The Ordered Weighted $\ell_1$ Norm: Atomic Formulation, Dual Norm, and Projections

Xiangrong Zeng, Student Member, IEEE, and Mário A. T. Figueiredo, Fellow, IEEE

Abstract—The ordered weighted $\ell_1$ norm (OWL) was recently proposed, with two different motivations: because of its good statistical properties as a sparsity promoting regularizer, and as generalization of the so-called octagonal shrinkage and clustering algorithm for regression (OSCAR). The OSCAR is a convex group-sparsity inducing regularizer, which does not require the prior specification of the group structure. Also recently, much interest has been raised by the atomic norm formulation of several regularizers, not only because it provides an new avenue for their theoretical characterization, but also because it is particularly well suited to a type of method known as conditional gradient (CG), or Frank-Wolfe, algorithm. In this paper, we derive the atomic formulation of the OWL and exploit this formulation to show how Tikhonov regularization schemes can be handled using state-of-the-art proximal splitting algorithms, while Ivanov regularization can be efficiently implemented via the Frank-Wolfe algorithm.

Index Terms—Group sparsity, atomic norm, Tikhonov regularization, Ivanov regularization, conditional gradient method, Frank-Wolfe algorithm.

I. INTRODUCTION

In signal processing and machine learning, in the context of sparse inference, much attention has been recently devoted, not only to standard sparsity (usually enforced/encouraged by the use of an $\ell_1$ regularizer, often called LASSO [1]), but also to regularizers that are able to yield structured/group sparsity. In fact, several regularizers that encourage group sparse solutions have been proposed in recent years, including the group LASSO (gLASSO) [2], the sparse gLASSO (sgLASSO) [3], the fused LASSO (fLASSO) [4], the elastic net (EN) [5], the octagonal shrinkage and clustering algorithm for regression (OSCAR) [6], and Mário A. T. Figueiredo, Student Member, IEEE, to mention the, arguably, best known examples (see a comprehensive review by Bach et al. [7]).

The gLASSO, as well as its several variants and descendants [7], require the prior specification of the structure of the groups, which is a strong requirement in many applications. The fLASSO, although not relying on a predefined group structure, does depend on a given order of the variables, making it unsuitable for variable selection/grouping in machine learning problems, namely linear regression or classification; in these problems, the order of the variables is usually meaningless, thus any regularizer should be invariant under permutations of these variables. In contrast, both the EN and the OSCAR were proposed for regression problems and are neither attached to any specific ordering of the variables nor to previous knowledge of the group structure.

The OSCAR regularizer (which has been shown to outperform EN in feature grouping [8]) consists of the $\ell_1$ norm plus a sum of pairwise $\ell_\infty$ penalties, which simultaneously encourage the components to be sparse and equal in magnitude, respectively. The OSCAR regularization problem can be efficiently solved by several state-of-the-art proximal splitting algorithms [9], such as the well known FISTA [10], TwiIST [11], or SpaRSA [12].

Recently, a regularizer that contains OSCAR (as well as the $\ell_1$ and $\ell_\infty$ norms) as a special case was proposed [13], [14]. In this paper, we refer to that regularizer as ordered weighted $\ell_1$ (OWL), and we will show that it is indeed a norm. Whereas in [13], OWL was proposed because of its good properties in terms of controlling the false discovery rate for variable selection, in [14] it was motivated as a generalization of OSCAR, for its ability to cluster/group regression variables. Very recently, the statistical performance of OWL regularization was analyzed, showing that its adequacy to deal with regression problems where the design matrix includes highly correlated columns [15].

The proximity operator of the OWL norm can be computed efficiently, with the leading cost being that of sorting the components of its argument [13], [14]. This fact makes problems involving OWL regularization efficiently solvable using a variety of proximal splitting algorithms, as mentioned above. However, other computational tools for the OWL norm (such as projections on a OWL norm ball) and methods (such as conditional gradient) have not yet appeared in the literature. The main contributions of this paper are: the derivation of the atomic norm formulation of the OWL norm; the derivation of the dual of the OWL norm; efficient methods to project of balls of both the OWL norm and its dual. In particular, the atomic formulation opens the door to using the conditional gradient (CG, also known as Frank-Wolfe [16]) algorithm to deal with problems where the regularization is formalized as an upper bound on the OWL norm (the so-called Ivanov formulation [17]).

Notation

Lower-case bold letters, e.g., $x, y$, denote (column) vectors, their transposes are $x^T, y^T$, and the $i$-th and $j$-th components are written as $x_i$ and $y_j$. Matrices are written in upper case bold, e.g., $A, B$. We use $|x|$ to denote the vector with the absolute values of the components of $x$. For some vector $x$, $x_{[i]}$ is its $i$-th largest component (i.e., for $x \in \mathbb{R}^n$,...
where $x[1] \geq x[2] \geq \cdots \geq x[n]$, with ties broken by some arbitrary rule); consequently, $|x|_i$ is the $i$-th largest component of $x$ in magnitude. The vector obtained by sorting (in non-increasing order) the components of $x$ is denoted as $x_i$, thus $|x|_i$ denotes the vector obtained by sorting the components of $x$ in non-increasing order of magnitude. Finally, $P(x)$ is a permutation matrix (thus $P(x)^{-1} = P(x)^T$) that sorts the components of $x$ in non-increasing order, i.e., $x_j = P(x) x$; naturally, $P(|x|)$ is a permutation matrix that sorts the components of $x$ in non-increasing order of magnitude.

