Multidimensional shrinkage-thresholding operator and Group LASSO penalties

Arnau Tibau Puig, Ami Wiesel, Gilles Fleury and Alfred O. Hero III

Abstract

The scalar shrinkage-thresholding operator is a key ingredient in variable selection algorithms arising in wavelet denoising, JPEG2000 image compression and predictive analysis of gene microarray data. In these applications, the decision to select a scalar variable is given as the solution of a scalar sparsity penalized quadratic optimization.

In some other applications, one seeks to select multidimensional variables. In this work, we present a natural multidimensional extension of the scalar shrinkage thresholding operator. Similarly to the scalar case, the threshold is determined by the minimization of a convex quadratic form plus an Euclidean norm penalty, however, here the optimization is performed over a domain of dimension $N \geq 1$. The solution to this convex optimization problem is called the multidimensional shrinkage threshold operator (MSTO). The MSTO reduces to the scalar case in the special case of $N = 1$. In the general case of $N > 1$ the optimal MSTO shrinkage can be found through a simple convex line search. We give an efficient algorithm for solving this line search and show that our method to evaluate the MSTO outperforms other state-of-the art optimization approaches. We present several illustrative applications of the MSTO in the context of Group LASSO penalized estimation.

Index Terms

Shrinkage-Thresholding Operator, Group LASSO regression, ℓ_2 penalized least squares, proximity operator.

The work of A. O. Hero III, G. Fleury and A. Tibau Puig was partially supported by the Digiteo DANSE project. The work of A. Wiesel was supported by a Marie Curie Outgoing International Fellowship within the 7th European Community Framework Programme. Part of this work was presented in the IEEE/SP 15th Workshop on Statistical Signal Processing, August 2009.

A. Tibau Puig and A. O. Hero III are with the Department of Electrical Engineering and Computer Science, University of Michigan (e-mail: atibaup, hero@umich.edu). A. Wiesel is with the School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel. G. Fleury and A. Tibau Puig are with the E3S - Supélec Systems Sciences / Signal Processing and Electronic Systems Department, Supélec, France

I. Introduction

The scalar shrinkage-thresholding operator is central to variable selection algorithms such as Iterative Thresholding [1] for image deblurring [2], wavelet-based deconvolution [3] or predictive analysis of gene expression microarrays [4].

In this paper, we introduce a multidimensional generalization of the scalar shrinkage thresholding operator. We define this operator as the minimization of a convex quadratic form plus a (non-squared) Euclidean norm penalty. We analyze this non-differentiable optimization problem and discuss its properties. In particular, in analogy to the scalar shrinkage operator, we show that this generalization yields a Multidimensional Shrinkage Thresholding Operator (MSTO) which takes a vector as an input and shrinks it or thresholds it depending on its Euclidean norm. Our results rely on a reformulation of the problem as a constrained quadratic problem with a conic constraint. Using conic duality theory, we transform this multidimensional optimization problem into a simple line search which can be efficiently implemented. We propose a simple algorithm to evaluate the MSTO and show by simulations that it outperforms other state-of-the-art algorithms.

In the second part of the paper, we discuss applications of the MSTO to three estimation problems. First, we consider the Euclidean norm penalized least squares and show, using the MSTO formulation, that this problem leads to a solution which is either the zero vector or the ridge-penalized least squares solution where the optimal shrinkage is chosen through a line search.

Next, we address Group LASSO penalized estimation with disjoint groups. This class of problems appears in many signal processing applications where the structure of the problem suggests enforcing a group-sparse estimate rather than a simple sparse estimate. Examples of this situation occur in spectrum cartography for cognitive radio [5], jointly-sparse signal recovery [6], regression with grouped variables [7] or source localization [8]. We show how the MSTO arises naturally in a block-descent algorithm for Group LASSO Linear Regression.

Finally, we show how the operator composition of MSTOs corresponds to the proximity operators for tree-structured Group LASSO penalties [9], [10], where the groups overlap in a hierarchical manner. Proximity operators can be understood as a generalization of convex projection operators and are a fundamental component of large-scale algorithms for non-differentiable convex problems [11], [1], [2].

This paper is organized as follows. In Section 2, we first define the MSTO and introduce our main theoretical result. Second, we discuss how to efficiently evaluate the MSTO. In Section 3 we illustrate applications of the MSTO in several statistical signal processing problems. We present numerical experiments in Section 4.

