# THE ATOMIC NORM FORMULATION OF OSCAR REGULARIZATION WITH APPLICATION TO THE FRANK-WOLFE ALGORITHM

*Xiangrong Zeng and Mário A. T. Figueiredo*

Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, Portugal

## ABSTRACT

This paper proposes atomic norm formulation of *octagonal shrinkage and clustering algorithm for regression* (OSCAR) regularization. The OSCAR regularizer can be reformulated using a *decreasing weighted sorted $\ell_1$ (DWSL1) norm* (which is shown to be convex). We also show how, by exploiting an atomic norm formulation, the Ivanov regularization scheme involving the OSCAR regularizer can be handled using the Frank-Wolfe (also known as conditional gradient) method.

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

## 1. 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 notions of structured/group sparsity. Several group-sparsity-inducing regularizers 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 several others not listed here due to space limitations (see a comprehensive review by Bach *et al* [7]). However, the gLASSO (and its many variants and descendants [7]) require prior knowledge about the structure of the groups, which is a too strong requirement in many applications. The fLASSO depends on a given order of variables, making it much better suited to signal processing applications than to variable selection and grouping in machine learning problems, such as regression or classification, where the order of the variables is usually meaningless, and any regularizer should be invariant under permutations of these variables. In contrast, the EN and OSCAR approaches were proposed for regression problems and are not attached to any specific ordering of the variables or to previous knowledge about group structure.

This work was partially supported by Fundao para a Ciłncia e Tecnologia, grants PEst-OE/EEI/LA0008/2013 and PTDC/EEI-PRO/1470/2012. Xiangrong Zeng is partially supported by grant SFRH/BD/75991/2011.

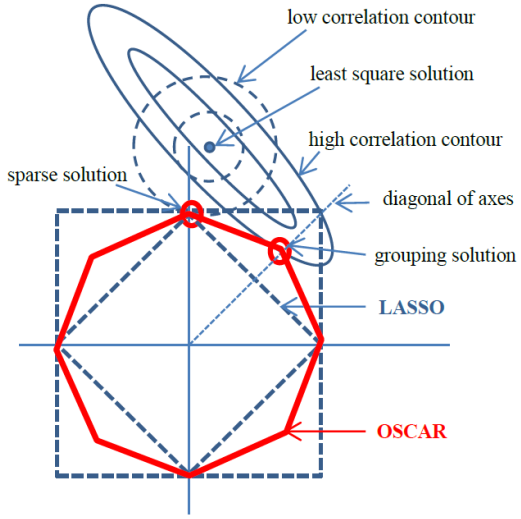


**Fig. 1**. Illustration of the OSCAR regularization.

The OSCAR regularizer (which has been shown to outperform EN in feature grouping [8]) is defined as

$$\phi_{\text{OSCAR}}^{\lambda_1,\lambda_2}(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}, \qquad (1)$$

where $\lambda_1$ and $\lambda_2$ are non-negative parameters (which, in practice, can be obtained, for example, using *cross validation*) [8]; the $\ell_1$ norm and the pairwise $\ell_\infty$ penalty simultaneously encourage the components to be sparse and equal in magnitude, respectively. Level curves of the OSCAR and LASSO regularizers are shown in Figure 1.

The Tikhonov regularization formulation for a regression problem with design matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, under OSCAR regularization, has the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \phi_{\text{OSCAR}}^{\lambda_1,\lambda_2}(\mathbf{x}) \qquad (2)$$

and can be efficiently solved in [9] by several state-of-the-art proximal splitting algorithms, such as the well known FISTA [10], TwIST [11], SpaRSA [12], ADMM [13], SBM [14], and PADMM [15].

As pointed out before [16], [17], it may happen that components with small magnitude that should be shrunk to zero by the $\ell_1$ norm are also penalized by the pairwise $\ell_\infty$ term, which may prevent accurate grouping; moreover, components with large magnitude that should simply be grouped by the pairwise $\ell_\infty$ norm are also shrunk by the $\ell_1$ norm. To overcome these drawbacks, we previously proposed the *SPARsity-and-Clustering* (SPARC) regularizer [16], [17], where the cardinality of the support of the solution is restricted and the pairwise $\ell_\infty$ penalty is applied only to the non-zero elements. The rationale behind this regularizer is that it enforces $K$-sparsity and encourages the non-zero components (and only those) to be pair-wise equal in magnitude.

