HAL
open science

Faster and still safe: combining screening techniques and structured dictionaries to accelerate the Lasso

Cassio Fraga Dantas, Rémi Gribonval

To cite this version:

Cassio Fraga Dantas, Rémi Gribonval. Faster and still safe: combining screening techniques and structured dictionaries to accelerate the Lasso. ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing, Apr 2018, Calgary, AB, Canada. pp.4069-4073, 10.1109/ICASSP.2018.8461514 . hal-01706392

HAL Id: hal-01706392

https://inria.hal.science/hal-01706392v1

Submitted on 11 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FASTER AND STILL SAFE: COMBINING SCREENING TECHNIQUES AND STRUCTURED DICTIONARIES TO ACCELERATE THE LASSO

Cássio F. Dantas, Rémi Gribonval

INRIA Rennes-Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes, France

ABSTRACT

Accelerating the solution of the Lasso problem becomes crucial when scaling to very high dimensional data. In this paper, we propose a way to combine two existing acceleration techniques: safe screening tests, which simplify the problem by eliminating useless dictionary atoms; and the use of structured dictionaries which are faster to operate with. A structured approximation of the true dictionary is used at the initial stage of the optimization, and we show how to define screening tests which are still safe despite the approximation error. In particular, we extend a state-of-the-art screening test, the GAP SAFE sphere test, to this new setting. The practical interest of the proposed methodology is demonstrated by considerable reductions in simulation time.

***Index Terms*—** Lasso, safe screening, structured dictionaries, sparsity.

1. INTRODUCTION

Sparsity-constrained linear inverse problems have been found useful for numerous applications in different areas such as statistics, signal processing and machine learning. The goal is to approximate an N-dimensional input vector $\mathbf{y}$ as a linear combination of a few columns (*atoms*) of a dictionary matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_K] \in \mathbb{R}^{N\times K}$, i.e. $y \approx \mathbf{X}\boldsymbol{\beta}$.

Among many possible formulations for achieving this task, the l_1-regularized least-squares, referred to as Lasso, is one of the most commonly adopted. It consists in finding a sparse coefficient vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^K$, solution of the following primal optimization problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \underbrace{\frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1}_{P(\boldsymbol{\beta})} \tag{1}$$

where the parameter $\lambda > 0$ controls the trade-off between the data fidelity and sparsity of the solution and $P(\boldsymbol{\beta})$ is the primal objective. It is also useful – as we will see in Section 2 – to define its dual formulation:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Delta_{\mathbf{X}}}{\operatorname{argmax}} \underbrace{\frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\left\|\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda}\right\|_2^2}_{D(\boldsymbol{\theta})} \tag{2}$$

denoting $\Delta_{\mathbf{X}} = \{\boldsymbol{\theta} \in \mathbb{R}^N : \|\mathbf{X}^T\boldsymbol{\theta}\|_\infty \leq 1\}$ the dual feasible set and $D(\boldsymbol{\theta})$ the dual objective. The dual and primal solutions ($\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$) are linked through the relation $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \lambda\hat{\boldsymbol{\theta}}$.

Due to its convex cost function, fast solvers with strong theoretical guarantees are available. Nevertheless, for large scale problems such methods may become computationally prohibitive and, for this reason, accelerating techniques are still an intense research topic.

In this paper we demonstrate how to combine two of such techniques:

1) Safe screening tests [1–5] allow to safely eliminate inactive dictionary atoms (those associated to zero entries in the solution vector $\hat{\boldsymbol{\beta}}$) before having complete knowledge of the solution $\hat{\boldsymbol{\beta}}$, with minor computational overhead. See Section 2 for more details.

2) Structured dictionaries [6–10] provide faster matrix-vector multiplications, which dominate the cost of typical iterative optimization algorithms for the Lasso, such as the iterative soft-thresholding algorithms (ISTA [11], FISTA [12]). Different types of structure have been proposed in the literature, e.g. product of sparse matrices [8], composition of circular convolutions [7], sums of Kronecker products [10], among others.

