Enhancing safe screening rules with adaptive thresholding for non-overlaping group sparse norm regularized problems

Hector Chahuara and Paul Rodriguez
Department of Electrical Engineering, Pontificia Universidad Católica del Perú, Lima, Perú
Email: {hector.chahuara, prodrig}@pucp.edu.pe

***Abstract*—Sparsity is an often desired property in machine learning and signal processing problems. Recently, techniques such as screening rules were proposed to exploit sparsity in order to diminish the computational requirements of large and huge-scale optimization problems. Nevertheless, existing methods provide rough estimations of the solution support discarding only a few entries in the solution, thus limiting the desired computational savings. In this paper, we propose a simple and computationally cheap modification for safe screening rules based on automatic thresholding and the observation that the screening metric has a distribution that, for practical purposes, can be considered unimodal. The proposed method is evaluated for MEG / EEG source imaging and image classification. Computational results indicate that the proposed screening scheme outperforms the safe method costing only minor losses in accuracy and yields approximate speedups of up to 167.59 for MEG / EEG source imaging, and up to 2.12 for image classification.**

***Index Terms*—Sparsity, screening rules, adaptive thresholding**

I. Introduction

Machine learning models are the most used solutions for a wide range of applications in academia and industry. Nevertheless, the computational requirements to implement said models are exorbitant for real applications. On the other hand, while sparsity-inducing regularization is often used as a mean to prevent overfitting, exploiting it to achieve computational savings remains an area in development.

Screening rules [1]–[11] are a family of methods that use sparsity for problem size reduction in sparse optimization problems by exploiting the structure of the original and its associated dual problems. These techniques have demonstrated to be effective for feature elimination by estimating and tracking the support of the solutions of optimization problems. In general, screening techniques arise as a consequence of the generalized Kuhn-Tucker theorem i.e. Karush-Kuhn-Tucker (KKT) conditions [12]. For an optimization problem with a sparsity-inducing regularization and matrix system $\mathbf{X} \in \mathbb{R}^{m\times n}$, its associated primal-dual pair can be expressed as

$$\hat{\mathbf{\Omega}}^{(\lambda)} \in \underset{\mathbf{\Omega}\in\mathbb{R}^{n\times p}}{\operatorname{argmin}} \; P_\lambda(\mathbf{\Omega}) := f(\mathbf{X\Omega}) + \lambda \cdot g(\mathbf{\Omega}) \tag{1a}$$

$$\hat{\mathbf{\Theta}}^{(\lambda)} \in \underset{\mathbf{\Theta}\in\mathbb{R}^{m\times p}}{\operatorname{argmax}} \; D_\lambda(\mathbf{\Theta}) := -f^*(-\lambda\cdot\mathbf{\Theta}) - \lambda\cdot g^*\left(\mathbf{X}^T\mathbf{\Theta}\right), \tag{1b}$$

where f and g are convex functions (f is L-smooth and g is non-smooth and separable) with Fenchel conjugates f^* and g^*, respectively, and $\hat{\mathbf{\Omega}}^{(\lambda)}$ and $\hat{\mathbf{\Theta}}^{(\lambda)}$ are the primal and dual solutions. The KKT conditions for this pair of problems are

$$\mathbf{X}^T\hat{\mathbf{\Theta}}^{(\lambda)} \in \partial g\left(\hat{\mathbf{\Omega}}^{(\lambda)}\right) \quad \text{(2a)} \qquad \hat{\mathbf{\Theta}}^{(\lambda)} = -\frac{\nabla f\left(\mathbf{X}\hat{\mathbf{\Omega}}^{(\lambda)}\right)}{\lambda}. \quad \text{(2b)}$$

The screening rule stem for discarding the k-th feature from (2a), and is expressed as

$$\mathbf{\Gamma}_k = \left\|\mathbf{X}_{:,k}^T\hat{\mathbf{\Theta}}^{(\lambda)}\right\| < 1 \rightarrow \hat{\mathbf{\Omega}}_{k,:}^{(\lambda)} = \mathbf{0}_p. \tag{3}$$

