Dynamic Screening with Approximate Dictionaries

Cassio Fraga Dantas, Rémi Gribonval

Dynamic Screening with Approximate Dictionaries

Cássio F. DANTAS, Rémi GRIBONVAL

INRIA Rennes-Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes, France
cassio.fraga-dantas@inria.fr, remi.gribonval@inria.fr

Résumé – Différentes stratégies pour accélérer la résolution du problème Lasso ont été proposées dernièrement. Notamment, les règles de *screening*, qui réduisent la dimensionalité du problème en permettent l'élimination de variables inutiles. Une autre technique consiste à approcher le dictionnaire par une matrice structurée plus rapide à manipuler. Cet article propose une façon de concilier ces deux stratégies. D'abord, on montre comment obtenir des règles de screening sûres vis-à-vis du problème exact en manipulant un dictionnaire approché. Ensuite, on adapte une règle de screening existante à ce nouveau cadre et on définit une procédure générale pour bien combiner les avantages des deux approches. Des réductions significatives de complexité ont été observées par rapport au screening isolé.

Abstract – Various strategies to accelerate the Lasso optimization have been recently proposed. Among them, screening rules provide a way to safely eliminate inactive variables, thus reducing the problem's dimensionality. Another line of work consists in replacing the dictionary matrix by a structured approximation of it, which is faster to manipulate. This paper proposes a method to conciliate both strategies. First, we show how to obtain safe screening rules for the exact problem while manipulating an approximate dictionary. We then adapt an existing screening rule to this new framework and define a general procedure to leverage the advantages of both strategies. Significant complexity reductions are obtained in comparison to screening rules alone.

1 Introduction

The ℓ_1-regularized least squares, referred to as Lasso, is a ubiquitous tool for variable selection in the context of underdetermined linear regression problems. By denoting $\mathbf{y} \in \mathbb{R}^N$ the observation vector and $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_K] \in \mathbb{R}^{N \times K}$ the design matrix (or dictionary), the Lasso problem consists in finding a sparse solution $\boldsymbol{\beta} \in \mathbb{R}^K$ of the following optimization problem :

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

where the parameter $\lambda > 0$ controls the trade-off between the data fidelity and sparsity of the solution. We suppose $\lambda \leq \lambda_{\text{max}} = \|\mathbf{X}^T\mathbf{y}\|_\infty$, since otherwise $\mathbf{0} \in \mathbb{R}^K$ is the unique solution.

This papers aims at combining two of the main approaches for accelerating the resolution of such sparsity-inducing optimization problems : 1) Screening techniques [1–5] provide safe rules for identifying inactive dictionary atoms on the optimum of a certain Lasso instance ; 2) Structured dictionaries [6, 7] lead to complexity savings on matrix-vector products, which are repeatedly performed on iterative thresholding optimization algorithms for the Lasso.

The overall idea is the following : starting the iterative Lasso optimization by using a structured approximation of the dictionary ($\tilde{\mathbf{X}}$) to take advantage of its reduced multiplication cost, and as the algorithm approaches the solution (and/or a considerable portion of the dictionary atoms have been screened out) switching back to the original dictionary.

A mandatory step for achieving this goal is determining how to obtain safe rules with respect to the Lasso problem (1) by manipulating an approximate version of the dictionary ($\tilde{\mathbf{X}}$).

In Section 2, we briefly recall the screening method, which we extend to approximate dictionaries in Section 3. The resulting optimization algorithm and simulation results are presented in Sections 4 and 5.

2 Screening tests

Screening tests rely on the dual formulation of the Lasso :

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\left\|\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda}\right\|_2^2 \quad \text{s.t.} \quad \|\mathbf{X}^T\boldsymbol{\theta}\|_\infty \leq 1. \tag{2}$$

where the dual solution $\hat{\boldsymbol{\theta}}$ is linked to a primal solution $\hat{\boldsymbol{\beta}}$ by the following relation $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \lambda\hat{\boldsymbol{\theta}}$.