In real applications, the dictionary matrix may not fit the desired structure. A possible solution is to find a structured approximation $\tilde{\mathbf{X}}$ of the original dictionary $\mathbf{X}$, such that

$$\mathbf{X} = \tilde{\mathbf{X}} + \mathbf{E} \tag{3}$$

where the matrix $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_K] \in \mathbb{R}^{N\times K}$ captures the approximation error.

The idea is to start the iterative optimization by manipulating the fast structured approximation $\tilde{\mathbf{X}}$ to take advantage of its reduced multiplication cost while the solution is still coarsely estimated. As the algorithm approaches the solution, a better approximation of $\mathbf{X}$ and eventually the true dictionary $\mathbf{X}$ is used in order to ensure convergence to the right solution.

In [13], we showed how to obtain safe screening tests while manipulating an approximate version of the dictionary matrix and extended a particular screening test called DST1 [3] (Dynamic Spherical Test) to this new setting. In this work, we extend a more complex test called GAP SAFE [4] which is the current state-of-the-art in terms of screening capabilities. Besides that, we present experimental results on running time measurements while in the previous work [13] only theoretical complexity results were available.

This paper is organized as follows: the screening methodology is briefly reviewed in Section 2 and extended to approximate dictionaries in Section 3. The proposed algorithm is presented in Section 4 and simulation results in Section 5.

2. REMINDERS ON SAFE SCREENING

Only a small subset of the dictionary atoms – the ones corresponding to the support of the solution $\hat{\boldsymbol{\beta}}$ – are actually used in the reconstruction $\mathbf{X}\hat{\boldsymbol{\beta}}$ of the input signal $\mathbf{y}$. The other atoms, referred to as *inactive*, could be removed with no impact on the problem's solution while significantly simplifying it.

2.1. Screening tests

Although the support of $\hat{\boldsymbol{\beta}}$ is not known beforehand, *safe screening tests*, first proposed in [1], provide a way to identify inactive atoms with certainty before completely solving problem (1).

Given a *safe region* $\mathcal{R}$ containing the dual solution $\hat{\boldsymbol{\theta}}$, the *screening test* $\mu_{\mathcal{R}}(\mathbf{x}_j)$ on the atom $\mathbf{x}_j$ is defined as follows

$$\mu_{\mathcal{R}}(\mathbf{x}_j) := \sup_{\boldsymbol{\theta} \in \mathcal{R}} |\mathbf{x}_j^T \boldsymbol{\theta}| < 1 \quad \Longrightarrow \quad \hat{\boldsymbol{\beta}}_j = 0. \tag{4}$$

By evaluating $\mu_{\mathcal{R}}(\mathbf{x}_j)$ for all $j \in \{1, \ldots, K\}$, we are able to partition the atoms into a rejection set $\mathcal{A}^{\mathrm{c}}$ that gathers the indexes of all eliminated (surely inactive) atoms and its complementary, the (potentially) active set $\mathcal{A}$

$$\begin{aligned} \mathcal{A}^{\mathrm{c}} &= \{j \in \{1, \ldots, K\} : \mu_{\mathcal{R}}(\mathbf{x}_j) < 1\}, \\ \mathcal{A} &= \{j \in \{1, \ldots, K\} : \mu_{\mathcal{R}}(\mathbf{x}_j) \geq 1\}. \end{aligned} \tag{5}$$

Sphere tests In particular, when $\mathcal{R}$ is a closed ℓ_2-ball with center $\mathbf{c}$ and radius r, denoted $B(\mathbf{c}, r) = \{\mathbf{z} : \|\mathbf{z} - \mathbf{c}\|_2 \leq r\}$, the test has a closed form

$$\mu_{B(\mathbf{c},r)}(\mathbf{x}_j) = |\mathbf{x}_j^T \mathbf{c}| + r\|\mathbf{x}_j\|_2 < 1 \implies \hat{\boldsymbol{\beta}}_j = 0 \tag{6}$$

2.2. Static vs. Dynamic screening

Screening tests rely on the computation of safe regions. Two types of screening tests can be identified: 1) Static tests, which are performed once and for all before the optimization begins. 2) Dynamic tests, which are repeatedly applied during an iterative optimization algorithm leveraging its current solution estimate ($\boldsymbol{\beta}_t$ at iteration t) to gradually narrow the safe region and increase the number of eliminated atoms.

