Safe Screening - Old and New

Author: Thu-Le Tran

Abstract. This report provides a comprehensive review of the development of safe screening techniques during the last 15 years.

1. Introduction

Safe screening is a technique to reduce the dimension of optimization problem, so that one can solve the reduced problem faster.

In the following, we mentioned some landmarks in this field.

Safe screening was first introduced in around 2010 by El Ghaoui et. al. [1]. They introduced the method for sparse machine learning problems including lasso, sparse logistic regression and sparse SVM. The common characteristic of these problems is that they are L1 norm regularization convex optimization problem, here L1 norm encourage the sparsity. Although the proposed methods aim to reduce the dimension of the problems, it is referred to as *Feature Elimination* due to the machine learning context.

Another landmark appearred when Ogawa et. al. [2] introduced the *Sample Elimination* for sparse (soft) Support Vector Machine (SVM) for discarding the non-support vectors in SVM models. In the language of optimization, this is equivalent to a dimensionality reduction for the dual problem of SVM.

Furthermore, Shibagaki et. al. [3] propose the safe screening techniques for simultaneous sample and feature elimination for sparse SVM.

[**LE:** Safe screening also has three paradisms:
- Static screening
- Sequential screening
- Dynamic screening

]

2. Lasso - Feature Elimination

[4]

3. SVM - Sample Elimination

[4]

4. Possible Research Directions

The general research direction is to study the safe screening for the following class of problems:

$$\min_{x\in\mathbb{R}^n} \quad f(Ax) + g(x) \tag{1}$$

Usually, to study the safe feature screening for problem (1) , one typically takes the following steps/ answering the following questions:

a) What is the structure of $x^\star$, what is $X^\star$?
b) If one can identify a partial structure of $x^\star$, say a set $\widetilde{X}$ containing $X^\star$, how to reduce the dimension of (1) if we know $\widetilde{X}$? Does this reduced problem lead to more efficient numerical resolution?
c) How to derive the safe feature screening for (1) ? i.e. under what conditions, we can assert that $x^\star \in \widetilde{X}$? In this general stage, one typically has the following steps:
   - We first derive the optimality condition of $x^\star$ and $u^\star$

   $$u^\star \in U^\star \Longleftrightarrow x^\star \in X^\star \tag{2}$$

   - We use the optimality condition to derive a safe screening of the form:

   $$\text{If } u^\star \in \tilde{U} \Longrightarrow x^\star \in \widetilde{X} \tag{3}$$

   We call $\tilde{U}$ the safe dual reion.
   - Given $\tilde{U}$, determine if $u^\star \in \tilde{U}$ is called safe screening test. In practice, one needs this test should be easy to do.

Note that in safe sample screening, we do the same step, but for the dual problem of (1) .

In this section, we will discuss some potential research directions of safe feasture screening corresponds to different choice of regularizations:

- Variational Denoising : $g(x) = \|Gx\|_1$ for some matrix G
- Sparse machine learning: $g(x)$ is some separable gauge
- k-support norm regression: $g(x)$ is a gauge of convex hull of k-sparse L2-norm.

4.1. Sparse Machine Learning Models

[4]

4.2. Sparse Optimal Transport

Optimal Transports and its limitations. Optimal Transport (OT) is a powerful technique for comparing the distributions, i.e. it defines a distance between measures. The most commonly use method is network flow. However, this method is quite slow. Another approach is to use regularization with negative entropy. The problem now can be solved verfy using Sinkhorn algorithms. However, the limitation of this method is that the optimal transportation plan is dense, not sparse.

To enforce the sparsity the L2 norm regularization of transportation plan has been used. Here, one use L2 norm since in this case, the dual problem can be solved efficiently.

In [5], the authors, in addition to L2 norm regularization function, they added the k-sparse constraint for each columns of the transportation plan. The optimal plan obtained by solving this problem is indeed very sparse.

In this note, we are going to apply safe screening for this sparse OT problem.

4.3. Variational denoising

4.3.1. A brief overview

Variational Denoising is a special class of more general class of regression problems.

- The general problem is Generalized Lasso [6]: $\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|b-Ax\|_2^2 + \lambda\|Dx\|_1$
- A special class of generalized Lasso is graph fused lasso [7]: $\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|b-Ax\|_2^2 + \lambda\|Gx\|_1$

where G is a matrix associated a non-directed graph $\mathcal{G} = (V, E)$ with $|V| = n$ vertices and $|E| = p$ edges. Matrix $G \in \mathbb{R}^{p\times n}$ in this case is defined as

$$\text{If } e_i = (v_j, v_k) \Longrightarrow [G]_{ij} = 1 \text{ and } [G]_{ik} = -1. \tag{4}$$

In this case,

$$\|Gx\|_1 = \sum_{e=(v_j,v_k)\in E} |x_j - x_k| \tag{5}$$

the total variation of vector $x \in \mathbb{R}^n$ on $\mathcal{G}$. In [7], the authors exploited the fact that the graph can be decomposed into trail (paths). By using ADMM method for solving graph fused lasso, the sub-problems are 1D fused lasso.

