^2}{\alpha}, & \sqrt{\mu} \geq \frac{|b|}{\alpha}, \end{cases}$$

establishing the result for Case I.
Case II $\boldsymbol{\alpha} = 0$. Note that since $b \neq 0$, $\phi(b, 0) + \mu \cdot 0 = \infty + 0 = \infty$, and therefore, the minimizer of H over $[0, 1]$ is not 0. There are two options: (i) the minimizer of H over $[0, 1]$ is $\tilde{u} = 1$ and this occurs when $H'(1) \leq 0 \Leftrightarrow \sqrt{\mu} \leq |b|$; the corresponding function value is $\varphi_{b,0}(\mu) = H(1) = \phi(b, 1) + \mu = b^2 + \mu$; (ii) the minimizer of H over $[0, 1]$ is attained at $\tilde{u} \in (0, 1)$, and in this case $H'(\tilde{u}) = 0$, meaning $\tilde{u} = \frac{|b|}{\sqrt{\mu}}$ with corresponding value $\varphi_{b,0}(\mu) = H(\tilde{u}) = \phi(b, \tilde{u}) + \mu\tilde{u} = 2|b|\sqrt{\mu}$. To conclude, we obtained that in the case $\alpha = 0$,

$$\varphi_{b,0}(\mu) = \begin{cases} 2|b|\sqrt{\mu} & \sqrt{\mu} > |b|, \\ b^2 + \mu & \sqrt{\mu} \leq |b|, \end{cases}$$

which fits formula (2.10).

The expression for for the derivative of $\varphi_{b,\alpha}$ readily follows from the formula of $\varphi_{b,\alpha}$.
(c) Follows directly from the expressions for $\varphi_{b,\alpha}$ given in parts (a) and (b). □

For the sake of evaluating the function $\mathcal{S}_k$ at a given point $\mathbf{x} \in \mathbb{R}^n$, it is enough to focus on the structure of $\varphi_{b,\alpha}$ for the specific case $\alpha = 0$. This is done in Corollary 2.1 below. The general case with $\alpha \neq 0$ will be used later on in Sect. 3 in order to compute the proximal operator of positive scalar multiples of $\mathcal{S}_k$.

Corollary 2.1 *Let $\mathbf{x} \in \mathbb{R}^n$ and $i \in [n]$. Then*

(a) for any $\mu \geq 0$,

$$\varphi_{x_i,0}(\mu) = \begin{cases} x_i^2 + \mu & \sqrt{\mu} \leq |x_i|, \\ 2|x_i|\sqrt{\mu} & \sqrt{\mu} > |x_i|, \end{cases} \tag{2.12}$$

and for any $\mu > 0$,

$$\varphi'_{x_i,0}(\mu) = \min\left\{\frac{|x_i|}{\sqrt{\mu}}, 1\right\}; \tag{2.13}$$

(b) the right derivative of q at 0 is given by $q'_+(0) = \|\mathbf{x}\|_0 - k$.

Proof *(a)* Invoking Lemma 2.4(a,b), we obtain that $\varphi_{x_i,0}$ has the form (2.12) and that if $\mu > 0$, $\varphi'_{x_i,0}$ is given by

$$\varphi'_{x_i,0}(\mu) = \begin{cases} 1 & \sqrt{\mu} \leq |x_i|, \\ \frac{|x_i|}{\sqrt{\mu}} & \sqrt{\mu} > |x_i|, \end{cases} = \min\left\{\frac{|x_i|}{\sqrt{\mu}}, 1\right\}.$$

(b) By the definition of q (see (2.9)),

$$q'_+(0) = \sum_{i=1}^{n} (\varphi_{x_i,0})'_+(0) - k = \|\mathbf{x}\|_0 - k,$$

where the last equality utilizes Lemma 2.4(c). □

The following lemma shows that the sparse envelope function at a given $\mathbf{x} \in \mathbb{R}^n$ is a special sum of squared ℓ_2 and ℓ_1 norms, where the former is computed on the components $\mathbf{x}$ with magnitude above a certain threshold, and the latter is computed on the remaining values. The threshold is dictated by a root of a monotone one-dimensional function. To present the lemma, we denote by $x_{\langle i \rangle}$ the component of $\mathbf{x}$ with the ith largest absolute value, meaning in particular that $|x_{\langle 1 \rangle}| \geq |x_{\langle 2 \rangle}| \geq \ldots |x_{\langle n \rangle}|$.

Lemma 2.5 *($\mathcal{S}_k$ as a sum of squared ℓ_1 and ℓ_2 norms) Let $\mathbf{x} \in \mathbb{R}^n$. Then*

$$\mathcal{S}_k(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{N_\mathbf{x}} x_{\langle i \rangle}^2 + \frac{1}{2(k - N_\mathbf{x})}\left(\sum_{i=N_\mathbf{x}+1}^{n} |x_{\langle i \rangle}|\right)^2, \tag{2.14}$$

where $N_\mathbf{x}$ is determined as follows:

(a) if $\|\mathbf{x}\|_0 \leq k$, then $N_\mathbf{x} = k - 1$, and consequently

$$\mathcal{S}_k(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2;$$

(b) if $\|\mathbf{x}\|_0 > k$, then $N_\mathbf{x} \in \{0, 1, \ldots, k-1\}$ is defined as

$$N_\mathbf{x} = \max\left\{i \in [n] : |x_{\langle i \rangle}| \geq \frac{1}{\tilde{\eta}}\right\}, \tag{2.15}$$

where $\tilde{\eta}$ is a root of the equation

$$g_\mathbf{x}(\eta) \equiv \sum_{i=1}^{n} \min\{|x_i|\eta, 1\} - k = 0 \tag{2.16}$$

over $[0, \infty)$.