However, the SPARC regularizer is non-convex while the OSCAR regularizer is convex, and the convexity is the necessary aspect for the atomic norm [18] whose favorable facial structure makes it a useful convex heuristic to recover simple models. In this paper, the OSCAR regularizer is reformulated as a decreasing weighted sorted $\ell_1$ (DWSL1) norm, and its atomic norm formulation and dual norm are exploited. Furthermore, we address the so-called Ivanov-type regularization scheme [19],

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\left\|\mathbf{y}-\mathbf{A}\mathbf{x}\right\|_2^2, \quad \text{subject to} \quad \phi_{\text{OSCAR}}^{\lambda_1,\lambda_2}(\mathbf{x}) \leq \epsilon \tag{3}$$

($\epsilon$ is a positive parameter) using atomic norm tools. In particular, we show how to tackle problem (3) using the Frank-wolfe (or conditional gradient) algorithm, with the help of the atomic norm formulation of $\phi_{\text{OSCAR}}^{\lambda_1,\lambda_2}$.

## 2. REFORMULATION OF THE OSCAR REGULARIZER

Before proceeding, let us define the decreasing weighted sorted $\ell_1$ (DWSL1) norm as

$$\Gamma_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}, \quad \Gamma_{\mathbf{w}}(\mathbf{z}) = \left\|\mathbf{w}\odot\grave{\mathbf{z}}\right\|_1, \tag{4}$$

where $\odot$ denotes the component-wise product, $\grave{\mathbf{z}}$ is the vector obtained from $\mathbf{z}$ by sorting its entries in decreasing order of magnitude (with ties broken by an arbitrary fixed rule) and $\mathbf{w}$ is vector of weights such that its components form a non-increasing sequence:

$$w_1 \geq w_2 \geq \cdots \geq w_d.$$

Let $\mathbf{P}(\mathbf{z})$ be the permutation matrix that sorts $\mathbf{z}$ into $\grave{\mathbf{z}}$, *i.e.*,

$$\grave{\mathbf{z}} = \mathbf{P}(\mathbf{z})\,\mathbf{z}, \tag{5}$$

which, of course, satisfies $\big(\mathbf{P}(\mathbf{z})\big)^{-1} = \big(\mathbf{P}(\mathbf{z})\big)^T$. Then, we can write

$$\Gamma_{\mathbf{w}}(\mathbf{z}) = \left\|\mathbf{w}\odot\big(\mathbf{P}(\mathbf{z})\,\mathbf{z}\big)\right\|_1 = \left\|\big((\mathbf{P}(\mathbf{z}))^T\,\mathbf{w}\big)\odot\mathbf{z}\right\|_1. \tag{6}$$

The convexity of $\Gamma_{\mathbf{w}}$ and the fact that it is a norm are given by the following two lemmas. We should point out that similar results were very recently proved (using different arguments) in [20].

**Lemma 1** *$\Gamma_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ is a convex function.*

*Proof:* Let $\mathbf{u},\mathbf{v}\in\mathbb{R}^d$, $\theta\in[0,1]$, $\mathbf{z}=\theta\mathbf{u}+(1-\theta)\,\mathbf{v}$, then $\Gamma(\mathbf{u}) = \left\|\mathbf{w}\odot\grave{\mathbf{u}}\right\|_1, \Gamma(\mathbf{v}) = \left\|\mathbf{w}\odot\grave{\mathbf{v}}\right\|_1, \Gamma(\mathbf{z}) = \left\|\mathbf{w}\odot\grave{\mathbf{z}}\right\|_1$, where $\grave{\mathbf{u}} = \mathbf{P}(\mathbf{u})\,\mathbf{u}$, $\grave{\mathbf{v}} = \mathbf{P}(\mathbf{v})\,\mathbf{v}$, and $\grave{\mathbf{z}} = \mathbf{P}(\mathbf{z})\,\mathbf{z}$. Thus,