- In case that $\mathcal{G}$ is a path, the problem is called (1D) fused lasso (embed in $\mathbb{R}$). In this case, assuming that the vertices has been sorted, i.e. $v_1 \leq v_2 \leq .... \leq v_n$, then

$$\|Gx\|_1 = \sum_{i=2}^{n} |x_i - x_{i-1}|. \tag{6}$$

Here $G \in \mathbb{R}^{(n-1)\times n}$. Solving fused lasso when $A = I$ can be done in linear time.

- Variational denoising (or 2D fused lasso) corresponds to the case of graph fused lasso, when $\mathcal{G}$ is a grid. This problem has been used to determine the constant pieces of 2D images.

> **New Idea**
>
> The safe screening method has been developed for generalized lasso [6] and fused lasso [8], but graph lasso and 2D fused lasso. Solving these problem (with improvement in a comparison with generalized lasso) can be a promissing research direction.

4.3.2. Review of safe screening for fused lasso

4.4. k-support norm regression

Some papers related to the safe screening w.r.t. the gauge function includes:
a) General piecewise function: [9]
b) General gauge function: [10]
c) slope/oscar norm regularization: [11]. Note that oscar norm is a special case of slope norm

For a review of sparse convex regularized problems with different regularization function, please refer to [4].

> **New Idea**

k-support norm regularization has been introduced in [12], [13] (for regression) [5] (for regularized Optimal Transport). But the screening rule for these norms has not been investigated. Note that k-support norm is a kind of (not exactly) group norm.

Furthermore, the slope norm and and the dual k-support norm can be viewed in the same framework as follows: Let us define the *decreasing operator*:

$$x^{\downarrow} := \left(x_{[i]}, x_{[2]}, ..., x_{[n]}\right) \tag{7}$$

where $[\cdot]$ is a permutation on $\{1, ..., n\}$ such that the absolute value of entries of x is non-increasing:

$$|x|_{[1]} \geq |x|_{[2]} \geq ... \geq |x|_{[n]}. \tag{8}$$

We also denote the top-k projection as

$$P_k x^{\downarrow} = \left(x_{[1]}, x_{[2]}, ..., x_{[k]}\right) \tag{9}$$

The dual k-support norm is defined as the Euclidean norm of top-k entries of decreasing map of x:

$$\|x\|_k^{\downarrow} := \left\|P_k x^{\downarrow}\right\|_2 \tag{10}$$

The slope norm w.r.t. to parameters $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ is defined as:

$$\|x\|_{\text{SLOPE}} := \langle \lambda, |x^{\downarrow}| \rangle \tag{11}$$

It is not hard to see that (10) and (11) are special cases function g defined as:

$$g(x) = \kappa\left(x^{\downarrow}\right). \tag{12}$$

where κ is a gauge function.

Bibliography

[1] L. E. Ghaoui, V. Viallon, and T. Rabbani, "Safe Feature Elimination for the LASSO and Sparse Supervised Learning Problems."

[2] K. Ogawa, Y. Suzuki, and I. Takeuchi, "Safe Screening of Non-Support Vectors in Pathwise SVM Computation."

[3] A. Shibagaki, M. Karasuyama, K. Hatano, and I. Takeuchi, "Simultaneous Safe Screening of Features and Samples in Doubly Sparse Modeling."

[4] F. Bach, "Optimization with Sparsity-Inducing Penalties," *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2011.

[5] T. Liu, J. Puigcerver, and M. Blondel, "Sparsity-Constrained Optimal Transport."

[6] S. Ren, S. Huang, J. Ye, and X. Qian, "Safe Feature Screening for Generalized LASSO," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2992–3006, Dec. 2018.

[7] W. Tansey and J. G. Scott, "A Fast and Flexible Algorithm for the Graph-Fused Lasso."

[8] J. Wang, W. Fan, and J. Ye, "Fused Lasso Screening Rules via the Monotonicity of Subdifferentials," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1806–1820, Sep. 2015.

[9] T. B. Johnson and C. Guestrin, "Unified Methods for Exploiting Piecewise Linear Structure in Convex Optimization," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2016.

[10] Y. Sun and F. Bach, "Safe Screening for the Generalized Conditional Gradient Method."

[11] C. Elvira and C. Herzet, "Safe Rules for the Identification of Zeros in the Solution of the SLOPE Problem," *SIAM Journal on Mathematics of Data Science*, vol. 5, no. 1, pp. 147–173, 2023.

[12] A. Argyriou, R. Foygel, and N. Srebro, "Sparse Prediction with the K-Support Norm," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2012.

[13] A. M. McDonald, M. Pontil, and D. Stamos, "Spectral K-Support Norm Regularization," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2014.