The following notation is used. Boldface upper case letters denote matrices, boldface lower case letters denote column vectors, and standard lower case letters denote scalars. The superscripts ^{T} and $\dagger$ denote the transpose and the matrix pseudoinverse operators, respectively. $\boldsymbol{\theta}_S$ denotes the subvector constructed from the indices in S. Given a symmetric matrix $\mathbf{X}$, $\xi_i\left(\mathbf{X}\right)$ refers to its i-th eigenvalue and $\mathbf{X} \succeq 0$ denotes semi-positive definiteness. Given a matrix $\mathbf{X}$, $\mathcal{R}\left(\mathbf{X}\right)$ and $\mathbf{X}_{S,T}$ denote its range and the submatrix constructed from the indices in S and T. $\mathbf{I}$ is the identity matrix. We define the second order cone K as [12]:

$$K = \left\{ \begin{bmatrix} \mathbf{x} \\ t \end{bmatrix} \in \mathbb{R}^{N+1} : ||\mathbf{x}||_2 \leq t \right\}, \tag{1}$$

and $\mathbf{z} \succeq_K 0$ means that $\mathbf{z} \in K$.

II. Multidimensional Shrinkage-Thresholding Operator (MSTO)

The scalar shrinkage-thresholding operator is usually defined as:

$$\begin{aligned} \mathcal{T}_{\lambda,h}\left(g\right) &:= \arg\min_x \frac{1}{2}hx^2 + gx + \lambda|x| \\ &= \begin{cases} -\frac{|g|-\lambda}{h}\text{sign}(g) & \text{if } |g| > \lambda \\ 0 & \text{otherwise.} \end{cases}, \end{aligned} \tag{2}$$

where $h, \lambda > 0$ and $g \in \mathbb{R}$. This operator takes a scalar g as an input and thresholds or shrinks its magnitude. A natural generalization is the following *Multidimensional Shrinkage Thresholding Operator* (MSTO):

$$\mathcal{T}_{\lambda,\mathbf{H}}\left(\mathbf{g}\right) \quad := \arg\min_{\mathbf{x}} \tfrac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x} + \lambda\|\mathbf{x}\|_2, \tag{3}$$

where $\mathbf{H} \succeq 0$ is an $N \times N$ matrix, $\lambda > 0$ and $\mathbf{g} \in \mathbb{R}^N$. This is a convex optimization problem and can be cast as a standard Second Order Cone Program (SOCP) [13]:

$$\begin{aligned} \min \quad & \tfrac{1}{2}t_1 + \lambda t_2 + \mathbf{g}^T\mathbf{x} \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{V}^T\mathbf{x} \\ \frac{t_1-1}{2} \\ \frac{t_1+1}{2} \end{bmatrix} \succeq_K 0, \quad \begin{bmatrix} \mathbf{x} \\ t_2 \end{bmatrix} \succeq_K 0, \end{aligned} \tag{4}$$

where $\mathbf{V}$ is such that $\mathbf{H} = \mathbf{V}\mathbf{V}^T$. SOCPs can be solved efficiently using interior point methods [13]. The next theorem shows that, as in the scalar case, the MSTO shrinks or thresholds the norm of the input vector $\mathbf{g}$ and that the corresponding SOCP (4) can be solved using a simple line search.

Theorem 1: Let $\mathbf{H} \succeq 0$, $\mathbf{g} \in \mathcal{R}(\mathbf{H})$ and $\lambda > 0$. The optimal value of the N-dimensional, non-differentiable problem:

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x} + \lambda\|\mathbf{x}\|_2, \tag{5}$$

is equal to the optimal value of the convex line-search:

$$\min_{\eta \geq 0} \quad \left(1 - \tfrac{1}{2}\mathbf{g}^T\mathbf{B}^{-1}(\eta)\,\mathbf{g}\right)\eta, \tag{6}$$

where:

$$\mathbf{B}(\eta) := \eta\mathbf{H} + \frac{\lambda^2}{2}\mathbf{I}. \tag{7}$$

Furthermore, their solutions are related by:

$$\mathcal{T}_{\lambda,\mathbf{H}}(\mathbf{g}) = \begin{cases} -\eta^*\mathbf{B}^{-1}(\eta^*)\,\mathbf{g} & \text{if } \|\mathbf{g}\|_2 > \lambda \\ \mathbf{0} & \text{otherwise,} \end{cases} \tag{8}$$

where $\eta^* \geq 0$ is the solution to (6).

