Linear Regression with Strongly Correlated Designs

Using Ordered Weighted $\ell_1$ (OWL) Regularization

Mário A. T. Figueiredo

Instituto de Telecomunicações and
Instituto Superior Técnico, Universidade de Lisboa
Portugal

Joint work with Robert Nowak (U Wisconsin, USA)
**Linear regression**: classical problem in statistics, machine learning, signal processing, with countless applications.

Observations: \( y = A \mathbf{x} + \mathbf{n} \)
Linear regression: classical problem in statistics, machine learning, signal processing, with countless applications.

Observations: \( y = Ax + n \)

- **Design matrix**: \( A = [a_1, a_2, ..., a_p] \in \mathbb{R}^{n \times p} \);
**Linear regression**: classical problem in statistics, machine learning, signal processing, with countless applications.

Observations: \( y = A \mathbf{x} + \mathbf{n} \)

- **Design matrix**: \( A = [a_1, a_2, \ldots, a_p] \in \mathbb{R}^{n \times p} \);
- **Regression coefficients**: \( \mathbf{x} \in \mathbb{R}^p \);
**Linear regression**: classical problem in statistics, machine learning, signal processing, with countless applications.

Observations: \( y = A \mathbf{x} + \mathbf{n} \)

- **Design matrix**: \( A = [a_1, a_2, ..., a_p] \in \mathbb{R}^{n \times p} \);
- **Regression coefficients**: \( \mathbf{x} \in \mathbb{R}^p \);
- **Noise** (or random perturbations): \( \mathbf{n} \in \mathbb{R}^n \);
**Linear regression**: classical problem in statistics, machine learning, signal processing, with countless applications.

**Observations**: \[ y = A \mathbf{x} + \mathbf{n} \]

- **Design matrix**: \( A = [a_1, a_2, \ldots, a_p] \in \mathbb{R}^{n \times p} \);
- **Regression coefficients**: \( \mathbf{x} \in \mathbb{R}^p \);
- **Noise** (or random perturbations): \( \mathbf{n} \in \mathbb{R}^n \);
- **Goal**: estimate \( \mathbf{x} \), from \( y \) and \( A \).
Regularized linear regression (classical criteria):

\[
\hat{x} = \arg \min_x \frac{1}{2} \| A x - y \|_2^2 + \lambda R(x)
\]
Regularized linear regression (classical criteria):

\[
\hat{x} = \text{arg min}_x \frac{1}{2} \| A x - y \|_2^2 + \lambda R(x)
\]

- \( R(x) = \| x \|_2^2 \) \( \Rightarrow \) \( \hat{x} = (A^T A + \lambda I)^{-1} A^T y \);
  ridge regression, (Hoerl and Kennard, 1970)

- Sparsity! (variable selection)
Regularization, Sparsity, and Variable Selection

Regularized linear regression (classical criteria):

\[ \hat{x} = \arg \min_x \frac{1}{2} \| A x - y \|_2^2 + \lambda R(x) \]

- \( R(x) = \| x \|_2^2 \) \( \Rightarrow \) \( \hat{x} = (A^T A + \lambda I)^{-1} A^T y; \) ridge regression, (Hoerl and Kennard, 1970)

- \( R(x) = \| x \|_1; \) LASSO (Tibshirani, 1996), basis pursuit denoising (Chen et al., 1995)

**Sparsity!** (variable selection)
Group/Structured Sparsity

Promote certain sparsity patterns (usually groups)
Group/Structured Sparsity

Promote certain sparsity patterns (usually groups)

**Group sparsity:** discard/keep entire *groups* of features \((\text{Bach et al., 2012})\)

- *density* inside each group
- *sparsity* with respect to the groups which are selected
- choice of groups: prior knowledge about the intended *sparsity patterns*
Group/Structured Sparsity

Promote certain sparsity patterns (usually groups)

**Group sparsity:** discard/keep entire *groups* of features (Bach et al., 2012)

- **density** inside each group
- **sparsity** with respect to the groups which are selected
- choice of groups: prior knowledge about the intended *sparsity patterns*

Yields statistical gains if the assumption is correct (Huang and Zhang, 2010; Stojnic et al., 2009)
Group/Structured Sparsity

Promote certain sparsity patterns (usually groups)

**Group sparsity**: discard/keep entire *groups* of features (Bach et al., 2012)

- **density** inside each group
- **sparsity** with respect to the groups which are selected
- choice of groups: prior knowledge about the intended *sparsity patterns*

Yields statistical gains if the assumption is correct (Huang and Zhang, 2010; Stojnic et al., 2009)

Many applications:

- feature template selection (Martins et al., 2011)
- multi-task learning (Caruana, 1997; Obozinski et al., 2010)
- multiple kernel learning (Bach, 2008)
- learning the structure of graphical models (Schmidt and Murphy, 2010)
Goal: identify all the covariates (e.g., genes, voxels,...) that are relevant in some problem/task
**Goal:** identify all the covariates (e.g., genes, voxels,...) that are relevant in some problem/task