Optimality conditions (KKT) at the dual solution $\hat{\boldsymbol{\theta}}$ imply that every dictionary atom for which $|\mathbf{x}_j^T\hat{\boldsymbol{\theta}}| < 1$ is not used on a Lasso solution and is referred to as *inactive*. Screening tests consist in using this fact to identify as many inactive atoms as possible before even having full knowledge of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$.

Since the dual problem optimal solution $\hat{\boldsymbol{\theta}}$ is not known, the scalar products $\mathbf{x}_j^T\hat{\boldsymbol{\theta}}$ cannot be evaluated. The idea is to identify a region $\mathcal{R}$, often called *safe region*, which is guaranteed to contain the optimal $\hat{\boldsymbol{\theta}}$. If for all $\boldsymbol{\theta} \in \mathcal{R}$ the inequality $|\mathbf{x}_j^T\boldsymbol{\theta}| < 1$ holds, then we can ensure that $\mathbf{x}_j$ is inactive. This sufficient condition can be expressed as the *screening test*, $\mu_{\mathcal{R}}(\mathbf{x}_j)$:

$$\mu_{\mathcal{R}}(\mathbf{x}_j) = \max_{\boldsymbol{\theta} \in \mathcal{R}} |\mathbf{x}_j^T\boldsymbol{\theta}| < 1 \quad \Longrightarrow \quad \hat{\boldsymbol{\beta}}_j = 0. \tag{3}$$

So, in practice, for each dictionary atom $\mathbf{x}_j$, we compute the test $\mu_{\mathcal{R}}(\mathbf{x}_j)$, and, depending on the result, we eliminate or not the atom. Formally, we are able to partition the atoms into a (potentially) active set

$$\mathcal{A} = \{j \in \{1, \dots, K\} : \mu_{\mathcal{R}}(\mathbf{x}_j) \geq 1\}, \tag{4}$$

and its complementary, the rejection set $\mathcal{A}^c$, that gathers the indexes of the eliminated atoms. Note that this is not a heuristic, in the sense that it will never lead to false rejections, hence its common denominations : *safe tests* or *safe rules*.

The region $\mathcal{R}$ might have different forms. The two most common in the literature are spheres [1–4] and domes [3,4] (i.e. intersection between a sphere and one or more half spaces).

Sphere tests In particular, when the safe region $\mathcal{R}$ 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

$$\mu_{B(\mathbf{c},r)}(\mathbf{x}_j) = |\mathbf{x}_j^T \mathbf{c}| + r\|\mathbf{x}_j\|_2 < 1 \implies \hat{\boldsymbol{\beta}}_j = 0 \quad (5)$$

The screening test should be designed so as not to entail a considerable computational overhead, after all, the goal is to reduce the cost of the Lasso resolution. Keep in mind that the test in (5) has to be repeated K times (one for each atom). Considering all tests, the calculation of the term $|\mathbf{x}_j^T \mathbf{c}|$ requires a matrix-vector multiplication $\mathbf{X}^T\mathbf{c}$, which might be costly. For this reason, screening techniques in the literature generally try to define the region (center and radius) so as to reuse calculations previously performed in the optimization algorithm.

Safe regions In this section, we recall ways to define a region which surely includes the dual solution $\hat{\boldsymbol{\theta}}$ of problem (2).

As iterative algorithms are often employed to solve the Lasso, the safe regions can be refined as the algorithm progresses. The associated tests are referred to as *dynamic*, as opposed to *static* tests in which a safe region is defined before the optimization begins and the screening is performed once and for all.

Note that the solution $\hat{\boldsymbol{\theta}}$ is the projection of $\mathbf{y}/\lambda$ on the feasible set $\{\boldsymbol{\theta} : \|\mathbf{X}^T\boldsymbol{\theta}\|_\infty \leq 1\}$ implying that, if a feasible point $\boldsymbol{\theta}_F$ is known, then $\hat{\boldsymbol{\theta}}$ can't be further away from $\mathbf{y}/\lambda$ than $\boldsymbol{\theta}_F$ in the ℓ_2 sense. This leads to the basic ℓ_2-spherical bound with center $\mathbf{c} = \mathbf{y}/\lambda$ and radius $r = \|\boldsymbol{\theta}_F - \mathbf{y}/\lambda\|_2$. Now the task comes down to determining a feasible point $\boldsymbol{\theta}_F$.