Proof: Since $\mathbf{H} \succeq 0$ and $\|\cdot\|_2$ is a norm, it follows that $\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x}$ and $\|\mathbf{x}\|_2$ are convex functions of $\mathbf{x}$. Also, (5) is equivalent to the following quadratic program with a second order conic constraint:

$$\begin{aligned} \min_{x,t} \quad & \tfrac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x} + t \\ \text{s.t.} \quad & \begin{bmatrix} -\lambda\mathbf{x} \\ -t \end{bmatrix} \preceq_K 0. \end{aligned} \tag{9}$$

Slater's condition for generalized inequalities is verified and strong duality holds. Since K is self-dual, the conic dual can be written as ([12], Section 5.9.1):

$$\max q(\mathbf{u}, \mu) \quad \text{s.t.} \quad \begin{bmatrix} \mathbf{u} \\ \mu \end{bmatrix} \succeq_K 0, \tag{10}$$

where the dual function is defined as

$$q(\mathbf{u}, \mu) = \min_{x,t} \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x} + t - \mathbf{u}^T(\lambda\mathbf{x}) - \mu t. \tag{11}$$

This inner minimization is unbounded in t unless $\mu = 1$ and in $\mathbf{x}$ unless $\mathbf{u} \in \mathcal{R}(\mathbf{H})$. Otherwise, its optimum satisfies:

$$\mathbf{x} = -\mathbf{H}^\dagger(\mathbf{g} - \lambda\mathbf{u}). \tag{12}$$

Plugging (12) in (10), and using the fact that a non differentiable dual conic constraint $\begin{bmatrix} \mathbf{u}^T, & 1 \end{bmatrix}^T \succeq_K 0$ is equivalent to a standard quadratic constraint $\|\mathbf{u}\|_2^2 \leq 1$, we obtain the following dual concave maximization:

$$\max_{\|\mathbf{u}\|_2^2\leq 1, \mathbf{u}\in\mathcal{R}(\mathbf{H})} -\frac{1}{2}\left(\mathbf{g}-\lambda\mathbf{u}\right)^T\mathbf{H}^{\dagger}\left(\mathbf{g}-\lambda\mathbf{u}\right). \tag{13}$$

The standard Lagrange dual of this problem is:

$$\min_{\eta\geq 0}\max_{\mathbf{u}\in\mathcal{R}(\mathbf{H})} -\frac{1}{2}\left(\mathbf{g}-\lambda\mathbf{u}\right)^T\mathbf{H}^{\dagger}\left(\mathbf{g}-\lambda\mathbf{u}\right)-\eta\left(\mathbf{u}^T\mathbf{u}-1\right). \tag{14}$$

Since $\mathbf{H}\succeq 0$ and $\mathbf{H}^{\dagger}\mathbf{g}\in\mathcal{R}\left(\mathbf{H}^{\dagger}\right)$, the inner maximization is a simple quadratic problem in $\mathbf{u}$ with solution:

$$\mathbf{u}=\frac{\lambda}{2}\mathbf{B}^{-1}\left(\eta\right)\mathbf{g}, \tag{15}$$

where $\mathbf{B}\left(\eta\right)$ is defined in (7). This leads to the following line search over the Lagrange multiplier η:

$$\min_{\eta\geq 0}\left(1-\frac{1}{2}\mathbf{g}^T\mathbf{B}^{-1}\left(\eta\right)\mathbf{g}\right)\eta, \tag{16}$$

which proves the equivalence between (5) and (6) and is convex by Lagrange's duality properties.