Screening rule (3) is not applicable since $\hat{\mathbf{\Theta}}^{(\lambda)}$ is not available, so instead the safe screening rule, which guarantees to only discard non-contributing features and that stems from assuming a region $\mathcal{R}$ (known as safe region) that contains $\hat{\mathbf{\Theta}}^{(\lambda)}$, is used

$$\mathbf{\Psi}_k = \max_{\mathbf{\Theta}\in\mathcal{R}} \left\|\mathbf{X}_{:,k}^T\mathbf{\Theta}\right\| < 1 \rightarrow \hat{\mathbf{\Omega}}_{k,:}^{(\lambda)} = \mathbf{0}_p, \tag{4}$$

where $\mathbf{\Psi}_k$ is a quantity that does not depend on $\hat{\mathbf{\Theta}}^{(\lambda)}$. Nevertheless, since typically $\mathbf{\Gamma}_k \leq \mathbf{\Psi}_k$, this rule can predict false non-zeros when used to estimate the true support of the solution.

In this paper, we propose a method based on automatic thresholding to improve the support estimation obtained from safe screening when the regularization function is a sparsity-promoting non-overlapping group norm. This class of problems encompasses multitask and multiclass learning models. The way it interacts with the other screening computation steps is detailed in the Figure 1. The proposed method considers that (3) can be generalized as

$$\mathbf{\Psi}_k < t \rightarrow \hat{\mathbf{\Omega}}_{k,:}^{(\lambda)} = \mathbf{0}_p, \tag{5}$$

where t is a threshold that controls the number of discarded features. The threshold computation is based on the interpreting $\mathbf{\Gamma}$ as a measure of the contribution of each one of the features / atoms that has an unimodal distribution, an observation which, to our knowledge, has not been exploited in existing screening schemes.

The proposed method was built on top of static safe screening rules, thus works as a preprocessing step (discards features prior to optimization time), and was embedded onto the accelerated proximal gradient (APG) method [13] and tested for the identifying active brain regions and image classification tasks. The remainder of this document is structured as follows. Previous works on screening rules are briefly presented in Section II. Section III presents our proposed approach. Computational experiments and results are presented in Section IV. Finally, the conclusions are stated in Section V.

II. Previous related work

Two main classes of screening methods can be distinguished depending on whether they ensure or not only non-contributing features to the solution are discarded. Safe screening rules are those that guarantee to only eliminate features non-contributing features, and are based on constructing a region $\mathcal{R}$ that contains $\hat{\mathbf{\Theta}}^{(\lambda)}$. The first safe rules to be proposed were aimed at the class of ℓ_1-norm regularized optimization problems [1], [3] as a preprocessing step to optimization solvers. These rules were expanded to be applicable to constrained problems [4] and group sparse regularized problems [14]. Further improvements include the formulation of the sequential safe screening rules [1] that exploit the idea of refining the estimated support for each hyperparameter value in a warm start sequence, and the dynamic safe screening rules [9], [10] that allowed for the estimated support to be refined at optimization time. Finally, other

This research was supported by the Army Research Office (ARO) under Grant W911NF-22-1-0296.

979-8-3503-3959-8/23/$31.00 ©2023 IEEE

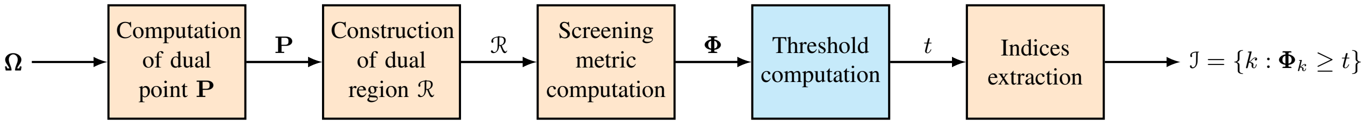
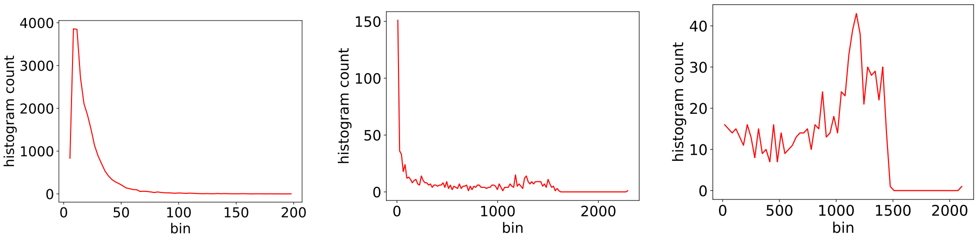

$\boldsymbol{\Omega}$
Computation of dual point $\mathbf{P}$
$\mathbf{P}$
Construction of dual region $\mathcal{R}$
$\mathcal{R}$
Screening metric computation
$\boldsymbol{\Phi}$
Threshold computation
t
Indices extraction
$\mathcal{I} = \{k : \boldsymbol{\Phi}_k \geq t\}$

Fig. 1: Workflow for the use of screening rules for support estimation. In cyan is highlighted the block that is the focus of this work.