Proof Suppose that $\|\mathbf{x}\|_0 \leq k$. In this case, by Corollary 2.1(b), $q'_+(0) = \|\mathbf{x}\|_0 - k \leq 0$, and therefore, by the concavity of q, it follows that 0 is a maximizer of q, and consequently, by (2.8),

$$2\mathcal{S}_k(\mathbf{x}) = q(0) = \sum_{i=1}^{n} \varphi_{x_i,0}(0) = \sum_{i=1}^{n} x_i^2 = \|\mathbf{x}\|_2^2,$$

establishing part (a). To prove part (b), suppose that $\|\mathbf{x}\|_0 > k$. By Corollary 2.1(b), $q'_+(0) = \|\mathbf{x}\|_0 - k > 0$, and therefore a maximizer of the dual function q, which necessarily exists by Lemma 2.3, must be a positive number. Recall that $q(\mu) = \sum_{i=1}^{n} \varphi_{x_i,0}(\mu) - k\mu$, and that by Corollary 2.1(a), $\varphi_{x_i,0}$ is differentiable over $(0, \infty)$ and hence $\tilde{\mu} > 0$ is a maximizer of q if and only if

$$q'(\tilde{\mu}) = \underbrace{\sum_{i=1}^{n} \varphi'_{x_i,0}(\tilde{\mu})}_{g_{\mathbf{x}}(\tilde{\mu})} - k = 0,$$

meaning (see Corollary 2.1(a)) if and only if

$$\tilde{g}_{\mathbf{x}}(\tilde{\mu}) \equiv \sum_{i=1}^{n} \min\left\{\frac{|x_i|}{\sqrt{\tilde{\mu}}}, 1\right\} - k = 0. \tag{2.17}$$

Define

$$N_{\mathbf{x}} = \max\{i \in [n] : |x_{\langle i \rangle}| \geq \sqrt{\tilde{\mu}}\}.$$

Then (2.17) translates to

$$N_{\mathbf{x}} + \sum_{i=N_{\mathbf{x}}+1}^{n} \frac{|x_{\langle i \rangle}|}{\sqrt{\tilde{\mu}}} = k. \tag{2.18}$$

Consequently, $N_{\mathbf{x}} \leq k$ since

$$N_{\mathbf{x}} - k = -\sum_{i=N_{\mathbf{x}}+1}^{n} \frac{|x_{\langle i \rangle}|}{\sqrt{\tilde{\mu}}} \leq 0.$$

In addition, $N_{\mathbf{x}}$ must be different than k, since otherwise, by (2.18),

$$\frac{1}{\sqrt{\tilde{\mu}}} \sum_{i=k+1}^{n} |x_{\langle i \rangle}| = 0,$$

which is a contradiction to the condition that $\|\mathbf{x}\|_0 > k$. We thus obtained that $N_{\mathbf{x}} \in \{0, 1, \ldots, k-1\}$. By (2.18), $\tilde{\mu} \equiv \left(\frac{\sum_{i=N_{\mathbf{x}}+1}^{n} |x_{\langle i \rangle}|}{k - N_{\mathbf{x}}}\right)^2$, and consequently, by (2.8), we have

$$\begin{aligned} 2\mathcal{S}_k(\mathbf{x}) &= q(\tilde{\mu}) \overset{(2.6)}{=} \sum_{i=1}^{N_{\mathbf{x}}} \varphi_{x_i,0}(\tilde{\mu}) + \sum_{i=N_{\mathbf{x}}+1}^{n} \varphi_{x_i,0}(\tilde{\mu}) - k\tilde{\mu} \\ &\overset{(2.12)}{=} \sum_{i=1}^{N_{\mathbf{x}}} (x_{\langle i \rangle}^2 + \tilde{\mu}) + \sum_{i=N_{\mathbf{x}}+1}^{n} 2|x_{\langle i \rangle}|\sqrt{\tilde{\mu}} - k\tilde{\mu} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N_{\mathbf{x}}} x_{\langle i\rangle}^2 + (N_{\mathbf{x}} - k)\tilde{\mu} + 2\sqrt{\tilde{\mu}} \sum_{i=N_{\mathbf{x}}+1}^{n} |x_{\langle i\rangle}| \\
&= \sum_{i=1}^{N_{\mathbf{x}}} x_{\langle i\rangle}^2 + (N_{\mathbf{x}} - k)\frac{1}{(k - N_{\mathbf{x}})^2}\left(\sum_{i=N_{\mathbf{x}}+1}^{n} |x_{\langle i\rangle}|\right)^2 + \frac{2}{k - N_{\mathbf{x}}}\left(\sum_{i=N_{\mathbf{x}}+1}^{n} |x_{\langle i\rangle}|\right)^2 \\
&= \sum_{i=1}^{N_{\mathbf{x}}} x_{\langle i\rangle}^2 + \frac{1}{k - N_{\mathbf{x}}}\left(\sum_{i=N_{\mathbf{x}}+1}^{n} |x_{\langle i\rangle}|\right)^2.
\end{aligned}
$$

Finally, making the change of variables $\eta = \frac{1}{\sqrt{\mu}}$, $\tilde{\eta} = \frac{1}{\sqrt{\tilde{\mu}}}$, we obtain the expression (2.15) for $N_{\mathbf{x}}$ and that $\tilde{\eta}(= \sqrt{\tilde{\mu}})$ is a root of $g_{\mathbf{x}}(\eta) = \tilde{g}_{\mathbf{x}}\left(\frac{1}{\eta^2}\right)$. □

Remark 2.2 If $k = 1$, then by Lemma 2.5, since $k - 1 = 0$, it follows that $N_{\mathbf{x}} = 0$ regardless of the value of $\mathbf{x}$, and consequently

$$\mathcal{S}_k(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_1^2.$$

Remark 2.3 The formula (2.14) was already established in [1, Proposition 2.1], but with a different expression for the choice of $N_{\mathbf{x}}$. Specifically, the choice of $N_{\mathbf{x}}$ in [1] requires the sorting of the elements of $\mathbf{x}$, meaning $O(n \log n)$ operations. In contrast, the choice of $N_{\mathbf{x}}$ in terms of the root of the one-dimensional function $g_{\mathbf{x}}$ given in (2.16) will be the key to proving an expected $O(n)$ complexity of the evaluation of both the function value and proximal operator of the sparse envelope function.

