Ordered Weighted ℓ_1 Regularized Regression with Strongly Correlated Covariates: Theoretical Aspects

Mário A. T. Figueiredo
Instituto de Telecomunicações
Instituto Superior Técnico
Universidade de Lisboa, Portugal

Robert D. Nowak
Depart. of Electrical and Computer Engineering
University of Wisconsin, Madison, USA

Abstract

This paper studies the *ordered weighted ℓ_1* (OWL) family of regularizers for sparse linear regression with strongly correlated covariates. We prove sufficient conditions for clustering correlated covariates, extending and qualitatively strengthening previous results for a particular member of the OWL family: OSCAR (*octagonal shrinkage and clustering algorithm for regression*). We derive error bounds for OWL with correlated Gaussian covariates: for cases in which clusters of covariates are strongly (even perfectly) correlated, but covariates in different clusters are uncorrelated, we show that if the true p-dimensional signal involves only s clusters, then $O(s \log p)$ samples suffice to accurately estimate it, regardless of the number of coefficients within the clusters. Since the estimation of s-sparse signals with completely independent covariates also requires $O(s \log p)$ measurements, this shows that by using OWL regularization, we pay no price (in the number of measurements) for the presence of strongly correlated covariates.

1 Introduction

In high-dimensional linear regression problems, it is likely that several covariates (also referred to as predictors or variables) are highly correlated; *e.g.*, in gene expression data, it is common to find groups of highly co-regulated genes. Using standard sparsity-inducing regularization (ℓ_1, also known as *LASSO* (Tibshirani, 1996)) in such scenarios is known to be unsatisfactory, as it leads to the selection of one of a set of highly correlated covariates, or an arbitrary convex combination thereof. For engineering purposes and/or scientific interpretability, it is often desirable to explicitly identify **all** of the covariates that are relevant for modelling the data, rather than just a subset thereof. Several approaches have been proposed to deal with this type of problems (Bühlmann et al., 2013; Genovese et al., 2012; Jia and Yu, 2010; Meinshausen and Bülhmann, 2010; Shah and Samworth, 2013; Shen and Huang, 2010; Zou and Hastie, 2005), the best known of which is arguably the *elastic net* (EN) introduced by Zou and Hastie (2005).

This paper is motivated by the OSCAR (*octagonal shrinkage and clustering algorithm for regression*), proposed by Bondell and Reich (2007) to address regression problems with correlated covariates, which has been shown to perform well in practice, but lacks a theoretical characterization of its error performance. The regularizer underlying OSCAR was recently shown to belong to the more general family of the *ordered weighted ℓ_1* (OWL) norms (Bogdan et al., 2013; Zeng and Figueiredo, 2014a), which also includes the ℓ_1 and ℓ_∞ norms. The goal of this paper is to provide a theoretical characterization of linear regression under OWL regularization, in the presence of highly correlated covariates. Our main contributions are the following.

a) We prove sufficient conditions for exact covariate clustering, considerably extending the results by Bondell and Reich (2007). In particular, for the squared error loss, our result holds for the general OWL family and, more importantly, under qualitatively weaker conditions: whereas our result shows that OWL can cluster groups of more than 2 covariates, the proof by Bondell and Reich (2007) explicitly excludes that case. Furthermore, we also give clustering conditions under the absolute error loss, which we believe are novel.

b) We derive error bounds for OWL regularization with correlated Gaussian covariates. For cases in which clusters of covariates are strongly (even perfectly) correlated, but covariates in different clusters are uncorrelated, we show that if the true p-dimensional signal involves only s clusters, then $O(s \log p)$ samples suffice to accurately estimate it, regardless of the number of coefficients within

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41. Copyright 2016 by the authors.

clusters. Since estimating s-sparse vectors with independent variables requires just as many measurements, this shows that by using OWL regularization no price is paid (in terms of the number of measurements) for the presence of those strongly correlated covariates.

This paper includes no experimental results, as its goal is to theoretically charaterize OWL regularization. The particular case of OSCAR was experimentally studied by Bondell and Reich (2007) and Zhong and Kwok (2012); the main conclusion from their experiments is not that OSCAR clearly outperforms EN in terms of accuracy, but that while typically requiring fewer degrees of freedom due to its exact clustering behavior, it is still competitive with EN. In other words, their claim is not that OSCAR achieves higher accuracy, but that its ability to identify clusters of correlated covariates improves interpretability. In very recent work on machine translation, Clark (2015) found that the ability of OSCAR to cluster coefficients brings a significant performance gain. This paper provides theoretical support to these experimental observations.

Notation

We denote (column) vectors by lower-case bold letters, *e.g.*, $\boldsymbol{x}$, $\boldsymbol{y}$, their transposes by $\boldsymbol{x}^T$, $\boldsymbol{y}^T$, the corresponding i-th and j-th components as x_i and y_j, and matrices by upper case bold letters, *e.g.*, $\boldsymbol{A}$, $\boldsymbol{B}$. A vector with all components equal to 1 is written as $\mathbf{1}$, and $|\boldsymbol{x}|$ is the vector with the absolute values of the components of $\boldsymbol{x}$. For $\boldsymbol{x} \in \mathbb{R}^p$, $x_{[i]}$ is its i-th largest component (*i.e.*, $x_{[1]} \geq x_{[2]} \geq \cdots \geq x_{[p]}$), and $\boldsymbol{x}_\downarrow$ is the vector obtained by sorting the components of $\boldsymbol{x}$ in non-increasing order. Finally, given $\boldsymbol{w} \in \mathbb{R}^p_+$, such that $w_1 \geq w_2 \geq ... \geq w_p \geq 0$, $\Delta_{\boldsymbol{w}} = \min\{w_l - w_{l+1},\ l = 1, ..., p-1\}$ is the minimum gap between consecutive components of $\boldsymbol{w}$ and $\bar{w} = \|\boldsymbol{w}\|_1/p$ is their average.

1.1 Definitions and Problem Formulation

The OWL norm (Bogdan et al., 2014; Zeng and Figueiredo, 2014a), is defined as

$$\Omega_{\boldsymbol{w}}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i\, |x|_{[i]} = \boldsymbol{w}^T |\boldsymbol{x}|_\downarrow, \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^p_+$ is a vector of weights, such that $w_1 \geq w_2 \geq \cdots \geq w_p \geq 0$ and $w_1 > 0$. Clearly, $\Omega_{\boldsymbol{w}}$ satisfies $w_1\|\boldsymbol{x}\|_\infty \leq \Omega_{\boldsymbol{w}}(\boldsymbol{x}) \leq w_1\|\boldsymbol{x}\|_1$ (with equalities if $w_2 = \cdots = w_p = 0$ or $w_1 = w_2 = \cdots = w_p$, respectively). It is also easy to show (using Chebyshev's sum inequality) that $\Omega_{\boldsymbol{w}}(\boldsymbol{x}) \geq \bar{w}\,\|\boldsymbol{x}\|_1$. The OSCAR regularizer (Bondell and Reich, 2007) is a special case of $\Omega_{\boldsymbol{w}}$, obtained by setting $w_i = \lambda_1 + \lambda_2\,(p-i)$, where $\lambda_1,\ \lambda_2 \geq 0$.