$$\begin{aligned}
\Gamma_{\mathbf{w}}(\mathbf{z}) &= \left\|\big(\mathbf{P}(\mathbf{z})\mathbf{z}\big)\odot\mathbf{w}\right\|_1 \\
&= \left\|\big(\mathbf{P}(\mathbf{z})(\theta\mathbf{u}+(1-\theta)\mathbf{v})\big)\odot\mathbf{w}\right\|_1 \\
&\leq \theta\left\|\big(\mathbf{P}(\mathbf{z})\mathbf{u}\big)\odot\mathbf{w}\right\|_1 + (1-\theta)\left\|\big(\mathbf{P}(\mathbf{z})\mathbf{v}\big)\odot\mathbf{w}\right\|_1 \\
&\leq \theta\left\|\mathbf{P}(\mathbf{u})\mathbf{u}\odot\mathbf{w}\right\|_1 + (1-\theta)\left\|\big(\mathbf{P}(\mathbf{v})\mathbf{v}\big)\odot\mathbf{w}\right\|_1 \\
&= \theta\,\Gamma_{\mathbf{w}}(\mathbf{u}) + (1-\theta)\,\Gamma_{\mathbf{w}}(\mathbf{v})
\end{aligned}$$

where the first inequality is simply the triangle inequality and the second one results from the following fact: if the entries of $\mathbf{b}$ form a non-increasing non-negative sequence, then $\left\|\mathbf{P}(\mathbf{c})\mathbf{a}\odot\mathbf{b}\right\|_1 \leq \left\|\mathbf{P}(\mathbf{a})\mathbf{a}\odot\mathbf{b}\right\|_1$, for any $\mathbf{c}$. ■

**Lemma 2** *If $w_1 > 0$, then $\Gamma_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ is a norm.*

*Proof:* The positive homogeneity of $\Gamma_{\mathbf{w}}$ (that is, that $\Gamma_{\mathbf{w}}(\alpha\mathbf{z}) = |\alpha|\,\Gamma_{\mathbf{w}}(\mathbf{z})$, for any $\alpha\in\mathbb{R}$) is obvious (and was in fact already used in the proof of Lemma 1). The triangle inequality results trivially from the convexity shown in Lemma 1, by taking $\theta=\frac{1}{2}$, combined with the positive homogeneity. Finally, we need to prove that $\Gamma_{\mathbf{w}}(\mathbf{z}) = 0 \Leftrightarrow \mathbf{z} = 0$; this is clearly true, if $w_1 > 0$, since it is clear that $w_1\|\mathbf{z}\|_1 \geq \Gamma_{\mathbf{w}}(\mathbf{z}) \geq w_1\|\mathbf{z}\|_\infty$. ■

Our motivation to consider $\Gamma_{\mathbf{w}}$ results from the fact that the OSCAR regularizer (as defined in (1)) is a particular case of this norm [8], that is,

$$\phi_{\text{OSCAR}}^{\lambda_1,\lambda_2}(\mathbf{x}) = \Gamma_{\mathbf{w}}(\mathbf{x}), \tag{7}$$

for $\mathbf{w} = [w_1, w_2, ..., w_n]^T$ given by

$$w_j = \lambda_1 + \lambda_2(n-j). \tag{8}$$

In the sequel, we assume that $\mathbf{w}$ is always as given by (8).

## 3. SOLVING IVANOV REGULARIZATION PROBLEM INVOLVING OSCAR REGULARIZER

The Tikhonov regularization formulation in (2) is one of several possible ways of using the OSCAR regularizer, and is the only one that has been previously considered [8] [9]. A common alternative (known as Ivanov regularization [19]) adopts a different criterion,

$$\begin{aligned}
&\min_{\mathbf{x}\in\mathbb{R}^n} \tfrac{1}{2}\|\mathbf{y}-\mathbf{A}\,\mathbf{x}\|_2^2 \\
&\text{subject to } \ \Gamma_{\mathbf{w}}(\mathbf{x}) \leq 1.
\end{aligned} \tag{9}$$

(notice that there is no loss of generality in taking 1 as the upper bound for $\Gamma_{\mathbf{w}}(\mathbf{x}) = \phi_{\text{OSCAR}}^{\lambda_1,\lambda_2}(\mathbf{x})$; any other other non-negative value can be absorbed by $\lambda_1$ and $\lambda_2$, equivalently by the weights $w_1, ..., w_n$). Due to the convexity of $\Gamma_{\mathbf{w}}$ (see Lemma 2), problems (9) and (2) are equivalent under additional mild conditions; however, it is sometimes more convenient to address one rather than the other.