The eigenvalues of $\mathbf{B}^{-1}\left(\eta\right)$ are real and can be characterized as:

$$\xi_i\left(\mathbf{B}^{-1}\left(\eta\right)\right)=\frac{1}{\eta\xi_i\left(\mathbf{H}\right)+\frac{\lambda^2}{2}}. \tag{17}$$

Since $\eta\geq 0$, $\xi_i\left(\mathbf{H}\right)\geq 0$ and $\lambda>0$, it holds that $0<\xi_i\left(\mathbf{B}^{-1}\left(\eta\right)\right)\leq\frac{2}{\lambda^2}$. Therefore, if $\|\mathbf{g}\|_2\leq\lambda$ then $\frac{1}{2}\mathbf{g}^T\mathbf{B}^{-1}\left(\eta\right)\mathbf{g}\leq 1$ and

$$\eta\left(1-\frac{1}{2}\mathbf{g}^T\mathbf{B}^{-1}\left(\eta\right)\mathbf{g}\right)\geq 0. \tag{18}$$

This implies that if $\|\mathbf{g}\|_2\leq\lambda$ the minimum in (16) is attained by choosing $\eta=0$. Plugging (15) into (12) yields (8). ■

A. *Evaluating the MSTO*

According to Theorem 1, evaluating the MSTO reduces to solving (6) when $\|\mathbf{g}\|_2>\lambda$. In the special case where $\mathbf{H}=k\mathbf{I}$ for some $k>0$, the optimality condition for (6) leads to a simple solution for its positive root:

$$\eta^*=\frac{\lambda}{2k}\left(\|\mathbf{g}\|_2-\lambda\right), \tag{19}$$

which yields the following closed form expression for the MSTO:

$$\mathcal{T}_{\lambda,k\mathbf{I}}\left(\mathbf{g}\right)=-\frac{1}{k}\left(\|\mathbf{g}\|_2-\lambda\right)_+\frac{\mathbf{g}}{\|\mathbf{g}\|_2}. \tag{20}$$

where $(x)_+ = \max(x, 0)$. This is equivalent to (2) if we define the multidimensional sign function as $\text{sign}(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ and coincides with the vectorial soft-threshold in [**?**]. If $\mathbf{H} \neq k\mathbf{I}$ and $\|\mathbf{g}\|_2 > \lambda$, evaluating the MSTO is non trivial and requires the numerical solution of the line-search in (6). In particular, we propose to use a Projected Newton approach with Goldstein step-length rule [14] which incorporates the advantages of second order methods while respecting the constraint $\eta \geq 0$ in (6). Let

$$w(\eta) \quad := \left(1 - \tfrac{1}{2}\mathbf{g}^T\mathbf{B}^{-1}(\eta)\,\mathbf{g}\right)\eta,$$

where $\mathbf{B}(\eta)$ is defined in (7). At iteration t, the Goldstein Projected Newton iteration for problem (6) is given by [14]:

$$\begin{aligned} \hat{\eta}^t &= \left(\eta^t - \frac{w'(\eta^t)}{w''(\eta^t)}\right)_+, \\ \eta^{t+1} &= \eta^t + \omega_n\left(\hat{\eta}^t - \eta^t\right), \end{aligned} \tag{21}$$

where $w'(\eta)$, $w''(\eta)$ are the first and second derivatives of $w(\eta)$ respectively. Letting $\delta \in (0, .5)$, the step length $\omega_n \in [0, 1]$ is determined according to the Goldstein scheme [14]:

$$\omega_n \in \begin{cases} \{0\} & \text{if } w'\left(\eta^t\right)\left(\eta^t - \hat{\eta}^t\right) = 0 \\ \{1\} & \text{if } w'\left(\eta^t\right)\left(\eta^t - \hat{\eta}^t\right) > 0, h\left(\eta^t, \hat{\eta}^t, 1\right) \geq \delta \\ \Omega_\delta\left(\eta^t, \hat{\eta}^t\right) & \text{if } w'\left(\eta^t\right)\left(\eta^t - \hat{\eta}^t\right) > 0, h\left(\eta^t, \hat{\eta}^t, 1\right) < \delta, \end{cases}$$

where $h(\eta, \hat{\eta}, \omega) = \frac{w(\eta) - w(\eta + \omega(\hat{\eta} - \eta))}{\omega w'(\eta)(\eta - \hat{\eta})}$ and $\Omega_\delta(\eta, \hat{\eta}) = \{\omega \in [0, 1], \delta \leq h(\eta, \hat{\eta}, \omega) \leq 1 - \delta\}$. Notice that for η^t close enough to the optimum, $\omega_n = 1$, which corresponds to the regular Newton regime. Here, $w'(\eta)$ and $w''(\eta)$ are given by the following formulae:

$$\begin{aligned} w'(\eta) &:= 1 - \tfrac{\lambda^2}{4}\mathbf{g}^T\mathbf{B}^{-2}(\eta)\,\mathbf{g}, \\ w''(\eta) &:= \tfrac{\lambda^2}{2}\mathbf{g}^T\mathbf{C}(\eta)\,\mathbf{g}, \end{aligned} \tag{22}$$

where $\mathbf{C}(\eta) := \mathbf{B}^{-3}(\eta)\,\mathbf{H}$. Convergence analysis for this line-search technique is available in [14].