This paper studies OWL-regularized linear regression under the squared error and absolute error losses. We consider the classical unconstrained formulations

$$\min_{\boldsymbol{x}\in\mathbb{R}^p}\ \frac{1}{2}\|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda\,\Omega_{\boldsymbol{w}}(\boldsymbol{x}), \tag{2}$$

$$\min_{\boldsymbol{x}\in\mathbb{R}^p}\ \|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{y}\|_1 + \lambda\,\Omega_{\boldsymbol{w}}(\boldsymbol{x}), \tag{3}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ is the design matrix, as well as the following constrained formulations:

$$\min_{\boldsymbol{x}\in\mathbb{R}^p} \Omega_{\boldsymbol{w}}(\boldsymbol{x}) \ \text{ s.t. }\ \frac{1}{n}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 \ \leq\ \varepsilon^2, \tag{4}$$

$$\min_{\boldsymbol{x}\in\mathbb{R}^p} \Omega_{\boldsymbol{w}}(\boldsymbol{x}) \ \text{ s.t. }\ \frac{1}{n}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_1 \ \leq\ \varepsilon. \tag{5}$$

Notice that (since all the involved functions are convex) the constrained and unconstrained formulations are equivalent in the following sense (Lorenz and Worliczek, 2013): given a (non-zero) solution $\widehat{\boldsymbol{x}}$ of (4) (respectively, of (5)), there is a choice of λ that makes $\widehat{\boldsymbol{x}}$ also a solution of (2) (respectively, of (3)). Conversely, if $\widehat{\boldsymbol{x}}$ is the unique solution of (2) (respectively, of (3)), then $\widehat{\boldsymbol{x}}$ also solves (4) (respectively, (5)), with $\varepsilon^2 = \frac{1}{n}\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{y}\|_2^2$ (respectively, with $\varepsilon = \frac{1}{n}\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{y}\|_1$). Regardless of these equivalences, certain angles of analysis are more convenient in the unconstrained formulations (2)–(3), while others are more convenient in the constrained form (4)–(5). The equivalences mentioned above mean that results concerning the solutions of (2) and (3) are, in principle, translatable to results about the solutions of (4) and (5), and *vice-versa*.

On the algorithmic side, the key tool for solving regularization problems involving the OWL norm (such as (2) – (5)) is its Moreau proximity operator, which can be computed in $O(p\log p)$ operations (Bogdan et al., 2014; Zeng and Figueiredo, 2014b), the same being true about the Euclidean projection onto an OWL norm ball (Davis, 2015).

1.2 Preview of the Main Results and Related Work

The first of our two main results (detailed in Section 2) gives sufficient conditions for OWL regularization to cluster strongly correlated covariates, in the sense that the coefficient estimates associated with such covariates are exactly equal (in magnitude). Our result for the squared error loss significantly extends and strengthens the main theorem for OSCAR presented by Bondell and Reich (2007), since our proof involves qualitatively weaker conditions and applies to the general OWL family. Furthermore, the result for the absolute error loss is, as far as we know, novel.

Our second main result (presented in Section 3) is a finite sample bound for formulations (4)–(5). To the best of our knowledge, these are the first finite sample error bounds for sparse regression with strongly correlated columns in the design matrix. To preview this result, consider the following special case (generalized below): assume we observe

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star\ +\ \boldsymbol{\nu}\,, \tag{6}$$

where $\boldsymbol{x}^\star \in \mathbb{R}^p$ is s-sparse (i.e., at most s nonzero components) and $\boldsymbol{\nu} \in \mathbb{R}^n$ is the measurement error, with

$\|\boldsymbol{\nu}\|_1/n \leq \varepsilon$, and about which we make no other assumptions. The design matrix $\boldsymbol{A}$ is Gaussian distributed; for the purposes of this introduction, assume the entries in each column of $\boldsymbol{A}$ are i.i.d. $\mathcal{N}(0,1)$, but different columns may be strongly correlated. Specifically, assume there are groups of identical columns, but those in different groups are uncorrelated; this models scenarios where groups of covariates are perfectly correlated, but uncorrelated with all others. Since certain columns of $\boldsymbol{A}$ are identical, in general there may be many sparse vectors $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x}^\star$. Among these, let $\bar{\boldsymbol{x}}^\star$ denote the vector with identical coefficients for replicated columns (i.e., if two columns of $\boldsymbol{A}$ are identical, so are the corresponding coefficients in $\bar{\boldsymbol{x}}^\star$).

Theorem 1.1 below states that the sufficient number of measurements n to estimate an s-sparse signal, at a given precision, grows like $n \sim s \log p$, agreeing with well-known bounds for sparse recovery under stronger conditions on $\boldsymbol{A}$, *e.g.*, restricted isometry property, incoherence, or fully i.i.d. measurements (Candès et al., 2006; Donoho, 2006; Haupt and Nowak, 2006; Vershynin, 2014). This also shows that, by using OWL, we pay no price (in terms of number of measurements) for the colinearity of some columns of $\boldsymbol{A}$.

Theorem 1.1. *Let $\boldsymbol{y}$, $\boldsymbol{A}$, $\boldsymbol{x}^\star, \boldsymbol{w}$, and $\Delta_{\boldsymbol{w}}$ be as defined above and assume $\Delta_{\boldsymbol{w}} > 0$. Let $\widehat{\boldsymbol{x}}$ be a solution to either of the two problems (4) or (5). Then,*

(i) for every pair of columns such that $\boldsymbol{a}_i = \boldsymbol{a}_j$ (respectively, $\boldsymbol{a}_i = -\boldsymbol{a}_j$), we have $\widehat{x}_i = \widehat{x}_j$ (respectively, $\widehat{x}_i = -\widehat{x}_j$);

(ii) the solution $\widehat{\boldsymbol{x}}$ satisfies (where the expectation is w.r.t. the random $\boldsymbol{A}$):

$$\mathbb{E}\,\|\widehat{\boldsymbol{x}} - \bar{\boldsymbol{x}}^\star\|_2 \leq \sqrt{8\pi}\left(\sqrt{32}\,\|\boldsymbol{x}^\star\|_2\,\frac{w_1}{\bar{w}}\sqrt{\frac{s\log p}{n}} + \varepsilon\right). \tag{7}$$

Theorem 1.1 (i) shows that OWL regularization automatically identifies and groups the colinear columns in $\boldsymbol{A}$. As mentioned above, in general there may be many sparse $\boldsymbol{x}$ yielding the same $\boldsymbol{A}\boldsymbol{x}$. This is where OWL becomes important: its solution includes all the colinear columns associated with the model, rather than an arbitrary subset thereof. This result is proved in Section 2, and generalized to cases where columns are not necessarily identical, but correlated enough.

Part (ii) of Theorem 1.1 is proved in Section 3, and also generalized to strongly correlated (rather than identical) covariates. (Notice that the factor $w_1/\bar{w}$ in (7) is typically small; *e.g.*, for OSCAR, $\bar{w} = \lambda_1 + \lambda_2(p-1)/2$ and $w_1/\bar{w} \leq 2$, whereas for ℓ_1, $w_1/\bar{w} = 1$.)

It is worth mentioning that these bounds for OWL are fundamentally different than those obtained for the LASSO with correlated Gaussian designs (Raskutti et al., 2010), which do not cover the case of exactly replicated columns, and essentially require a full-rank design matrix.

Although OWL does bear similarity to the *elastic net* (EN), in the sense that they both aim at handling highly correlated covariates, OWL yields exact covariate clustering, whereas EN does not. In terms of theoretical analysis, the consistency results for the EN proved by Jia and Yu (2010) are asymptotic, not finite sample bounds, and require the so-called *elastic irrepresentability condition* (EIC), which is stronger than our assumptions. We are not aware of finite sample error bounds for EN that come close to those for OWL that we prove in this paper. Finally, although our bounds are for Gaussian designs, generalization to the sub-Gaussian case can be obtained using the tools proposed by Vershynin (2014). We thus feel that our assumptions are less restrictive and more relevant than the EIC.

It is also interesting to observe that the error bound in (7) is essentially the same holding for group-LASSO (Rao et al., 2012), assuming the groups are known a priori rather than automatically identified.

Finally, a particular member of the OWL family (for a specific choice of the weights $\boldsymbol{w}$) was recently studied, namely in in terms of *false discovery rate* (FDR) control, adaptivity, and asymptotic minimaxity (Bogdan et al., 2013, 2014; Su and Candès, 2015); however, these results are only for orthogonal or uncorrelated covariates, which is not the scenario to which this paper is devoted.

We conclude this section with a simple toy example (Fig. 1) illustrating the qualitatively different behaviour of OWL and LASSO regularization. In this example, $p = 100$, $n = 10$, and $\boldsymbol{x}^\star$ has 20 non-zero components in 2 groups of size 10, with the corresponding columns of $\boldsymbol{A}$ being highly correlated. Clearly, $n = 10$ is insufficient to allow LASSO to recover $\boldsymbol{x}^\star$, which is 20-sparse, while OWL successfully recovers its structure.

2 OWL Clustering

This section studies the clustering behaviour of OWL, extending the results of Bondell and Reich (2007) in several ways: for the squared error loss, our results apply to the more general case of OWL and, more importantly, hold under qualitatively weaker conditions; the result for the absolute error case is novel.