II. OWL: Atomic Formulation and Dual Norm

A. The OWL Norm

The ordered weighted $\ell_1$ (OWL) norm [14], denoted as $\Omega_w : \mathbb{R}^n \to \mathbb{R}_+$, is defined as

$$\Omega_w(x) = \sum_{i=1}^{n} |x|_i \quad w_i = w^T |x|_i$$

where $w$ is a vector of non-increasing weights, i.e., satisfying $w_1 \geq w_2 \geq \cdots \geq w_n$.

Notice that if $w_1 = \cdots = w_n = 1$, then $\Omega_w(x) = \|x\|_1$, whereas taking $w_1 = 1$, and $w_i = \cdots = w_n = 0$, yields $\Omega_w(x) = \|x\|_\infty$. It is also clear that $w_1 \|x\|_\infty \leq \Omega_w(x) \leq w_1 \|x\|_1$. Finally, choosing

$$w_i = \lambda_1 + \lambda_2 (n-i) \quad \text{for} \quad i = 1, \ldots, n$$

where $\lambda_1, \lambda_2$ are non-negative parameters, makes the OWL norm become the OSCAR regularizer [9], that is,

$$\Omega_w(x) = \lambda_1 \|x\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}.$$  (3)

The fact that OWL is a norm was proved in [14, 13].

B. Atomic Norm Formulation of $\Omega_w(x)$

Consider a set $A \subset \mathbb{R}^n$ (the collection of so-called atoms), which is compact, centrally symmetric about the origin (i.e., $a \in \text{conv}(A) \Rightarrow -a \in \text{conv}(A)$), and $\text{conv}(A)$ contains a ball of radius $\varepsilon$ around the origin, for some $\varepsilon > 0$ [13]. Then, the atomic norm of some $x \in \mathbb{R}^n$ induced by $A$ is defined as

$$\|x\|_A = \inf \{ t \geq 0 : x \in t \text{conv}(A) \}.$$  (4)

For instance, taking $A = \{ \pm e_i \}$ (the set of all the vector with one component equal to $+1$ or $-1$ and all the others equal to zero, which has cardinality $|A| = 2n$) yields $\|x\|_A = \|x\|_1$, whereas for $A = \{ -1, -1 \}^n$ (which has cardinality $|A| = 2^n$), we obtain $\|x\|_A = \|x\|_\infty$. The $\ell_2$ norm is recovered if $A$ is the (infinite) set of all unit norm vectors.

Atomic norms can also be defined for matrices and other mathematical objects, and have recently been the focus of much research interest (see the work of Chandrasekaran et al [18] and Jaggi [16], and references therein). In convex analysis [19], the gauge function of a convex set $C$ is defined as

$$\gamma(x|C) = \inf \{ t \geq 0 : x \in t C \},$$

thus the atomic norm is simply $\|x\|_A = \gamma(x|\text{conv}(A))$.

Next, we discuss the atomic formulation of $\Omega_w(x)$. Obviously, due to the central symmetry property, we can focus of the non-negative orthant of $\mathbb{R}^n$, where we claim that the atomic set is given (in the general case) by

$$\tilde{B} = \bigcup_{i=1}^{n} \tilde{B}_i$$

where

$$\tilde{B}_1 = \left\{ \begin{array}{ccc} \tau_1 & 0 & 0 \\ 0 & \tau_1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \tau_1 \end{array} \right\},$$

$$\tilde{B}_{n-1} = \left\{ \begin{array}{ccc} \tau_{n-1} & \tau_{n-1} & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \tau_{n-1} \end{array} \right\},$$

$$\tilde{B}_n = \left\{ \begin{array}{ccc} \tau_n & \vdots & \vdots \\ \vdots & \tau_n & \vdots \\ \vdots & \vdots & \tau_n \end{array} \right\},$$

where

$$\tau_i = \left( \sum_{j=1}^{i} w_j \right)^{-1}.$$  (7)

In words, $\tilde{B}_i$ contains all the $n!/(i! (n-i)!)$ vectors with $i$ components equal to $\tau_i$ and the remaining ones equal to zero. Since the $\tilde{B}_i$ are mutually disjoint, the total number of atoms in the non-negative orthant is

$$\bigcup_{i=1}^{n} \tilde{B}_i = \sum_{i=1}^{n} \binom{n}{i} = 2^n - 1.$$

To cover all the orthants of $\mathbb{R}^n$ in a centrally symmetric fashion, we need to consider all the possible sign configurations of the non-zeros of each atom of each subset $\tilde{B}_i$. We denote the resulting sets as $B_i$, that is,

$$B_1 = \left\{ \begin{array}{ccc} \tau_1 & -\tau_1 & 0 \\ 0 & \tau_1 & -\tau_1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & -\tau_1 \end{array} \right\},$$

$$B_n = \left\{ \begin{array}{ccc} \tau_n & \tau_n & \vdots \\ \vdots & \tau_n & \vdots \\ \tau_n & \vdots & \vdots \end{array} \right\}.$$
In conclusion, we claim (and will prove below) that the atomic set underlying the OWL norm is given by

$$A = \bigcup_{i=1}^{n} B_i.$$  \hfill (9)