A classical algorithm that has recently seen a revival of interest to address problems of the form (9) is the *conditional gradient method* (CGM, also know as the Frank-Wolfe method [21], [22]). Although there are three main variants of CGM [22], the generic CGM for (9) is as follows (denoting $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\,\mathbf{x}\|_2^2$ and $\mathcal{D} = \{\mathbf{x} \in \mathbb{R}^n : \Gamma_{\mathbf{w}}(\mathbf{x}) \leq 1\}$):

**Algorithm** *CGM for (9)*

1. Set $i = 0$ and $\mathbf{x}_0 \in \mathcal{D}$.
2. **repeat**
3. $\quad \mathbf{d}_i = \arg\min_{\mathbf{d}\in\mathcal{D}} \langle \mathbf{d}, \nabla f(\mathbf{x}_i)\rangle$
4. $\quad \gamma_i = \frac{2}{i+2}$
5. $\quad \mathbf{x}_{i+1} = (1-\gamma_i)\mathbf{x}_i + \gamma_i \mathbf{d}_i$
6. $\quad i \leftarrow i + 1$
7. **until** some stopping criterion is satisfied.

The CGM is particularly convenient when the regularizer is a so-called *atomic norm* [22]. We will now show how $\Gamma_{\mathbf{w}}(\mathbf{x})$ can be written as an atomic norm and how that can be exploited to efficiently implement the CGM.

### 3.1. Atomic Norm Formulation of $\Gamma_{\mathbf{w}}(\mathbf{x})$

Let $\mathcal{A} \subset \mathbb{R}^n$ (a collection of *atoms*), such that $\text{conv}(\mathcal{A})$ is compact, centrally symmetric about the origin (*i.e.*, $\mathbf{a} \in \text{conv}(\mathcal{A}) \Rightarrow -\mathbf{a} \in \text{conv}(\mathcal{A})$), and $\text{conv}(\mathcal{A})$ contains a ball of radius $\epsilon$ around the origin, for some $\epsilon > 0$ [18]. Then, the *atomic norm* of some $\mathbf{x} \in \mathbb{R}^n$ induced by $\mathcal{A}$ is defined as

$$\|\mathbf{x}\|_{\mathcal{A}} = \inf\{t > 0 : \mathbf{x} \in t\,\text{conv}(\mathcal{A})\}. \tag{10}$$

For instance, taking $\mathcal{A} = \{\pm\mathbf{e}_i\}$ (the set of all the vector with one component equal to $+1$ or $-1$ and all the others equal to zero, which has cardinality $2\,n$) yields $\|\mathbf{x}\|_{\mathcal{A}} = \|\mathbf{x}\|_1$, whereas for $\mathcal{A} = \{-1, +1\}^n$, we obtain $\|\mathbf{x}\|_{\mathcal{A}} = \|\mathbf{x}\|_\infty$. The $\ell_2$ norm is recovered if $\mathcal{A}$ 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 considerable research interest (see the work of Chandrasekaran et al [18] and Jaggi [22], and references therein).

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

$$\check{\mathcal{B}} = \bigcup_{i=1}^{n} \check{\mathcal{B}}_i \tag{11}$$

where

$$\check{\mathcal{B}}_1 = \left\{ \begin{bmatrix} \tau_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \tau_1 \\ \vdots \\ 0 \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \tau_1 \end{bmatrix} \right\},$$

$$\check{\mathcal{B}}_2 = \left\{ \begin{bmatrix} \tau_2 \\ \tau_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_2 \\ 0 \\ \tau_2 \\ 0 \\ \vdots \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \tau_2 \\ \tau_2 \end{bmatrix} \right\},$$