2.2 Solving the 1D problem

2.2.1 Bisection

Consider the case where $\|\mathbf{x}\|_0 > k$. The function $g_{\mathbf{x}}$ given in (2.16) is continuous and nondecreasing. We can also describe two values for which $g_{\mathbf{x}}$ has opposite signs, implying that the one-dimensional problem can be solved by simple root-finding procedures such as bisection. To describe the point for which $g_{\mathbf{x}}$ is positive, denote $q = \|\mathbf{x}\|_0$, and observe that in our notation $|x_{\langle q\rangle}|$ is the minimal absolute value among all the nonzero components of $\mathbf{x}$. Since $\min\left\{\frac{|x_i|}{|x_{\langle q\rangle}|}, 1\right\} = 1$ if $x_i \neq 0$ and 0 otherwise, it follows that

$$g_{\mathbf{x}}\left(\frac{1}{|x_{\langle q\rangle}|}\right) = \sum_{i=1}^{n} \min\left\{\frac{|x_i|}{|x_{\langle q\rangle}|}, 1\right\} - k = \|\mathbf{x}\|_0 - k > 0.$$

In addition, if we denote $\gamma = \max\left\{\frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_\infty k}, 1\right\}$, then

$$g_{\mathbf{x}}\left(\frac{1}{\gamma\|\mathbf{x}\|_\infty}\right) = \sum_{i=1}^{n} \min\left\{\frac{|x_i|}{\gamma\|\mathbf{x}\|_\infty}, 1\right\} - k \overset{\gamma\geq 1}{=} \sum_{i=1}^{n} \frac{|x_i|}{\gamma\|\mathbf{x}\|_\infty} - k = \frac{\|\mathbf{x}\|_1}{\gamma\|\mathbf{x}\|_\infty} - k \overset{\gamma \geq \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_\infty k}}{\leq} 0.$$

Therefore, the search for the root of $g_{\mathbf{x}}$ can be over the bounded interval $[\underbrace{(\gamma\|\mathbf{x}\|_\infty)^{-1}}_{\ell}, \underbrace{|x_{\langle q\rangle}|^{-1}}_{u}]$.

The simplest approach for numerically solving the equation will be to employ a bisection procedure. The worst case number of iterations of the bisection method is $O(\log(\frac{u-\ell}{\varepsilon}))$, where ε is the required accuracy. Since a single evaluation of $g_{\mathbf{x}}$ at any point requires $O(n)$ amount of elementary operations, the overall computational effort is $O(\log(\frac{u-\ell}{\varepsilon})n)$.

Remark 2.4 Since $g_{\mathbf{x}}$ can also be written as $g_{\mathbf{x}}(\eta) = \sum_{j:x_j\neq 0} \min\{|x_j|\eta, 1\} - k$, then given that we know beforehand the locations of the nonzero elements in $\mathbf{x}$, the evaluation of $g_{\mathbf{x}}$ can be done in $O(\|\mathbf{x}\|_0)$ operations, meaning that the overall computational effort can be reduced to $O(\log(\frac{u-\ell}{\varepsilon})\|\mathbf{x}\|_0)$.

2.2.2 Randomized root search

A second option for solving the one-dimensional problem (2.16) would be to construct a method similar to randomized median finding [6, Section 9.2] that exploits the fact that the function is a sum of piecewise linear functions, each with a single breakpoint. The presented algorithm is a simple extension of the randomized algorithm from [10] used to solve the one-dimensional problem arising in the computation of the orthogonal projection onto the l_1-ball, see also [4] for a similar approach for finding a root of a different one-dimensional function. The exact formulation of the one-dimensional problem we consider in this paper is as follows.

Problem 1D-G
Input: $\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \boldsymbol{\gamma} \in \mathbb{R}^m$, $\delta \in \mathbb{R}$.
Functional representation of the input:

$$F(\eta) = \sum_{j=1}^{m} F_j(\eta) - \delta, \text{ where } F_j(\eta) = \begin{cases} \alpha_j^1\eta + \beta_j^1, & \eta \leq \gamma_j, \\ \alpha_j^2\eta + \beta_j^2, & \eta > \gamma_j. \end{cases} \tag{2.19}$$

Assumptions: (a) $\alpha_j^1\gamma_j + \beta_j^1 = \alpha_j^2\gamma_j + \beta_j^2$ for any $j \in [m]$. [this condition ensures continuity of F_j.]
(b) F is nondecreasing and has a root.
Output: a point $\eta^* \in \mathbb{R}$ for which $F(\eta^*) = 0$

The randomized approach will lead to an algorithm with an expected amount of computations that is at most linear in m (number of piecewise linear functions with a single breakpoint), and the number of computations will not depend on any tolerance, or on a size of an initial interval, like in the bisection method. To use the randomized method, we will exploit properties (a) and (b) described in the assumptions of problem 1D-G.

The algorithm is based on the following observation: suppose that we randomly choose $p \in [m]$, and that this index p satisfies $F(\gamma_p) < 0$. Define $\Omega = \{j : \gamma_j > \gamma_p\}$. Then by the monotonicity and continuity of F, the function has a root in (γ_p, ∞) and for any $j \notin \Omega$, by the structure F_j, it holds that $F_j(\eta) = \alpha_j^2\eta + \beta_j^2$ for any $\eta > \gamma_p$ (since $\eta > \gamma_p \geq \gamma_j$), implying that for any $\eta > \gamma_p$,

$$F(\eta) = \sum_{j\in\Omega} F_j(\eta) + \tilde{\alpha}\eta + \tilde{\beta}, \tag{2.20}$$

where $\tilde{\alpha} = \sum_{j\notin\Omega} \alpha_j^2$, $\tilde{\beta} = \sum_{j\notin\Omega} \beta_j^2 - \delta$. A similar argument, shows that if $F(\gamma_p) > 0$, then (2.20) holds for any $\eta < \gamma_p$ with $\Omega = \{j : \gamma_j < \gamma_p\}$, $\tilde{\alpha} = \sum_{j\notin\Omega} \alpha_j^1$ and $\tilde{\beta} = \sum_{j\notin\Omega} \beta_j^1 - \delta$.

The above observation implies that no matter what is the sign of $F(\gamma_p)$, from this point onwards, in order to evaluate the function F on values from the relevant intervals ((γ_p, ∞) or $(-\infty, \gamma_p)$), the solution algorithm can just keep the values of $\tilde{\alpha}$ and $\tilde{\beta}$ and take into account only the indices in Ω (disregarding the indices in $[m] \setminus \Omega$). If for a certain index p, $F(\gamma_p) = 0$, the algorithm stops and returns $\eta^* = \gamma_p$. Otherwise, at a certain point, $\Omega = \emptyset$, and the root of the function is the root of the affine function $\tilde{\alpha}\eta + \tilde{\beta}$, meaning $\eta^* = -\frac{\tilde{\beta}}{\tilde{\alpha}}$ (we assume that all arithmetic can be performed exactly).