(a) MEG / EEG data set (b) MNIST data set (c) Fashion-MNIST data set

Fig. 2: Observed histograms of $\boldsymbol{\Phi}$ (at $\lambda = 10^{-2} \cdot \lambda_{\text{max}}$) for the data sets used in the tasks of study in this work

developments explored the choice for the region $\mathcal{R}$ e.g. ball (safe sphere) [1], [4], [5], ellipsoid [7], dome [8], [15], among others.

Strong screening rules do not guarantee to only discard non-contributing features. Because of that, they are deemed too aggressive and thus need a check on the KKT conditions to reinsert erroneously discarded features. These were initially proposed for ℓ_1-norm regularized problems in [2], expanded to other problems such as the Sorted L-One Penalized Estimation (SLOPE) [11], and, recently, combined with safe screening for LASSO-type problems [16].

III. Proposed Method

A. Distribution of the screening metric

It is important noticing that the dual problem (1b) consists on a data fidelity term that fits the solution to the observed information, and a regularization term that imposes a dependency to the matrix $\mathbf{X}$ of the problem. Then, the dual solution $\hat{\boldsymbol{\Theta}}^{(\lambda)}$ corresponds to what the observed information should be for the primal solution $\hat{\boldsymbol{\Omega}}^{(\lambda)}$ i.e. a sort of correct information. Then, $\boldsymbol{\Gamma}$ is the correlation between each one of the features and the correct information, so it can be interpreted as a measure of how much a feature contributes to the solution $\hat{\boldsymbol{\Omega}}_k^{(\lambda)}$.

In the context of sparse optimization, non-contributing features reach low values of $\boldsymbol{\Gamma}$ (since the correct information does not depend on non-contributing features). Moreover, if the primal solution $\hat{\boldsymbol{\Omega}}^{(\lambda)}$ is highly sparse, there is a high number of non-contributing features and their corresponding values in $\boldsymbol{\Gamma}$ will be low. So, under this conditions, the distribution of $\boldsymbol{\Gamma}$ can be assumed to be unimodal with the modal value being the value corresponding to non-contributing features.

In practice, $\boldsymbol{\Psi}$ is used instead of $\boldsymbol{\Gamma}$ to formulate computationally applicable screening rules. It has been demonstrated [6] that the k-th value of $\boldsymbol{\Psi}$ is bounded according to

$$\boldsymbol{\Gamma}_k \leq \boldsymbol{\Psi}_k \leq \boldsymbol{\Gamma}_k + \|\mathbf{X}_{k,:}\| \cdot \operatorname{diam}(\mathcal{R}) \,, \tag{6}$$

where $\operatorname{diam}(\mathcal{R})$ is the diameter of the region $\mathcal{R}$ i.e. the distance between the two farthest points in $\mathcal{R}$. On the other hand, the computationally applicable rules often arise from using a known dual point $\mathbf{P} \in \mathcal{R}$ into (10) to arrive to a quantities that are feasible to be computed. These quantities form the actual screening metric $\boldsymbol{\Phi}$, for which the k-th component is often computed as

$$\boldsymbol{\Phi}_k = \|\mathbf{X}_{k,:}\mathbf{P}\| + \|\mathbf{X}_{k,:}\| \cdot \|\boldsymbol{\Theta} - \mathbf{P}\| \,, \tag{7}$$

where $\boldsymbol{\Psi}_k \leq \boldsymbol{\Phi}_k$. By considering

$$\|\mathbf{X}_{k,:}\mathbf{P}\| \leq \boldsymbol{\Psi}_k \qquad \|\boldsymbol{\Theta} - \mathbf{P}\| \leq \operatorname{diam}(\mathcal{R}) \qquad D = \max_k \ \|\mathbf{X}_{k,:}\| \,,$$

and when used with (6) into (7), $\boldsymbol{\Phi}_k$ can be bounded as

$$\boldsymbol{\Gamma}_k \leq \boldsymbol{\Phi}_k \leq \boldsymbol{\Gamma}_k + 2 \cdot D \cdot \operatorname{diam}(\mathcal{R}) \,. \tag{9}$$