2.1 Squared Error Loss

The following theorem shows that criterion (2) *clusters* (*i.e.*, yields coefficient estimates of equal magnitude) the columns that are correlated enough.

Theorem 2.1. *Let $\widehat{\boldsymbol{x}}$ be a solution of (2), and $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ be*

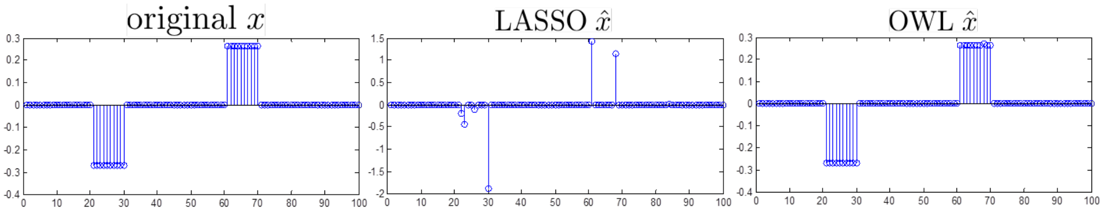

Figure 1: Toy example illustrating the qualitatively different behaviour of OWL and LASSO.

two columns of $\boldsymbol{A}$*. Then,*

$$\textbf{(a)} \quad \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 < \Delta_{\boldsymbol{w}}/\|\boldsymbol{y}\|_2 \quad \Rightarrow \quad \widehat{x}_i = \widehat{x}_j \qquad (8)$$

$$\textbf{(b)} \quad \|\boldsymbol{a}_i + \boldsymbol{a}_j\|_2 < \Delta_{\boldsymbol{w}}/\|\boldsymbol{y}\|_2 \quad \Rightarrow \quad \widehat{x}_i = -\widehat{x}_j. \qquad (9)$$

Clearly, part (b) of the theorem is a simple corollary of part (a), which results from flipping the signs of either $\boldsymbol{a}_i$ or $\boldsymbol{a}_j$ and the corresponding coefficient. Notice that if two columns are identical, $\boldsymbol{a}_i = \boldsymbol{a}_j$, or symmetrical, $\boldsymbol{a}_i = -\boldsymbol{a}_j$, any $\Delta_{\boldsymbol{w}} > 0$ is sufficient to guarantee that these two columns are clustered, *i.e.*, the corresponding coefficient estimates are equal in magnitude.

The following corollary addresses the case where the columns of $\boldsymbol{A}$ have zero mean and unit norm (as is common practice), and results from denoting $\rho_{ij} = \boldsymbol{a}_i^T\boldsymbol{a}_j$ as the sample correlation between the i-th and j-th covariates, in which case $\|\boldsymbol{a}_i \pm \boldsymbol{b}_j\|_2 = \sqrt{2 \pm 2\rho_{ij}}$.

Corollary 2.1. *Let the columns of* $\boldsymbol{A}$ *satisfy* $\mathbf{1}^T\boldsymbol{a}_k = 0$ *and* $\|\boldsymbol{a}_k\|_2 = 1$*, for* $k = 1, ..., p$*. Denote* $\rho_{ij} = \boldsymbol{a}_i^T\boldsymbol{a}_j \in [-1, 1]$*. Then,*

$$\textbf{(a)} \quad \sqrt{2 - 2\,\rho_{ij}} < \Delta_{\boldsymbol{w}}/\|\boldsymbol{y}\|_2 \quad \Rightarrow \quad \widehat{x}_i = \widehat{x}_j \qquad (10)$$

$$\textbf{(b)} \quad \sqrt{2 + 2\,\rho_{ij}} < \Delta_{\boldsymbol{w}}/\|\boldsymbol{y}\|_2 \quad \Rightarrow \quad \widehat{x}_i = -\widehat{x}_j. \qquad (11)$$

Corollary 2.1 recovers the main theorem of Bondell and Reich (2007), since in the case of OSCAR, $\Delta_{\boldsymbol{w}} = \lambda_2$. However, our result holds under qualitatively weaker conditions: unlike in their proof, we do not require that both $\widehat{x}_i$ and $\widehat{x}_j$ are different from zero and from all other $\widehat{x}_k$, for $k \neq i, j$, neither that $\boldsymbol{A}$ is such that $\widehat{x}_k \geq 0$, for all k. Furthermore, our results apply to the general family of OWL regularizers, of which OSCAR is only a particular case.

2.2 Absolute Error Loss

The following theorem parallels Theorem 2.1, now for absolute error loss regression (eq. (3)).

Theorem 2.2. *Let* $\widehat{\boldsymbol{x}}$ *be a solution of* (3)*. Then*

$$\textbf{(a)} \quad \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_1 < \Delta_{\boldsymbol{w}} \quad \Rightarrow \quad \widehat{x}_i = \widehat{x}_j \qquad (12)$$

$$\textbf{(b)} \quad \|\boldsymbol{a}_i + \boldsymbol{a}_j\|_1 < \Delta_{\boldsymbol{w}} \quad \Rightarrow \quad \widehat{x}_i = -\widehat{x}_j \qquad (13)$$

As above, part (b) of the theorem results directly from part (a). Under the normalization assumptions used in Corollary 2.1, another (weaker) sufficient condition can be obtained, which depends on the sample correlations, as stated in the following corollary (the proof of which simply amounts to using the well-known inequality $\|\boldsymbol{a}\|_1 \leq \sqrt{n}\|\boldsymbol{a}\|_2$ together with the assumed column normalization):

Corollary 2.2. *Let* $\widehat{\boldsymbol{x}}$ *be any minimizer of the objective function in* (3) *and assume the columns of* $\boldsymbol{A}$ *are normalized, that is,* $\mathbf{1}^T\boldsymbol{a}_k = 0$ *and* $\|\boldsymbol{a}_k\|_2 = 1$*, for* $i = k, ..., p$*. As above, let* $\rho_{ij} = \boldsymbol{a}_i^T\boldsymbol{a}_j$*. Then,*

$$\textbf{(a)} \quad \sqrt{n(2 - 2\,\rho_{ij})} < \Delta_{\boldsymbol{w}} \quad \Rightarrow \quad \widehat{x}_i = \widehat{x}_j \qquad (14)$$

$$\textbf{(b)} \quad \sqrt{n(2 + 2\,\rho_{ij})} < \Delta_{\boldsymbol{w}} \quad \Rightarrow \quad \widehat{x}_i = -\widehat{x}_j. \qquad (15)$$

2.3 Proofs of Theorems 2.1 and 2.2

The proofs of Theorems 2.1 and 2.2 are based on the following two new lemmas about the OWL norm (the proofs of which are provided in the supplementary material).

Lemma 2.1. *Consider a vector* $\boldsymbol{x} \in \mathbb{R}_+^p$ *and two of its components* x_i *and* x_j*, such that* $x_i > x_j$ *(if they exist). Let* $\boldsymbol{z} \in \mathbb{R}_+^p$ *be obtained by applying to* $\boldsymbol{x}$ *a so-called Pigou-Dalton*[1] *transfer of size* $\varepsilon \in \big(0, (x_i - x_j)/2\big)$*, that is:* $z_i = x_i - \varepsilon$*,* $z_j = x_j + \varepsilon$*, and* $z_k = x_k$*, for* $k \neq i, j$*. Then,*

$$\Omega_{\boldsymbol{w}}(\boldsymbol{x}) - \Omega_{\boldsymbol{w}}(\boldsymbol{z}) \geq \Delta_{\boldsymbol{w}}\,\varepsilon. \qquad (16)$$

Lemma 2.2. *Consider a vector* $\boldsymbol{x} \in \mathbb{R}_+^p$ *and two of its non-zero components* x_i *and* x_j *(if they exist). Let now* $\boldsymbol{z} \in \mathbb{R}_+^p$ *be obtained by subtracting* $\varepsilon \in \big(0, \min\{x_i, x_j\}\big)$ *from* x_i *and* x_j*, that is:* $z_i = x_i - \varepsilon$*,* $z_j = x_j - \varepsilon$*, and* $z_k = x_k$*, for* $k \neq i, j$*. Then,* (16) *also holds.*

It is worth pointing out that what Lemma 2.1 states about the OWL norm $\Omega_{\boldsymbol{w}}$ can be seen as a property of *strong Schur convexity*, which, as far as we know, didn't exist in the literature on majorization theory and Schur convexity (Marshall et al., 2011). For more details about this observation, which is tangential to the topic of this paper, but potentially useful in other contexts, see Appendix A.

