# IS ORDERED WEIGHTED $\ell_1$ REGULARIZED REGRESSION ROBUST TO ADVERSARIAL PERTURBATION ? A CASE STUDY ON OSCAR

Pin-Yu Chen*, Bhanukiran Vinzamuri*, Sijia Liu (*equal contribution)

IBM Research

## Objectives

In this work, we investigate the robustness of sparse regression models with strongly correlated covariates to adversarially designed measurement noises. Specifically, we consider the family of ordered weighted $\ell_1$ (OWL) [1] regularized regression methods and study the case of OSCAR [2] (octagonal shrinkage clustering algorithm for regression) in the adversarial setting.

## Introduction

The OWL family of regularizers is a widely adopted method for sparse regression with strongly correlated covariates. It is worth mentioning that the octagonal shrinkage and clustering algorithm for regression [2], which is called as OSCAR, is in fact a special case of the OWL regularizer [3]. OSCAR is known to be more effective in identifying feature groups (i.e., strongly correlated covariates) than other feature selection methods such as LASSO.

$$\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} w_i |x|_i^{\downarrow}. \tag{1}$$

The OSCAR regularizer [2] is a special case of the OWL norm in (1) when $w_i = \lambda_1 + \lambda_2(p-i)$, where $\lambda_1, \lambda_2 \geq 0$.

We consider the OWL-regularized linear regression problem taking the following form:

$$\text{Minimize}_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\Omega_{\mathbf{w}}(\mathbf{x}), \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^n$ is the vector of $n$ noisy measurements, $\mathbf{A} \in \mathbb{R}^{n\times p}$ is the design matrix, and $\lambda \geq 0$ is the regularization parameter of the OWL norm.

$$\begin{aligned} &\text{Maximize}_{\boldsymbol{\nu}\in\mathbb{R}^n} \ \|\mathbf{x}(\boldsymbol{\nu}) - \mathbf{x}^*\|_2^2 \\ &\text{subject to } \|\boldsymbol{\nu}\|_1/n \leq \epsilon, \\ &\mathbf{x}(\boldsymbol{\nu}) = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\Omega_{\mathbf{w}}(\mathbf{x}). \end{aligned} \tag{3}$$

Our adversarial formulation studies the robustness of OWL-regularized regression in the worst-case scenario by exploring the space of constrained measurement noise to maximize the feature group identification loss in (3). In our setting, we assume the adversary has access to the ground-truth vector $\mathbf{x}^*$ so (3) can be written as

$$\mathbf{x}(\boldsymbol{\nu}) = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{A}(\mathbf{x}^* - \mathbf{x}) + \boldsymbol{\nu}\|_2^2 + \lambda\Omega_{\mathbf{w}}(\mathbf{x}). \tag{4}$$

$$\begin{aligned} \mathbf{x}(\boldsymbol{\nu}) &= \text{Prox}_{\text{OSCAR-APO}}\left(\mathbf{u}^* - \mathbf{A}^T(\mathbf{A}\mathbf{u}^* - \mathbf{y})/\alpha^*\right) \\ &= \text{Prox}_{\text{OSCAR-APO}}\left(\mathbf{u}^* - \mathbf{A}^T(\mathbf{A}\mathbf{u}^* - \mathbf{A}\mathbf{x}^* - \boldsymbol{\nu})/\alpha^*\right), \end{aligned} \tag{5}$$

where $\text{Prox}_{\text{OSCAR-APO}}(\cdot)$ is defined in [3]. With the method of Lagrange multipliers, we are interested in solving the following alternative optimization problem

$$\text{Minimize}_{\boldsymbol{\nu}\in\mathbb{R}^n} -\|\mathbf{x}(\boldsymbol{\nu}) - \mathbf{x}^*\|_2^2 + \frac{\gamma}{n}\|\boldsymbol{\nu}\|_1, \tag{6}$$

where $\gamma > 0$ is a tunable regularization coefficient such that the solution $\boldsymbol{\nu}^*$ to (6) will satisfy the norm constraint $\|\boldsymbol{\nu}^*\|_1/n \leq \epsilon$.

## Main Experimental Result

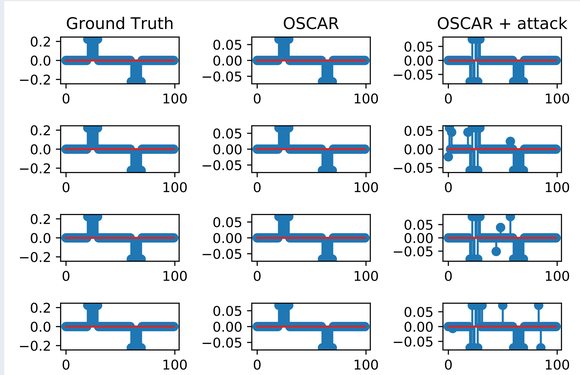


Figure: Assessing effect of our proposed adversarial attack on OSCAR (rightmost column) by varying the attack strength $\epsilon$ with $\epsilon = 0.05$ (top row), $\epsilon = 0.1$ (second row), $\epsilon = 0.2$ (third row) and $\epsilon = 0.3$ (fourth row). In each column, the ground truth refers to $\mathbf{x}^*$, OSCAR refers to the regression results without noise, and OSCAR + Attack refers to the regression results against our designed adversarial noises. The feature groups can be adversarially misaligned even for small $\epsilon$.