2.3. GAP SAFE sphere region

The screening test should be designed to entail minor computational overhead. Considering that the test in (6) has to be repeated for each one of the K atoms, the calculation of the term $|\mathbf{x}_j^T \mathbf{c}|$ comes down to a matrix-vector multiplication $\mathbf{X}^T \mathbf{c}$, which might be costly. For this reason, the safe region (center and radius) besides being as small as possible should also be defined so as to reuse previous calculations from the optimization algorithm.

A state-of-the-art dynamic safe spherical region was proposed in [4].

Theorem 1 (GAP SAFE sphere)**.** *For any* $(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t) \in \mathbb{R}^K \times \Delta_{\mathbf{X}}$, *denoting* $G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t) = P(\boldsymbol{\beta}_t) - D(\boldsymbol{\theta}_t)$ *the duality gap at iteration* t*, we have*

$$\hat{\boldsymbol{\theta}} \in B\left(\mathbf{c} = \boldsymbol{\theta}_t, r = \frac{1}{\lambda}\sqrt{2G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)}\right) \tag{7}$$

Proof. See [4], Theorem 2. □

Although this region is provably safe for any $\boldsymbol{\theta}_t \in \Delta_{\mathbf{X}}$ (dual feasible point), the authors, following [1], propose to use a scaled version of the current residual $\boldsymbol{\rho}_t = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}_t$:

$$\boldsymbol{\theta}_t = \alpha_t \boldsymbol{\rho}_t, \qquad \alpha_t = \left[\frac{\mathbf{y}^T \boldsymbol{\rho}_t}{\lambda \|\boldsymbol{\rho}_t\|_2^2}\right]_{-\frac{1}{\|\mathbf{X}^T \boldsymbol{\rho}_t\|_\infty}}^{\frac{1}{\|\mathbf{X}^T \boldsymbol{\rho}_t\|_\infty}}, \tag{8}$$

denoting $[z]_a^b := \min(\max(z, a), b)$ the projection of the scalar z onto the segment $[a, b]$.

The safe test that uses the region in (7) is also easily computable since the duality gap is often calculated as a stopping criterion and the matrix-vector product $\mathbf{X}^T \mathbf{c}$ becomes $\mathbf{X}^T \boldsymbol{\rho}_t$ which is part of the update of any gradient-based algorithm.

3. SCREENING WITH APPROXIMATE DICTIONARIES

In order to combine the screening techniques with the use of a fast approximate dictionary it is imperative to derive screening tests that manipulate $\tilde{\mathbf{X}}$ but remain safe with respect to the original problem (1), i.e. with respect to the atoms of $\mathbf{X}$.

3.1. Safe sphere tests with approximate dictionaries

For a given spherical safe region, one cannot just apply the test (6) to the approximate atoms, that is to test whether $\mu_{B(\mathbf{c},r)}(\tilde{\mathbf{x}}_j) < 1$, because the screening would be performed with respect to the approximate atoms, not the original ones, as desired. Intuitively, a certain "security margin" is required to account for the approximation error. Supposing $B(\mathbf{c}, r)$ a safe sphere (to be determined), we want to define a test $\tilde{\mu}_{B(\mathbf{c},r)}(\tilde{\mathbf{x}}_j)$ on the atoms $\tilde{\mathbf{x}}_j$ which is safe (with respect to $\mathbf{x}_j$), i.e $\tilde{\mu}_{B(\mathbf{c},r)}(\tilde{\mathbf{x}}_j) < 1 \implies \hat{\boldsymbol{\beta}}_j = 0$. In [13], we have shown that the following test is appropriate.

$$\tilde{\mu}_{B(\mathbf{c},r)}(\tilde{\mathbf{x}}_j) = |\tilde{\mathbf{x}}_j^T \mathbf{c}| + \|\mathbf{e}_j\|_2 \|\mathbf{c}\|_2 + r\|\mathbf{x}_j\|_2 \tag{9}$$

In [13] we extended the Dynamic Safe Test (DST1) [3] to the use of approximate dictionaries. In the next section, we extend the state-of-the-art dynamic test GAP SAFE [4], which was shown to considerably outperform DST1 in terms of screening performance.