**Problem:** with highly correlated covariates, LASSO may select an arbitrary subset thereof
Goal: identify all the covariates (e.g., genes, voxels,...) that are relevant in some problem/task

Problem: with highly correlated covariates, LASSO may select an arbitrary subset thereof

Group regularizers may solve this problem, but require a priori knowledge of group structure
Variable Selection and Grouping

- **Goal**: identify all the covariates (e.g., genes, voxels,...) that are relevant in some problem/task

- **Problem**: with highly correlated covariates, LASSO may select an arbitrary subset thereof

- **Group regularizers** may solve this problem, but require *a priori* knowledge of group structure

- **Alternatives** (without predefined groups):
  - Elastic net (EN)  
    (Zou and Hastie, 2005; De Mol et al., 2009)
Goal: identify all the covariates (e.g., genes, voxels,...) that are relevant in some problem/task

Problem: with highly correlated covariates, LASSO may select an arbitrary subset thereof

Group regularizers may solve this problem, but require a priori knowledge of group structure

Alternatives (without predefined groups):

- Elastic net (EN)  
  (Zou and Hastie, 2005; De Mol et al., 2009)

- Cluster lasso  
  (Bühlmann et al., 2013)
Goal: identify all the covariates (e.g., genes, voxels,...) that are relevant in some problem/task

Problem: with highly correlated covariates, LASSO may select an arbitrary subset thereof

Group regularizers may solve this problem, but require a priori knowledge of group structure

Alternatives (without predefined groups):

- Elastic net (EN)  
  (Zou and Hastie, 2005; De Mol et al., 2009)

- Cluster lasso  
  (Bühlmann et al., 2013)

- Octagonal shrinkage and clustering algorithm for regression (OSCAR)  
  (Bondell and Reich, 2007; Zhong and Kwok, 2012)
Elastic Net (EN) and OSCAR

Goal of **EN**: including groups of correlated variables.

Goal of **OSCAR**: grouping correlated variables.

\[
\text{Elastic net: } R(x) = \lambda_1 \|x\|_1 + \lambda_2 \|x\|_2^2
\]

\[
\text{OSCAR: } R(x) = \lambda_1 \|x\|_1 + \lambda_2 \sum_{i<j} \max\{\|x_i\|, \|x_j\|\}
\]

OSCAR is competitive with EN, LASSO, ridge, in terms of MSE; OSCAR yields explicit variable grouping (Bondell and Reich, 2007)
Elastic Net (EN) and OSCAR

Goal of EN: including groups of correlated variables.

Goal of OSCAR: grouping correlated variables.

Elastic net:

\[ R(x) = \lambda_1 \|x\|_1 + \lambda_2 \|x\|_2^2 \]
Elastic Net (EN) and OSCAR

Goal of **EN**: including groups of correlated variables.

Goal of **OSCAR**: grouping correlated variables.

- **Elastic net:**
  \[ R(\mathbf{x}) = \lambda_1 \| \mathbf{x} \|_1 + \lambda_2 \| \mathbf{x} \|_2^2 \]

- **OSCAR:**
  \[ R(\mathbf{x}) = \lambda_1 \| \mathbf{x} \|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\} \]
Elastic Net (EN) and OSCAR

Goal of **EN**: including groups of correlated variables.

Goal of **OSCAR**: grouping correlated variables.

- **Elastic net**:
  \[ R(x) = \lambda_1 \|x\|_1 + \lambda_2 \|x\|_2^2 \]

- **OSCAR**:
  \[ R(x) = \lambda_1 \|x\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\} \]

OSCAR is competitive with EN, LASSO, ridge, in terms of MSE; OSCAR yields explicit variable grouping (Bondell and Reich, 2007).
Elastic Net (EN) and OSCAR

Goal of **EN**: including groups of correlated variables.

Goal of **OSCAR**: grouping correlated variables.

- **Elastic net**: 
  \[ R(x) = \lambda_1 \|x\|_1 + \lambda_2 \|x\|_2^2 \]

- **OSCAR**: 
  \[ R(x) = \lambda_1 \|x\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\} \]

OSCAR is competitive with EN, LASSO, ridge, in terms of MSE;

OSCAR yields explicit variable grouping (Bondell and Reich, 2007)
Some OSCAR Results on Synthetic Data

From (Bondell and Reich, 2007)