Finally, if $\hat{\boldsymbol{\Omega}}^{(\lambda)}$ is sparse, then $\boldsymbol{\Gamma}$ can be assumed to have unimodal distribution with mode $\boldsymbol{\Gamma}_{\text{mode}}$. In turn, the distribution of $\boldsymbol{\Phi}$ can be considered approximately unimodal with mode $\boldsymbol{\Phi}_{\text{mode}}$ such that

$$\boldsymbol{\Gamma}_{\text{mode}} \leq \boldsymbol{\Phi}_{\text{mode}} \leq \boldsymbol{\Gamma}_{\text{mode}} + 2 \cdot D \cdot \operatorname{diam}(\mathcal{R}) \,, \tag{10}$$

i.e. the mode is of $\boldsymbol{\Phi}$ is bounded. Observations made on the screening metric, that are listed in Figure 2, corroborate that the unimodality assumption on the screening metric is valid.

B. Threshold computation

In order to compute the threshold t, we assume that $\boldsymbol{\Phi}$ has an unimodal distribution, then, its histogram, from which an idealized though faithful representation can be observed in Figure 3, can be analyzed to formulate a rule for the computation of t. A basic criteria would be to take a threshold greater than $\boldsymbol{\Phi}_{\text{mode}}$ to discard with great probability the values associated with non-contributing features. Nevertheless, in practice, this strategy can be either too conservative or too aggressive depending on the location of the threshold for safe screening (which is equal to one).

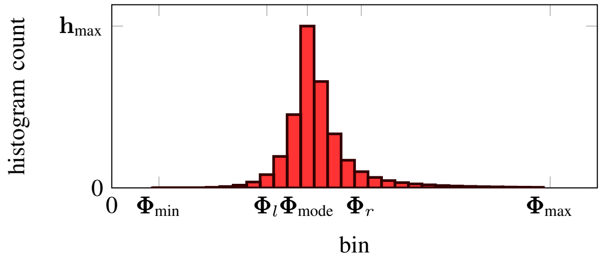


Fig. 3: Idealized representation of the histogram of $\boldsymbol{\Phi}$

The following values can be extracted from the histogram of $\boldsymbol{\Phi}$

$$\boldsymbol{\Phi}_{\text{min}} \leq \boldsymbol{\Phi}_l \leq \boldsymbol{\Phi}_{\text{mode}} \leq \boldsymbol{\Phi}_r \leq \boldsymbol{\Phi}_{\text{max}}, \tag{11}$$

TABLE I: Mathematical elements used to apply safe screening rules for the cases of study: MEG / EEG experiment modeled as multitask LASSO, and image classification modeled as sparse multinomial logistic regression. It is worth noting that in this table, $\mathbf{\Omega}, \mathbf{\Theta}, \mathbf{Z}$ are matrices ($\mathbf{Z} \in \mathbb{R}^{m \times n}$ to compute of λ_{max} and L for sparse multinomial logistic regression), and $\mathbf{X}$ is the matrix of the system.

	Multitask LASSO	**Sparse multinomial logistic regression**
$f(\mathbf{\Omega})$	$\frac{1}{2}\Vert\mathbf{\Omega}-\mathbf{Z}\Vert_F^2$	$-\sum_k\sum_l\left[\mathbf{Z}\odot\textbf{softmax}(\mathbf{\Omega})\right]_{k,l}$
$f^*(\mathbf{\Theta})$	$\frac{1}{2}\Vert\mathbf{\Theta}+\mathbf{Z}\Vert_F^2-\frac{1}{2}\Vert\mathbf{Z}\Vert_F^2$	$\sum_k \mathbf{NH}\left(\mathbf{\Theta}_{k,:}+\mathbf{Z}_{k,:}\right)$
$\nabla f(\mathbf{\Omega})$	$\mathbf{\Omega}-\mathbf{Z}$	$\textbf{softmax}(\mathbf{\Omega})-\mathbf{Z}$
$g(\mathbf{\Omega})$	$\Vert\mathbf{\Omega}\Vert_{1,2}$	$\Vert\mathbf{\Omega}\Vert_{1,2}$
λ_{max}	$\left\Vert\mathbf{X}^T\mathbf{Z}\right\Vert_{\infty,2}$	$\left\Vert\mathbf{X}^T\left(\frac{1}{n}\cdot\mathbf{1}_{m\times n}-\mathbf{Z}\right)\right\Vert_{\infty,2}$
L	1	$\frac{n-1}{n}$