The proof of Theorem 2.1 also uses a basic result in convex analysis relating minimizers of a convex function with its

[1] The Pigou-Dalton (a.k.a. Robin Hood) transfer is a fundamental quantity used in the mathematical study of economic inequality (Dalton, 1920; Pigou, 1912).

directional derivatives (Rockafellar, 1970). Given a proper function f, its *directional derivative* at $\boldsymbol{x} \in \mathrm{dom}(f)$ (*i.e.*, $f(\boldsymbol{x}) \neq \infty$), in the direction $\boldsymbol{u}$, is defined as

$$f'(\boldsymbol{x};\boldsymbol{u}) = \lim_{\alpha\rightarrow 0^+} \big(f(\boldsymbol{x}+\alpha\boldsymbol{u}) - f(\boldsymbol{x})\big)/\alpha.$$

Lemma 2.3. *Let f be a real-valued, proper, convex function, and $\boldsymbol{x} \in dom(f)$. Then, $\boldsymbol{x} \in \arg\min f$, if and only if $f'(\boldsymbol{x};\boldsymbol{u}) \geq 0$, for any $\boldsymbol{u}$.*

Proof. (of Theorem 2.1) Let $L_2(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{y}\|_2^2$ and $f(\boldsymbol{x}) = L_2(\boldsymbol{x}) + \Omega_{\boldsymbol{w}}(\boldsymbol{x})$ (*i.e.*, the objective function in (2)). Assume that the condition $\|\boldsymbol{y}\|_2\,\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 < \Delta_{\boldsymbol{w}}$ is satisfied for some pair of columns and consider some $\widehat{\boldsymbol{x}}$ such that $\widehat{x}_i \neq \widehat{x}_j$ (w.l.o.g., let $\widehat{x}_i > \widehat{x}_j$). The directional derivative of L_2 at $\widehat{\boldsymbol{x}}$, in the direction $\boldsymbol{u}$, where $u_i = -1$, $u_j = 1$, and $u_k = 0$, for $k \neq i,j$, is

$$\begin{aligned} & L_2'(\widehat{\boldsymbol{x}};\boldsymbol{u}) \\ & \quad = \lim_{\alpha\rightarrow 0^+} \frac{\|\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}} + \alpha(\boldsymbol{a}_i - \boldsymbol{a}_j)\|_2^2 - \|\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}}\|_2^2}{2\,\alpha} \\ & \quad = \boldsymbol{g}^T(\boldsymbol{a}_i - \boldsymbol{a}_j), \end{aligned} \tag{17}$$

where $\boldsymbol{g} = \boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}}$. Consider now the directional derivative of $\Omega_{\boldsymbol{w}}$ at $\widehat{\boldsymbol{x}}$, in the same direction $\boldsymbol{u}$:

$$\Omega_{\boldsymbol{w}}'(\widehat{\boldsymbol{x}};\boldsymbol{u}) = \lim_{\alpha\rightarrow 0^+} \frac{\Omega_{\boldsymbol{w}}(\widehat{\boldsymbol{x}} + \alpha\boldsymbol{u}) - \Omega_{\boldsymbol{w}}(\widehat{\boldsymbol{x}})}{\alpha}.$$

If $\widehat{x}_i$ and $\widehat{x}_j$ are both non-negative or non-positive, $|\widehat{\boldsymbol{x}}+\alpha\boldsymbol{u}|$ corresponds to a Pigou-Dalton transfer of size α applied to $|\widehat{\boldsymbol{x}}|$, thus Lemma 2.1 (recalling $\Omega_{\boldsymbol{w}}(\boldsymbol{v}) = \Omega_{\boldsymbol{w}}(|\boldsymbol{v}|)$) guarantees that

$$\begin{aligned} \Omega_{\boldsymbol{w}}'(\widehat{\boldsymbol{x}};\boldsymbol{u}) & = \lim_{\alpha\rightarrow 0^+} \frac{\Omega_{\boldsymbol{w}}(|\widehat{\boldsymbol{x}} + \alpha\boldsymbol{u}|) - \Omega_{\boldsymbol{w}}(|\widehat{\boldsymbol{x}}|)}{\alpha} && (18) \\ & \leq \lim_{\alpha\rightarrow 0^+} \frac{-\Delta_{\boldsymbol{w}}\,\alpha}{\alpha} \; = \; -\Delta_{\boldsymbol{w}}. && (19) \end{aligned}$$

If $\mathrm{sign}(\widehat{x}_i)\,\mathrm{sign}(\widehat{x}_j) = -1$, then $|\widehat{\boldsymbol{x}} + \alpha\boldsymbol{u}|$ corresponds to subtracting α from both $|\widehat{x}_i|$ and $|\widehat{x}_j|$, thus Lemma 2.2 also yields (19). Finally, adding (17) and (19), and using the Cauchy-Schwarz inequality,

$$\begin{aligned} f'(\boldsymbol{x};\boldsymbol{u}) & \leq \boldsymbol{g}^T(\boldsymbol{a}_i - \boldsymbol{a}_j) - \Delta_{\boldsymbol{w}} \\ & \leq \|\boldsymbol{g}\|_2\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 - \Delta_{\boldsymbol{w}} \\ & \leq \|\boldsymbol{y}\|_2\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 - \Delta_{\boldsymbol{w}} \; < \; 0, \end{aligned}$$

(after noticing $\|\boldsymbol{g}\|_2 \leq \|\boldsymbol{y}\|_2$), showing that $\widehat{\boldsymbol{x}}$ is not a minimizer. □

The following lemma (proved in the supplementary material) will be used in proving Theorem 2.2.

Lemma 2.4. *Let $L_1(\boldsymbol{x}) = \|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{y}\|_1$, consider any $\boldsymbol{x}$ and two of its components, x_i and x_j, and define $\boldsymbol{v}$ according to $v_i = x_i - \varepsilon$, $v_j = x_j + \varepsilon$, for some $\varepsilon \in \mathbb{R}$, and $v_k = x_k$, for $k \neq i,j$. Then, $L_1(\boldsymbol{v}) - L_1(\boldsymbol{x}) \leq |\varepsilon|\;\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_1$.*

Proof. (of Theorem 2.2) Assume the condition $\Delta_{\boldsymbol{w}} > \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_1$ is satisfied, and that $\widehat{\boldsymbol{x}}$ is a solution of (3). To prove that $\mathrm{sign}(\widehat{x}_i) = \mathrm{sign}(\widehat{x}_j)$, suppose that $\mathrm{sign}(\widehat{x}_i) \neq \mathrm{sign}(\widehat{x}_j)$, which implies that at least one of $\widehat{x}_i$ or $\widehat{x}_j$ is non-zero; without loss of generality, let $\widehat{x}_i > 0$, thus $\widehat{x}_j \leq 0$. We need to consider two cases: $\widehat{x}_j < 0$ and $\widehat{x}_j = 0$.