<table>
<thead>
<tr>
<th>Example</th>
<th>Med. MSE (Std. Err.)</th>
<th>MSE 10th perc.</th>
<th>MSE 90th perc.</th>
<th>Med. Df</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ridge 2.31 (0.18)</td>
<td>0.98</td>
<td>4.25</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Lasso   1.92 (0.16)</td>
<td>0.68</td>
<td>4.02</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Elastic Net 1.64 (0.13)</td>
<td>0.49</td>
<td>3.26</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Oscar    1.68 (0.13)</td>
<td>0.52</td>
<td>3.34</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Ridge   2.94 (0.18)</td>
<td>1.36</td>
<td>4.63</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Lasso   2.72 (0.24)</td>
<td>0.98</td>
<td>5.50</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Elastic Net 2.59 (0.21)</td>
<td>0.95</td>
<td>5.45</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Oscar    2.51 (0.22)</td>
<td>0.96</td>
<td>5.06</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>Ridge   1.48 (0.17)</td>
<td>0.56</td>
<td>3.39</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Lasso   2.94 (0.21)</td>
<td>1.39</td>
<td>5.34</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Elastic Net 2.24 (0.17)</td>
<td>1.02</td>
<td>4.05</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Oscar    1.44 (0.19)</td>
<td>0.51</td>
<td>3.61</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>Ridge   27.4 (1.17)</td>
<td>21.2</td>
<td>36.3</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Lasso   45.4 (1.52)</td>
<td>32.0</td>
<td>56.4</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Elastic Net 34.4 (1.72)</td>
<td>24.0</td>
<td>45.3</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Oscar    25.9 (1.26)</td>
<td>19.1</td>
<td>38.1</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>Ridge   70.2 (3.05)</td>
<td>41.8</td>
<td>103.6</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Lasso   64.7 (3.03)</td>
<td>27.6</td>
<td>116.5</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Elastic Net 40.7 (3.40)</td>
<td>17.3</td>
<td>94.2</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Oscar    51.8 (2.92)</td>
<td>14.8</td>
<td>96.3</td>
<td>12</td>
</tr>
</tbody>
</table>
Generalizing OSCAR: The OWL

OSCAR: \[ R_{\text{OSCAR}}^{\lambda_1, \lambda_2}(x) = \lambda_1 \|x\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\} \]
Generalizing OSCAR: The OWL

OSCAR: \[ R_{OSCAR}^{\lambda_1, \lambda_2}(x) = \lambda_1 \|x\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\} \]

\[ = \sum_{i=1}^{p} \left( \lambda_1 + \lambda_2 (p - i) \right) |x|_{[i]}, \]

where \[ |x|_{[1]} \geq |x|_{[2]} \geq \cdots \geq |x|_{[p]} \] (sorted entries of \( |x| \)).
Generalizing OSCAR: The OWL

OSCAR: \[ R^\lambda_1,\lambda_2_{\text{OSCAR}}(x) = \lambda_1 \|x\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\} \]

\[ = \sum_{i=1}^{p} (\lambda_1 + \lambda_2 (p - i)) |x|_{[i]}, \]

where \[ |x|_{[1]} \geq |x|_{[2]} \geq \cdots \geq |x|_{[p]} \] (sorted entries of \(|x|\)).

The ordered weighted \( \ell_1 \) (OWL) norm

\[ \Omega_w(x) = \sum_{i=1}^{p} w_i |x|_{[i]} \]

where \( w_1 \geq w_2 \geq \cdots \geq w_p \geq 0 \)
Generalizing OSCAR: The OWL

**OSCAR:**

\[
R_{\text{OSCAR}}^{\lambda_1,\lambda_2}(x) = \lambda_1 \|x\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}
\]

\[
= \sum_{i=1}^{p} \left( \lambda_1 + \lambda_2(p-i) \right) |x|_{[i]},
\]

where \(|x|_{[1]} \geq |x|_{[2]} \geq \cdots \geq |x|_{[p]}\) (sorted entries of \(|x|\)).

The **ordered weighted \(\ell_1\) (OWL) norm**

\[
\Omega_w(x) = \sum_{i=1}^{p} w_i |x|_{[i]} = w^T |x|_{\downarrow}
\]

where \(w_1 \geq w_2 \geq \cdots \geq w_p \geq 0\) and \(|x|_{\downarrow} = \left[ |x|_{[1]}, |x|_{[2]}, \ldots, |x|_{[p]} \right]^T\)
Toy example

\[ A \in \mathbb{R}^{10 \times 30} \]

every column has 3 replicates

\[ x^* \text{ generating } y = Ax^* \]

\[
\hat{x} = \arg \min \|x\|_1 \\
\text{subject to } \frac{1}{n} \|y - Ax\|_2^2 \leq \varepsilon
\]

\[
\hat{x} = \arg \min \Omega_w(x) \\
\text{subject to } \frac{1}{n} \|y - Ax\|_2^2 \leq \varepsilon
\]
The ordered weighted $\ell_1$ (OWL) norm

$$\Omega_w(x) = \sum_{i=1}^{p} w_i |x|_i = w^T |x|_\downarrow$$
The ordered weighted $\ell_1$ (OWL) norm

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

Proposed independently by:

- Bogdan et al. (2013), for false discovery rate (FDR) control in variable selection with weakly correlated covariates