Algorithm 1: Randomized Root Search

Input: $\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2, \boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \boldsymbol{\gamma} \in \mathbb{R}^m, \delta \in \mathbb{R}$

Output: $\eta^* \in \mathbb{R}$ for which $F(\eta^*) = 0$, where $F = \sum_{j=1}^n F_j - \delta$ with F_j given in (2.19).

Initialization: $\Omega = [m], \tilde{\alpha} = 0, \tilde{\beta} = -\delta$.

General step:

while $\Omega \neq \emptyset$

  pick $p \in \Omega$ at random

  compute $F(\gamma_p) = \tilde{\alpha}\gamma_p + \tilde{\beta} + \sum_{j\in\Omega} F_j(\gamma_p)$

  if $(F(\gamma_p) < 0)$

    $A \leftarrow \{j \in \Omega : \gamma_j > \gamma_p\}$

    $\tilde{\alpha} \leftarrow \tilde{\alpha} + \sum_{j\in\Omega\setminus A} \alpha_j^2$, $\tilde{\beta} \leftarrow \tilde{\beta} + \sum_{j\in\Omega\setminus A} \beta_j^2$

    $\Omega \leftarrow A$

  elseif $F(\gamma_p) > 0$

    $A \leftarrow \{j \in \Omega : \gamma_j < \gamma_p\}$

    $\tilde{\alpha} \leftarrow \tilde{\alpha} + \sum_{j\in\Omega\setminus A} \alpha_j^1$, $\tilde{\beta} \leftarrow \tilde{\beta} + \sum_{j\in\Omega\setminus A} \beta_j^1$

    $\Omega \leftarrow A$

  elseif $F(\gamma_p) = 0$

    return $\eta^* = \gamma_p$.

  end if

end while

return $\eta^* = -\frac{\tilde{\beta}}{\tilde{\alpha}}$

The complexity analysis of Algorithm 1 is essentially identical to the complexity analysis of the randomized median finding algorithm [6, Section 9.2], and therefore the expected amount of iterations is $O(m)$.

Solving problem (2.16): In the case where $F = g_{\mathbf{x}}$, we can write F as

$$F(\eta) = \sum_{i\in I(\mathbf{x})} \min\{|x_i|\eta, 1\} - k,$$

where $I(\mathbf{x}) = \{i : x_i \neq 0\}$. Denote $I(\mathbf{x}) = \{i_1, i_2, \ldots, i_q\}$ where $i_1 < i_2 < \cdots < i_q$ ($q = \|\mathbf{x}\|_0$). We can thus define in this case $F_j(\eta) = \begin{cases} |x_{i_j}|\eta, & \eta \leq \frac{1}{|x_{i_j}|}, \\ 1, & \eta > \frac{1}{|x_{i_j}|}, \end{cases}$ $j \in [q]$, and consequently solve the one-dimensional problem (2.16) by employing Algorithm 1 with input

$$\alpha_j^1 = |x_{i_j}|, \beta_j^1 = 0, \alpha_j^2 = 0, \beta_j^2 = 1, \gamma_j = \frac{1}{|x_{i_j}|}, \delta = k, j \in [q].$$

The algorithm requires an expected amount of $O(n)$ (or even just $O(\|\mathbf{x}\|_0)$) operations.

3 Proximal mapping of the sparse envelope function

In this section we will show how to efficiently compute the proximal operator of positive scalar multiples of $\mathcal{S}_k$. The ability to perform such an operation efficiently opens the possibility of using proximal-based methods for solving optimization problems involving the sparse envelope function. For example, FISTA [3] can be employed to solve the composite model (1.1) in the case where f is convex and smooth. We begin with the following lemma that shows that the proximal operator can be determined in terms of the optimal solution of a convex problem that resembles the optimization problem defined in Lemma 2.2 for computing the value of $\mathcal{S}_k$.

Lemma 3.1 *Let $\lambda > 0$ and $\mathbf{x} \in \mathbb{R}^n$. Then $\mathbf{w} = \text{prox}_{\lambda \mathcal{S}_k}(\mathbf{x})$ is given by*

$$w_i = \frac{x_i u_i}{\lambda + u_i}, \quad i \in [n], \tag{3.1}$$

where $(u_1, u_2, \ldots, u_n)^T$ is the optimal solution of the problem

$$\min_{\mathbf{u} \in D_k} \sum_{i=1}^{n} \phi(x_i, \lambda + u_i). \tag{3.2}$$

Proof By definition,

$$\mathbf{w} = \text{prox}_{\lambda \mathcal{S}_k}(\mathbf{x}) = \underset{\mathbf{z}}{\text{argmin}} \left\{ \lambda S_k(\mathbf{z}) + \frac{1}{2} \|\mathbf{z} - \mathbf{x}\|_2^2 \right\}.$$

Using Lemma 2.2, we thus need to solve (reversing the order of minimizations with respect to $\mathbf{z}, \mathbf{u}$):

$$\min_{\mathbf{u} \in D_k} \min_{\mathbf{z}} \left\{ \Phi(\mathbf{z}, \mathbf{u}, \mathbf{x}) \equiv \frac{\lambda}{2} \sum_{i=1}^{n} \phi(z_i, u_i) + \frac{1}{2} (z_i - x_i)^2 \right\}.$$

Solving for $\mathbf{z}$, we get that for any $i \in [n]$, if $u_i > 0$, then $\frac{\lambda \bar{z}_i}{u_i} + \bar{z}_i - x_i = 0$, meaning that

$$\bar{z}_i \equiv w_i = \frac{x_i u_i}{\lambda + u_i}, \tag{3.3}$$

where $\mathbf{u}$ is the minimizer of the problem $\min_{\mathbf{u} \in D_k} \Phi(\mathbf{u}, \bar{\mathbf{z}}, \mathbf{x})$. Equation (3.3) also holds when $u_i = 0$, since in that case, $\bar{z}_i = 0$. Plugging the expression (3.3) for $\bar{\mathbf{z}}$ in Φ, yields (using the "convention" that $0/0 = 0$)

$$\begin{aligned} \Phi(\mathbf{u}, \bar{\mathbf{z}}, \mathbf{x}) &= \frac{1}{2} \sum_{i=1}^{n} \left(\lambda \frac{\bar{z}_i^2}{u_i} + (\bar{z}_i - x_i)^2 \right) \\ &= \frac{1}{2} \sum_{i=1}^{n} \left(\frac{\lambda x_i^2 u_i^2}{u_i (\lambda + u_i)^2} + \frac{\lambda^2 x_i^2}{(\lambda + u_i)^2} \right) = \frac{\lambda}{2} \sum_{i=1}^{n} \frac{x_i^2}{\lambda + u_i} = \frac{\lambda}{2} \sum_{i=1}^{n} \phi(x_i, \lambda + u_i), \end{aligned}$$

which proves the desired claim. □

Assigning a Lagrange multiplier for the inequality constraint $\mathbf{e}^T \mathbf{u} \leq k$ in problem (3.2), we obtain the Lagrangian function

$$L(\mathbf{u}, \mu) = \sum_{i=1}^{n} (\phi(x_i, \lambda + u_i) + \mu u_i) - k\mu.$$