Notice that, since each element of $B_i$ contains $i$ non-zero components, the cardinality of this atomic set (in the general case) is

$$|A| = \sum_{i=1}^{n} \binom{n}{i} 2^i = 3^n - 1,$$  \hfill (10)

again because the sets $B_i$ are all mutually disjoint.

The atomic set presented above is for a general OWL norm, i.e., for the case where the components of $w$ constitute a strictly decreasing positive sequence. Several special cases require fewer atoms. For example, if $w_j = \lambda$, for $j = 1, \ldots, n$, we recover the standard $\ell_1$ norm. In that case, notice that $\tau_i = (i \lambda)^{-1}$, thus it is easy to show that $B_j \subset \text{conv}(B_1)$, for $j = 2, \ldots, n$, thus the atomic set can be reduced to $B_1$, which is well known to yield the $\ell_1$ norm \cite{18}. If $w_1 = \lambda$ and $w_j = 0$, for $j = 2, \ldots, n$, we recover the $\ell_\infty$ norm. In that case, we have $\tau_1 = 1/\lambda$, for all $i = 1, \ldots, n$, and it is easy to show that $B_j \subset \text{conv}(B_n)$, for $j = 1, \ldots, n - 1$, thus the atomic set is reduced to $B_n$.

Notice also that, in the general case (as defined in the previous paragraph), $A$ is a minimal set of atoms, that is, there is no set $A'$ strictly contained in $A$ such that $\text{conv}(A') = \text{conv}(A)$.

Next, we prove that $\|x\|_A$ is equivalent to $\Omega_w(x)$.

**Theorem 1:** Let $\|x\|_A$ be defined as in (4), with $A$ given by (10), and $\Omega_w(x)$ be as defined in (1). Then, for any $x \in \mathbb{R}^n$, $\|x\|_A = \Omega_w(x)$.

**Proof:** Since $\|x\|_A$ and $\Omega_w(x)$ are both homogeneous (they are norms), it suffices to show that $\|x\|_A = \Omega_w(x)$, for any $x$ such that $\|x\|_A = 1$, i.e., such that $x$ is at the boundary of $\text{conv}(A)$. Furthermore, since $\|x\|_A$ and $\Omega_w(x)$ are (in addition to homogeneous) also invariant w.r.t. permutations of the components of its argument and w.r.t. sign changes, we can focus on the first orthant and assume without loss of generality that $x$ belongs to the following convex cone:

$$\mathcal{T} = \{x \in \mathbb{R}^n : x_1 \geq x_2 \geq \cdots \geq x_n \geq 0\}. \hfill (11)$$

If $x \in \mathcal{T}$ and $\|x\|_A = 1$, then $x$ belongs to the intersection of the boundary of $\text{conv}(A)$ with $\mathcal{T}$, which implies that $x$ can be written as a convex combination of the $n$ elements of $A$ that belong to $\mathcal{T}$. Formally, there exist $\theta_1, \theta_2, \ldots, \theta_n \in [0, 1]$ and $\sum_{i=1}^{n} \theta_i = 1$, such that

$$x = \theta_1 \begin{bmatrix} \tau_1 \\ 0 \end{bmatrix} + \cdots + \theta_n \begin{bmatrix} \tau_n \\ 0 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \theta_i \tau_i \\ \sum_{i=2}^{n} \theta_i \tau_i \end{bmatrix}.$$

that is, the components of $x = [x_1, \ldots, x_n]^T$ are given by

$$x_k = \sum_{i=k}^{n} \theta_i \tau_i = \sum_{i=k}^{n} \theta_i \left( \sum_{j=1}^{i} w_j \right)^{-1}.$$ 

Then, computing $\Omega_w(x)$ yields

$$\Omega_w(x) = \sum_{k=1}^{n} w_k x_k = \sum_{k=1}^{n} \sum_{i=k}^{n} w_k \theta_i \left( \sum_{j=1}^{i} w_j \right)^{-1}. \hfill (12)$$

Now, noticing the the double summation $\sum_{k=1}^{n} \sum_{i=k}^{n}$ is equivalent to $\sum_{k=1}^{n} \sum_{i=k}^{n}$, we obtain

$$\Omega_w(x) = \sum_{i=1}^{n} \theta_i \left( \sum_{j=1}^{i} w_j \right)^{-1} \sum_{k=1}^{n} w_k = \sum_{i=1}^{n} \theta_i = 1, \hfill (13)$$

confirming that $\Omega_w(x) = 1$.

