Fast OSCAR and OWL Regression via Safe Screening Rules

Runxue Bao[1], Bin Gu[2], Heng Huang[1,2]

[1]Electrical and Computer Engineering, University of Pittsburgh, PA, United States
[2]JD Finance America Corporation, Mountain View, CA, United States

ICML 2020

Outline

Background

Screening Rule for OWL Regression

Algorithms and Theoretical Analysis

Experiments

OWL Regression

We consider the linear regression with Ordered Weighted L_1-Norm (OWL) [5, 11] as:

$$\min_{\beta} P_{\lambda}(\beta) := \frac{1}{2}\|y - X\beta\|_2^2 + \sum_{i=1}^{d} \lambda_i |\beta|_{[i]}, \tag{1}$$

where $X \in \mathbb{R}^{n\times d}$ is the design matrix, $y \in \mathbb{R}^d$ is the measurement vector, λ is a non-negative vector of d non-increasing weights and $|\beta|_{[1]} \geq |\beta|_{[2]} \geq \cdots |\beta|_{[d]}$ are the ordered coefficients in absolute value.

Note (1) is a general form of many learning problems:

- Lasso: $\lambda_1 = \lambda_2 = \cdots = \lambda_d$,
- L_∞-norm regression: $\lambda_1 > 0$ and $\lambda_2 = \cdots = \lambda_d = 0$,
- OSCAR [2]: $\lambda_i = \alpha_1 + \alpha_2(d - i), i = 1, 2, \cdots, d$.

OWL Regression

- OWL Ball [12]:

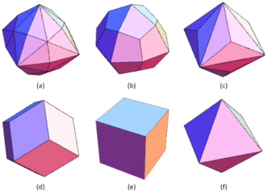


Figure: Illustration of OWL ball in $\mathbb{R}^3$ with different weights.

- Properties:
  - simultaneously promote the sparsity and clustering without any prior information,
  - achieve the minimax estimation from the estimation side,
  - control the false discovery rate from the testing side.

OWL Regression

Optimization Algorithms:

- Accelerated proximal gradient descent algorithm (APGD) [1],

$$\operatorname{prox}(y, \lambda) := \underset{x \in \mathbb{R}^d}{\arg\min} \frac{1}{2}\|y - x\|_2^2 + \sum\nolimits_{i=1}^{d} \lambda_i |x|_{[i]}, \tag{2}$$

- Stochastic proximal gradient descent algorithm with variance reduction (SPGD) [10],
- Suffering high computation costs and memory burden in the high-dimensional setting.

Screening Rules

Screening rules [4]: identify inactive features whose parameter must be zeros at the optimum.

Optimality conditions for Lasso:

$$\begin{aligned} x_i^\top \theta^* + \lambda \operatorname{sign}(\beta_i^*) &= 0, \quad \text{if } \beta_i^* \neq 0, & (3) \\ |x_i^\top \theta^*| &\leq \lambda, \quad \text{if } \beta_i^* = 0. & (4) \end{aligned}$$

Screening rules for Lasso:

$$|x_i^\top \theta^*| < \lambda \Rightarrow \beta_i^* = 0. \tag{5}$$

Motivation

- OWL regression suffers huge computational costs in practice,
- Sparsity is all around in OWL regression,
- Screening is widely used to accelerate the training of sparse learning models such as SVM [6], Sparse SVM [9, 13], Lasso [3], Sparse-group Lasso [7], Proximal Weighted Lasso [8].

Background

Screening Rule for OWL Regression

Algorithms and Theoretical Analysis

Experiments

Screening Test

The dual of OWL regression:

$$\max_{\theta} D_{\lambda}(\theta) := -\frac{1}{2}\|\theta\|_2^2 - \theta^{\top} y, \tag{6}$$

$$s.t. \quad |X^{\top}\theta| \preceq \lambda_{r(\beta)}, \tag{7}$$

where $\lambda_{r(\beta)}$ is the vector of $\lambda_{r(\beta_i)}$ and $\preceq$ means the conditions are satisfied element-wisely.