- If $\widehat{x}_j < 0$, take an alternative solution $\boldsymbol{v}$, with $v_i = \widehat{x}_i - \varepsilon$, $v_j = \widehat{x}_j + \varepsilon$, where $\varepsilon \in (0, \min\{\widehat{x}_i, -\widehat{x}_j\})$, and $v_k = \widehat{x}_k$, for $k \neq i,j$. Since $\widehat{x}_i > 0$ and $\widehat{x}_j < 0$, we have $|v_i| = |\widehat{x}_i| - \varepsilon$ and $|v_j| = |\widehat{x}_j| - \varepsilon$, thus Lemma 2.2 yields
$$\Omega_{\boldsymbol{w}}(\widehat{\boldsymbol{x}}) - \Omega_{\boldsymbol{w}}(\boldsymbol{v}) = \Omega_{\boldsymbol{w}}(|\widehat{\boldsymbol{x}}|) - \Omega_{\boldsymbol{w}}(|\boldsymbol{v}|) \geq \Delta_{\boldsymbol{w}}\,\varepsilon. \tag{20}$$
Combining this inequality with Lemma 2.4 contradicts the optimality of $\widehat{\boldsymbol{x}}$, since
$$\begin{aligned} & L_1(\boldsymbol{v}) + \Omega_{\boldsymbol{w}}(\boldsymbol{v}) - \big(L_1(\widehat{\boldsymbol{x}}) + \Omega_{\boldsymbol{w}}(\widehat{\boldsymbol{x}})\big) \\ & \qquad \leq \big(\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_1 - \Delta_{\boldsymbol{w}}\big)\,\varepsilon \; < \; 0. \end{aligned} \tag{21}$$
- If $\widehat{x}_j = 0$, follow the same argument with $\varepsilon \in (0, x_i/2)$, *i.e.*, $v_i = \widehat{x}_i - \varepsilon$ and $v_j = \widehat{x}_j + \varepsilon = \varepsilon$. Since, in this case, $|v_i| = |\widehat{x}_i| - \varepsilon$ and $|v_j| = |\widehat{x}_j| + \varepsilon$, Lemma 2.1 also yields inequality (20), which combined with Lemma 2.4 contradicts again the optimality of $\widehat{\boldsymbol{x}}$.

Knowing $\mathrm{sign}(\widehat{x}_i) = \mathrm{sign}(\widehat{x}_j)$, let $\widehat{\boldsymbol{x}}$ be a solution of (3) such that $\widehat{x}_i \neq \widehat{x}_j$; without loss of generality, consider that both $\widehat{x}_i$ and $\widehat{x}_j$ are non-negative, and that $\widehat{x}_i > \widehat{x}_j$. Consider an alternative solution $\boldsymbol{u}$ such that $u_i = \widehat{x}_i - \varepsilon$, $u_j = \widehat{x}_j + \varepsilon$, for some $\varepsilon \in (0, (x_i - x_j)/2)$, and $u_k = \widehat{x}_k$, for $k \neq i,j$. Combining Lemmas 2.1 and 2.4, yields precisely the same inequality as in (21), contradicting the optimality of $\widehat{\boldsymbol{x}}$, thus concluding the proof. □

3 OWL Error Bounds

Consider the observation model in (6) and the other assumptions about $\boldsymbol{x}$ and $\boldsymbol{\nu}$ made in Section 1.1. Moreover, assume that $\boldsymbol{x}^\star \in \mathbb{R}^p$ satisfies $\|\boldsymbol{x}^\star\|_1 \leq \sqrt{s}\,\|\boldsymbol{x}^\star\|_2$; this is true, *e.g.*, if $\boldsymbol{x}^\star$ is s-sparse. At the heart of our analysis is the following model for correlated measurement matrices. Assume that the rows of $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ are i.i.d. $\mathcal{N}(\boldsymbol{0}, \boldsymbol{C}^T\boldsymbol{C})$, *i.e.*, its columns are not necessarily independent. Let matrix $\boldsymbol{C}$ be $r \times p$, with $n \leq r \leq p$, so that $\mathrm{rank}(\boldsymbol{C}) \leq r$. Note that $\boldsymbol{A}$ can be written as $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$, where $\boldsymbol{B} \in \mathbb{R}^{n\times r}$ has i.i.d. $\mathcal{N}(0,1)$ entries. The role of $\boldsymbol{C}$ is to mix, or even replicate, columns of $\boldsymbol{B}$. Figure 2 illustrates this in a case where every column is one of three identical replicates.

3.1 General OWL Error Bound

The main result of this section is stated in the following theorem, the proof of which is based on the techniques introduced by Vershynin (2014). We also present a corollary for the particular case where $\boldsymbol{C}$ simply replicates columns of $\boldsymbol{B}$, *i.e.*, if $\boldsymbol{A}$ includes groups of identical columns; this

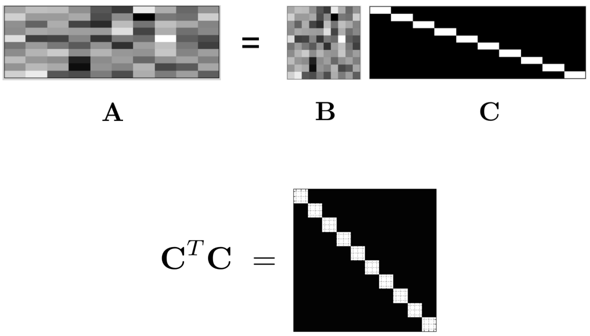

Figure 2: Matrix $\boldsymbol{A} = \boldsymbol{BC}$ with 10 groups of 3 identical replicates and the corresponding covariance matrix $\boldsymbol{C}^T\boldsymbol{C}$.

corollary is shown to imply part (ii) of Theorem 1.1. All expectations are with respect to the Gaussian distribution of the design matrix $\boldsymbol{A}$.

Theorem 3.1. *Let $\boldsymbol{y}$, $\boldsymbol{A}$, $\boldsymbol{x}^\star$, ε, and $\boldsymbol{w}$ be as defined above, and let $\widehat{\boldsymbol{x}}$ be a solution to one of the optimization problems (4) or (5). Then,*

$$\mathbb{E}\sqrt{(\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star)^T\boldsymbol{C}^T\boldsymbol{C}(\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star)} \quad (22)$$

$$\leq \quad \sqrt{8\pi}\left(4\sqrt{2}\min_{\ell=1,2}\|\boldsymbol{C}\|_\ell\,\|\boldsymbol{x}^\star\|_2\,\frac{w_1}{\bar{w}}\sqrt{\frac{s\log p}{n}}+\varepsilon\right),$$

where (recall) $\bar{w} = p^{-1}\|\boldsymbol{w}\|_1$ and $\min_{\ell=1,2}\|\boldsymbol{C}\|_\ell$ is the min of the matrix 1-norm and 2-norm of $\boldsymbol{C}$.

To help understand this bound, consider a special case recovering known results. If $r = p$ and $\boldsymbol{C} = \boldsymbol{I}$, matrix $\boldsymbol{A}$ is i.i.d. $\mathcal{N}(0,1)$, which corresponds to the well known compressive sensing case. The bound (22) recovers the usual type of result in this situation (Vershynin, 2014), *i.e.*,

$$\mathbb{E}\|\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star\|_2 = O\Big(\|\boldsymbol{x}^\star\|_2\sqrt{\frac{s\log p}{n}}\Big).$$

Notably, in this setting the OWL error bound is the same (up to small constant factors) as that of LASSO. Slightly more generally, if $\boldsymbol{C} \neq \boldsymbol{I}$, but has full rank, and $\lambda_{\max}$ and $\lambda_{\min} > 0$ are its largest and smallest singular values, then using the $\|\boldsymbol{C}\|_2$ factor in (22) yields bounds similar to those proved by Raskutti et al. (2010):

$$\mathbb{E}\|\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star\|_2 = O\Big(\frac{\lambda_{\max}}{\lambda_{\min}}\,\|\boldsymbol{x}^\star\|_2\sqrt{\frac{s\log p}{n}}\Big),$$

Since for $w_1 = w_2 = \cdots = w_p$, $\Omega_{\boldsymbol{w}}(\boldsymbol{x}) = w_1\|\boldsymbol{x}\|_1$, (22) also holds for the LASSO.

The bound in (22) is more novel and interesting if $\boldsymbol{C}$ is rank deficient. Consider $r < p$, with $\boldsymbol{C}$ leading to exactly replicated columns, as in Fig. 2. In this case, the covariance $\boldsymbol{C}^T\boldsymbol{C}$ has a block diagonal structure (Fig. 2), with each block, corresponding to a group of replicated columns, being equal to a rank-1 matrix with all entries equal. In this case, $(\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star)^T\boldsymbol{C}^T\boldsymbol{C}(\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star)$ is the sum of squared errors between the averages of $\widehat{\boldsymbol{x}}$ and $\boldsymbol{x}^\star$ within each group. This is very reasonable because both $\boldsymbol{A}\widehat{\boldsymbol{x}}$ and $\boldsymbol{A}\boldsymbol{x}^\star$ are functions of only these averages, since the columns corresponding to the groups are identical. Recall that Theorems 2.1 and 2.2 imply that $\widehat{\boldsymbol{x}}$ is constant-valued in each group of replicated columns. Also, in this case $\|\boldsymbol{C}\|_1 = 1$, whereas $\|\boldsymbol{C}\|_2$ is equal to the square-root of the largest group size. Theorem 1.1 follows directly from these observations.