**C. Dual Norm of $\Omega_w$**

We will now show that the dual norm of $\Omega_w$, which by definition is given by

$$\Omega_w^*(x) = \max_{\Omega_w(u) \leq 1} \langle u, x \rangle,$$  \hfill (14)

can be obtained via the atomic formulation, that is,

$$\|x\|_A^* = \max_{\|u\|_A \leq 1} \langle u, x \rangle = \max_{\|u\|_{\text{conv}(A)}} \langle u, x \rangle = \max_{a \in A} \langle a, x \rangle,$$  \hfill (15)

where the third equality results from the well-known fundamental result in linear programming according to which the maximum (and minimum) of a linear function over a bounded closed convex polygonal region is attained at least at one of the region’s vertices.

Let $x_k \in \mathbb{R}^k$ be a sub-vector of $x \in \mathbb{R}^n$, consisting of the $k$ largest (in magnitude) elements of $x$ (naturally, $\|x_1\|_1 = \|x\|_\infty$ and $\|x_{(n)}\|_1 = \|x_1\|$). Then, we have

$$\max_{a \in B_1} \langle a, x \rangle = \tau_1 \|x_1\|_1 = \tau_1 \|x\|_\infty$$

$$\max_{a \in B_2} \langle a, x \rangle = \tau_2 \|x_2\|_1$$

$$\vdots$$

$$\max_{a \in B_{n-1}} \langle a, x \rangle = \tau_{n-1} \|x_{(n-1)}\|_1$$

$$\max_{a \in B_n} \langle a, x \rangle = \tau_n \|x_n\|_1 = \tau_n \|x\|_1$$

Combining (14), (15), and (16), provides the proof of the following lemma:
and their dual norms (in red bold): (a) the $\ell_1$ norm and its dual norm (the $\ell_\infty$ norm); (b), (c) and (d): the OWL norm with different choices of $w$, and their dual norms; (e) the $\ell_\infty$ norm and its dual norm (the $\ell_1$ norm).

Lemma 1: The dual norm of $\Omega_w$ is given by
\[
\|x\|_A^* (x) = \max \left\{ \gamma_k \|x(k)\|_1, k = 1, \ldots, n \right\}.
\] (17)

It is interesting to notice that the atoms underlying $\|x\|_A^*$ can also be obtained. Figure 1 shows balls (in $\mathbb{R}^2$) of the $\ell_1$ and $\ell_\infty$ norms and other OWL norms, and of their dual norms. A deeper study of the atomic formulation of these dual norms is left for future work.

III. SOLVING REGULARIZATION PROBLEMS

A. Regularization Formulations

In regularization theory, there are three standard formulations, depending on how the regularizer (here, an atomic norm) and the data-fidelity term (here, simply the least squares cost function) are combined to achieve a balance between the two goals [17]:

1) Tikhonov regularization (referred to as OWL-T)
\[
\min_x \frac{1}{2} \|y - Bx\|_2^2 + \tau \|x\|_A
\] (18)

2) Morozov regularization (referred to as OWL-M)
\[
\min_x \|x\|_A, \quad \text{s.t. } \|y - Bx\|_2 \leq \delta
\] (19)

3) Ivanov regularization (referred to as OWL-I)
\[
\min_x \frac{1}{2} \|y - Bx\|_2^2, \quad \text{s.t. } \|x\|_A \leq \varepsilon
\] (20)

where $\tau$, $\delta$, and $\varepsilon$ are regularization parameters. Since they are convex, these three formulations are equivalent, under mild conditions, in the sense that it is possible (though usually not easy) to adjust the regularization parameters such that the solutions are the same. However, when addressing a specific problem, it is often more convenient to use one or another of these formulations, usually because it may be easier to adjust the corresponding regularization parameter.

B. Key Computational Ingredients

Before addressing the regularization formulations just mentioned, we first define some notations that will be useful below. Let $G_\varepsilon = \{x : \|x\|_A \leq \varepsilon\}$ and $C_\xi = \{x : \|x\|_A^* \leq \xi\}$ be the atomic norm ball and dual atomic norm ball, respectively, with the given radius. The key computational ingredients to address these problems are the Euclidean projectors onto $G_\varepsilon$ and $C_\xi$, as well as the Moreau proximity operators [19], [20] of $\|\cdot\|_A$ and its dual $\|\cdot\|_A^*$. In this paper, we will also focus on the conditional gradient (CG, also known as Frank-Wolfe) algorithm, briefly reviewed in the next subsection, as an efficient tool to address OWL-I.

1) The Conditional Gradient Method: Consider a problem of the form
\[
\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to } x \in \mathcal{D},
\] (21)

where $f$ is convex and continuously differentiable and $\mathcal{D}$ is a compact convex set.