Utilizing the definition of $\varphi_{b,\alpha}$ as given in (2.7), we can deduce that the dual objective function is given by

$$Q(\mu) \equiv \min_{\mathbf{u}:0\le\mathbf{u}\le\mathbf{e}} L(\mathbf{u},\mu) = \sum_{i=1}^{n} \varphi_{x_i,\lambda}(\mu) - k\mu.$$

Therefore, the dual of problem (3.2) is the maximization problem

$$\max\{Q(\mu) : \mu \ge 0\}. \tag{3.4}$$

Remark 3.1 A direct consequence of Lemma 2.4 is that if $\tilde{\mu} > 0$, the function $\mathbf{u} \mapsto L(\mathbf{u}, \tilde{\mu})$ has a unique minimizer over $\{\mathbf{u} : \mathbf{0} \le \mathbf{u} \le \mathbf{e}\}$ given by $u_i = \varphi'_{x_i,\lambda}(\tilde{\mu})$.

The next theorem shows how the proximal operator of the sparse envelope function reduces to a one-dimensional search.

Theorem 3.1 *Let $\mathbf{x} \in \mathbb{R}^n$ and $\lambda > 0$;*

(a) if $\|\mathbf{x}\|_0 \le k$, then $\operatorname{prox}_{\lambda \mathcal{S}_k}(\mathbf{x}) = \frac{1}{\lambda+1}\mathbf{x}$;
(b) if $\|\mathbf{x}\|_0 > k$, then $\mathbf{w} = \operatorname{prox}_{\lambda \mathcal{S}_k}(\mathbf{x})$ is given by

$$w_i = \frac{x_i u_i}{\lambda + u_i}, \quad i = 1, 2, \ldots, n, \tag{3.5}$$

where $u_i = u_i(\tilde{\eta})$ with $u_i(\cdot)$ being defined as[4]

$$u_i(\eta) \equiv \begin{cases} 0, & \eta \le \frac{\lambda}{|x_i|}, \\ |x_i|\eta - \lambda, & \frac{\lambda}{|x_i|} < \eta < \frac{\lambda+1}{|x_i|}, \\ 1, & \eta \ge \frac{\lambda+1}{|x_i|}, \end{cases} \quad i = 1, 2, \ldots, n, \tag{3.6}$$

and $\tilde{\eta}$ is a root of the function

$$h_{\mathbf{x}}(\eta) \equiv \sum_{i=1}^{n} u_i(\eta) - k, \tag{3.7}$$

which is nondecreasing and satisfies

$$h_{\mathbf{x}}\left(\frac{\lambda}{\|\mathbf{x}\|_\infty}\right) < 0, h_{\mathbf{x}}\left(\frac{\lambda+1}{|x_{\langle q\rangle}|}\right) > 0, q = \|\mathbf{x}\|_0.$$

Proof First note that by Lemma 2.4(c) it follows that

$$Q'_+(0) = \sum_{i=1}^{n} (\varphi_{x_i,\lambda})'_+(0) - k = \|\mathbf{x}\|_0 - k. \tag{3.8}$$

(a) Suppose that $\|\mathbf{x}\|_0 \le k$. Then by (3.8),

$$Q'_+(0) = \|\mathbf{x}\|_0 - k \le 0.$$

and therefore, since Q is concave over $\mathbb{R}_+$ (being a dual objective function), it follows that 0 is a maximizer of Q, and thus

$$\max_{\mu\in\mathbb{R}_+} Q(\mu) = Q(0) = \sum_{i=1}^{n} \varphi_{x_i,\lambda}(0) - k \cdot 0 = \sum_{i=1}^{n} \frac{x_i^2}{\lambda+1}.$$

[4] If $x_i = 0$, then the formula (3.6) implies that $u_i(\eta) = 0$ for all $\eta \ge 0$.

On the other hand, plugging into the objective function of the primal problem (3.2) the feasible choice

$$u_i = I(x_i) \equiv \begin{cases} 1 & x_i \neq 0, \\ 0, & x_i = 0, \end{cases} \tag{3.9}$$

we obtain the same function value as the one of the dual problem:

$$\begin{aligned} \sum_{i=1}^{n} \frac{x_i^2}{\lambda + I(x_i)} &= \sum_{i:x_i \neq 0} \frac{x_i^2}{\lambda + 1} + \sum_{i:x_i = 0} \frac{x_i^2}{\lambda + 0} \\ &= \sum_{i:x_i \neq 0} \frac{x_i^2}{\lambda + 1} + \sum_{i:x_i = 0} \frac{x_i^2}{\lambda + 1} = \sum_{i=1}^{n} \frac{x_i^2}{\lambda + 1} = Q(0), \end{aligned}$$

implying, by weak duality, that $\mathbf{u}$ given by (3.9) is the optimal solution of problem (3.2), meaning that by Lemma 3.1,

$$\mathrm{prox}_{\lambda \mathcal{S}_k}(\mathbf{x}) = \frac{\mathbf{x}}{\lambda + 1}.$$