Proof. (Theorem 1.1 (ii)) Recall that $\bar{\boldsymbol{x}}^\star$ satisfies $\boldsymbol{A}\bar{\boldsymbol{x}}^\star = \boldsymbol{A}\boldsymbol{x}^\star$ and that if two columns of $\boldsymbol{A}$ are identical, then so are the corresponding components of $\bar{\boldsymbol{x}}^\star$. Because of the group structure of $\bar{\boldsymbol{x}}^\star$ and $\widehat{\boldsymbol{x}}$ (assuming strictly decreasing weights), and the special form of $\boldsymbol{C}$ in the case of exactly replicated columns,

$$\|\widehat{\boldsymbol{x}}-\bar{\boldsymbol{x}}^\star\|_2^2 \;\leq\; (\widehat{\boldsymbol{x}}-\bar{\boldsymbol{x}}^\star)^T\boldsymbol{C}^T\boldsymbol{C}(\widehat{\boldsymbol{x}}-\bar{\boldsymbol{x}}^\star)\,,$$

from which (7) results. □

More generally, if $\boldsymbol{C}$ is *approximately* like that of Fig. 2 (each column is *approximately* 1-sparse), then the same reasoning and interpretation apply *approximately*. For example, if each column of $\boldsymbol{C}$ is sufficiently close to one of the canonical unit vectors, then Theorems 2.1–2.2 imply that $\widehat{\boldsymbol{x}}$ is constant on each group of (nearly) replicated columns, effectively averaging the corresponding columns in the prediction $\boldsymbol{A}\widehat{\boldsymbol{x}}$, helping to mitigate the effects of noise in these features and improving prediction.

The bound (22) in Theorem 3.1 holds for both (4) and (5). In fact, since $\frac{1}{n}\|\boldsymbol{Ax}-\boldsymbol{y}\|_2^2 \leq \varepsilon^2$ implies $\frac{1}{n}\|\boldsymbol{Ax}-\boldsymbol{y}\|_1 \leq \varepsilon$, the ℓ_1 constraint is less restrictive. In both cases, Theorem 3.1 shows that the number of samples sufficient to estimate an s-sparse signal with a given precision grows like $n \sim s\log p$; this agrees with well-known sample complexity bounds for sparse recovery under stronger assumptions, such as the restricted isometry property or i.i.d. measurements (Candès et al., 2006; Donoho, 2006; Haupt and Nowak, 2006; Vershynin, 2014). In the case of groups of (nearly) replicated columns, the number of samples grows linearly with the number of nonzero groups, rather than the total number of nonzero components in $\bar{\boldsymbol{x}}^\star$. This is where OWL regularization becomes important, by selecting all colinear columns associated with the model; i.e., if the columns are colinear (or correlated enough), the OWL solution selects a representation including all the columns associated with the sparse model, rather than a subset.

3.2 Proof of Theorem 3.1

The proof of Theorem 3.1 is based on the approach developed by Vershynin (2014). The key ingredient is the so-called *general M^* bound* (Vershynin, 2014, Theorem 5.1), which applies to the case where $\boldsymbol{A}$ is i.i.d. Gaussian ($\boldsymbol{C} = \boldsymbol{I}$, in our set-up). The following theorem extends that bound to cover our model $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C}$, for general $\boldsymbol{C}$.

Theorem 3.2 (extended general M^* bound)**.** *Let $\mathcal{T}$ be a bounded subset of $\mathbb{R}^p$, $\boldsymbol{B} \in \mathbb{R}^{n\times r}$ an i.i.d. Gaussian matrix, $\boldsymbol{C} \in \mathbb{R}^{r\times p}$ a fixed matrix, and $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{C} \in \mathbb{R}^{n\times p}$. Fix $\varepsilon \geq 0$ and consider the set*

$$\mathcal{T}_\varepsilon \;:=\; \left\{\boldsymbol{u} \in \mathcal{T} \,:\, \|\boldsymbol{A}\boldsymbol{u}\|_1/n \leq \varepsilon\right\}. \tag{23}$$

Then, with $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I})$ being a standard Gaussian random vector in $\mathbb{R}^q$,

$$\begin{aligned}&\mathbb{E} \sup_{\boldsymbol{u}\in\mathcal{T}_\varepsilon} \left(\boldsymbol{u}^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{u}\right)^{1/2}\\ &\leq\; \sqrt{\frac{8\pi}{n}}\,\mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{T}} |\langle \boldsymbol{C}^T\boldsymbol{g}, \boldsymbol{u}\rangle| \;+\; \varepsilon\sqrt{\frac{\pi}{2}}.\end{aligned} \tag{24}$$

The proof (in the supplementary materials) is based on symmetrization and contraction inequalities, modifying the proof by Vershynin (2014) to account for $\boldsymbol{C}$. Theorem 3.2 can be used to derive error bounds for estimating signals known to belong to some subset (sparsity is a special case considered below). Let $\mathcal{K} \subset \mathbb{R}^p$ be given and suppose that we observe $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{\nu}$, with $\frac{1}{n}\|\boldsymbol{\nu}\|_1 \leq \varepsilon$, where $\boldsymbol{x}^\star \in \mathcal{K}$. Recall the *Minkowski gauge* of $\mathcal{K}$, defined as

$$\|\boldsymbol{x}\|_{\mathcal{K}} \;=\; \inf\{\lambda > 0 \,:\, \lambda^{-1}\boldsymbol{x} \in \mathcal{K}\},$$

which is a norm if $\mathcal{K}$ is a compact and origin-symmetric convex set with non-empty interior (Rockafellar, 1970). The following theorem, which extends one by Vershynin (2014, Theorem 6.2), is then used to prove Theorem 3.1.

Theorem 3.3. *Let $\widehat{\boldsymbol{x}} \in \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_{\mathcal{K}}$, subject to $\frac{1}{n}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_1 \leq \varepsilon$, then*

$$\begin{aligned}&\mathbb{E}\sup_{\boldsymbol{x}^\star\in\mathcal{K}} \left\{(\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star)^T\boldsymbol{C}^T\boldsymbol{C}(\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star)\right\}^{1/2}\\ &\leq\; \sqrt{8\pi}\left(\frac{\mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{K}-\mathcal{K}} |\langle \boldsymbol{C}^T\boldsymbol{g}, \boldsymbol{u}\rangle|}{\sqrt{n}} + \varepsilon\right).\end{aligned} \tag{25}$$

Proof. The constraint guarantees that $\frac{1}{n}\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{y}\|_1 \leq \varepsilon$, whereas by assumption, $\frac{1}{n}\|\boldsymbol{A}\boldsymbol{x}^\star - \boldsymbol{y}\|_1 = \frac{1}{n}\|\boldsymbol{\nu}\|_1 \leq \varepsilon$. Thus, $\|\widehat{\boldsymbol{x}}\|_{\mathcal{K}} \leq \|\boldsymbol{x}^\star\|_{\mathcal{K}} \leq 1$, since $\boldsymbol{x}^\star \in \mathcal{K}$. The inequality $\|\widehat{\boldsymbol{x}}\|_{\mathcal{K}} \leq 1$ implies that $\widehat{\boldsymbol{x}} \in \mathcal{K}$.