The static test in [1] is obtained by taking $\boldsymbol{\theta}_F = \mathbf{y}/\lambda_{\max}$ whose feasibility follows directly from the definition of $\lambda_{\max}$.

A dynamic safe region can be obtained by defining the feasible point at iteration t, $\boldsymbol{\theta}_t$, proportional to the current residuals $\boldsymbol{\rho}_t = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}_t$ [3]. We denote $[z]_a^b := \min(\max(z, a), b)$ the projection of the scalar z onto the segment $[a, b]$.

$$\begin{cases} \boldsymbol{\theta}_t = \alpha_t \boldsymbol{\rho}_t, \\ \alpha_t = \left[\dfrac{\mathbf{y}^T\boldsymbol{\rho}_t}{\lambda\|\boldsymbol{\rho}_t\|_2^2}\right]_{-\frac{1}{\|\mathbf{X}^T\boldsymbol{\rho}_t\|_\infty}}^{\frac{1}{\|\mathbf{X}^T\boldsymbol{\rho}_t\|_\infty}} \end{cases} \quad (6)$$

The resulting spherical region $B(\mathbf{y}/\lambda, \|\boldsymbol{\theta}_t - \mathbf{y}/\lambda\|_2)$ gives rise to the dynamic spherical test (DST1) introduced by [3].

3 Extending screening tests

Suppose that an approximate version $\tilde{\mathbf{X}}$ of the dictionary $\mathbf{X}$ is available (e.g. for faster matrix-vector product), such that

$$\mathbf{X} = \tilde{\mathbf{X}} + \mathbf{E}, \quad (7)$$

where $\mathbf{E}$ is the approximation error matrix. Each atom (column) $\tilde{\mathbf{x}}_j$ of $\tilde{\mathbf{X}}$ is thus a "distorted" version of the original atom $\mathbf{x}_j$, that is $\mathbf{x}_j = \tilde{\mathbf{x}}_j + \mathbf{e}_j$.

The question then arises : is it possible to provide safe tests with respect to the original Lasso problem (1) by manipulating $\tilde{\mathbf{X}}$ instead of $\mathbf{X}$?

Sphere tests with approximate dictionaries If a spherical safe region is given, one cannot simply apply the original test (5) to the approximate atoms, that is $\mu_{B(\mathbf{c},r)}(\tilde{\mathbf{x}}_j)$. It is intuitive to imagine that a certain "security margin" should be added to the test in (5) to account for the approximation error. By substituting (7) in (5), we obtain

$$\begin{aligned} \mu_{B(\mathbf{c},r)}(\mathbf{x}_j) &= |(\tilde{\mathbf{x}}_j + \mathbf{e}_j)^T\mathbf{c}| + r\|\mathbf{x}_j\|_2 \\ &\leq |\tilde{\mathbf{x}}_j^T\mathbf{c}| + \|\mathbf{e}_j\|_2\|\mathbf{c}\|_2 + r\|\mathbf{x}_j\|_2. \end{aligned} \quad (8)$$

Both $\|\mathbf{x}_j\|_2$ and $\|\mathbf{e}_j\|_2$ can be precalculated and stored in memory, leading to the definition of the following test on $\tilde{\mathbf{x}}_j$

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

Clearly, it is a safe test, since

$$\left(\mu_{B(\mathbf{c},r)}(\mathbf{x}_j) \leq\right) \tilde{\mu}_{B(\mathbf{c},r)}(\tilde{\mathbf{x}}_j) < 1 \implies \hat{\boldsymbol{\beta}}_j = 0 \quad (10)$$

Safe regions with approximate dictionaries Although manipulating an approximate version of the Lasso problem (1), we seek to define regions which are safe with respect to the *exact* problem (i.e. contain the dual solution $\hat{\boldsymbol{\theta}}$ of the *exact* dual problem (2) and *not necessarily* the dual solution of its approximate version), because we want the variable elimination to be done with respect to the *exact* atoms.