(b) Since $\|\mathbf{x}\|_0 - k > 0$, it follows by (3.8) that

$$Q'_+(0) = \|\mathbf{x}\|_0 - k > 0,$$

and thus all maximizers of Q over $\mathbb{R}_+$ are positive. The existence of a maximizer for the dual problem (3.4) follows by the strong duality theorem. By Lemma 2.4(a,b) the function q is differentiable over the positive numbers, and thus $\tilde{\mu} > 0$ is a maximizer of Q iff $Q'(\tilde{\mu}) = 0$, meaning if and only if $\tilde{\mu} > 0$ is a solution of the equation

$$Q'(\mu) = \sum_{i=1}^{n} \varphi'_{x_i,\lambda}(\mu) - k = 0. \tag{3.10}$$

By Remark 3.1, the function $\mathbf{u} \mapsto L(\mathbf{u}, \tilde{\mu})$ has a unique minimizer over $\{\mathbf{u} : \mathbf{0} \leq \mathbf{u} \leq \mathbf{e}\}$ given by

$$u_i = \varphi'_{x_i,\lambda}(\tilde{\mu}) = \begin{cases} 1, & \sqrt{\tilde{\mu}} \leq \frac{|x_i|}{\lambda+1}, \\ \frac{|x_i|}{\sqrt{\tilde{\mu}}} - \lambda, & \frac{|x_i|}{\lambda+1} < \sqrt{\tilde{\mu}} < \frac{|x_i|}{\lambda}, \\ 0, & \sqrt{\tilde{\mu}} \geq \frac{|x_i|}{\lambda} \end{cases}$$

By (3.10), $\mathbf{u}$ satisfies the primal constraint $\sum_{i=1}^{n} u_i \leq k$ (actually, it is satisfied as an equality), and thus by strongly duality, it is the optimal solution of problem (3.2). Consequently, by Lemma 3.1 it follows that $\mathbf{w}$ given by (3.5) is equal to $\mathrm{prox}_{\lambda \mathcal{S}_k}(\mathbf{x})$.

Making the change of variables $\eta = \frac{1}{\sqrt{\mu}}$, $\tilde{\eta} = \frac{1}{\sqrt{\tilde{\mu}}}$, we obtain that Eq. (3.10) is transformed into the equation $h_{\mathbf{x}}(\eta) = 0$, where

$$h_{\mathbf{x}}(\eta) = Q'\left(\frac{1}{\eta^2}\right), \tag{3.11}$$

and that the relation $u_i = \varphi'_{x_i,\lambda}(\tilde{\mu})$ becomes $u_i = u_i(\tilde{\eta})$ with $u_i(\cdot)$ defined in (3.6). Also, since Q', as a derivative of the concave dual function is nonincreasing, it follows that $h_{\mathbf{x}}$ defined by the relation (3.11) is nondecreasing. Finally, denoting $q = \|\mathbf{x}\|_0$, it holds that

$$h_{\mathbf{x}}\left(\frac{\lambda}{\|\mathbf{x}\|_\infty}\right) = 0 - k = -k < 0,$$

$$h_{\mathbf{x}}\left(\frac{\lambda+1}{|x_{\langle q\rangle}|}\right)=\sum_{i:x_i\neq 0}u_i\left(\frac{\lambda+1}{|x_{\langle q\rangle}|}\right)-k=\|\mathbf{x}\|_0-k>0.$$

□

We thus conclude from Theorem 3.1 that, much like the problem of computing the sparse envelope function itself, evaluation of the proximal mapping of the sparse envelope function boils down to a one-dimensional search problem (in the case $\|\mathbf{x}\|_0 > k$) that can be solved in expected linear time using randomized root search.

Randomized Root Search The function $h_{\mathbf{x}}$ as represented in (3.7) is not of the form required by the randomized root search method since the functions $u_i(\cdot)$ (given in (3.6)) have two breakpoints each. Consequently, randomized root search cannot be employed directly, but fortunately, it is possible to represent $h_{\mathbf{x}}$ as the sum of $2n$ continuous piecewise linear functions with a single breakpoint. To explain the transformation, note that the functions $u_i(\cdot)$ in the case where $x_i \neq 0$ are of the form

$$G(\eta)=\begin{cases}0 & \eta\leq\alpha,\\ m\eta+n, & \alpha<\eta<\beta,\\ 1, & \eta\geq\beta.\end{cases}\tag{3.12}$$

where $m\in\mathbb{R}\setminus\{0\}$, $n\in\mathbb{R}$, $\alpha<\beta$ and

$$\alpha=-\frac{n}{m},\ \beta=\frac{1-n}{m}.$$

The above relations are necessary to ensure that G is continuous. The main observation established in Lemma 3.2 and illsutated in Figure 1 is that G can be decomposed into two continuous piecewise linear functions with a single breakpoint.

Lemma 3.2 *Let G be given in (3.12) where $m\in\mathbb{R}\setminus\{0\}$, $n\in\mathbb{R}$, $\alpha=-\frac{n}{m}$ and $\beta=\frac{1-n}{m}$. Then for any $\eta\in\mathbb{R}$,*

$$G(\eta)=\frac{1}{2}G_1(\eta)+\frac{1}{2}G_2(\eta),$$

where

$$G_1(\eta)=m|\eta-\alpha|,\ G_2(\eta)=1-m|\eta-\beta|.$$

Proof The proof is technical and split into three cases:

Case I $\boldsymbol{\eta\leq\alpha}$. In this case, $G_1(\eta)=m(\alpha-\eta)$, $G_2(\eta)=1-m(\beta-\eta)$, and thus,

$$\begin{aligned}G_1(\eta)+G_2(\eta)&=m(\alpha-\eta)+1-m(\beta-\eta)\\&=1+m(\alpha-\beta)=1+m\cdot\left(-\frac{1}{m}\right)=0=2G(\eta).\end{aligned}$$

Case II $\boldsymbol{\alpha<\eta<\beta}$. Here $G_1(\eta)=m(\eta-\alpha)$, $G_2(\eta)=1-m(\beta-\eta)$, and consequently,

$$\begin{aligned}G_1(\eta)+G_2(\eta)&=m(\eta-\alpha)+1-m(\beta-\eta)=2m\eta+1-m(\alpha+\beta)\\&=2m\eta+1-m\frac{1-2n}{m}=2m\eta+2n=2G(\eta).\end{aligned}$$