3.2. GAP SAFE sphere with approximate dictionaries

Now that we have a safe test $\tilde{\mu}$ depending only on $\tilde{\mathbf{X}}$, we need to determine the safe region, i.e. $\mathbf{c}$ and r, as a function of $\tilde{\mathbf{X}}$ not $\mathbf{X}$.

A first step to be able to reuse the results in Theorem 1 is to obtain a dual feasible point $\tilde{\boldsymbol{\theta}}_t \in \Delta_{\mathbf{X}}$ (beware, we are interested in $\Delta_{\mathbf{X}}$ not $\Delta_{\tilde{\mathbf{X}}}$). Note that $\boldsymbol{\theta}_t$ cannot be calculated as in (8) since it depends on $\mathbf{X}$. However, the following dual point proportional to $\tilde{\boldsymbol{\rho}} = \mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}$ can be proven feasible (see [13] for a proof):

$$\tilde{\boldsymbol{\theta}}_t = \tilde{\alpha}_t \tilde{\boldsymbol{\rho}}_t, \quad \tilde{\alpha}_t = \left[\frac{\mathbf{y}^T \tilde{\boldsymbol{\rho}}_t}{\lambda \|\tilde{\boldsymbol{\rho}}_t\|_2^2}\right]_{-\frac{1}{\max_j\left(|\tilde{\mathbf{x}}_j^T \tilde{\boldsymbol{\rho}}_t| + \|\mathbf{e}_j\|_2 \|\tilde{\boldsymbol{\rho}}_t\|_2\right)}}^{\frac{1}{\max_j\left(|\tilde{\mathbf{x}}_j^T \tilde{\boldsymbol{\rho}}_t| + \|\mathbf{e}_j\|_2 \|\tilde{\boldsymbol{\rho}}_t\|_2\right)}} \tag{10}$$

Now, to construct a safe sphere with center $\tilde{\boldsymbol{\theta}}_t$ we employ the following reasoning (illustrated in Figure 1):

i) $\tilde{\boldsymbol{\theta}}_t$ is feasible with respect to $\mathbf{X}$, i.e. $\tilde{\boldsymbol{\theta}}_t \in \Delta_{\mathbf{X}}$.

ii) A GAP safe sphere (Thm. 1) would pick $r' = \frac{1}{\lambda}\sqrt{2G(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t)} = \frac{1}{\lambda}\sqrt{2(P(\boldsymbol{\beta}_t) - D(\tilde{\boldsymbol{\theta}}_t))}$, but it cannot be calculated since $P(\boldsymbol{\beta}_t)$ depends on $\mathbf{X}$.

iii) Instead, we calculate a *modified primal*

$$\tilde{P}(\boldsymbol{\beta}_t) = \|\tilde{\mathbf{X}}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{11}$$

Sphere test	Center	Radius
GAP SAFE	$\mathbf{c} = \boldsymbol{\theta}_t$	$r = \frac{1}{\lambda}\sqrt{2G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)}$
Extended GAP SAFE	$\tilde{\mathbf{c}} = \tilde{\boldsymbol{\theta}}_t$	$\tilde{r} = \frac{1}{\lambda}\sqrt{2\tilde{G}(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t) + 2\delta'}$

Table 1: Sphere center and radius for (extended) GAP SAFE. See Equations (7) and (8), (10) and (13).

which gives rise to a *modified duality gap*

$$\tilde{G}(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t) = \tilde{P}(\boldsymbol{\beta}_t) - D(\tilde{\boldsymbol{\theta}}_t). \tag{12}$$

We then define a radius $\tilde{r}$ as a function of $\tilde{G}(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t)$ such that $\tilde{r} \geq r'$ and, consequently, $B(\tilde{\boldsymbol{\theta}}_t, \tilde{r})$ is safe.