The CG is a classical algorithm (originally due to Frank and Wolfe [21]) for problems of the form (21), which has recently sparked a revival of interest [16]. Although there are three main variants of CG [16], the generic CG for (21) is as follows:

Algorithm Conditional Gradient
1. Set $i = 0$ and $x_0 \in \mathcal{D}$.
2. repeat
3. $d_i = \arg\min_{d \in \mathcal{D}} \langle d, \nabla f(x_i) \rangle$
4. $\gamma_i = \frac{2}{i+\tau}$
5. $x_{i+1} = (1 - \gamma_i)x_i + \gamma_id_i$
6. $i \leftarrow i + 1$
7. until some stopping criterion is satisfied.

The key step of this algorithm is finding the conditional gradient $d_i$ (see line 3 above), which becomes particularly convenient when $\mathcal{D}$ is an atomic norm ball [16].

2) Projection Onto $G_\varepsilon$: The projection of some $v \in \mathbb{R}^n$ onto $G_\varepsilon$ is defined as
\[
\text{proj}_{G_\varepsilon}(v) = \arg\min_{x \in G_\varepsilon} \|v - x\|_2^2,
\] (22)

which is actually the OWL-I problem (see (20)) with $B = I$, and could be addressed using the CG algorithm, by setting $f(x) = \frac{1}{2} \|v - x\|_2^2$ and $\mathcal{D} = G_\varepsilon$. However, (22) can be efficiently computed as shown in the following lemma.

Lemma 2: The Euclidean projection onto $G_\varepsilon$ is given by
\[
\text{proj}_{G_\varepsilon}(v) = \text{sign}(v) \circ P_T(|v|) \text{proj}_{L_\varepsilon}(P(|v|) |v|),
\] (23)

where
\[
L_\varepsilon = \{x : \sum_{i=1}^n w_i |x_i| \leq \varepsilon\}
\]
is a $w$-weighted $\ell_1$ ball of radius $\varepsilon$, and $T$ is defined as (11).

Proof: It is obvious that since $\Omega_w$ is insensitive to the signs of its argument, the signs of $\text{proj}_{G_\varepsilon}(v)$ will match those of its argument; consequently, we can assume without loss of generality that $v$ is in the first orthant $\mathbb{R}^n_+$. The proof proceeds by noticing that any vector $x \in \mathbb{R}^n_+$ can be written (maybe in a non-unique way) as $x = Q u$, where $Q$ is a permutation
matrix and $u \in \mathcal{T}$ (see (11)). We can thus reformulate (22) as
\[
\text{proj}_{\mathcal{L}_c \cap \mathcal{T}}(v) = Q \hat{u} \text{ with }
\]
\[
(\hat{u}, \hat{Q}) = \arg \min_{u \in \mathcal{L}_c \cap \mathcal{T}} \min_{Q \in S_n} \|v - Qu\|^2_2,
\]
(24)
where $S_n$ is the so-called symmetric group, i.e., the set of all permutations of $n$ symbols. Since $Q^TQ = I$, the inner minimization can be written as
\[
\min_{Q \in S_n} \|v - Qu\|^2_2 = \|v\|^2_2 + \|u\|^2_2 - 2 \max_{Q \in S_n} u^TQ^Tv.
\]
(25)
Noticing that $u \in \mathcal{T}$ implies that $u_\perp = u$, and invoking the classical Hardy-Littlewood-Pólya inequality\(^1\) shows that the optimum is attained for $Q = P(|v|)^T$. The outer minimization in (24) then becomes
\[
\hat{u} = \arg \min_{u \in \mathcal{L}_c \cap \mathcal{T}} \|P(|v|)|v| - u\|^2_2 = \text{proj}_{\mathcal{L}_c \cap \mathcal{T}}(P(|v|)|v|).
\]
(26)
Finally, considering the sign of $v$, and combing (26) with $Q = P(|v|)^T$ yields (22), completing the proof.

Notice that the projection onto $\mathcal{L}_c \cap \mathcal{T}$ doesn’t seem trivial to compute, and will be left for future work. In fact, this projection can be written (in the first orthant) as
\[
\text{proj}_{\mathcal{L}_c \cap \mathcal{T}}(z) = \arg \min_{x \in \mathcal{C}_c} \frac{1}{2} \|z - x\|^2_2
\]
\[
\text{s.t. } \sum_{i=1}^n u_i x_i \leq \varepsilon
\]
\[
x_1 \geq x_2 \geq \cdots \geq x_n,
\]
which a projection with an isotonicity constraint.

3) Projection Onto $\mathcal{C}_\xi$: The projection onto the dual norm ball $\mathcal{C}_\xi$ is given by
\[
\text{proj}_{\mathcal{C}_\xi}(v) = \arg \min_{x \in \mathcal{C}_\xi} \|v - x\|^2_2
\]
\[
= \xi \arg \min_{x \in \mathcal{C}_\xi} \left\| \frac{1}{\xi} v - x \right\|^2_2
\]
\[
= \xi \arg \min_{x \in \mathcal{C}_\xi} \frac{1}{2} \|v - x\|^2_2.
\]
(28)
As suggested in (18), (28) can, in general, be solved using a cutting plane method or ellipsoid method\(^2\). A more efficient alternative is to invoke Moreau’s theorem to write
\[
\text{proj}_{\mathcal{C}_\xi}(v) = v - \text{prox}_{\xi \|\cdot\|_A}(v),
\]
(29)
where $\text{prox}_{\xi \|\cdot\|_A}$ is the proximity operator of $\xi \|\cdot\|_A$ (defined below in (30)), for which fast $O(n \log n)$ algorithms exist\(^3\),\(^4\).