Case III $\boldsymbol{\eta\geq\beta}$. Here $G_1(\eta)=m(\eta-\alpha)$ and $G_2(\eta)=1-m(\eta-\beta)$, and hence

$$G_1(\eta)+G_2(\eta)=m(\eta-\alpha)+1-m(\eta-\beta)=1-m(\alpha-\beta)$$

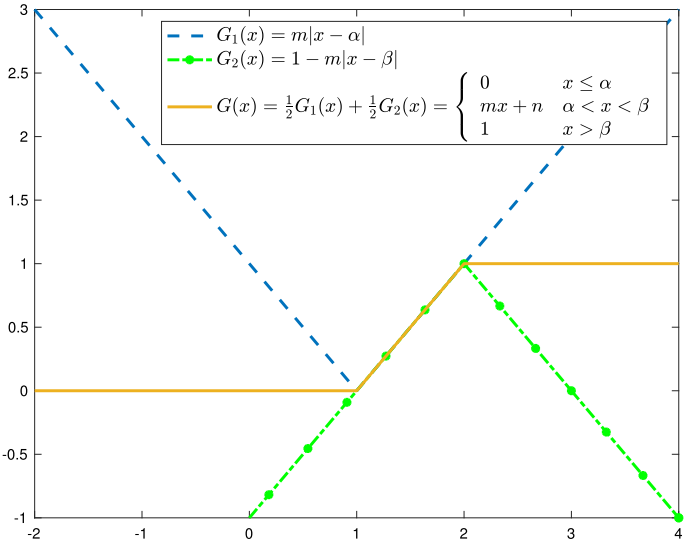

Fig. 1 The function G in solid black lines has two breakpoints, and it can be decomposed into to piecewise linear functions with a single breakpoint (one convex and one concave)

$$= 1 - m\left(-\frac{1}{m}\right) = 2 = 2G(\eta).$$

□

Using Lemma 3.2, we conclude that $h_{\mathbf{x}}$ has the representation (recalling that $u_i(\eta) \equiv 0$ whenever $x_i = 0$)

$$h_{\mathbf{x}}(\eta) = \frac{1}{2}\sum_{i:x_i\neq 0} v_i(\eta) + \frac{1}{2}\sum_{i:x_i\neq 0} w_i(\eta) - k,$$

where

$$v_i(\eta) \equiv |\eta|x_i| - \lambda|\,, w_i(\eta) \equiv 1 - |\eta|x_i| - (\lambda+1)|\,, i = 1, 2, \ldots, n.$$

Thus, employing the randomized root search method with the $2\|\mathbf{x}\|_0$ functions v_i, w_i, as input, we obtain that the root of $h_{\mathbf{x}}$ can be found in $O(\|\mathbf{x}\|_0)$ computational effort. Since $\|\mathbf{x}\|_0 \le n$, this also establishes the $O(n)$ complexity result.

4 Application to sparse support vector machines

As described in the introduction, a possible formulation of the sparse SVM problem is given by

$$\begin{aligned} \min\ & \tfrac{1}{2}\|\mathbf{w}\|_2^2 + C\mathbf{e}^T\boldsymbol{\xi} \\ \text{s.t.}\ & \|\mathbf{w}\|_0 \le k, \\ & \mathbf{Y}(\mathbf{X}\mathbf{w} + \beta\mathbf{e}) \ge \mathbf{e} - \boldsymbol{\xi}, \\ & \boldsymbol{\xi} \ge \mathbf{0}. \end{aligned}$$

Obviously, for any parameter $\lambda \in [0, 1]$, the last problem is the same as

$$\begin{array}{ll} \min & \frac{1-\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + C\mathbf{e}^T\boldsymbol{\xi} \\ \text{s.t.} & \|\mathbf{w}\|_0 \le k, \\ & \mathbf{Y}(\mathbf{X}\mathbf{w} + \beta\mathbf{e}) \ge \mathbf{e} - \boldsymbol{\xi}, \\ & \boldsymbol{\xi} \ge \mathbf{0}. \end{array} \tag{4.1}$$

Recalling the definition of s_k [Eq. (2.1)], the last problem can be rewritten as

$$\begin{array}{ll} \min & \frac{1-\lambda}{2}\|\mathbf{w}\|_2^2 + \lambda s_k(\mathbf{x}) + C\mathbf{e}^T\boldsymbol{\xi} \\ \text{s.t.} & \mathbf{Y}(\mathbf{X}\mathbf{w} + \beta\mathbf{e}) \ge \mathbf{e} - \boldsymbol{\xi}, \\ & \boldsymbol{\xi} \ge \mathbf{0}. \end{array} \tag{P}$$

Since the above is a nonconvex problem that is difficult to tackle, we consider a relaxation constructed by replacing s_k with its convex biconjugate $\mathcal{S}_k$.

$$\begin{array}{ll} \min & \frac{1-\lambda}{2}\|\mathbf{w}\|_2^2 + \lambda \mathcal{S}_k(\mathbf{x}) + C\mathbf{e}^T\boldsymbol{\xi} \\ \text{s.t.} & \mathbf{Y}(\mathbf{X}\mathbf{w} + \beta\mathbf{e}) \ge \mathbf{e} - \boldsymbol{\xi}, \\ & \boldsymbol{\xi} \ge \mathbf{0}. \end{array} \tag{R}$$

Obviously, since $\mathcal{S}_k$ is an underestimator of s_k, it follows that val(R) $\le$ val(P). The choice parameter λ controls the trade-off between the tightness of the relaxation and the strong convexity parameter of the objective function w.r.t. $\mathbf{w}$: large λ means tighter relaxation and small strong convexity parameter and small λ means looser relaxation with large strong convexity parameter. To solve problem (R), we suggest to construct the dual problem, which much like the dual of the original SVM problem (1.2) is much easier to handle [7]. As we will see, even though the sparse envelope function has no explicit expression, it is possible to write the dual of (R) in terms of its Moreau envelope, which can be computed via the proximal mapping. We begin with the construction of the Lagrangian (with $\boldsymbol{\alpha} \in \mathbb{R}^n_+$ being the dual variables vector)