- Zeng and Figueiredo (2014), generalizing OSCAR, for variable grouping with strongly correlated covariates
The OWL Norm

The ordered weighted $\ell_1$ (OWL) norm

$$
\Omega_w(x) = \sum_{i=1}^{p} w_i |x_i| = w^T |x|_{\downarrow}
$$

- Proposed independently by:
  - Bogdan et al. (2013), for false discovery rate (FDR) control in variable selection with weakly correlated covariates
  - Zeng and Figueiredo (2014), generalizing OSCAR, for variable grouping with strongly correlated covariates

- Remaining of the talk focuses on the OWL
  - Part I: covariate clustering analysis
  - Part II: statistical analysis
Some Properties of the OWL

The **ordered weighted $\ell_1$ (OWL) norm**

$$\Omega_w(x) = \sum_{i=1}^{p} w_i |x|_{[i]} = w^T |x|_\downarrow$$
Some Properties of the OWL

The ordered weighted $\ell_1$ (OWL) norm

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

- $\Omega_w : \mathbb{R}^p \rightarrow \mathbb{R}_+$ is indeed a norm, iff $w_1 > 0$. 

Relationship with $\ell_1$ norm:

$$\bar{w} \parallel x \parallel_1 \leq \Omega_w(x) \leq w_1 \parallel x \parallel_1;$$

where $\bar{w} = \frac{1}{p} \sum_{i=1}^{p} w_i$, with equalities if $w_1 = w_2 = \cdots = w_p$. 

Obviously, $\Omega_w(x) \geq w_1 \parallel x \parallel_\infty$ (equality if $w_2 = w_3 = \cdots = w_p = 0$). 

Proximity operator ($\Omega(p \log p)$), projection onto an OWL-ball ($\Omega(p \log p)$), atomic formulation are all known (yesterday's poster).
Some Properties of the OWL

The **ordered weighted** \( \ell_1 \) (OWL) norm

\[
\Omega_w(x) = \sum_{i=1}^{p} w_i |x|_i = w^T |x|_{\downarrow}
\]

- \( \Omega_w : \mathbb{R}^p \rightarrow \mathbb{R}_+ \) is indeed a norm, iff \( w_1 > 0 \).

- Relationship with \( \ell_1 \)

\[
\bar{w} \|x\|_1 \leq \Omega_w(x) \leq w_1 \|x\|_1;
\]

where \( \bar{w} = \frac{1}{p} \sum_{i=1}^{p} w_i \), with equalities if \( w_1 = w_2 = \cdots = w_p \).
Some Properties of the OWL

The **ordered weighted** $\ell_1$ (OWL) norm

$$\Omega_w(x) = \sum_{i=1}^{p} w_i |x[i] = w^T \downarrow x$$

- $\Omega_w : \mathbb{R}^p \to \mathbb{R}_+$ is indeed a norm, iff $w_1 > 0$.

- Relationship with $\ell_1$

  $$\bar{w} \|x\|_1 \leq \Omega_w(x) \leq w_1 \|x\|_1;$$

  where $\bar{w} = \frac{1}{p} \sum_{i=1}^{p} w_i$, with equalities if $w_1 = w_2 = \cdots = w_p$.

- Obviously, $\Omega_w(x) \geq w_1 \|x\|_\infty$ (equality if $w_2 = w_3 = \cdots = w_p = 0$).
Some Properties of the OWL

The **ordered weighted** $\ell_1$ (OWL) norm

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

- $\Omega_w : \mathbb{R}^p \rightarrow \mathbb{R}_+$ is indeed a norm, iff $w_1 > 0$.

- Relationship with $\ell_1$

  $$\bar{w} \|x\|_1 \leq \Omega_w(x) \leq w_1 \|x\|_1;$$

  where $\bar{w} = \frac{1}{p} \sum_{i=1}^{p} w_i$, with equalities if $w_1 = w_2 = \cdots = w_p$.

- Obviously, $\Omega_w(x) \geq w_1 \|x\|_\infty$ (equality if $w_2 = w_3 = \cdots = w_p = 0$).

- **Proximity operator** $(O(p \log p))$, **projection** onto an OWL-ball $(O(p \log p))$, **atomic** formulation are all known (yesterday’s poster).

M. Figueiredo (IT, IST, U Lisboa)
Atoms

\[
\begin{bmatrix}
0 \\
\frac{1}{w_1}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\frac{1}{w_1 + w_2} \\
\frac{1}{w_1 + w_2}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\frac{1}{w_1} \\
0
\end{bmatrix}
\]
Part I: Clustering Analysis
Definition (Majorization (Marshall et al., 2011))

Consider $x, y \in \mathbb{R}^p$. It is said that $x$ majorizes $y$, denoted $x \succ y$, if

$$
\sum_{i=1}^{p} x_i = \sum_{i=1}^{p} y_i \quad \text{and} \quad \sum_{i=1}^{j} x[i] \geq \sum_{i=1}^{j} y[i], \quad \text{for } j = 1, \ldots, p - 1. \quad (1)
$$
Majorization and Schur Convexity