4) Proximity Operators: The proximity operator of $\xi \|\cdot\|_A$ is defined as
\[
\text{prox}_{\xi \|\cdot\|_A}(v) = \arg \min_{x \in \mathbb{R}^n} \left( \|x\|_A + \frac{1}{2} \|x - v\|^2 \right).
\]
(30)
According to Moreau’s theorem\(^5\),
\[
\text{prox}_{\xi \|\cdot\|_A}(v) = v - \text{proj}_{\mathcal{C}_1}(v).
\]
(31)
Although this formula is interesting, it may not be computationally relevant, since there are direct efficient ways to compute this proximity operator\(^6\),\(^7\).

In the same vein as above, the proximity operator of $\|\cdot\|^p_2$ can be computed by
\[
\text{prox}_{\|\cdot\|^p_2}(v) = v - \text{proj}_{\mathcal{C}_1}(v).
\]
(32)
As above shown, the projection onto $\mathcal{G}_c$ is given by (23).

C. Tackling OWL-T Via Proximal Splitting Algorithms

With an efficient algorithm to compute $\text{prox}_{\|\cdot\|_A}$, the OWL-T formulation can be tackled by any state-of-the-art proximal splitting algorithm (PSA), such as the well-known FISTA\(^8\), TwIST\(^9\), or SpaRSA\(^10\). Notice that for solving OWL-M, the CSALSA algorithm\(^11\), as an instance of the alternating direction method of multipliers (ADMM) family\(^12\), can avoid the (hard) projection on the ellipsoid in the constraint and uses a (simple) projection on an Euclidean ball, thanks to the use of variable splitting.

D. Tackling OWL-I Via CG

The OWL-I problem (20) perfectly fits the form (21), with $f(x) = \frac{1}{2} \|y - Bx\|^2_2$, and $D = \mathcal{G}_\xi$. Denoting $h = (-\nabla f(x_0)) = B^T(y - Bx_0)$, line 3 of the CG algorithm becomes
\[
d_i = \arg \min_{x \in \mathcal{C}_\xi} \langle x, \nabla f(x_0) \rangle
\]
\[
= \arg \max_{x \in \text{Conv}(A)} \langle x, h \rangle
\]
\[
= \varepsilon \arg \max_{a \in A} \langle a, h \rangle.
\]
(33)
The final maximization problem in (33) can be solved by the following three steps:
\[
\hat{s}^* = \text{sign}(h)
\]
\[
k^* = \arg \max_{k \in \{1, \ldots, n\}} \{ \tau_k \|h(k)\|_1 \}
\]
\[
d_i = \varepsilon \hat{s}^* \odot \arg \max_{a \in B_{k^*}} \langle a, |h| \rangle,
\]
(34)
where $|h|$ is the vector with the magnitudes of the components of $h$. Notice that the computational cost of each step is dominated by the $O(n \log n)$ cost of sorting the elements of $|h|$ to obtain the several $h(k)$.

IV. EXPERIMENTS: COMPARING PSA VS CG

In this section, we compare the efficiency of PSA versus that of CG, in solving linear regression problems under OWL regularization. In particular, we focus on the OSCAR regularizer\(^6\), which is a particular instance of the OWL norm (see Section II-A). We consider the OWL-T formulation, solved by FISTA\(^10\),\(^9\), and the OWL-I scheme (which is the Ivanov form of OSCAR, referred to as AtomicOSCAR) solved by the CG algorithm. We report experiments (using MATLAB on a 64-bit Windows 7 computer, with an Intel Core i7 3.07 GHz processor and 6.0 GB of RAM) on a benchmark synthetic dataset and a real dataset.
experiments reported below, we set to a common random sample of a standard Gaussian. In the $z_1 = 100$ to $\tau = 1$ keeping $B$. Breast cancer data

algorithm is faster than the OSCAR solved by FISTA. amongst some threshold. Figure 2 shows the elapsed time of the two algorithms, showing that the AtomicOSCAR solved by the CG is roughly 4 times faster than the OSCAR solved by FISTA, but obtain similar accuracies. Regarding the most time-consuming operations of these two algorithms, which are computing $\prox_{\|\cdot\|_A}$ and computing $\prox_{\|\cdot\|_A}$, respectively, the latter is much faster than the former, which explains the fast speed of the AtomicOSCAR solved by CG.