where $\mathbf{\Phi}_r$ is a value corresponding to the threshold for unimodal data that can be computed by using techniques such as [17], and $\mathbf{\Phi}_l$ can be computed similarly for the left part of the histogram.

The proposed criteria for the computation of t takes into account the shape of the histogram of $\mathbf{\Phi}$, and is crafted such that the resulting value of t is greater than one (to achieve more computational savings than safe screening). Before analyzing the histogram of $\mathbf{\Phi}$ to formulate a rule for the computation of t, it is important noting that if $\mathbf{\Phi}_{\text{max}} \leq 1$, then the solution of the problem, denoted by $\hat{\mathbf{\Omega}}^{(\lambda)}$, is all-zero, since safe thresholding should discard all features. For that reason, in this case, it is not necessary to compute t. Then, the following cases for the computation of t can be distinguished:

- $\mathbf{\Phi}_{\text{mode}} \leq 1 < \mathbf{\Phi}_{\text{max}}$: This indicates that $\hat{\mathbf{\Omega}}^{(\lambda)}$ is highly sparse, then t can be chosen in $\left[\max\left(1, \mathbf{\Phi}_r\right), \mathbf{\Phi}_{\text{max}}\right[$. In practice, for $0 \leq \alpha_0 < 1$, the following rule can be applied
$$t = \alpha_0 \cdot \max\left(1, \mathbf{\Phi}_r\right) + \left(1 - \alpha_0\right) \cdot \mathbf{\Phi}_{\text{max}}. \tag{12}$$
- $1 < \mathbf{\Phi}_{\text{mode}}$ and the histogram of $\mathbf{\Phi}$ is not left skewed: This indicates that $\hat{\mathbf{\Omega}}^{(\lambda)}$ is moderately sparse. If the histogram of $\mathbf{\Phi}$ is right skewed, then an aggressive strategy can be applied. It was observed that t can be reasonably chosen in $\left[\mathbf{\Phi}_{\text{mode}}, \mathbf{\Phi}_h\right]$. Then, by considering $0 \leq \alpha_1 \leq 1$, t can be chosen as
$$t = \alpha_1 \cdot \mathbf{\Phi}_{\text{mode}} + \left(1 - \alpha_1\right) \cdot \mathbf{\Phi}_r. \tag{13}$$
On the other hand, if the histogram of $\mathbf{\Phi}$ is also not right skewed, t can be reasonably chosen in $\left]\max\left(1, \mathbf{\Phi}_l\right), \mathbf{\Phi}_{\text{mode}}\right]$. Then, by considering $0 < \alpha_2 \leq 1$, t can be chosen as
$$t = \alpha_2 \cdot \max\left(1, \mathbf{\Phi}_l\right) + \left(1 - \alpha_2\right) \cdot \mathbf{\Phi}_{\text{mode}}. \tag{14}$$
- $\mathbf{\Phi}_l \leq 1 < \mathbf{\Phi}_{\text{mode}}$ and the histogram of $\mathbf{\Phi}$ is left skewed: This indicates that $\hat{\mathbf{\Omega}}^{(\lambda)}$ could be moderately sparse, and then $\left]1, \mathbf{\Phi}_{\text{mode}}\right]$ can be considered a reasonable range for t. In practice, it suffices to apply, for $0 < \alpha_3 \leq 1$, the following rule
$$t = \alpha_3 + \left(1 - \alpha_3\right) \cdot \mathbf{\Phi}_{\text{mode}}. \tag{15}$$
- $1 < \mathbf{\Phi}_l$ and the histogram of $\mathbf{\Phi}$ is left skewed: This indicates that $\hat{\mathbf{\Omega}}^{(\lambda)}$ is not sparse or lowly sparse. Considering the latter case, then t can be reasonably chosen in $\left]\max\left(1, \mathbf{\Phi}_{\text{min}}\right), \mathbf{\Phi}_l\right]$. In practice, it suffices to apply, for $0 < \alpha_4 \leq 1$, the following rule
$$t = \alpha_4 \cdot \max\left(1, \mathbf{\Phi}_{\text{min}}\right) + \left(1 - \alpha_4\right) \cdot \mathbf{\Phi}_l. \tag{16}$$