The screening condition for each variable in OWL regression as:

$$|x_i^{\top}\theta^*| < \lambda_{r(\beta_i^*)} \Rightarrow \beta_i^* = 0, \tag{8}$$

where θ^* is the optimum of the dual.

Screening Test

Challenges:

- Existing screening rules are limited to separable penalties while OWL penalty is non-separable. Thus, all the hyperparameters for each variable in OWL regression is unfixed,
- How to derive an efficient screening rule with the numerous hyperparameters.

Objective: To screen as many variables whose coefficients should be zero as possible by

- constructing a small and safe region for the left term of (8) with the unknown dual optimum,
- exploring the unknown order structure for the right term of (8) with the primal optimum.

Upper Bound for the Left Term

By the triangle inequality, we have:

$$|x_i^\top \theta^*| \le |x_i^\top \theta| + \|x_i\| \|\theta^* - \theta\|, \tag{9}$$

Suppose θ and θ^* are any feasible and the optimum of the dual respectively, we have:

$$\|\theta - \theta^*\| \le \sqrt{2G(\beta, \theta)}. \tag{10}$$

where $G(\beta, \theta) = P(\beta) - D(\theta)$ is the intermediate duality gap.

Hence, we can derive the screening test with the upper bound of the left term as:

$$|x_i^\top \theta| + \|x_i\| \sqrt{2G(\beta, \theta)} < \lambda_{r(\beta_i^*)}. \tag{11}$$

The intermediate duality gap can be computed by β and θ.

Iterative Strategy for the Screening Rule

We can do screening test first as:

$$|x_i^\top \theta| + \|x_i\| \sqrt{2G(\beta, \theta)} < \lambda_d \Rightarrow \beta_i^* = 0. \tag{12}$$

Thus, we can partition the variables into an active set $\mathcal{A}$ of the variables that cannot be removed and an inactive set $\mathcal{A}'$ as the complementary set of $\mathcal{A}$.

Suppose $\mathcal{A}$ has m active features at iteration k, we can assign an arbitrary permutation of $d - m$ smallest parameters to the screened coefficients. Thus, the order of these variables is known. Then, we can derive the order of the new screened variables further by doing screening test as:

$$|x_i^\top \theta| + \|x_i\| \sqrt{2G(\beta, \theta)} < \lambda_m \Rightarrow \beta_i^* = 0. \tag{13}$$

At each iteration, we repeat the screening test until $\mathcal{A}$ keeps unchanged.

Safe Screening Rule with Iterative Strategy

Algorithm 1 Safe Screening Rule for OWL Regression with Iterative Strategy

```
Input: $\mathcal{A}, \lambda, \beta_k, \theta_k, G(\beta_k, \theta_k), X$.
while $\mathcal{A}$ still changes do
    Do the screening test based on (13).
    Update $\mathcal{A}$.
end while
Output: New active set $\mathcal{A}$.
```

Property:

The iterative screening rule we proposed is guaranteed to be safe for Algorithm 1 and the whole training process of OWL regression.

Background

Screening Rule for OWL Regression

Algorithms and Theoretical Analysis

Experiments

Screening Rule in Proximal Gradient Algorithms

Algorithm 2 Accelerated Proximal Gradient Descent Algorithm with Safe Screening Rules

```
Input: $\beta^0, b^1 = \beta^0, t_0 = 1$.
for $k = 1, 2, \cdots$ do
    Compute dual $\theta$ and duality gap.
    Compute active set $\mathcal{A}$ based on Algorithm 1.
    $\beta^k = \mathrm{prox}_{t_k,\lambda}(b^k - t_k X^\top (X b^k - y))$.
    $t_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4t_k^2})$.
    $b^{k+1} = \beta^k + \frac{t_k - 1}{t_{k+1}}(\beta^k - \beta^{k-1})$.
end for
Output: Coefficient $\beta$.
```