$$\vdots$$

$$\check{\mathcal{B}}_{n-1} = \left\{ \begin{bmatrix} \tau_{n-1} \\ \tau_{n-1} \\ \vdots \\ \tau_{n-1} \\ 0 \end{bmatrix} \begin{bmatrix} \tau_{n-1} \\ \vdots \\ \tau_{n-1} \\ 0 \\ \tau_{n-1} \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ \tau_{n-1} \\ \tau_{n-1} \\ \vdots \\ \tau_{n-1} \end{bmatrix} \right\}$$

and

$$\check{\mathcal{B}}_n = \left\{ [\tau_n, \tau_n, \cdots, \tau_n]^T \right\}, \tag{12}$$

where

$$\begin{aligned} \tau_i = \left( \sum_{j=1}^{i} w_j \right)^{-1} &= \left( \sum_{j=1}^{i} [\lambda_1 + \lambda_2 (n-j)] \right)^{-1} \\ &= \left( \lambda_1 i + \lambda_2\, i \left( n - \tfrac{i+1}{2} \right) \right)^{-1}. \end{aligned} \tag{13}$$

Notice that $\left|\check{\mathcal{B}}_i\right| = \binom{n}{i} = n!/(i!\,(n-i)!)$, for $i = 1, \cdots, n$, thus the total number of atoms in the first orthant is

$$\left| \bigcup_{i=1}^{n} \check{\mathcal{B}}_i \right| = \sum_{i=1}^{n} \binom{n}{i} = 2^n - 1,$$

since all the $\check{\mathcal{B}}_i$ are mutually disjoint.

To cover all the orthants, we consider all the possible sign configurations of the non-zeros of each atom of each subset $\check{\mathcal{B}}_i$. We denote the resulting sets as $\mathcal{B}_i$; for example,

$$\mathcal{B}_1 = \left\{ \begin{bmatrix} \tau_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} -\tau_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \tau_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -\tau_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots \right\},$$

$$\mathcal{B}_2 = \left\{ \begin{bmatrix} \tau_2 \\ \tau_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} -\tau_2 \\ \tau_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_2 \\ -\tau_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} -\tau_2 \\ -\tau_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots \right\},$$

and so on. Consequently, since each element of $\check{\mathcal{B}}_i$ contains $i$ non-zero components, the cardinality of the complete atomic set (in the general case) is

$$|\mathcal{A}| = \sum_{i=1}^{n} \binom{n}{i} 2^i = 3^n - 1.$$

Notice that if $\lambda_2 = 0$, thus $w_i = \lambda_1$, we recover the $\ell_1$ norm; in this case, $\tau_i = (i\,\lambda_1)^{-1}$, thus $\check{\mathcal{B}}_j \subset \text{conv}(\check{\mathcal{B}}_1)$, for $j = 2, ..., n$, and $\mathcal{A}$ reduces to $\mathcal{B}_1$ [18].

Next, we prove that $\|\mathbf{x}\|_{\mathcal{A}}$ is equivalent to $\Gamma_{\mathbf{w}}(\mathbf{x})$.

**Lemma 3** *For any* $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\|_{\mathcal{A}} = \Gamma_{\mathbf{w}}(\mathbf{x})$.

*Proof:* Since $\|\mathbf{x}\|_{\mathcal{A}}$ and $\Gamma_{\mathbf{w}}(\mathbf{x})$ are obviously homogeneous, it suffices to show that $\|\mathbf{x}\|_{\mathcal{A}} = 1 \Leftrightarrow \Gamma_{\mathbf{w}}(\mathbf{x}) = 1$. Since $\|\mathbf{x}\|_{\mathcal{A}}$ and $\Gamma_{\mathbf{w}}(\mathbf{x})$ are (in addition to homogeneous) also invariant w.r.t. permutations of the components of its argument, we can assume without loss of generality that the components of $\mathbf{x}$ satisfy $x_1 \geq x_2 \geq \cdots x_n \geq 0$. If $\|\mathbf{x}\|_{\mathcal{A}} = 1$, then $\mathbf{x} \in \text{conv}(\mathcal{A})$, *i.e.*, there exist $\theta_1, \theta_2, \cdots, \theta_n \in [0, 1]$ and $\sum_{i=1}^{n} \theta_i = 1$, such that $\mathbf{x} = \sum_{i=1}^{n} \theta_i \mathbf{b}_i$, where