We now show how to obtain dual feasible points like (6) using the approximate dictionary $\tilde{\mathbf{X}}$. A feasible point requires that $\|\mathbf{X}^T\boldsymbol{\theta}_F\|_\infty = \max_j\left(|\mathbf{x}_j^T\boldsymbol{\theta}_F|\right) \leq 1$. If we suppose, once again, a feasible point in the form $\tilde{\boldsymbol{\theta}}_t = \tilde{\alpha}_t\tilde{\boldsymbol{\rho}}_t$ proportional to the current residuals $\tilde{\boldsymbol{\rho}}_t = \mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_t$, we have

$$|\mathbf{x}_j^T\tilde{\boldsymbol{\theta}}_t| \leq |\tilde{\alpha}_t|\left(|\tilde{\mathbf{x}}_j^T\tilde{\boldsymbol{\rho}}_t| + \|\mathbf{e}_j\|_2\|\tilde{\boldsymbol{\rho}}_t\|_2\right). \quad (11)$$

Therefore $\tilde{\boldsymbol{\theta}}_t$ is a feasible dual point for the original problem, i.e. $|\mathbf{x}_j^T\tilde{\boldsymbol{\theta}}_t| \leq 1\ \forall j$, as soon as

$$|\tilde{\alpha}_t| \leq \frac{1}{\max_j\left(|\tilde{\mathbf{x}}_j^T\tilde{\boldsymbol{\rho}}_t| + \|\mathbf{e}_j\|_2\|\tilde{\boldsymbol{\rho}}_t\|_2\right)} \quad (12)$$

which leads to the following definition for $\tilde{\boldsymbol{\theta}}_t$

$$\begin{cases} \tilde{\boldsymbol{\theta}}_t = \tilde{\alpha}_t\tilde{\boldsymbol{\rho}}_t, \\ \tilde{\alpha}_t = \left[\dfrac{\mathbf{y}^T\boldsymbol{\rho}_t}{\lambda\|\boldsymbol{\rho}_t\|_2^2}\right]_{-\frac{1}{\max_j\left(|\tilde{\mathbf{x}}_j^T\tilde{\boldsymbol{\rho}}_t| + \|\mathbf{e}_j\|_2\|\tilde{\boldsymbol{\rho}}_t\|_2\right)}}^{\frac{1}{\max_j\left(|\tilde{\mathbf{x}}_j^T\tilde{\boldsymbol{\rho}}_t| + \|\mathbf{e}_j\|_2\|\tilde{\boldsymbol{\rho}}_t\|_2\right)}} \end{cases} \quad (13)$$

This implies that $B(\mathbf{y}/\lambda, \|\tilde{\boldsymbol{\theta}}_t - \mathbf{y}/\lambda\|_2)$ is a safe region for the original problem. Combining it with the approximate test (9), we obtain a safe test analogous to DST1 that uses an approximate version of the dictionary instead (we will call it A-DST1).

Revisiting the intuition that a certain "security margin" would be necessary to a safe rule that uses an approximate dictionary, we can identify two locations where a margin was added : to adapt the test $\mu_{B(\mathbf{c},r)}$, and to calculate a feasible point $\tilde{\boldsymbol{\theta}}_t$.

Naturally, these "relaxed" rules often lead to fewer screened atoms, as will be illustrated in Section 5.

4 Algorithm and complexity

Algorithm 1 implements an iterative soft-thresholding (ISTA) optimization technique for the Lasso problem combined with a dynamic screening using an approximate dictionary. 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 $\mathcal{A}$. Similarly, $\boldsymbol{\beta}_{[\mathcal{A}]}$ is a vector containing the elements of $\boldsymbol{\beta}$ indexed by $\mathcal{A}$. The step-size L_t is set using the backtracking strategy as described in [8]. In practice, the screening can be performed at regular intervals instead of every iteration.