Definition (Majorization (Marshall et al., 2011))

Consider $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. It is said that $\mathbf{x}$ majorizes $\mathbf{y}$, denoted $\mathbf{x} \succ \mathbf{y}$, if

$$\sum_{i=1}^{p} x_i = \sum_{i=1}^{p} y_i \quad \text{and} \quad \sum_{i=1}^{j} x[i] \geq \sum_{i=1}^{j} y[i], \quad \text{for } j = 1, \ldots, p - 1. \quad (1)$$

Examples: $(4, 0, 0, 0) \succ (3, 1, 0, 0) \succ (2, 1, 1, 0) \succ (1, 1, 1, 1)$
Majorization and Schur Convexity

Definition (Majorization (Marshall et al., 2011))

Consider $x, y \in \mathbb{R}^p$. It is said that $x$ majorizes $y$, denoted $x \succ y$, if

$$
\sum_{i=1}^{p} x_i = \sum_{i=1}^{p} y_i \quad \text{and} \quad \sum_{i=1}^{j} x[i] \geq \sum_{i=1}^{j} y[i], \quad \text{for } j = 1, \ldots, p - 1. \quad (1)
$$

Examples: $(4, 0, 0, 0) \succ (3, 1, 0, 0) \succ (2, 1, 1, 0) \succ (1, 1, 1, 1)$

Definition (Schur-convexity (Marshall et al., 2011))

Let $A \subseteq \mathbb{R}^p$; a function $f : A \rightarrow \mathbb{R}$ is Schur-convex in $A$ if,

$$
\forall x, y \in A, \quad x \succ y \Rightarrow f(x) \geq f(y),
$$

M. Figueiredo (IT, IST, U Lisboa) Ordered Weighted $\ell_1$ (OWL) SPARS’2015 14 / 27
Majorization and Schur Convexity

**Definition (Majorization (Marshall et al., 2011))**

Consider $x, y \in \mathbb{R}^p$. It is said that $x$ majorizes $y$, denoted $x \succ y$, if

\[
\sum_{i=1}^{p} x_i = \sum_{i=1}^{p} y_i \quad \text{and} \quad \sum_{i=1}^{j} x[i] \geq \sum_{i=1}^{j} y[i], \quad \text{for } j = 1, \ldots, p - 1.
\]  

(1)

Examples: \((4, 0, 0, 0) \succ (3, 1, 0, 0) \succ (2, 1, 1, 0) \succ (1, 1, 1, 1)\)

**Definition (Schur-convexity (Marshall et al., 2011))**

Let $A \subseteq \mathbb{R}^P$; a function $f : A \rightarrow \mathbb{R}$ is Schur-convex in $A$ if,

\[
\forall x, y \in A, \ x \succ y \ \Rightarrow \ f(x) \geq f(y),
\]

and strictly Schur-convex, if the second inequality is strict when $x$ is not a permutation of $y$. 
Definition (Pigou-Dalton transfer (Marshall et al., 2011))

Consider \( \mathbf{x} \in \mathbb{R}_+^p \) and two components, \( x_i, x_j \), s.t. \( x_i > x_j \). We say that \( \mathbf{y} \) (\( \mathbf{y} \prec \mathbf{x} \)) results from a Pigou-Dalton transfer of size \( \varepsilon \in (0, (x_i - x_j)/2) \) if

\[
y_i = x_i - \varepsilon, \quad y_j = x_j + \varepsilon, \quad y_k = x_k, \quad \text{for } k \neq i, j.
\]
Definition (Pigou-Dalton transfer (Marshall et al., 2011))

Consider $x \in \mathbb{R}^p_+$ and two components, $x_i, x_j$, s.t. $x_i > x_j$. We say that $y$ ($y \prec x$) results from a Pigou-Dalton transfer of size $\varepsilon \in (0, (x_i - x_j)/2)$ if

$$y_i = x_i - \varepsilon, \quad y_j = x_j + \varepsilon, \quad y_k = x_k, \quad \text{for } k \neq i, j.$$

The Pigou-Dalton transfer (a.k.a. Robin-Hood transfer) is used in the study of measures of economic inequality (Dalton, 1920; Pigou, 1912).
Strong Schur Convexity

**Definition (Pigou-Dalton transfer (Marshall et al., 2011))**

Consider $\mathbf{x} \in \mathbb{R}^p_+$ and two components, $x_i, x_j$, s.t. $x_i > x_j$. We say that $\mathbf{y}$ ($\mathbf{y} \prec \mathbf{x}$) results from a Pigou-Dalton transfer of size $\varepsilon \in (0, (x_i - x_j)/2)$ if

$$y_i = x_i - \varepsilon, \quad y_j = x_j + \varepsilon, \quad y_k = x_k, \quad \text{for } k \neq i, j.$$ 

The Pigou-Dalton transfer (a.k.a. Robin-Hood transfer) is used in the study of measures of economic inequality (Dalton, 1920; Pigou, 1912).