$$\mathbf{b}_i = [\underbrace{\tau_i \ \cdots \tau_i}_{i} \ \underbrace{0 \ \cdots \ 0}_{n-i}]^T \in \check{\mathcal{B}}_i,$$

that is

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

Consequently, the components of this $\mathbf{x}$ 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 the $\Gamma_{\mathbf{w}}$ norm of this $\mathbf{x}$ yields

$$\Gamma_{\mathbf{w}}(\mathbf{x}) = \sum_{k=1}^{n} w_k x_k = \sum_{k=1}^{n} w_k \left[ \sum_{i=k}^{n} \theta_i \left( \sum_{j=1}^{i} w_j \right)^{-1} \right]$$
$$= w_1 \left( \frac{\theta_1}{w_1} + \frac{\theta_2}{w_1 + w_2} + \cdots + \frac{\theta_n}{w_1 + w_2 + \cdots + w_n} \right)$$
$$+ w_2 \left( \frac{\theta_2}{w_1 + w_2} + \cdots + \frac{\theta_n}{w_1 + w_2 + \cdots + w_n} \right) + \cdots +$$
$$w_{n-1} \left( \frac{\theta_{n-1}}{w_1 + w_2 + \cdots + w_{n-1}} + \frac{\theta_n}{w_1 + w_2 + \cdots + w_n} \right)$$
$$+ w_n \left( \frac{\theta_n}{w_1 + w_2 + \cdots + w_n} \right)$$
$$= \frac{\theta_1 w_1}{w_1} + \frac{\theta_2 (w_1 + w_2)}{\dot{w}_1 + w_2} + \cdots + \frac{\theta_{n-1} (w_1 + w_2 + \cdots + w_{n-1})}{w_1 + w_2 + \cdots + w_{n-1}}$$
$$+ \frac{\theta_n (w_1 + w_2 + \cdots + w_n)}{w_1 + w_2 + \cdots + w_n} = \theta_1 + \theta_2 + \cdots + \theta_n = 1,$$

which shows that $\|\mathbf{x}\|_{\mathcal{A}} = 1 \Rightarrow \Gamma_{\mathbf{w}}(\mathbf{x}) = 1$. The reverse implication is also easy to show, and actually it is just the reverse process of above proof. ■

### 3.2. Dual Norm of $\Gamma_{\mathbf{w}}$

We will now show that the dual norm of $\Gamma_{\mathbf{w}}$, defined as

$$\Gamma_{\mathbf{w}}^{*}(\mathbf{x}) = \max_{\Gamma_{\mathbf{w}}(\mathbf{u}) \leq 1} \langle \mathbf{u}, \mathbf{x} \rangle \tag{14}$$

can be obtained via the atomic formulation, that is,

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

Let $\mathbf{x}_{(k)} \in \mathbb{R}^k$ be a sub-vector of $\mathbf{x} \in \mathbb{R}^n$, consisting of the $k$ largest (in magnitude) elements of $\mathbf{x}$ (naturally, $\left\|\mathbf{x}_{(1)}\right\|_1 = |x_1| = \|\mathbf{x}\|_\infty$ and $\left\|\mathbf{x}_{(n)}\right\|_1 = \|\mathbf{x}\|_1$). Then, we have

$$\begin{aligned} \max_{\mathbf{a} \in \mathcal{B}_1} \langle \mathbf{a}, \mathbf{x} \rangle &= \tau_1 \|\mathbf{x}_{(1)}\|_1 \ = \ \tau_1 \|\mathbf{x}\|_\infty \\ \max_{\mathbf{a} \in \mathcal{B}_2} \langle \mathbf{a}, \mathbf{x} \rangle &= \tau_2 \left\|\mathbf{x}_{(2)}\right\|_1 \\ &\vdots \\ \max_{\mathbf{a} \in \mathcal{B}_{n-1}} \langle \mathbf{a}, \mathbf{x} \rangle &= \tau_{n-1} \left\|\mathbf{x}_{(n-1)}\right\|_1 \\ \max_{\mathbf{a} \in \mathcal{B}_n} \langle \mathbf{a}, \mathbf{x} \rangle &= \tau_n \left\|\mathbf{x}_{(n)}\right\|_1 \ = \ \tau_n \|\mathbf{x}\|_1 \end{aligned} \tag{16}$$