Theorem 2 (Extended GAP SAFE sphere)**.** *For any* $(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t) \in \mathbb{R}^K \times \Delta_{\mathbf{X}}$*, denoting* $\tilde{G}(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t) = \tilde{P}(\boldsymbol{\beta}_t) - D(\tilde{\boldsymbol{\theta}}_t)$ *the "modified" duality gap at iteration* t *and* $\|\cdot\|$ *the operator norm, we have*

$$\hat{\boldsymbol{\theta}} \in B\left(\tilde{\mathbf{c}} = \tilde{\boldsymbol{\theta}}_t, \tilde{r} = \frac{1}{\lambda}\sqrt{2\tilde{G}(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t) + 2\delta'}\right) \tag{13}$$

$$\textit{with} \qquad \delta' = \|\tilde{\boldsymbol{\rho}}_t\|_2 \|\mathbf{E}\| \|\boldsymbol{\beta}_t\|_2 + \frac{1}{2}\|\mathbf{E}\|^2 \|\boldsymbol{\beta}_t\|_2^2$$

Proof. By replacing the relation between $\mathbf{X}$ and $\tilde{\mathbf{X}}$ (eq. (3)) on the expression of the primal objective $P(\boldsymbol{\beta}_t)$, we obtain

$$\begin{aligned} P(\boldsymbol{\beta}_t) &= \frac{1}{2}\|(\tilde{\mathbf{X}} + \mathbf{E})\boldsymbol{\beta}_t - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}_t\|_1 \\ &= \tilde{P}(\boldsymbol{\beta}_t) + \frac{1}{2}\|\mathbf{E}\boldsymbol{\beta}_t\|_2^2 - \tilde{\boldsymbol{\rho}}_t^T(\mathbf{E}\boldsymbol{\beta}_t), \end{aligned}$$

which implies that

$$\tilde{G}(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t) - G(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t) = \underbrace{\tilde{\boldsymbol{\rho}}_t^T(\mathbf{E}\boldsymbol{\beta}_t) - \frac{1}{2}\|\mathbf{E}\boldsymbol{\beta}_t\|_2^2}_{-\delta}. \tag{14}$$

The right-hand side can be seen as a security margin δ which, when added to the $\tilde{G}(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t)$, makes it equal to $G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)$ and, thus, safe.

In practice, however, the calculation of δ is too computationally demanding since it requires the matrix-vector product $\mathbf{E}\boldsymbol{\beta}_t$. To avoid it, we adopt the following margin instead:

$$\delta' = \|\tilde{\boldsymbol{\rho}}_t\|_2 \|\mathbf{E}\| \|\boldsymbol{\beta}_t\|_2 + \frac{1}{2}\|\mathbf{E}\|^2 \|\boldsymbol{\beta}_t\|_2^2 \geq \delta \tag{15}$$

where we used the fact that $\|\mathbf{E}\boldsymbol{\beta}_t\|_2 \leq \|\mathbf{E}\| \|\boldsymbol{\beta}_t\|_2$ and $\|\mathbf{E}\|$ can be precalculated. Other bounds could also be used, e.g. $\|\mathbf{E}\boldsymbol{\beta}_t\|_2 \leq \|\boldsymbol{\beta}_t\|_1 \max_j(\|\mathbf{e}_j\|_2)$.

Since $\delta' \geq \delta$, then $\tilde{G}(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t) + \delta' \geq G(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t)$ and $\tilde{r} \geq r'$. Given that $B(\tilde{\boldsymbol{\theta}}_t, r')$ defines a safe sphere (from Theorem 1) $B(\tilde{\boldsymbol{\theta}}_t, r')$ also does. □

Table 1 summarizes the resulting extended GAP SAFE sphere region in comparison to the original one. Combining it with the test in (9), we obtain the extended GAP SAFE sphere test: $\tilde{\mu}_{B(\tilde{\mathbf{c}}, \tilde{r})}(\tilde{\mathbf{x}}_j) < 1 \implies \hat{\boldsymbol{\beta}}_j = 0$.

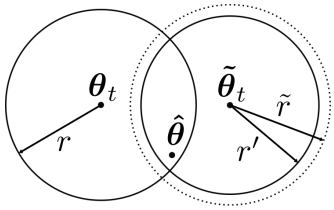

Fig. 1: GAP safe spheres centered at $\boldsymbol{\theta}_t$ and $\tilde{\boldsymbol{\theta}}_t$ (solid lines) and extended GAP sphere (dotted line) with a larger radius $\tilde{r}$.