III. Applications

In this section, we consider promising applications for the MSTO.

A. Linear regression with ℓ_2 norm penalty

Given a vector of n observations $\mathbf{y}$ and an $n \times N$ design matrix $\mathbf{X}$, we consider the following class of problems:

$$\min_{\theta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^r + \lambda\|\boldsymbol{\theta}\|_2^q. \tag{23}$$

Depending on r and q, this problem specializes to ridge regression ($r = 2$, $q = 2$), robust least-squares ($r = 1$, $q = 1$) [Theorem 3.2, [15]] or ℓ_2-penalized least squares ($r = 2$, $q = 1$). The following corollary of Theorem 1 characterizes the solution of the latter.

Corollary 2: The solution to the ℓ_2-penalized least squares

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2, \tag{24}$$

is:

$$\hat{\boldsymbol{\theta}} = \begin{cases} \left(\mathbf{X}^T\mathbf{X} + \epsilon\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y} & \text{if } \|\mathbf{X}^T\mathbf{y}\|_2 > \frac{\lambda}{2} \\ \mathbf{0} & \text{otherwise,} \end{cases} \tag{25}$$

where the shrinkage parameter $\epsilon = \frac{\lambda^2}{4\eta^*}$ is such that $\eta^* > 0$ solves:

$$\min_{\eta > 0} \left(1 - \mathbf{y}^T\mathbf{X}\left(\eta\mathbf{X}^T\mathbf{X} + \frac{\lambda^2}{4}\right)^{-1}\mathbf{X}^T\mathbf{y}\right)\eta. \tag{26}$$

In the special case where $\mathbf{X}$ is orthogonal ($2\mathbf{X}^T\mathbf{X} = k\mathbf{I}$) then (23) has the closed form solution (25) with $\epsilon = \frac{\lambda k}{2(k\|\mathbf{y}\|_2 - \lambda)}$.
The proof of this Corollary follows immediately from Theorem 1 by observing that $\hat{\boldsymbol{\theta}} = \mathcal{T}_{\lambda, 2\mathbf{X}^T\mathbf{X}}\left(-2\mathbf{X}^T\mathbf{y}\right)$.

B. Block-wise optimization for Group LASSO Linear Regression

Given $\mathbf{X}$, $\mathbf{y}$ as in the previous section and q *disjoint* groups of indices $G_i \subseteq \{1, ..., N\}$ satisfying $\cup_i G_i = \{1, ..., N\}$, the Group LASSO linear regression problem [7] is defined as:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in R^N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \sum_{i=1}^{q} \lambda_i \|\boldsymbol{\theta}_{G_i}\|_2, \tag{27}$$

where λ_i are fixed penalty parameters which we assume known. For an arbitrary design matrix $\mathbf{X}$, problem (27) can be solved using a Block Coordinate Descent (BCD) algorithm. The main idea of the BCD method is to iteratively solve (27) for each block G_i, letting the parameters corresponding to the other blocks remain fixed. Defining $\mathbf{H} = 2\mathbf{X}^T\mathbf{X}$, $\mathbf{g} = -2\mathbf{X}^T\mathbf{y}$ and using the MSTO operator (3) we can obtain the following update rule for each block G_i at iteration t:

$$\boldsymbol{\theta}_{G_i}^t \leftarrow \mathcal{T}_{\lambda_i, \mathbf{H}_{G_i,G_i}}\left(\boldsymbol{\theta}_{\bar{G}_i}^{t-1}\mathbf{H}_{\bar{G}_i,G_i} + \mathbf{g}_{G_i}\right), \tag{28}$$

where $\bar{G}_i$ is the complementary set of indices with respect to G_i. Convergence of this algorithm is guaranteed for this cost function [16].