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

1: **Initialize :** $t = 0$, $\mathcal{A}_0 = \{1, \dots, K\}$, $\tilde{\mathbf{X}}_0 = \tilde{\mathbf{X}}$, $\boldsymbol{\beta}_0 = \mathbf{0}$
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 (13)
8: $\mathcal{A}_{t+1} \leftarrow \{j \in \mathcal{A}_t : \tilde{\mu}_{B(\mathbf{y}/\lambda, \|\tilde{\boldsymbol{\theta}}_t - \mathbf{y}/\lambda\|_2)}(\tilde{\mathbf{x}}_j) \geq 1\}$
9: $\tilde{\mathbf{X}}_{t+1} \leftarrow (\tilde{\mathbf{X}}_t)_{[\mathcal{A}_{t+1}]}, \quad \boldsymbol{\beta}_{t+1} \leftarrow (\boldsymbol{\beta}_{t+1})_{[\mathcal{A}_{t+1}]}$
10: $t \leftarrow t + 1$
11: **end while**
12: —— *Switch to original* $\mathbf{X}$ ——
13: Repeat loop in lines 2–10 until convergence using $\tilde{\mathbf{X}}_t = \mathbf{X}_{[\mathcal{A}_t]}$ and $\mu_{B(\mathbf{y}/\lambda, \|\boldsymbol{\theta}_t - \mathbf{y}/\lambda\|_2)}(\mathbf{x}_j)$ with $\boldsymbol{\theta}_t$ set using (6).

If the optimization loop (lines 2 to 8) was carried out until convergence, the solution of the approximate problem would be obtained. That's why, at some point, it is necessary to switch back to the dictionary $\mathbf{X}$ which guarantees the convergence to a solution of the original Lasso (1). At this point, the screening obtained with $\tilde{\mathbf{X}}$ can be safely applied to $\mathbf{X}$. For now, we do not specify any particular criterion to define the switching moment. This topic is further discussed in Section 4.2.

4.1 Complexity analysis

The screening test introduces only a negligible computational overhead because it relies primarily on the matrix-vector multiplications performed in the optimization algorithm update (namely $\tilde{\mathbf{X}}^T \tilde{\boldsymbol{\rho}}$, line 4) or that can be precalculated ($\tilde{\mathbf{X}}^T \mathbf{y}$). The other required calculations for the conventional screening add up to a total of $4N + 2|\mathcal{A}_t|$ operations (see [3] for more details).

Considering that the norms $\|\mathbf{x}_j\|_2$, $\|\mathbf{e}_j\|_2$ and the products $\|\mathbf{e}_j\|_2 \|\mathbf{c}\|_2$ (with $\mathbf{c} = \mathbf{y}/\lambda$) are precalculated, the approximate screening entails an additional cost of only $|\mathcal{A}_t|$ operations due to the products $\|\mathbf{e}_j\|_2 \|\boldsymbol{\rho}_t\|_2$ in the calculation of $\tilde{\alpha}_t$.

As desired, the screening represents a rather low overhead $\mathcal{O}(|\mathcal{A}_t| + N)$, compared to the optimization update which costs $\mathcal{O}(|\mathcal{A}_t| N)$. Without screening, this cost raises to $\mathcal{O}(KN)$.

In Table 1 we show the number of operations of a complete iteration in Algorithm 1 (ISTA update + screening), following [3]. We denote $\text{flops}_{\mathbf{X}}$ the iteration cost with the conventional screening, $\text{flops}_{\tilde{\mathbf{X}}}$ with the approximate screening and flops_N without screening. The *Relative Complexity* (RC) [6] quantifies the proportional complexity reduction entailed by the approximate dictionary, meaning that its multiplication by a vector costs $\text{RC} \times NK$ with $0 < \text{RC} \leq 1$ instead of NK. To simplify the analysis, we neglect the fact that screening may even further reduce the multiplication cost of the approximate dictionary.