Combining (14), (15), and (16), yields the following lemma:

**Lemma 4** *The dual norm of* $\Gamma_{\mathbf{w}}$ *is given by*

$$\Gamma_{\mathbf{w}}^{*}(\mathbf{x}) = \max \left\{ \tau_k \left\|\mathbf{x}_{(k)}\right\|_1, k = 1, \cdots, n \right\}. \tag{17}$$

### 3.3. Solving $\Gamma_{\mathbf{w}}(\mathbf{x})$ Constrained Problems by Conditional Gradient Method

We will now address in detail the problem of tackling (9) using the CGM (or Frank-Wolfe method [21], [22]), as already introduced above, exploiting the atomic formulation of $\Gamma_{\mathbf{w}}$.

The key step of the CGM is finding the conditional gradient $\mathbf{d}_i$ (see line 3 of the CGM presented above). Using the atomic norm formulation, denoting $\mathbf{g} = (-\nabla f(\mathbf{x}_i))$, and recalling that $\mathcal{D} = \{\mathbf{x} \in \mathbb{R}^n : \Gamma_{\mathbf{w}}(\mathbf{x}) \leq 1\}$), we have

$$\begin{aligned} \mathbf{d}_i &= \arg\max_{\mathbf{x} \in \mathcal{D}} \langle \mathbf{x}, \mathbf{g} \rangle \\ &= \arg \max_{\mathbf{x} \in \text{conv}(\mathcal{A})} \langle \mathbf{x}, \mathbf{g} \rangle \\ &= \arg\max_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{a}, \mathbf{g} \rangle. \end{aligned} \tag{18}$$

The final maximization problem in (18) can be solved by the following three steps:

$$\begin{aligned} \mathbf{s} &= \text{sign}(\mathbf{g}) \\ k^* &= \arg \max_{k \in \{1, \cdots, n\}} \left\{ \tau_k \left\|\mathbf{g}_{(k)}\right\|_1 \right\} \\ \mathbf{d}_i &= \mathbf{s} \odot \arg \max_{\mathbf{a} \in \check{\mathcal{B}}_{k^*}} \langle \mathbf{a}, |\mathbf{g}| \rangle, \end{aligned} \tag{19}$$

where $|\mathbf{g}|$ is the vector with the magnitudes of the components of $\mathbf{g}$. The cost of implementing the conditional gradient step is dominated by the $O(n \log n)$ cost of sorting the elements of $|\mathbf{g}|$ (once per iteration) to obtain the several $\mathbf{g}_{(k)}$.

## 4. CONCLUSIONS

We have proposed an atomic norm formulation of *octagonal shrinkage and clustering algorithm for regression* (OSCAR) for feature selection and grouping. We showed that the OSCAR regularizer can also be reformulated as a *decreasing weighted sorted* $\ell_1$ (DWSL1) norm and as an atomic norm. Using the atomic norm formulation, we showed how to tackle the the Ivanov regularization scheme with the OSCAR regularizer via the conditional gradient method.

## 5. 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 proximal splitting algorithms," *arXiv preprint arXiv:1309.6301*, 2013.

[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] 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.

[14] T. Goldstein and S. Osher, "The split bregman method for l1-regularized problems," *SIAM Journal on Imaging Sciences*, vol. 2, pp. 323–343, 2009.

[15] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, pp. 120–145, 2011.

[16] X. Zeng and M. A. T. Figueiredo, "Sparsity and clustering regularization for regression," in *Workshop on Signal processing with Adaptive Sparse Structured Representations*, 2013.

[17] X. Zeng and M. A. T. Figueiredo, "A novel sparsity and clustering regularization," in *19th Portuguese Conference on Pattern Recognition*, 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] D. Lorenz and N. Worliczek, "Necessary conditions for variational regularization schemes," *Inverse Problems*, vol. 29, 2013.

[20] M. Bogdana, E. Bergb, W. Suc, and E. J. Candès, "Statistical estimation and testing via the sorted l1 norm," *arXiv preprint arXiv:1310.1969*, 2013.

[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] 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.