C. MSTO in proximity operators for tree-structured penalties

The proximity operator of a (possibly non-differentiable) convex function $\Omega(\mathbf{x})$ is defined as [17], [11]:

$$\mathcal{P}_{\Omega}(\mathbf{g}) := \arg\min_{\mathbf{x}} \frac{1}{2} ||\mathbf{x}-\mathbf{g}||_2^2 + \Omega(\mathbf{x}).$$

Proximity operators are the main ingredient of proximal algorithms [11], [2], which arise in LASSO and Group LASSO penalized linear regression [1], [2], [9], collaborative sparse modeling [18] and hierarchical dictionary learning [10]. In these applications, proximal algorithms can be understood as a generalization of quasi-Newton methods to non-differentiable convex problems. An important example is the Iterative Thresholding procedure [1], [2] which solves:

$$\min_{\mathbf{x}} \mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{g}^T\mathbf{x} + \Omega(\mathbf{x}),$$

by generating the sequence

$$\mathbf{x}^{t+1} \leftarrow \mathcal{P}_{k\Omega}\left(\mathbf{x}^t - k\left(2\mathbf{H}\mathbf{x}^t + \mathbf{g}\right)\right),$$

for an appropriate $k > 0$.

The proximity operator of $\Omega(\mathbf{x}) = ||\mathbf{x}||_2$ is given by the orthogonal MSTO (20). In this section we show that the composition of orthogonal MSTOs corresponds to the proximity operator of Group LASSO penalties with (possibly overlapping) hierarchical groups. Given $\lambda > 0$, q groups of indices $G_i \subseteq \{1, ..., N\}$ and a partial order $\mathcal{O} = (o_1, ..., o_q)$ of the groups such that $G_{o_{i+1}} \cap G_{o_i} \neq \emptyset$ only if $G_{o_i} \subseteq G_{o_{i+1}}$ we consider the following function:

$$\Gamma(\mathbf{x}) = \lambda \sum_{i=1}^{q} \left|\left|\mathbf{x}_{G_{o_i}}\right|\right|_2. \tag{29}$$

It can be shown [10] that:

$$\mathcal{P}_{\Gamma}(\mathbf{g}) = \bigcirc_{i=1}^{q}\left(\mathcal{T}_{G_{o_i},\lambda,\mathbf{I}}\right)(\mathbf{g}), \tag{30}$$

where $\bigcirc$ is the composition operator and:

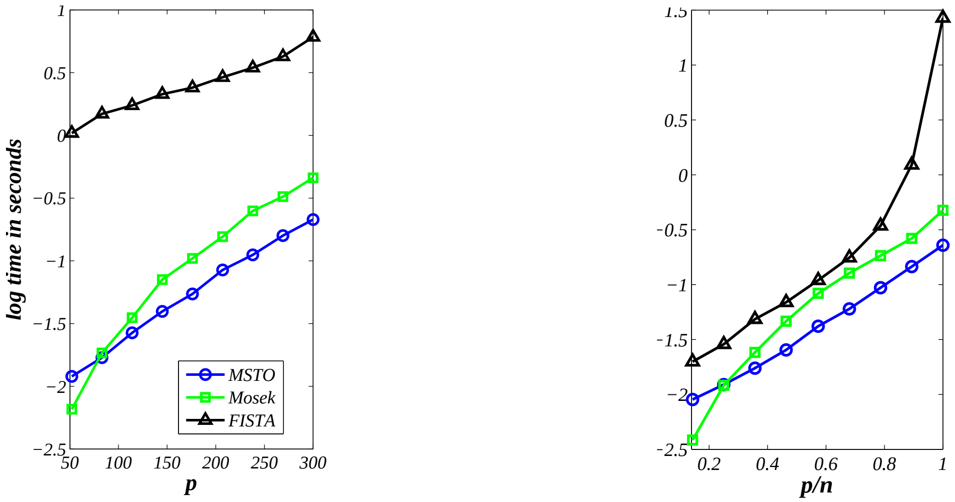
$$\mathcal{T}_{s,\lambda,\mathbf{I}}(\mathbf{g}) \quad := \arg\min_{\mathbf{x}} \tfrac{1}{2}\mathbf{x}^T\mathbf{x} + \mathbf{g}^T\mathbf{x} + \lambda\|\mathbf{x}_s\|_2, \tag{31}$$

where $s \subseteq \{1, \cdots, N\}$. It is clear that $\left[\mathcal{T}_{s,\lambda,\mathbf{I}}(\mathbf{g})\right]_s = \mathcal{T}_{\lambda,\mathbf{I}}(\mathbf{g}_s)$ and $\left[\mathcal{T}_{s,\lambda,\mathbf{I}}(\mathbf{g})\right]_i = \mathbf{g}_i$ for $i \notin s$.