C. Feature correction

Since the value of t is greater than one (which is the value for the safe screening threshold), our proposed method might discard relevant features. In the same way to the strong screening rules [2], this undesired effect can be mitigated by performing a check on the KKT conditions to reinsert erroneously discarded contributing features in the optimization procedure.

IV. Experiments and results

A. Experiments settings

Computational tests were carried on an Intel i7-2600K (32 GB RAM, 8 MB cache, 3.4 GHz). The proposed method was embedded into the APG solver, and will be evaluated using the following tasks:

- MEG / EEG source imaging: Modeled as multitask LASSO. The results of this task will be assessed using relative error for reconstruction of measurements.
- Image classification: Modeled as sparse multinomial logistic regression. The results of this task will be assessed using the accuracy in training and test set.

All these tasks can be modeled as optimization problems of the form (1a). The proposed method was tested for static gap safe screening with ball as the safe region [18]. The elements needed to perform gap safe screening for each one of the cases of study are detailed in Table I i.e. loss and regularization functions, gradient and dual of the loss function, λ_{max} (minimum value of the hyperparameter λ for which the solution of the problem is all-zero), and Lipschitz constant L. Other mathematical elements used in Table I include the softmax function which is defined, for a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, as

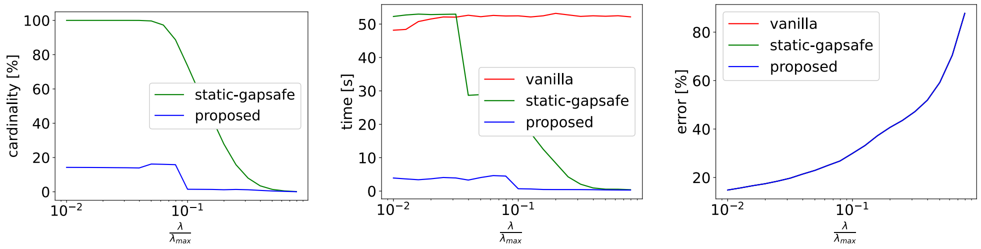
$$\textbf{softmax}(\mathbf{X}) = \begin{bmatrix} \frac{\textbf{exp}\left(\mathbf{X}_{0,:}\right)^T}{\sum_{k=0}^{n-1} \textbf{exp}\left(\mathbf{X}_{0,k}\right)} & \cdots & \frac{\textbf{exp}\left(\mathbf{X}_{m-1,:}\right)^T}{\sum_{k=0}^{n-1} \textbf{exp}\left(\mathbf{X}_{m-1,k}\right)} \end{bmatrix}^T, \tag{17}$$

and is part of the sparse multinomial logistic regression which models image classification in this work, and **NH** that represents the negative entropy function and is defined, for a vector $\mathbf{x} \in \mathbb{R}^n$, as

$$\mathbf{NH}(\mathbf{x}) = \begin{cases} \sum_{k=0}^{n-1} \mathbf{x}_k \cdot \log\left(\mathbf{x}_k\right) & \text{if } \sum_{k=0}^{n-1} \mathbf{x}_k = 1 \text{ and } \mathbf{x}_k > 0, \\ +\infty & \text{if otherwise.} \end{cases} \tag{18}$$

B. Computational results

Cardinality (percentage of detected active features), processing time (in seconds), and quality metrics are measured for a grid of 20 hyperparameter values ($10^k \cdot \lambda_{\text{max}}$ with $k = -2, -1.9, \ldots, -0.1$). It is important to mention that the horizontal axis in the Figures 4, 5 and 6 is the relative hyperparameter i.e. λ normalized with respect to λ_{max}. Three versions are tested for each one of the cases of study: optimization without screening, with static gap safe screening rules [6], and with static gap safe screening enhanced with the proposed adaptive thresholding scheme, which are labeled as vanilla, static-gapsafe and proposed, respectively. In all the tests, for the proposed method we considered $\alpha_0 = 0.75$, and $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.25$.