A. Synthetic data

We consider a benchmark synthetic regression problem (also used in [1, 5, 6, 8]). The vector of observed responses is $y = B\mathbf{x}^* + \mathbf{w}$, where the true coefficient vector is

$$\mathbf{x}^* = \begin{bmatrix} 3, \ldots, 3, 0, \ldots, 0 \end{bmatrix}^T \in \mathbb{R}^n$$

(35)

the noise is Gaussian, $\mathbf{w} \sim \mathcal{N}(0, 15^2\mathbf{I})$, and the design matrix $B = [\mathbf{b}_1, \ldots, \mathbf{b}_n] \in \mathbb{R}^{m \times n}$ is generated as

$$\mathbf{b}_1 = \mathbf{z}_1 + \epsilon_i, \quad i = 1, \ldots, 0.1n;$$

$$\mathbf{b}_2 = \mathbf{z}_2 + \epsilon_i, \quad i = 0.1n + 1, \ldots, 0.2n;$$

$$\mathbf{b}_3 = \mathbf{z}_3 + \epsilon_i, \quad i = 0.2n + 1, \ldots, 0.3n;$$

$$\mathbf{b}_i \sim \mathcal{N}(0, 1), \quad i = 0.3n + 1, \ldots, n.$$

where the $\epsilon_i$ are independent and identically distributed zero-mean Gaussian vectors with variance 0.16, and $\mathbf{z}_1$, $\mathbf{z}_2$, and $\mathbf{z}_3$ are constant vectors, i.e. each has all its components equal to a common random sample of a standard Gaussian. In the experiments reported below, we set $m = n$ and vary $n$ from 100 to 6400.

Notice that, since we are using the OSCAR regularizer [3] in OWL-T, we don’t need to adjust $\lambda_1$, $\lambda_2$, and $\tau$ simultaneously, since $\tau$ multiplies both $\lambda_1$ and $\lambda_2$. Thus, for each $n$, we adjust $\lambda_1$ and $\lambda_2$ using a validation set of 400 samples, keeping $\tau = 1$. The same applies to the OWL-I formulation, with $\epsilon = 1$. Since we are comparing different formulations (OWL-T and OWL-I), the stopping criterion is a critical issue, which we sidestep as follows: we run FISTA and the CG algorithm until the prediction error on a test set falls below some threshold. Figure 2 shows the elapsed time of the two algorithms, showing that the AtomicOSCAR solved by the CG algorithm is faster than the OSCAR solved by FISTA.

B. Breast cancer data

We report experiments on the breast cancer dataset \(^1\), which contains 8141 genes in 295 tumors, where 300 genes are known to be most correlated with the responses. To reduce the class imbalance, we duplicate the positive samples twice, yielding a total of 451 samples. The resulting samples are randomly split into subsets with 100, 100, and 251 samples, for cross validation (CV) \(^1\), training, and testing, respectively.

The total times for training and CV (and average elapsed time of computing $\prox_{\|\cdot\|_A}$ and \(^3\)) in each iteration of FISTA and CG, respectively, as well as the test set accuracies, averaged over 50 repetitions, are shown in Table \(^1\) from which, we can see that the AtomicOSCAR solved by CG is roughly 4 times faster than the OSCAR solved by FISTA, but obtain similar accuracies. Regarding the most time-consuming operations of these two algorithms, which are computing $\prox_{\|\cdot\|_A}$ and computing $\prox_{\|\cdot\|_A}$, respectively, the latter is much faster than the former, which explains the fast speed of the AtomicOSCAR solved by CG.

V. CONCLUSIONS

We have derived the atomic norm formulation of the ordered weighted $\ell_1$ (OWL) norm. Using the atomic norm formulation, we showed how to tackle the Tikhonov and Ivanov regularization schemes, under OWL regularization, via proximal splitting algorithms and the conditional gradient method, respectively.

REFERENCES

[1] R. Tibshirani, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society (B), pp. 267–288, 1996.

[2] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” Journal of the Royal Statistical Society (B), vol. 68, pp. 49–67, 2005.

[3] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “The sparse-group lasso,” Journal of Computational and Graphical Statistics, 2012, to appear.

[4] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, “Sparsity and smoothness via the fused lasso,” Journal of the Royal Statistical Society (B), vol. 67, pp. 91–108, 2004.

[5] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” Journal of the Royal Statistical Society (B), vol. 67, pp. 301–320, 2005.

[6] H.D. Bondell and B.J. Reich, “Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR,” Biometrics, vol. 64, pp. 115–123, 2007.

[7] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, “Structured sparsity through convex optimization,” Statistical Science, vol. 27, no. 4, pp. 450–468, 2012.

[8] L.W. Zhong and J.T. Kwok, “Efficient sparse modeling with automatic feature grouping,” IEEE Transactions on Neural Networks and Learning Systems, vol. 23, pp. 1436–1447, 2012.

[9] X. Zeng and M.A.T. Figueiredo, “Solving OSCAR regularization problems by fast approximate proximal splitting algorithms,” Digital Signal Processing, vol. 31, pp. 124–135, 2014.

[10] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” SIAM Journal on Imaging Sciences, vol. 2, pp. 183–202, 2009.

[11] J. Barzilai and J.M. Borwein, “Two-point step size gradient methods,” IMA Journal of Numerical Analysis, vol. 8, pp. 141–148, 1988.