4. PROPOSED ALGORITHM

The extended screening test developed in Section 3 can be combined to any first-order iterative optimization technique. In Algorithm 1, we combine it to an iterative shrinkage-thresholding algorithm (ISTA [11]) but it could also be FISTA [12], TwIST [14], Chambolle-Pock [15] and others.

In the initial iterations, the approximate fast dictionary is used, until a *switching criterion* is met. An iteration consists in a conventional ISTA update (lines 4-5) with step-size L_t set by the backtracking rule [12], followed by the screening test evaluation (lines 7-8) and application (line 9). We denote $\mathrm{ST}_u(x) = \mathrm{sign}(x)(|x| - u)_+$ the soft-thresholding operation and $\mathbf{X}_{[\mathcal{A}]}$ a sub-matrix of $\mathbf{X}$ composed of the columns indexed by the elements of $\mathcal{A}$. Similarly, $\boldsymbol{\beta}_{[\mathcal{A}]}$ is a vector containing the entries of $\boldsymbol{\beta}$ indexed by the elements of $\mathcal{A}$.

Algorithm 1 $\hat{\boldsymbol{\beta}} = \text{FastDynamicScreening}(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y}, \lambda)$

1: **Initialize:** $t = 0$, $\mathcal{A}_0 = \{1, \ldots, K\}$, $\tilde{\mathbf{X}}_0 = \tilde{\mathbf{X}}$, $\boldsymbol{\beta}_0 = \mathbf{0}$
2: **while** switching criterion not met **do**
3: —— *ISTA update* ——
4: $\tilde{\boldsymbol{\rho}}_{t+1} \leftarrow \mathbf{y} - \tilde{\mathbf{X}}_t\boldsymbol{\beta}_t$
5: $\boldsymbol{\beta}_{t+1} \leftarrow \mathrm{ST}_{\lambda/L_t}(\boldsymbol{\beta}_t + \frac{1}{L_t}\tilde{\mathbf{X}}_t^T\tilde{\boldsymbol{\rho}}_{t+1})$
6: —— *Screening* ——
7: Set $\tilde{\boldsymbol{\theta}}_t$ using (10)
8: $\mathcal{A}_{t+1} \leftarrow \{j \in \mathcal{A}_t : \tilde{\mu}_{B(\tilde{\mathbf{c}}, \tilde{r})}(\tilde{\mathbf{x}}_j) \geq 1\}$
9: $\tilde{\mathbf{X}}_{t+1} \leftarrow (\tilde{\mathbf{X}}_t)_{[\mathcal{A}_{t+1}]}$, $\boldsymbol{\beta}_{t+1} \leftarrow (\boldsymbol{\beta}_{t+1})_{[\mathcal{A}_{t+1}]}$
10: $t \leftarrow t + 1$
11: **end while**
12: —— *Switch to original* $\mathbf{X}$ ——
13: Repeat loop in lines 2-10 until convergence using $\tilde{\mathbf{X}}_t \leftarrow \mathbf{X}_{[\mathcal{A}_t]}$ and $\mu_{B(\mathbf{c}, r)}(\mathbf{x}_j)$ with $\boldsymbol{\theta}_t$ set using (8).

The switching point is controlled by two components: 1) a threshold on the screening ratio $|\mathcal{A}_t|$ (cardinality of the set $\mathcal{A}_t$), above which manipulating the true dictionary $\mathbf{X}$ is already cheap enough; 2) a threshold on the convergence, proportional to the approximation error. After switching back to $\mathbf{X}$, the conventional screening test is used. The stopping criterion after switching is a threshold on the duality gap.

5. EXPERIMENTAL RESULTS

In this section, the proposed methodology is evaluated through simulation time comparisons on synthetic data. We measured the time needed to solve the Lasso using a standard ISTA algorithm compared to the same algorithm with dynamic screening (GAP SAFE sphere)

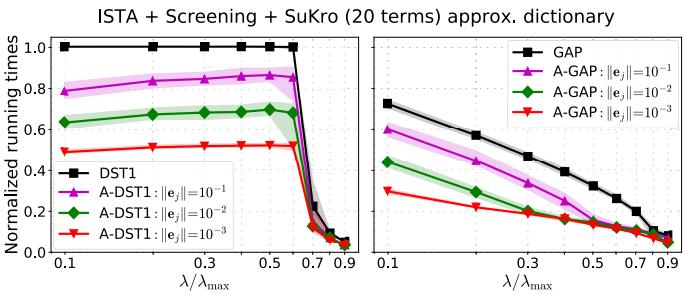

Fig. 2: Running times normalized with respect to ISTA without screening. Left: DST1 screening and its extended version A-DST1. Right: GAP SAFE and its extended version A-GAP.