**Definition (Strong Schur convexity (Figueiredo and Nowak, 2014))**

Function $f$ is $S$-strongly Schur-convex if there exists a constant $S > 0$, s.t.

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \varepsilon S,$$

whenever $\mathbf{y} \prec \mathbf{x}$ result from a Pigou-Dalton transfer of size $\varepsilon$ applied to $\mathbf{x}$. 
Consider \( \Omega_w \), with \( w_1 \geq w_2 \geq \cdots \geq x_p \geq 0 \), and let

\[
\Delta = \min\{w_1 - w_2, w_2 - w_3, \ldots, w_{p-1} - w_p\}.
\]

Then, \( \Omega_w \) is \( \Delta \)-strongly Schur-convex.
Strong Schur Convexity of $\Omega_w$ and Exact Grouping

**Lemma (Figueiredo and Nowak (2014))**

Consider $\Omega_w$, with $w_1 \geq w_2 \geq \cdots \geq x_p \geq 0$, and let

$$\Delta = \min \{w_1 - w_2, w_2 - w_3, \ldots, w_{p-1} - w_p\}.$$

Then, $\Omega_w$ is $\Delta$-strongly Schur-convex.

This lemma underlies the proof of the following theorem

**Theorem (Exact grouping (Figueiredo and Nowak, 2014))**

Let $\hat{x} \in \text{arg min } \frac{1}{2} \| y - Ax \|_2^2 + \Omega_w(x)$; then,

(i) $\| a_i - a_j \|_2 < \Delta / \| y \|_2 \Rightarrow \hat{x}_i = \hat{x}_j$

(ii) $\| a_i + a_j \|_2 < \Delta / \| y \|_2 \Rightarrow \hat{x}_i = -\hat{x}_j
Corollary (Standardized Columns (Figueiredo and Nowak, 2014))

Let $\hat{x} \in \text{arg min} \frac{1}{2} \| y - Ax \|_2^2 + \Omega_w(x)$, assume the columns of $A$ have zero-mean $1^T a_k = 0$ and unit norm $\| a_k \|_2 = 1$, and $\rho_{ij} = a_i^T a_j$. Then,

(i) $\sqrt{2 - 2 \rho_{ij}} < \Delta / \| y \|_2 \Rightarrow \hat{x}_i = \hat{x}_j$

(ii) $\sqrt{2 + 2 \rho_{ij}} < \Delta / \| y \|_2 \Rightarrow \hat{x}_i = -\hat{x}_j$
Corollary (Standardized Columns (Figueiredo and Nowak, 2014))

Let \( \hat{x} \in \arg\min \frac{1}{2}\|y - Ax\|_2^2 + \Omega_w(x) \), assume the columns of \( A \) have zero-mean \( 1^T a_k = 0 \) and unit norm \( \|a_k\|_2 = 1 \), and \( \rho_{ij} = a_i^T a_j \). Then,

(i) \( \sqrt{2 - 2\rho_{ij}} < \Delta/\|y\|_2 \) \( \Rightarrow \hat{x}_i = \hat{x}_j \)

(ii) \( \sqrt{2 + 2\rho_{ij}} < \Delta/\|y\|_2 \) \( \Rightarrow \hat{x}_i = -\hat{x}_j \)

- Recovers the theorem by Bondell and Reich (2007) for OSCAR (\( \Delta = \lambda_2 \)), but under much weaker conditions.
Corollary (Standardized Columns (Figueiredo and Nowak, 2014))

Let \( \hat{x} \in \arg\min \frac{1}{2} \| y - Ax \|_2^2 + \Omega_w(x) \), assume the columns of \( A \) have zero-mean \( 1^T a_k = 0 \) and unit norm \( \|a_k\|_2 = 1 \), and \( \rho_{ij} = a_i^T a_j \). Then,

(i) \( \sqrt{2 - 2 \rho_{ij}} < \Delta / \| y \|_2 \Rightarrow \hat{x}_i = \hat{x}_j \)

(ii) \( \sqrt{2 + 2 \rho_{ij}} < \Delta / \| y \|_2 \Rightarrow \hat{x}_i = -\hat{x}_j \)

- Recovers the theorem by Bondell and Reich (2007) for OSCAR (\( \Delta = \lambda_2 \)), but under much weaker conditions.