IV. Numerical Results

In this section we illustrate the advantage of evaluating the MSTO using our theoretical results. To this end, we compare the elapsed times to evaluate equation (3) using three different optimization methods. The first one, which we denote by MSTO in the figures, solves the dual problem in (6) using the projected Newton approach described in Sec. II-A. The second method uses an accelerated first order method named FISTA[1] [2] and the third method uses the commercial state-of-the-art SOCP solver Mosek®. Our experiment consists of solving problem (24) for randomly generated $\mathbf{X}$ and $\mathbf{y}$ where we control the conditioning of the matrix $\mathbf{H} = 2\mathbf{X}^T\mathbf{X}$ through the ratio p/n (where p is the number of columns and n is the number of rows of $\mathbf{X}$).

We show in Figure 1 the average elapsed times to achieve the same value of the objective function, as a function of the number of variables and the ratio p/n. Our algorithm clearly outperforms the other two over a large range of values of p and p/n.



(a) $p/n = 1$ fixed, varying p

(b) $n = 250$ fixed, varying p/n

Fig. 1. Comparison of Mosek®, MSTO and FISTA elapsed times for solving (24) while varying p (with $p/n = 1$ fixed, plot (a)) and varying p/n (with $n = 250$ fixed, plot (b)). MSTO is significantly faster than the other two under the conditions considered.

V. Conclusions

We have introduced the MSTO, a multidimensional generalization of the Scalar Shrinkage Thresholding Operator. Our main theoretical result shows that the MSTO is the solution of a convex problem that

[1] FISTA is implemented using backtracking and (20) to compute its corresponding shrinkage/thresholding update.

performs Shrinkage/Thresholding on the norm of its input and that it can be efficiently evaluated in large finite dimensional spaces. The MSTO appears naturally in many algorithms for non-linear estimation with Group LASSO-type penalties. We show by simulation that our theory yields an algorithm that outperforms state-of-the-art second-order cone program solvers.

REFERENCES

[1] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math*, vol. 57, no. 11, pp. 1413–1457, 2004.

[2] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci*, vol. 2, pp. 183–202, 2009.

[3] R.D. Nowak and M.A.T. Figueiredo, "Fast wavelet-based image deconvolution using the EM algorithm," in *Conference Record of the 35th Asilomar Conference*, 2001, vol. 1.

[4] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Class prediction by nearest shrunken centroids, with applications to DNA microarrays," *Statistical Science*, pp. 104–117, 2003.

[5] J.A. Bazerque, G. Mateos, and G.B. Giannakis, "Group-Lasso on Splines for Spectrum Cartography," *Arxiv preprint arXiv:1010.0274*, 2010.

[6] M.B. Wakin, M.F. Duarte, S. Sarvotham, D. Baron, and R.G. Baraniuk, "Recovery of jointly sparse signals from few random projections," in *Proc. Neural Inform. Processing Systems–NIPS*, 2005.

[7] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, vol. 68, pp. 49–67, 2006.

[8] D. Malioutov, M. Cetin, and A.S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 8 Part 2, pp. 3010–3022, 2005.

[9] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *Annals of Statistics*, vol. 37, no. 6A, pp. 3468–3497, 2009.

[10] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for sparse hierarchical dictionary learning," in *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[11] P.L. Combettes and V.R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, 2006.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press.

[13] M.S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra and its Applications*, vol. 284, no. 1-3, pp. 193–228, 1998.

[14] J.C. Dunn, "Newton's Method and the Goldstein Step-Length Rule for Constrained Minimization Problems," *SIAM Journal on Control and Optimization*, vol. 18, pp. 659, 1980.

[15] L. El Ghaoui and H. Lebret, "Robust solutions to least squares problems with uncertain data," *SIAM Journal Matrix Analysis and Applications*, October 1997.

[16] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001.

[17] J.J. Moreau, "Proximite et dualite dans un espace hilbertien," *Bull. Soc. Math. France*, vol. 93, no. 2, pp. 273–299, 1965.

[18] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. Eldar, "Collaborative hierarchical sparse modeling," in *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*. IEEE, 2010, pp. 1–6.