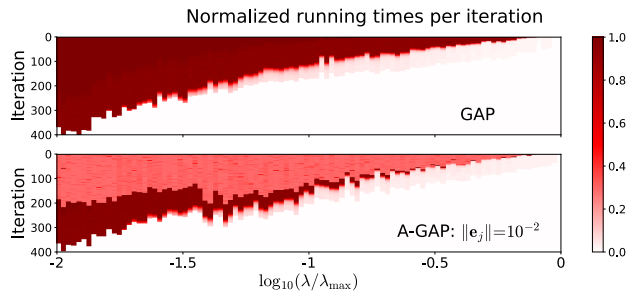

Fig. 3: Iteration times normalized w.r.t. the average iteration cost of ISTA without screening. Top: GAP SAFE sphere. Bottom: GAP SAFE sphere with approximate dictionary at initial iterations.

and the proposed technique (considering three different approximation errors).

In the experiments, we use a particular kind of fast structured dictionaries referred to as *SuKro* [10] which can be written as a sum of Kronecker products $\mathbf{X} = \sum_k \mathbf{A}_k \otimes \mathbf{B}_k$ and is particularly suited to two-dimensional signals (e.g. images). Its reduced multiplication cost comes essentially from the fact that the sub-matrices $\mathbf{A}_k$ and $\mathbf{B}_k$ are much smaller than $\mathbf{X}$ ($\sqrt{N} \times \sqrt{K}$ instead of $N \times K$).

The data dimension is set to $N = 2500$ and the number of atoms to $K = 10000$. We generate $\tilde{\mathbf{X}}$ as a SuKro dictionary with $k = 20$ terms $\mathbf{A}_k$ and $\mathbf{B}_k$ with size 50×100 and columns drawn i.i.d. uniformly on the unit sphere. The base dictionary $\mathbf{X}$ is then calculated by adding an error matrix $\mathbf{E}$ with columns drawn i.i.d zero-mean Gaussian all with identical ℓ_2-norms taking the values $\|\mathbf{e}_j\|_2 = \{10^{-1}, 10^{-2}, 10^{-3}\}$. The operator norm $\|\mathbf{E}\|$ is calculated and stored, taking values typically around $(2.99 \pm 3{\cdot}10^{-3})\|\mathbf{e}_j\|_2$.

An acceleration factor of 4 on matrix-vector multiplications was empirically measured for the SuKro dictionary compared to an unstructured one. We focus here on evaluating the influence of the approximation error, since the acceleration provided by the fast dictionary remains the same in all cases. In practice, higher approximation errors would rather lead to higher speedup factors. As a reference, in [8] accelerations of about 10 times are obtained with approximation errors around 10^{-2} for large MEG gain matrices, proving that the error-speedup compromises adopted here are quite realistic.

Unit-norm input data $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ is generated by using a sparse vector $\boldsymbol{\beta}$ with active set determined by a Bernoulli distribution with probability $p = 0.02$ and zero-mean Gaussian entries.

The switching criterion is composed by two thresholds:

1) Screening ratio: $|\mathcal{A}_t| < K/4$, denoting $|\mathcal{A}_t|$ the cardinality of the set $\mathcal{A}_t$ and 4 is the acceleration factor obtained by the fast dictionary in this case.
2) Convergence: $\tilde{G}(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t) < \max_j(\|\mathbf{e}_j\|_2)$.

After switching back to $\mathbf{X}$, the algorithm stops as soon as the duality gap is smaller than $5{\cdot}10^{-6}$.