- Similar results can be proved for the absolute error loss.
Part II: Statistical Analysis
Statistical Bounds

Scenario and assumptions

\[ y = A \times \star + n \|

\|

\|

1 \leq \sqrt{s} \|

\|

2 \leq \epsilon \text{ (no other assumptions on the noise)}

Rows of \( A \in \mathbb{R}^{n \times p} \) are i.i.d. \( \mathcal{N}(0, C^T C) \)

..equivalently, \( A = BC \), with rows of \( B \in \mathbb{R}^{n \times r} \) i.i.d. \( \mathcal{N}(0, I) \), and \( C \in \mathbb{R}^{r \times p} \)

Illustration (exactly replicated columns):

M. Figueiredo (IT, IST, U Lisboa)
Statistical Bounds

Scenario and assumptions

\[ y = A x^* + n \]
Statistical Bounds

Scenario and assumptions

- \( y = A x^* + n \)
- \( \|x^*\|_1 \leq \sqrt{s} \|x\|_2 \) (e.g., \( x^* \) is \( s \)-sparse)
Statistical Bounds

Scenario and assumptions

- $y = A x^* + n$
- $\|x^*\|_1 \leq \sqrt{s} \|x\|_2$ (e.g., $x^*$ is $s$-sparse)
- $\frac{1}{n} \|n\|_1 \leq \varepsilon$ (no other assumptions on the noise)
Statistical Bounds

Scenario and assumptions

- \( y = A x^* + n \)
- \( \|x^*\|_1 \leq \sqrt{s} \|x\|_2 \) (e.g., \( x^* \) is \( s \)-sparse)
- \( \frac{1}{n} \|n\|_1 \leq \varepsilon \) (no other assumptions on the noise)
- Rows of \( A \in \mathbb{R}^{n \times p} \) are i.i.d. \( \mathcal{N}(0, C^T C) \)
Statistical Bounds

Scenario and assumptions

- \( y = A \mathbf{x}^* + \mathbf{n} \)
- \( \|\mathbf{x}^*\|_1 \leq \sqrt{s} \|\mathbf{x}\|_2 \) (e.g., \( \mathbf{x}^* \) is \( s \)-sparse)
- \( \frac{1}{n} \|\mathbf{n}\|_1 \leq \varepsilon \) (no other assumptions on the noise)
- Rows of \( A \in \mathbb{R}^{n \times p} \) are i.i.d. \( \mathcal{N}(0, C^T C) \)
- ..equivalently, \( A = BC \), with rows of \( B \in \mathbb{R}^{n \times r} \) i.i.d. \( \mathcal{N}(0, I) \), and \( C \in \mathbb{R}^{r \times p} \)
- Illustration (exactly replicated columns):

\[
\begin{array}{ccc}
A & \quad = \quad & B \\
\end{array}
\]

Illustration (exactly replicated columns):
Another Illustration: Highly Correlated Groups of Columns

Notice that \( r \leq r \) Similar columns are contiguous only for visualization.
Another Illustration: Highly Correlated Groups of Columns

Notice that $\text{rank}(A) \leq r$

$A = B = C^{T}C$
Another Illustration: Highly Correlated Groups of Columns

- Notice that \( \text{rank}(A) \leq r \)
- Similar columns are contiguous only for visualization
Statistical Bound

Theorem (Figueiredo and Nowak (2014))

Let $y$, $A$, $x^*$, and $\varepsilon$ be as defined above, and $\hat{x}$ be a solution to one of the two following problems:

$$
\min_{x \in \mathbb{R}^p} \Omega_w(x) \quad \text{subject to} \quad \frac{1}{n} \|Ax - y\|_2^2 \leq \varepsilon^2
$$

$$
\min_{x \in \mathbb{R}^p} \Omega_w(x) \quad \text{subject to} \quad \frac{1}{n} \|Ax - y\|_1 \leq \varepsilon.
$$

Then (with $\gamma(C) = \min\{\|C\|_1, \|C\|_2\}$)

$$
\mathbb{E}\|C(\hat{x} - x^*)\|_2 \leq \sqrt{8\pi} \left( \sqrt{32} \gamma(C) \|x^*\|_2 \frac{w_1}{\bar{w}} \sqrt{s \log p} \frac{n}{n} + \varepsilon \right),
$$

Proof based on techniques and tools by Vershynin (2014).

Key step: extension of the general $M^\star$ bound for $A = BC$. 

M. Figueiredo (IT, IST, U Lisboa)
Statistical Bound

**Theorem (Figueiredo and Nowak (2014))**

Let $y$, $A$, $x^*$, and $\varepsilon$ be as defined above, and $\hat{x}$ be a solution to one of the two following problems:

\[
\begin{align*}
\min_{x \in \mathbb{R}^p} & \quad \Omega_w(x) \quad \text{subject to} \quad \frac{1}{n} \|Ax - y\|_2^2 \leq \varepsilon^2 \\
\min_{x \in \mathbb{R}^p} & \quad \Omega_w(x) \quad \text{subject to} \quad \frac{1}{n} \|Ax - y\|_1 \leq \varepsilon.
\end{align*}
\]

Then (with $\gamma(C) = \min\{\|C\|_1, \|C\|_2\}$)

\[
\mathbb{E} \|C(\hat{x} - x^*)\|_2 \leq \sqrt{8\pi} \left( \sqrt{32} \gamma(C) \|x^*\|_2 \frac{w_1}{\bar{w}} \sqrt{\frac{s \log p}{n}} + \varepsilon \right),
\]