Screening Rule in Proximal Gradient Algorithms

Algorithm 3 Stochastic Proximal Gradient Descent Algorithm with Safe Screening Rules

```
Input: $\beta^0, l$.
for $k = 1, 2, \cdots$ do
    Compute dual $\theta$ and duality gap.
    Compute active set $\mathcal{A}$ based on Algorithm 1.
    $\tilde{\beta} = \tilde{\beta}^{k-1}, \tilde{v} = \nabla F(\tilde{\beta}), \beta^0 = \tilde{\beta}$.
    for $t = 1, 2, \cdots, T$ do
        Pick mini-batch $I_t \subseteq X$ of size $l$.
        $v_t = (\nabla f_{I_t}(\beta^{t-1}) - \nabla f_{I_t}(\tilde{\beta}))/l + \tilde{v}$.
        $\beta^t = \mathrm{prox}_{\eta,\lambda}(\beta^{t-1} - \eta v_t)$.
    end for
    $\tilde{\beta}^k = \beta^T$.
end for
Output: Coefficient $\beta$.
```

Complexity Analysis

- Suppose the active set size for iteration k is d_k, the complexity of screening rule is $O(d_k)$.
- Algorithm 2 reduces the complexity $O(d(n + \log d))$ required by APGD to $O(d_k(n + \log d_k))$ for iteration k.
- Algorithm 3 reduces the complexity $O(d(n + Tl + T \log d))$ required by SPGD to $O(d_k(n + Tl + T \log d_k))$ for main loop k where T is inner loop size and l is mini-batch size.

Hence, in high-dimensional sparse learning, the computation costs of both APGD and SPGD are promising to be effectively reduced by our screening rule.

Theoretical Analysis

In terms of the convergence, we have:

Property:

Iterative algorithm Ψ with our screening rule to solve OWL regression converges to the optimum if Ψ converges to the optimum.

In terms of screening ability, we have:

Property:

θ converges to θ^* of the dual if β converges to β^* of the primal.

Property:

Based on the optimality conditions, we have that final active set $\mathcal{A}^*$ satisfies that $\min_{i \in A^*} |x_i^\top \theta^*| = \lambda_{|\mathcal{A}^*|}$. Then, as Ψ converges, there exists an iteration number $K_0 \in \mathbb{N}$ s.t. $\forall k \geq K_0$, any variable $j \notin \mathcal{A}^*$ is screened by our screening rule.

Background

Screening Rule for OWL Regression

Algorithms and Theoretical Analysis

Experiments

Experiments

Table: The real-world datasets used in the experiments.

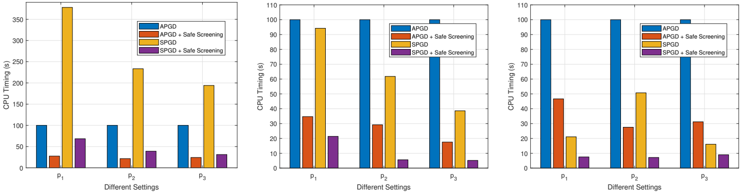
Dataset	Sample size	Attributes
Duke Breast Cancer	44	7129
Colon Cancer	62	2000
Cardiac Left	3360	1600
Cardiac Right	3360	1600
IndoorLoc Longitude	21048	529
Slice Localization	53500	386

Experiments

The compared algorithms:

- APGD: Accelerated proximal gradient descent [1].
- APGD + Screening: Accelerated proximal gradient descent with the safe screening rule.
- SPGD: Stochastic proximal gradient descent with variance reduction we adopt in [10].
- SPGD + Screening: Stochastic proximal gradient method with the safe screening rule.