Fig. 2 shows execution times normalized with respect to the standard ISTA algorithm without screening, as a function of $\lambda/\lambda_{\max}$[1]. The plotted values are the medians over 100 runs and the shaded areas contain the 25%-to-75% percentiles. The left graph shows the results obtained by the dynamic spherical test (DST1 [3]) and its extended version for approximate dictionaries (A-DST1 [13]), while the right graph presents the results obtained by the dynamic GAP SAFE sphere test [4] and its extended version (denoted A-GAP for Approximate GAP) for three different approximation errors.

First, notice that the GAP SAFE screening indeed leads to significant accelerations both with respect to no screening and to DST1. Even so, the proposed approach is still capable to further reduce the execution time, especially for lower regularizations $\lambda/\lambda_{\max}$.

Fig. 3 shows in a colormap the elapsed time per iteration normalized with respect to the average iteration cost of ISTA without screening. The results are presented as a function of the regularization level $\lambda/\lambda_{\max}$. Dark reds represent higher runtimes. The use of approximate fast dictionaries proves to be quite complementary to the screening tests, providing acceleration to the initial iterations of the optimization process (upper part of the graphs) when the screening tests are still ineffective. We can see that weakly regularized configurations (left part of the graphs) are particularly challenging. Not only do they require a greater amount of iterations to converge, but also the screening techniques struggle more to start acting. The fast approximate dictionary is particularly helpful in such scenarios.

6. CONCLUSION

In an effort to speed up the resolution of the Lasso problem especially for large scale scenarios, we proposed a method that combines screening tests and fast structured dictionaries on a first-order iterative optimization algorithm. We have shown how to extend a state-of-the art screening test to approximate dictionaries while keeping the safety of the test. Simulation results proves the effectiveness of the technique, leading to considerable running time reductions.

The proposed framework could also be extended to other sparsity-inducing inverse problems such as the Group-Lasso or the regularized logistic regression. Additional experiments with real datasets are a short-term perspective as well as handling multiple approximations of the dictionary with different error levels and complexity gains.

7. REFERENCES

[1] Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani, "Safe feature elimination in sparse supervised learning," *EECS Department, University of California, Berkeley, Tech. Rep*, 2010.

[2] Zhen James Xiang, Hao Xu, and Peter J Ramadge, "Learning sparse representations of high dimensional data on large scale dictionaries.," in *NIPS*, 2011, vol. 24, pp. 900–908.

[1] We denote $\lambda_{\max} := \|\mathbf{X}^T\mathbf{y}\|_\infty$. If $\lambda > \lambda_{\max}$ the primal solution is the zero vector.

[3] Antoine Bonnefoy, Valentin Emiya, Liva Ralaivola, and Rémi Gribonval, "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.

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

[5] Z. J. Xiang, Y. Wang, and P. J. Ramadge, "Screening tests for lasso problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2016.

[6] Ron Rubinstein, Michael Zibulevsky, and Michael Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *Signal Processing, IEEE Transactions on*, vol. 58, no. 3, pp. 1553–1564, 2010.

[7] Olivier Chabiron, Francois Malgouyres, Jean-Yves Tourneret, and Nicolas Dobigeon, "Toward fast transform learning," *International Journal of Computer Vision*, vol. 114, no. 2-3, pp. 195–216, 2015.

[8] L. Le Magoarou and R. Gribonval, "Flexible multilayer sparse approximations of matrices and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 688–700, June 2016.

[9] J. Sulam, B. Ophir, M. Zibulevsky, and M. Elad, "Trainlets: Dictionary learning in high dimensions," *IEEE Transactions on Signal Processing*, vol. 64, no. 12, pp. 3180–3193, June 2016.

[10] Cassio Dantas, Michele Nazareth da Costa, and Renato Lopes, "Learning dictionaries as a sum of kronecker products," *IEEE Signal Processing Letters*, 2017.

[11] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, no. 11, pp. 1413–1457, 2004.

[12] Amir Beck and Marc Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[13] Cassio F. Dantas and Rémi Gribonval, "Dynamic Screening with Approximate Dictionaries," in *XXVIème colloque GRETSI*, Juan-les-Pins, France, Sep 2017.

[14] J. M. Bioucas-Dias and M. A. T. Figueiredo, "A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2992–3004, Dec 2007.

[15] Antonin Chambolle and Thomas Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, May 2011.