[12] S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo, “Sparse reconstruction by separable approximation,” IEEE Transactions on Signal Processing, vol. 57, pp. 2479–2493, 2009.

[13] J. Bogdan, E. Berg, W. Su, and E. Candès, “Statistical estimation and testing via the ordered $\ell_1$ norm,” arXiv preprint http://arxiv.org/pdf/1310.1969v1.pdf, 2013.

[14] X. Zeng and M.A.T. Figueiredo, “Decreasing weighted sorted $\ell_1$ regularization,” IEEE Signal Processing Letters, vol. 21, pp. 1240–1244, 2014.

[15] M. Figueiredo and R. Nowak, “Sparse estimation with strongly correlated variables using ordered weighted $\ell_1$ regularization,” arXiv preprint arXiv:, 2014.

\(^1\)http://cbio.ensmp.fr/~ljacob/.

\(^3\)We have derived the atomic norm formulation of the ordered weighted $\ell_1$ (OWL) norm. Using the atomic norm formulation, we showed how to tackle the Tikhonov and Ivanov regularization schemes, under OWL regularization, via proximal splitting algorithms and the conditional gradient method, respectively.

\(^1\)We have derived the atomic norm formulation of the ordered weighted $\ell_1$ (OWL) norm. Using the atomic norm formulation, we showed how to tackle the Tikhonov and Ivanov regularization schemes, under OWL regularization, via proximal splitting algorithms and the conditional gradient method, respectively.

\(^1\)We have derived the atomic norm formulation of the ordered weighted $\ell_1$ (OWL) norm. Using the atomic norm formulation, we showed how to tackle the Tikhonov and Ivanov regularization schemes, under OWL regularization, via proximal splitting algorithms and the conditional gradient method, respectively.

\(^1\)We have derived the atomic norm formulation of the ordered weighted $\ell_1$ (OWL) norm. Using the atomic norm formulation, we showed how to tackle the Tikhonov and Ivanov regularization schemes, under OWL regularization, via proximal splitting algorithms and the conditional gradient method, respectively.

\(^1\)We have derived the atomic norm formulation of the ordered weighted $\ell_1$ (OWL) norm. Using the atomic norm formulation, we showed how to tackle the Tikhonov and Ivanov regularization schemes, under OWL regularization, via proximal splitting algorithms and the conditional gradient method, respectively.
### Table I
RESULTS OF TIME AND TEST ACCURACY

| Algorithms                        | Time (seconds) | Test accuracy |
|-----------------------------------|----------------|---------------|
|                                   | CV             | Training      |
| OSCAR solved by FISTA             | 2442.5612 (0.0288) | 4.0462 (0.0209) | 77.65 |
| AtomicOSCAR solved by CG          | 496.6427 (0.0012) | 1.0946 (0.0010) | 76.83 |

*For FISTA in CV, (a) averaged elapsed time of computing $\text{prox}_{\AA} \cdot$ in each iteration.*

*For FISTA in Training, (b) averaged elapsed time of computing $\text{prox}_{\AA} \cdot$ in each iteration.*

*For CG in CV, (c) averaged elapsed time of computing $\frac{1}{\lambda'}$ in each iteration.*

*For CG in Training, (d) averaged elapsed time of computing $\frac{1}{\lambda'}$ in each iteration.*

---

[16] M. Jaggi, “Revisiting Frank-Wolfe: Projection-free sparse convex optimization,” in Proceedings of the 30th International Conference on Machine Learning (ICML-13), 2013, pp. 427–435.

[17] D. Lorenz and N Worliczek, “Necessary conditions for variational regularization schemes,” Inverse Problems, vol. 29, 2013.

[18] V. Chandrasekaran, B. Recht, P. A Parrilo, and A. S Willsky, “The convex geometry of linear inverse problems,” Foundations of Computational Mathematics, vol. 12, no. 6, pp. 805–849, 2012.

[19] R. T. Rockafellar, Convex Analysis, Princeton University Press, 1970.

[20] H.H. Bauschke and P.L. Combettes, Convex analysis and monotone operator theory in Hilbert spaces, 2011.

[21] M. Frank and P. Wolfe, “An algorithm for quadratic programming,” Naval research logistics quarterly, vol. 3, no. 1-2, pp. 95–110, 1956.

[22] G. Hardy, J. Littlewood, and G. Pólya, Inequalities, Cambridge University Press, 1934.

[23] Y. Nesterov and I. Nesterov, Introductory lectures on convex optimization: A basic course, vol. 87, Springer, 2004.

[24] J.M. Bioucas-Dias and M.A.T. Figueiredo, “A new TwIST: two-step iterative shrinkage/thresholding algorithms for image restoration,” IEEE Transactions on Image Processing, vol. 16, pp. 2992–3004, 2007.

[25] M.V. Afonso, J.M. Bioucas-Dias, and M.A.T. Figueiredo, “An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems,” IEEE Transactions on Image Processing, vol. 20, pp. 681–695, 2011.

[26] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Foundations and Trends® in Machine Learning, vol. 3, pp. 1–122, 2011.