Running Time



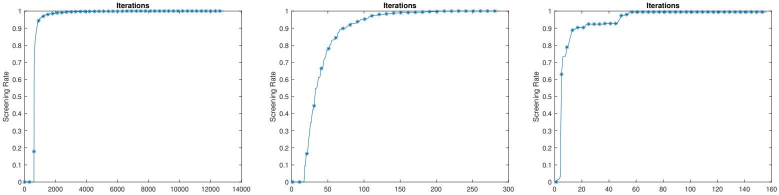
(a) Colon Cancer

(b) Cardiac Right

(c) Slice Localization

Figure: Average running time of different algorithms without and with safe screening rules under different settings.

Screening Rate



(a) Colon Cancer (b) Cardiac Right (c) Slice Localization

Figure: The screening rate of different datasets in the stochastic setting.

Experiments

Table: Prediction errors of different algorithms.

Dataset	APGD	APGD+Screening	SPGD	SPGD+Screening
Duke Breast Cancer	0.6523	**0.6523**	0.6523	**0.6523**
Colon Cancer	0.9453	**0.9453**	0.9453	**0.9453**
Cardiac Left	0.9453	**0.9453**	0.9453	**0.9453**
Cardiac Right	0.5276	**0.5276**	0.5276	**0.5276**
IndoorLoc Longitude	0.5531	**0.5531**	0.5531	**0.5531**
Slice Localization	0.6162	**0.6162**	0.6162	**0.6162**

Conclusion and Future Work

- We propose the first safe screening rule for linear regression with the family of OWL regularizers:
    - We effectively explore the order structure of the primal solution of the non-separable penalty via the iterative strategy,
    - We prove that our screening rule can be safely applied to existing optimization algorithms both in the batch and stochastic setting without any loss of accuracy,
    - The empirical performance shows the superiority of our algorithms with significant computational gain.

- Future work:
    - design faster algorithms to solve OWL models by avoiding more useless updates,
    - design safe screening rules for the models with nonconvex non-separable penalties.

M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès.
Slope—adaptive variable selection via convex optimization.
The annals of applied statistics, 9(3):667–698, 2015.

H. D. Bondell and B. J. Reich.
Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar.
Biometrics, 64(1):115–123, 2008.

O. Fercoq, A. Gramfort, and J. Salmon.
Mind the duality gap: safer rules for the lasso.
In *International Conference on Machine Learning*, pages 333–342, 2015.

T. R. Laurent El Ghaoui, Vivian Viallon.
Safe feature elimination in sparse supervised learning.
Pacific Journal of Optimization, 8:667–698, 2012.

M. lgorzata Bogdana, E. van den Bergb, W. Suc, and E. J. Candesc.

Statistical estimation and testing via the ordered l1 norm.
2013.

J. Liu, Z. Zhao, J. Wang, and J. Ye.
Safe screening with variational inequalities and its application to lasso.
arXiv preprint arXiv:1307.7577, 2013.

E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon.
Gap safe screening rules for sparse-group lasso.
In *Advances in Neural Information Processing Systems*, pages 388–396, 2016.

A. Rakotomamonjy, G. Gasso, and J. Salmon.
Screening rules for lasso with non-convex sparse regularizers.
In *International Conference on Machine Learning*, pages 5341–5350, 2019.

A. Shibagaki, M. Karasuyama, K. Hatano, and I. Takeuchi.
Simultaneous safe screening of features and samples in doubly sparse modeling.

In *International Conference on Machine Learning*, pages 1577–1586, 2016.

L. Xiao and T. Zhang.
A proximal stochastic gradient method with progressive variance reduction.
SIAM Journal on Optimization, 24(4):2057–2075, 2014.

X. Zeng and M. A. Figueiredo.
Decreasing weighted sorted l_1 regularization.
IEEE Signal Processing Letters, 21(10):1240–1244, 2014.

X. Zeng and M. A. Figueiredo.
The ordered weighted l1 norm: Atomic formulation, projections, and algorithms.
arXiv preprint arXiv:1409.4271, 2014.

W. Zhang, B. Hong, W. Liu, J. Ye, D. Cai, X. He, and J. Wang.
Scaling up sparse support vector machines by simultaneous feature and sample reduction.

In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4016–4025. JMLR. org, 2017.

Thank You!