- Proof based on techniques and tools by Vershynin (2014).
Statistical Bound

**Theorem (Figueiredo and Nowak (2014))**

Let $y$, $A$, $x^*$, and $\varepsilon$ be as defined above, and $\hat{x}$ be a solution to one of the two following problems:

\[
\min_{x \in \mathbb{R}^p} \Omega_w(x) \quad \text{subject to} \quad \frac{1}{n} \|Ax - y\|_2^2 \leq \varepsilon^2
\]

\[
\min_{x \in \mathbb{R}^p} \Omega_w(x) \quad \text{subject to} \quad \frac{1}{n} \|Ax - y\|_1 \leq \varepsilon.
\]

Then (with $\gamma(C) = \min\{\|C\|_1, \|C\|_2\}$)

\[
\mathbb{E}\|C(\hat{x} - x^*)\|_2 \leq \sqrt{8\pi} \left(\sqrt{32} \gamma(C) \|x^*\|_2 \frac{w_1}{w} \sqrt{\frac{s \log p}{n}} + \varepsilon\right),
\]

- Proof based on techniques and tools by Vershynin (2014).
- Key step: extension of the general $M^*$ bound for $A = BC$. 
Columns of $A$ are either identical or uncorrelated.
Statistical Bound: Insight From a Special Case

- Columns of $A$ are either identical or uncorrelated.
- Let $\bar{x}^*$ have identical components, for identical columns of $A$. 

$$\mathbb{E} \| \hat{x} - \bar{x}^* \|_2 \leq \sqrt{8 \pi} \left( 4 \sqrt{2} \| x^* \|_2 w \bar{w} \sqrt{s \log p} n + \varepsilon \right).$$

i.e., number of samples sufficient to achieve a given precision grows as $n \sim s \log p$ as in bounds with stronger assumptions, e.g., RIP or i.i.d. design (Candès et al., 2006; Candès and Tao, 2007; Donoho, 2006; Haupt and Nowak, 2006; Vershynin, 2014).

No price is paid for the colinearities in $A$. 

M. Figueiredo (IT, IST, U Lisboa)
Columns of $A$ are either identical or uncorrelated.

Let $\bar{x}^*$ have identical components, for identical columns of $A$.

In this case, the theorem claims that

$$E \| \hat{x} - \bar{x}^* \|_2 \leq \sqrt{8\pi} \left( 4\sqrt{2} \| x^* \|_2 \frac{w_1}{\bar{w}} \sqrt{\frac{s \log p}{n}} + \varepsilon \right).$$
Columns of $A$ are either identical or uncorrelated.

Let $\bar{x}^*$ have identical components, for identical columns of $A$.

In this case, the theorem claims that

$$\mathbb{E} \| \hat{x} - \bar{x}^* \|_2 \leq \sqrt{8\pi} \left( 4\sqrt{2} \| x^* \|_2 \frac{w_1}{\bar{w}} \sqrt{\frac{s \log p}{n} + \varepsilon} \right).$$

i.e., number of samples sufficient to achieve a given precision grows as

$$n \sim s \log p$$

as in bounds with stronger assumptions, e.g., RIP or i.i.d. design (Candès et al., 2006; Candès and Tao, 2007; Donoho, 2006; Haupt and Nowak, 2006; Vershynin, 2014)
Columns of $A$ are either identical or uncorrelated.

Let $\bar{x}^*$ have identical components, for identical columns of $A$.

In this case, the theorem claims that

$$\mathbb{E} \| \hat{x} - \bar{x}^* \|_2 \leq \sqrt{8\pi} \left( 4\sqrt{2} \|x^*\|_2 \frac{w_1}{\bar{w}} \sqrt{s \log p} \frac{n}{n} + \varepsilon \right).$$

i.e., number of samples sufficient to achieve a given precision grows as

$$n \sim s \log p$$

as in bounds with stronger assumptions, e.g., RIP or i.i.d. design (Candès et al., 2006; Candès and Tao, 2007; Donoho, 2006; Haupt and Nowak, 2006; Vershynin, 2014)

No price is paid for the colinearities in $A$
Conclusions

- **OSCAR**: a regularizer that aims at identifying groups of correlated variables in linear regression.
- OSCAR is a particular case of the **OWL** norm.
- Exact clustering properties of OWL regularization
- Statistical sample complexity bounds for OWL regularization with correlated designs
Conclusions

- **OSCAR**: a regularizer that aims at identifying groups of correlated variables in linear regression.
- OSCAR is a particular case of the **OWL** norm.
- Exact clustering properties of OWL regularization
- Statistical sample complexity bounds for OWL regularization with correlated designs
- Ongoing work: how to select the weights?
- Ongoing work: other losses, e.g. logistic, hinge,...
Thank you.