Next, apply Theorem 3.2 to the set $\mathcal{T} = \mathcal{K} - \mathcal{K}$, with 2ε instead of ε, yielding

$$\begin{aligned}&\mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{T}_{2\varepsilon}} \left(\boldsymbol{u}^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{u}\right)^{1/2}\\ &\leq\; \sqrt{2\pi/n}\,\mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{T}} |\langle \boldsymbol{C}^T\boldsymbol{g}, \boldsymbol{u}\rangle| \;+\; \sqrt{8\pi}\varepsilon\,,\end{aligned}$$

From here, all we need to show is that for any $\boldsymbol{x}^\star \in \mathcal{K}$, $\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star \in \mathcal{T}_{2\varepsilon}$. To see this, simply note that $\widehat{\boldsymbol{x}}, \boldsymbol{x}^\star \in \mathcal{K}$, so $\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star \in \mathcal{K} - \mathcal{K} = \mathcal{T}$. By the triangle inequality,

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{A}(\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star)\|_1 \;&=\; \frac{1}{n}\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{y} + \boldsymbol{\nu}\|_1\\ &\leq\; \frac{1}{n}\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{y}\|_1 + \frac{1}{n}\|\boldsymbol{\nu}\|_1 \;\leq\; 2\,\varepsilon,\end{aligned}$$

showing that $\boldsymbol{u} = \widehat{\boldsymbol{x}} - \boldsymbol{x}^\star \in \mathcal{T}_{2\varepsilon}$ (see (23)). □

Proof. (Theorem 3.1) Since $\boldsymbol{x}^\star$ is assumed to satisfy $\|\boldsymbol{x}^\star\|_1 \leq \sqrt{s}\|\boldsymbol{x}^\star\|_2$, we first need to construct an OWL ball that contains all $\boldsymbol{x} \in \mathbb{R}^p$ with $\|\boldsymbol{x}\|_1 \leq \sqrt{s}\|\boldsymbol{x}^\star\|_2$. Let

$$\mathcal{K} = \{\boldsymbol{x} \in \mathbb{R}^p \,:\, \Omega_{\boldsymbol{w}}(\boldsymbol{x}) \leq w_1\sqrt{s}\|\boldsymbol{x}^\star\|_2\}.$$

Because $\Omega_{\boldsymbol{w}}(\boldsymbol{x}) \leq w_1\|\boldsymbol{x}\|_1$, all vectors satisfying $\|\boldsymbol{x}\|_1 \leq \sqrt{s}\|\boldsymbol{x}^\star\|_2$ belong to $\mathcal{K}$. Also, because $\Omega_{\boldsymbol{w}}(\boldsymbol{x})$ is a norm, and $\mathcal{K}$ a ball thereof, $\|\boldsymbol{x}\|_{\mathcal{K}}$ is proportional to $\Omega_{\boldsymbol{w}}(\boldsymbol{x})$.

The quantity $\mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{K}-\mathcal{K}} |\langle \boldsymbol{C}^T\boldsymbol{g}, \boldsymbol{u}\rangle|$ in (25) is called the width of $\mathcal{K}$ and satisfies

$$\mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{K}-\mathcal{K}} |\langle \boldsymbol{C}^T\boldsymbol{g}, \boldsymbol{u}\rangle| \;=\; \mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{K}-\mathcal{K}} |\langle \boldsymbol{g}, \boldsymbol{C}\boldsymbol{u}\rangle|.$$

Noting that $\|\boldsymbol{C}\boldsymbol{u}\|_1 \leq \|\boldsymbol{C}\|_1\|\boldsymbol{u}\|_1 \leq \|\boldsymbol{C}\|_1\,\Omega_{\boldsymbol{w}}(\boldsymbol{u})/\bar{w}$ (see the first paragraph of Subsection 1.1), and the fact that the triangle inequality and the definition of $\mathcal{K}$ imply that, for any $\boldsymbol{u} \in \mathcal{K} - \mathcal{K}$, $\Omega_{\boldsymbol{w}}(\boldsymbol{u}) \leq 2w_1\sqrt{s}\|\boldsymbol{x}^\star\|_2$, yields

$$\|\boldsymbol{C}\boldsymbol{u}\|_1 \;\leq\; 2\,\|\boldsymbol{C}\|_1\,\frac{w_1}{\bar{w}}\,\sqrt{s}\,\|\boldsymbol{x}^\star\|_2 \;=:\; \rho\,. \tag{26}$$

The width can be then bounded as

$$\begin{aligned}\mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{K}-\mathcal{K}} |\langle \boldsymbol{g}, \boldsymbol{C}\boldsymbol{u}\rangle| \;&\leq\; \mathbb{E}\sup_{\{\boldsymbol{v}\,:\,\|\boldsymbol{v}\|_1\leq\rho\}} |\langle \boldsymbol{g}, \boldsymbol{v}\rangle|\\ &\leq\; \rho\,\mathbb{E}\max_{i=1,\ldots,r} |g_i|,\end{aligned} \tag{27}$$

where the second inequality results from the fact that the $\boldsymbol{v}$ achieving the supremum places all its mass ρ on the largest component of $\boldsymbol{g}$ (in magnitude). The classical Gaussian tail bound[2] together with the *union bound* yield $\mathbb{P}(\max_{i=1,\ldots,r} |g_i| > t) \leq r e^{-t^2/2}$; consequently,

$$\begin{aligned}\mathbb{E}\max_{i=1,\ldots,r} |g_i| \;&=\; \int_0^\infty \mathbb{P}\Big(\max_{i=1,\ldots,r} |g_i| > t\Big)\,dt\\ &\leq\; \sqrt{2\log r} + r\int_{\sqrt{2\log r}}^\infty e^{-t^2/2}\,dt\\ &\leq\; \sqrt{2\log r} + \sqrt{\pi/2}\\ &<\; 2\sqrt{2\log r},\end{aligned} \tag{28}$$

where the second inequality results from applying the Gaussian tail bound[2], and the third from assuming $r > 2$. Plugging ρ (as defined in (26)) back in, leads to

$$\mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{K}-\mathcal{K}} |\langle \boldsymbol{g}, \boldsymbol{C}\boldsymbol{u}\rangle| \;\leq\; 4\sqrt{2}\,\|\boldsymbol{C}\|_1\,\frac{w_1}{\bar{w}}\,\|\boldsymbol{x}^\star\|_2\,\sqrt{s\log r}.$$

[2] If $g \sim \mathcal{N}(0, 1)$, then $\mathbb{P}(g > t) \leq e^{-t^2/2}/2$.

We can also modify the argument above to obtain a different bound, in terms of $\|\boldsymbol{C}\|_2$ instead of $\|\boldsymbol{C}\|_1$, which can be tighter in certain cases. Recall that we must bound the width $\mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{K}-\mathcal{K}}|\langle\boldsymbol{C}^T\boldsymbol{g},\boldsymbol{u}\rangle|$ and that, as shown above (at the very beginning of the proof),

$$\|\boldsymbol{u}\|_1 \leq 2\frac{w_1}{\bar{w}}\sqrt{s}\,\|\boldsymbol{x}^\star\|_2 =: \rho'. \tag{29}$$

Consequently,

$$\begin{aligned}\mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{K}-\mathcal{K}}|\langle\boldsymbol{C}^T\boldsymbol{g},\boldsymbol{u}\rangle| &\leq \mathbb{E}\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_1\leq\rho'}|\langle\boldsymbol{C}^T\boldsymbol{g},\boldsymbol{v}\rangle|\\ &\leq \rho'\,\mathbb{E}\max_{i=1,\ldots,p}|\boldsymbol{c}_i^T\boldsymbol{g}|\,,\end{aligned}$$

where $\boldsymbol{c}_i$ is the i-th column of $\boldsymbol{C}$; the second inequality stems from the fact that the $\boldsymbol{v}$ achieving the supremum places its total mass ρ' on the largest component of $\boldsymbol{C}^T\boldsymbol{g}$ (in magnitude). Note that $\boldsymbol{c}_i^T\boldsymbol{g}\sim\mathcal{N}(0,\|\boldsymbol{c}_i\|_2^2)$ and that $\|\boldsymbol{c}_i\|_2\leq\|\boldsymbol{C}\|_2$, since $\boldsymbol{c}_i=\boldsymbol{C}\boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is a canonical unit vector. From this and the bounding argument in (28), it follows that

$$\begin{aligned}\mathbb{E}\max_{i=1,\ldots,p}|\boldsymbol{c}_i^T\boldsymbol{g}| &\leq \|\boldsymbol{C}\|_2\,\mathbb{E}\max_{i=1,\ldots,p}\left|\frac{\boldsymbol{c}_i^T\boldsymbol{g}}{\|\boldsymbol{c}_i\|_2}\right|\,,\\ &\leq 2\,\|\boldsymbol{C}\|_2\sqrt{2\log p}\,,\end{aligned}$$

where we assume $p>2$. Plugging ρ' (as defined in (29)) back in yields the bound

$$\mathbb{E}\sup_{\boldsymbol{u}\in\mathcal{K}-\mathcal{K}}|\langle\boldsymbol{C}^T\boldsymbol{g},\boldsymbol{u}\rangle|\leq 4\sqrt{2}\,\|\boldsymbol{C}\|_2\,\frac{w_1}{\bar{w}}\,\|\boldsymbol{x}^\star\|_2\,\sqrt{s\log p}.$$

Theorem 3.1 now follows directly from Theorem 3.3. □

4 Conclusion

In this paper, we have studied sparse linear regression with strongly correlated covariates under the recently proposed *ordered weighted* ℓ_1 (OWL) regularization, which generalizes the *octagonal shrinkage and clustering algorithm for regression* (OSCAR) (Bondell and Reich, 2007). We have proved sufficient conditions for OWL regularization to cluster the coefficient estimates, extending and qualitatively strengthening a previous result by Bondell and Reich (2007). We have also characterized the statistical performance of OWL regularization for generative models with clusters of strongly correlated covariates. Essentially, we have shown that, by using OWL regularization, no price is paid (in terms of the number of measurements) for the presence of strongly correlated covariates.