(a) Cardinality

(b) Processing time

(c) Relative error

Fig. 4: Computational results for the MEG / EEG source imaging experiment

1) MEG / EEG source imaging: Electroencephalography (EEG) and magnetoencephalography (MEG) are brain imaging modalities that allow to identify active brain regions. Typically, this problem consists on solving a multitask regression problem with squared loss where every task corresponds to a time instant. Nevertheless, it is valid to impose a temporal stationary assumption i.e. the recovered sources are identical during a short time interval, then, this task can be modeled as a multitask LASSO [19]. This experiment used a joint MEG / EEG data set (see [18] for more details) with number of sensors $n = 360$ (301 MEG and 59 EEG sensors), number of possible sources is $p = 22494$, and number of time instants $q = 20$.

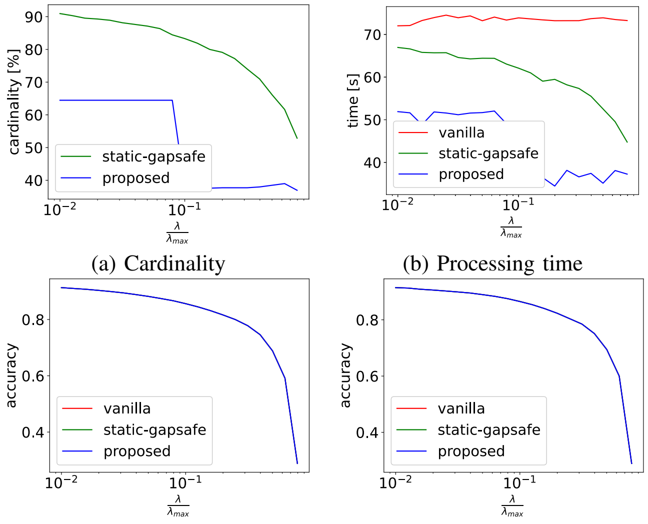
Computational results for this experiment can be observed in Figure 4. It is clear our method discards significantly more features than safe screening (more noticeable for small values of λ) while having no drop in quality in terms of relative error . In terms of speed, the proposed scheme yields a speedup of up to 167.59 approximately, and it is significantly faster than safe screening for small values of λ, thus our proposed method exploits better the sparsity of the solution than safe screening in this experiment.

2) Image classification: The image classification experiments use sparse multinomial logistic regression as the model, and applies it to the MNIST and the Fashion-MNIST data sets. The MNIST data set [20] consists on size-normalized and centered 28×28 grayscale images of 70000 handwritten digits organized in 10 classes. On the other hand, the Fashion-MNIST dataset [21] consists on 28×28 grayscale images of 70000 fashion products from 10 classes, with 7000 images per class. In both data sets, the training and test sets have 60000 and 10000 images, respectively.

Computational results for image classification using the MNIST and the Fashion-MNIST data sets can be observed in Figures 5 and 6, respectively. It is important mentioning that our proposed method discards more features than safe screening. Despite the potential unsafe behavior of our proposed method (for the chosen parameters), the accuracy results (in both training and test sets) indicate that our proposed technique only produces zero or minor losses, which means that mostly features with zero or small contribution were discarded. In terms of performance, our proposed method is faster than safe screening, and yields speedups of up to 1.5 and 2.12 for the MNIST and the Fashion-MNIST data sets, respectively.

V. Conclusions

A new method for feature screening based on the observation that the measure related to the contribution of each feature to the solution has unimodal distribution was proposed. To the best of our knowledge, our key observation, i.e. in praxis, the histogram of the screening metric $\mathbf{\Phi}$ is unimodal, has not been exploited before. The efficiency of our proposed screening method was demonstrated on experiments using medical and computer vision data. Experimental results suggest that our proposed method, despite the additional computational overhead it introduces over safe screening, improves the acceleration of optimization solvers targeting sparse regularization beyond what safe screening allows, while yielding either zero or only some minor losses in quality metrics.