## Algorithm

**Input:** $\mathbf{A}$, $\mathbf{x}^*$, $\mathbf{x}^*$, $\mathbf{w}$, $\mathbf{u}^*$, $\alpha^*$, $\epsilon$, $\{\eta_k\}$
**Output:** $\boldsymbol{\nu}^*$
**Initialization:** $k = 0$, $\gamma = \gamma_0$, $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_n)$ and $\boldsymbol{\nu}^{(0)} = \epsilon \cdot \mathbf{g}/\|\mathbf{g}\|_1$
**while** not converged **do**

1. $\mathbf{b} = \mathbf{u}^* - \mathbf{A}^T\left[\mathbf{A}\mathbf{u}^* - \mathbf{A}\mathbf{x}^* - \boldsymbol{\nu}^{(k)}\right]/\alpha^*$
2. Find the permutation $\mathbf{P}(\mathbf{b})$ s.t. $\mathbf{P}(\mathbf{b})|\mathbf{b}| = |\mathbf{b}|^{\downarrow}$
3. $\mathbf{w} = \mathbf{P}(\mathbf{b})^T\mathbf{w}$
4. $\frac{\partial f}{\partial \nu_i} = \sum_{j:|b_j|>w_j} -2\left(b_j - \text{sign}(b_j)w_j - x_j^*\right)\cdot\frac{A_{ij}}{\alpha^*}$ for all $i \in \{1,\ldots,n\}$
5. $\boldsymbol{\nu}^{(k+1)} = S_{\gamma/n}\left[\boldsymbol{\nu}^{(k)} - \eta_k\nabla f(\boldsymbol{\nu}^{(k)})\right]$
6. $k \leftarrow k+1$

**end while**
$\boldsymbol{\nu}^* \leftarrow \boldsymbol{\nu}^{(k)}$
**if** $\|\boldsymbol{\nu}^*\|_1/n > \epsilon$ **then**
Re-initialization with a larger $\gamma$ and redo the while loop
**end if**

## Discussion

In Figure 1, the x-axis represents the feature index and the y-axis represents the coefficient values. The left column represents the ground-truth $\mathbf{x}^*$ with two defined feature groups. The middle column shows the feature grouping obtained after running OSCAR algorithm in the noiseless setting. The right column shows how the grouping is adversely affected after our attack. We varied the noise budget $\epsilon$ from 0.05 to 0.3 to assess the effect of the attack. One can observe that although some grouped features are retained up to a certain degree, the true effect of the attack can be seen on the features which are misaligned from their original feature groups, even for relatively small $\epsilon$.

## Conclusion and Future Work

To study the robustness of OWL-regularized regression methods in the adversarial setting, we propose a novel formulation for finding norm-bounded adversarial perturbations in the measurement model and illustrates the pipeline of adversarial noise generation in the case of OSCAR with APO as its solver. In the adversarial setting, the experimental results show that our proposed approach can effectively craft adversarial noises that severely degrade the regression performance in identifying ground-truth grouped features, even in the regime of small noise budgets. Our results indicate the potential risk of lacking robustness to adversarial noises in the tested regression method. One possible extension of our approach is to devise adversary-resilient regression methods. Our future work also includes developing a generic framework for generating adversarial noises for the entire family of OWL-regularized regression methods.

## References

[1] Mario Figueiredo and Robert Nowak.
Ordered weighted $\ell_1$ regularized regression with strongly correlated covariates: Theoretical aspects.
In *Artificial Intelligence and Statistics*, pages 930–938, 2016.

[2] Howard D Bondell and Brian J Reich.
Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar.
*Biometrics*, 64(1):115–123, 2008.

[3] Xiangrong Zeng and Mario AT Figueiredo.
Solving oscar regularization problems by fast approximate proximal splitting algorithms.
*Digital Signal Processing*, 31:124–135, 2014.

[4] Xiangrong Zeng and Mário AT Figueiredo.
The ordered weighted $\ell_1$ norm: Atomic formulation, projections, and algorithms.
*arXiv preprint arXiv:1409.4271*, 2014.

## Contact Information

▪ Email: Pin-Yu.Chen@ibm.com