Future work will include the experimental evaluation of OWL regularization and its combination with other loss functions, such as logistic and hinge. An important open problem concerns the choice of the weight vector $\boldsymbol{w}$, in order to fully exploit the flexibility of the OWL family.

Appendix A: Strong Schur Convexity of $\Omega_{\boldsymbol{w}}$

This appendix briefly reviews the main concepts of majorization and Schur convexity (Marshall et al., 2011), and introduces a new notion of *strong Schur convexity*, showing that Lemma 2.1 is nothing but a statement about the strong Schur convexity of $\Omega_{\boldsymbol{w}}$.

Let $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^p$. Vector $\boldsymbol{y}$ is said to *majorize* $\boldsymbol{x}$ (denoted $\boldsymbol{y}\succ\boldsymbol{x}$) if

$$\mathbf{1}^T\boldsymbol{x}=\mathbf{1}^T\boldsymbol{y}\;\text{ and }\;\sum_{i=1}^k y_{[i]}\geq\sum_{i=1}^k x_{[i]},\;\text{ for } k=1,...,p-1.$$

Intuitively, $\boldsymbol{y}\succ\boldsymbol{x}$ if the two vectors have the same sum, and the components of $\boldsymbol{x}$ have a more homogenous distribution than those of $\boldsymbol{y}$. If $\boldsymbol{y}$ is a permutation of $\boldsymbol{x}$, then $\boldsymbol{y}\succ\boldsymbol{x}$ and $\boldsymbol{x}\succ\boldsymbol{y}$. The majorization relation is a *preorder* (*i.e.*, it is reflexive and transitive).

Let $\mathcal{A}\subseteq\mathbb{R}^p$. Function $\phi:\mathcal{A}\to\mathbb{R}$ is said to be *Schur-convex* on $\mathcal{A}$, if $\boldsymbol{y}\succ\boldsymbol{x}\Rightarrow\phi(\boldsymbol{y})\geq\phi(\boldsymbol{x})$. Furthermore, if

$$(\boldsymbol{y}\succ\boldsymbol{x})\wedge(\boldsymbol{y}\text{ is not a permutation of }\boldsymbol{x})\;\;\Rightarrow\phi(\boldsymbol{y})>\phi(\boldsymbol{x}),$$

then ϕ is said to be *strictly Schur-convex*. Intuitively, Schur-convex functions "prefer" (*i.e.*, yield lower values) vector arguments with more uniformly distributed components.

A definition of *strong Schur convexity* requires a measure of "amount of majorization" of $\boldsymbol{y}$ with respect to $\boldsymbol{x}$. A natural choice for this purpose is the so-called *Pigou-Dalton (a.k.a. Robin Hood) transfer* (Dalton, 1920; Marshall et al., 2011; Pigou, 1912). Specifically, given $\boldsymbol{y}$, consider two of its components y_i and y_j, such that $y_i>y_j$. We say that $\boldsymbol{y}$ $\varepsilon-$majorizes $\boldsymbol{x}$, denoted $\boldsymbol{y}\succ_\varepsilon\boldsymbol{x}$, if $\boldsymbol{x}$ results from a Pigou-Dalton transfer of size $\varepsilon\in(0,(y_i-y_j)/2]$ applied to $\boldsymbol{y}$, *i.e.*, $x_i=y_i-\varepsilon$, $x_j=y_j+\varepsilon$, and $x_k=y_k$, for $k\neq i,j$.

Based on the notion of $\varepsilon-$majorization introduced in the previous paragraph, we propose the following definition of *strong Schur convexity*. A function $\phi:\mathcal{A}\to\mathbb{R}$ is said to be $\delta-$strongly Schur convex on $\mathcal{A}$ if

$$\boldsymbol{y}\succ_\varepsilon\boldsymbol{x}\;\Rightarrow\;\phi(\boldsymbol{y})-\phi(\boldsymbol{x})\geq\delta\,\varepsilon.$$

Finally, given this definition of strong Schur convexity, it is clear that what Lemma 2.1 shows is that the OWL norm $\Omega_{\boldsymbol{w}}$ is $\Delta_{\boldsymbol{w}}$-strongly Schur convex on $\mathbb{R}^p_+$. In contrast, it is easy to show that neither the ℓ_1 norm nor the EN regularizer are strongly Schur convex.

References

M. Bogdan, E. van den Berg, W. Su, and E. Candès. Statistical estimation and testing via the ordered ℓ_1 norm. Technical report, http://arxiv.org/pdf/1310.1969v1.pdf, 2013.

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. Candès. SLOPE – adaptive variable selection via convex optimization. Technical report, arxiv.org/abs/1407.3824, 2014.

H. Bondell and B. Reich. Regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64:115–123, 2007.

P. Bühlmann, P. Rüttiman, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143:1835–1858, 2013.

E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509, 2006.

J. Clark. *Locally Non-Linear Learning via Feature Induction and Structured Regularization in Statistical Machine Translation*. PhD thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2015.

H. Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30:348–361, 1920.

D. Davis. An $o(n \log(n))$ algorithm for projecting onto the ordered weighted ℓ_1 norm ball. Technical report, arxiv.org/abs/1505.00870, 2015.

D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006.

C. Genovese, J. Jin, L. Wasserman, and Z. Yao. A comparison of the LASSO and marginal regression. *Journal of Machine Learning Research*, 13:2107–2143, 2012.

J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52:4036–4048, 2006.

J. Jia and B. Yu. On model selection consistency of the elastic net when $p \gg n$. *Statistica Sinica*, 20:595–611, 2010.

D. Lorenz and N. Worliczek. Necessary conditions for variational regularization schemes. *Inverse Problems*, 29(7): 075016, 2013.

A. Marshall, I. Olkin, and B. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, New York, 2011.

N. Meinshausen and P. Bülhmann. Stability selection. *Journal of the Royal Statistical Society (B)*, 72(4):417–473, 2010.

A. Pigou. *Wealth and Welfare*. Macmillan, London, 1912.

N. Rao, B. Recht, and R. Nowak. Tight measurement bounds for exact recovery of structured sparse signals. In *Proceedings of AISTATS*, 2012.

G. Raskutti, M. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

R. Shah and R. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society (B)*, 75(1):55–80, 2013.

X. Shen and H.-C. Huang. Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105:727–739, 2010.

W. Su and E. Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. Technical report, arxiv.org/abs/1503.08393, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (B)*, 58(1): 267–288, 1996.

R. Vershynin. Estimation in high dimensions: A geometric perspective. Technical report, http://arxiv.org/abs/1405.5103, 2014.

X. Zeng and M. Figueiredo. Decreasing weighted sorted ℓ_1 regularization. *IEEE Signal Processing Letters*, 21: 1240–1244, 2014a.

X. Zeng and M. Figueiredo. The ordered weighted ℓ_1 norm: atomic formulation, dual norm, and projections. Technical report, arxiv.org/abs/1409.4271, 2014b.

L. Zhong and J. Kwok. Efficient sparse modeling with automatic feature grouping. *IEEE Transactions on Neural Networks and Learning Systems*, 23:1436–1447, 2012.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 67(2):301–320, 2005.