$$\begin{aligned} L(\mathbf{w}, \beta, \boldsymbol{\xi}; \boldsymbol{\alpha}) &= \frac{1-\lambda}{2}\|\mathbf{w}\|_2^2 + \lambda\mathcal{S}_k(\mathbf{w}) + C\mathbf{e}^T\boldsymbol{\xi} - \boldsymbol{\alpha}^T\left[\mathbf{Y}(\mathbf{X}\mathbf{w} + \beta\mathbf{e}) - \mathbf{e} + \boldsymbol{\xi}\right] \\ &= \frac{1-\lambda}{2}\left\|\mathbf{w} - \frac{\mathbf{X}^T\mathbf{Y}\boldsymbol{\alpha}}{1-\lambda}\right\|_2^2 + \lambda\mathcal{S}_k(\mathbf{w}) - (\mathbf{e}^T\mathbf{Y}\boldsymbol{\alpha})\beta + (C\mathbf{e} - \boldsymbol{\alpha})^T\boldsymbol{\xi} \\ &\quad - \frac{1}{2(1-\lambda)}\|\mathbf{X}^T\mathbf{Y}\boldsymbol{\alpha}\|_2^2 + \mathbf{e}^T\boldsymbol{\alpha} \\ &= \lambda\left[\frac{1-\lambda}{2\lambda}\left\|\mathbf{w} - \frac{\mathbf{X}^T\mathbf{Y}\boldsymbol{\alpha}}{1-\lambda}\right\|_2^2 + \mathcal{S}_k(\mathbf{w})\right] - (\mathbf{e}^T\mathbf{Y}\boldsymbol{\alpha})\beta + (C\mathbf{e} - \boldsymbol{\alpha})^T\boldsymbol{\xi} \\ &\quad - \frac{1}{2(1-\lambda)}\|\mathbf{X}^T\mathbf{Y}\boldsymbol{\alpha}\|_2^2 + \mathbf{e}^T\boldsymbol{\alpha}. \end{aligned}$$

We recall that for a proper closed and convex function $h : \mathbb{R}^n \to (-\infty, \infty]$, the Moreau envelope [8] is given by $M_h^\mu(\mathbf{x}) \equiv \min_{\mathbf{u}}\left\{h(\mathbf{u}) + \frac{1}{2\mu}\|\mathbf{x} - \mathbf{u}\|_2^2\right\} = h(\text{prox}_{\mu h}(\mathbf{x})) + \frac{1}{2\mu}\|\mathbf{x} - \text{prox}_{\mu h}(\mathbf{x})\|_2^2$, and hence the dual function is given by

$$\min_{\mathbf{w},\beta,\boldsymbol{\xi}} L(\mathbf{w}, \beta, \boldsymbol{\xi}; \boldsymbol{\alpha}) = \lambda M_{\mathcal{S}_k}^{\frac{\lambda}{1-\lambda}}\left(\frac{\mathbf{X}^T\mathbf{Y}\boldsymbol{\alpha}}{1-\lambda}\right) - \frac{1}{2(1-\lambda)}\|\mathbf{X}^T\mathbf{Y}\boldsymbol{\alpha}\|_2^2 + \mathbf{e}^T\boldsymbol{\alpha}$$

in the case where $\mathbf{e}^T\mathbf{Y}\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\alpha} \le C\mathbf{e}$ and $-\infty$ otherwise. The minimizer of the above minimization problem provides the relation between the optimal primal and dual solutions:

$\mathbf{w} = \text{prox}_{\frac{\lambda}{1-\lambda}\mathcal{S}_k}\left(\frac{\mathbf{X}^T\mathbf{Y}\boldsymbol{\alpha}}{1-\lambda}\right)$. The obtained dual problem in minimization form is given by

$$\begin{aligned} \min\ & F(\boldsymbol{\alpha}) \equiv -\lambda M_{\mathcal{S}_k}^{\frac{\lambda}{1-\lambda}}\left(\frac{\mathbf{X}^T\mathbf{Y}\boldsymbol{\alpha}}{1-\lambda}\right) + \frac{1}{2(1-\lambda)}\|\mathbf{X}^T\mathbf{Y}\boldsymbol{\alpha}\|_2^2 - \mathbf{e}^T\boldsymbol{\alpha} \\ \text{s.t.}\ & \mathbf{e}^T\mathbf{Y}\boldsymbol{\alpha} = 0, \\ & \mathbf{0} \le \boldsymbol{\alpha} \le C\mathbf{e}. \end{aligned} \tag{DR}$$

Problem (DR) can be solved using accelerated gradient-based methods such as FISTA [3]. For that, we first notice that the objective function F is differentiable over $\mathbb{R}^n$ and that, using the basic properties of the Moreau envelope and the proximal mapping [2], the gradient ∇F is Lipschitz continuous with constant $\frac{\|\mathbf{X}\|_2^2}{1-\lambda}$ and given by

$$\nabla F(\boldsymbol{\alpha}) = \mathbf{X}\mathbf{X}\text{prox}_{\frac{\lambda}{1-\lambda}\mathcal{S}_k}\left(\frac{\mathbf{X}^T\mathbf{Y}\boldsymbol{\alpha}}{1-\lambda}\right) - \mathbf{e}.$$

In addition, the orthogonal projection onto the feasible set can be efficiently computed, see for example [2, Theorem 6.27].

5 Concluding remarks

In this paper we studied the sparse envelope function which is the biconjugate of the function $\mathbf{x} \mapsto \frac{1}{2}\|\mathbf{x}\|_2^2 + \delta_{\{\mathbf{x}:\|\mathbf{x}\|_0 \le k\}}$. This function serves as a natural regularizer in cases where both sparsity and grouping properties are expected. We have shown that both the function value and proximal operator of the sparse envelope function can be computed in linear time complexity in expectation. A natural future direction will be to consider different combinations of the ℓ_2 and ℓ_0 norms, as well as investigating general ℓ_p norms ($p \ge 0$).

Funding Funding was provided by Israel Science Foundation (Grant Number 92621).

References

1. Argyriou, A., Foygel, R., Srebro, N.: Sparse prediction with the k-support norm. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1457–1465. Curran Associates Inc., New York (2012)
2. Beck, A.: First-Order Methods in Optimization, Volume 25 of MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM). Mathematical Optimization Society, Philadelphia (2017)
3. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**(1), 183–202 (2009)
4. Brucker, P.: An $O(n)$ algorithm for quadratic knapsack problems. Oper. Res. Lett. **3**(3), 163–166 (1984)
5. 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)
6. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 3rd edn. MIT Press, Cambridge (2009)
7. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
8. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bull. Soc. Math. France **93**, 273–299 (1965)
9. Rockafellar, R.T.: Convex Analysis. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton (1970)
10. Singer,Y., Duchi, J., Shalev-Shwartz, S., Chandra, T.: Efficient projections onto the l1-ball for learning in high dimensions. In: Proceedings of the International Conference on Machine Learning (ICML) (2008)
11. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)

12. Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V.: Feature selection for svms. In: Advances in Neural Information Processing Systems 13, pp. 668–674. MIT Press, Cambridge (2001)
13. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. **67**(2), 301–320 (2005)