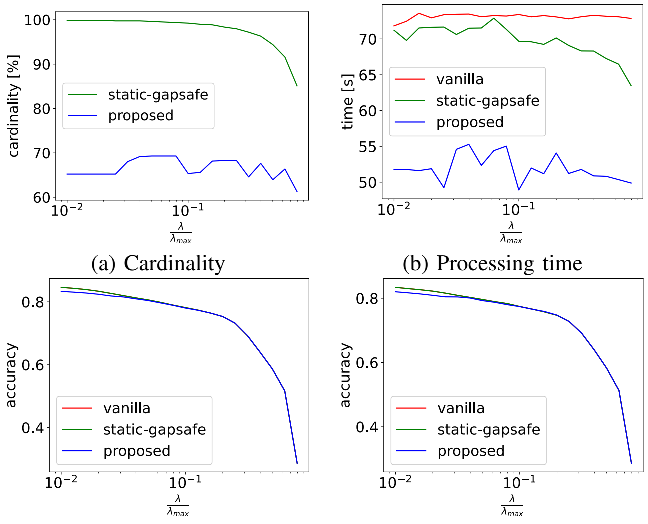
(a) Cardinality

(b) Processing time

(c) Accuracy in train set

(d) Accuracy in test set

Fig. 5: Computational results for the MNIST data set



(a) Cardinality

(b) Processing time

(c) Accuracy in train set

(d) Accuracy in test set

Fig. 6: Computational results for the Fashion-MNIST data set

References

[1] L. El Ghaoui, Vivian Viallon, and Tarek Rabbani, "Safe Feature Elimination in Sparse Supervised Learning," EECS Dept., University of California at Berkeley, Tech. Rep. UC/EECS-2010-126, September 2010.

[2] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 74, no. 2, pp. 245–266, 2012.

[3] O. Fercoq, A. Gramfort, and J. Salmon, "Mind the duality gap: safer rules for the lasso," in *ICML*, 2015.

[4] A. Raj, J. Olbrich, B. Gärtner, B. Schölkopf, and M. Jaggi, "Screening Rules for Convex Problems," 2016.

[5] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon, "Gap Safe Screening Rules for Sparsity Enforcing Penalties," *J. Mach. Learn. Res.*, vol. 18, no. 1, p. 4671–4703, Jan. 2017.

[6] E. Ndiaye, "Safe optimization algorithms for variable selection and hyperparameter tuning," Ph.D. dissertation, 10 2018.

[7] L. Dai and K. Pelckmans, "An ellipsoid based, two-stage screening test for BPDN," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, 2012, pp. 654–658.

[8] Z. J. Xiang and P. J. Ramadge, "Fast lasso screening tests based on correlations," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2137–2140, 2012.

[9] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval, "A dynamic screening principle for the lasso," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 6–10.

[10] ——, "Dynamic screening: Accelerating first-order algorithms for the lasso and group-lasso," *IEEE Transactions on Signal Processing*, vol. 63, no. 19, pp. 5121–5132, 2015.

[11] J. Larsson, M. Bogdan, and J. Wallin, "The Strong Screening Rule for SLOPE," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[12] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[13] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM J. Imaging Sciences*, vol. 2, pp. 183–202, 01 2009.

[14] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon, "GAP Safe screening rules for sparse multi-task and multi-class models," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 811–819.

[15] T.-L. Tran, C. Elvira, H.-P. Dang, and C. Herzet§, "Beyond GAP screening for Lasso by exploiting new dual cutting half-spaces," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 2056–2060.

[16] Y. Zeng, T. Yang, and P. Breheny, "Hybrid safe–strong rules for efficient optimization in lasso-type problems," *Computational Statistics Data Analysis*, vol. 153, p. 107063, 2021.

[17] P. L. Rosin, "Unimodal thresholding," *Pattern Recognition*, vol. 34, pp. 2083–2096, 2001.

[18] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon, "Gap safe screening rules for sparse multi-task and multi-class models," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 811–819.

[19] A. Gramfort, M. Kowalski, and M. Hämäläinen, "Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods," *Physics in Medicine Biology*, vol. 57, no. 7, p. 1937, mar 2012.

[20] Y. LeCun, C. Cortes, and C. Burges, "MNIST handwritten digit database," *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010.

[21] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms," 2017.