TABLE 1 – Complete iteration complexity

flops_N	$(K + \|\boldsymbol{\beta}_t\|_0)N + 4K + N$
$\text{flops}_{\mathbf{X}}$	$(\|\mathcal{A}_t\| + \|\boldsymbol{\beta}_t\|_0)N + 6\|\mathcal{A}_t\| + 5N$
$\text{flops}_{\tilde{\mathbf{X}}}$	$(\text{RC} \times K + \|\boldsymbol{\beta}_t\|_0)N + 7\|\mathcal{A}_t\| + 5N$

4.2 Switching strategy

At an early optimization stage, the complexity gain provided by the fast dictionary $\tilde{\mathbf{X}}$ is very appealing. However, as more atoms are eliminated by the screening, this advantage may gradually fade. At a given point, the number of active atoms $|\mathcal{A}_t|$ may become so small that the use of $\tilde{\mathbf{X}}$ does not pay off anymore. This inspires a first criterion, which consists in switching as soon as the approximate screening reaches the threshold

$$|\mathcal{A}_t| < \text{RC}\ K \tfrac{N}{N-1}. \tag{14}$$

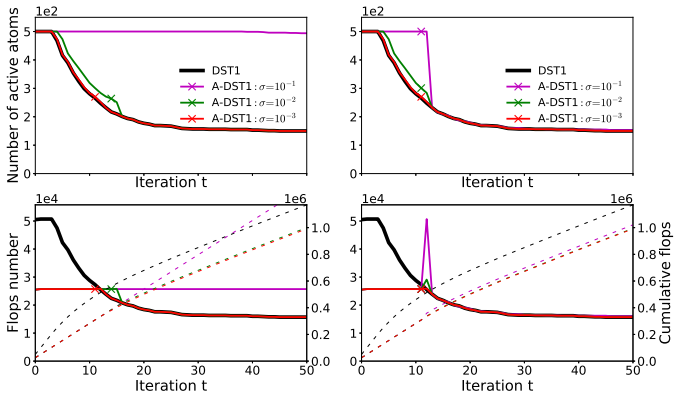
However, since the approximate screening rules lead to less atom eliminations than the conventional screening, such a switching can be delayed by some iterations (or even not happen at all, depending on the approximation error) as will be shown in Section 5 (Fig. 1a). We refer to this criterion as *naive criterion*.

As a more efficient heuristic, we propose to run two screening tests in parallel : a) the approximate test on the approximate atoms, $\tilde{\mu}_{\mathcal{R}}(\tilde{\mathbf{x}}_\mathbf{j})$, to screen the matrix-vector computations ; b) the *conventional* test on the approximate atoms, $\mu_{\mathcal{R}}(\tilde{\mathbf{x}}_\mathbf{j})$, whose screening level $|\mathcal{A}'_t|$ is used in equation (14) to decide when to switch. Although being unsafe with respect to the original problem in $\mathbf{X}$, it serves as a fairly good heuristic estimation of the original screening ratio. As will be shown in Section 5 (Fig. 1b), this new switching criterion, referred to as *improved criterion*, is much more robust to the approximation error. It has no impact on the safety of screening, which is performed with the first test, that is safe.

When λ is much smaller than λ_{max}, the screening level may remain forever above the threshold (14) simply because the Lasso solution $\hat{\boldsymbol{\beta}}$ (of both original and approximate problems) may not be sparse enough. Then, even with the approach just described, the switching would never take place. In order to avoid converging to the solution of the approximate problem, a convergence-based switching criterion can be used. The longer $\tilde{\mathbf{X}}$ is kept, the more the convergence to the truly sought solution might be delayed, especially for a high approximation error. Bearing that in mind, the following switching criterion, which leads to an earlier switching in the case of a higher approximation error, was set empirically as

$$\tfrac{|\mathcal{P}_t - \mathcal{P}_{t-1}|}{\mathcal{P}_t} < 10^{-2} \max_j(\|\mathbf{e}_j\|^2), \tag{15}$$

where $\mathcal{P}_t := \frac{1}{2}\|\tilde{\mathbf{X}}\boldsymbol{\beta}_t - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}_t\|_1$ is the primal objective of the approximate problem at iteration t.



(a) (b)

FIGURE 1 – Top : Evolution of number of active atoms with iterations, for $\lambda = 0.6\lambda_{\max}$. Bottom : flops per iteration. (a) naive switching (b) improved switching. 'x' indicates the switching point. Dashed lines show the cumulative number of flops.

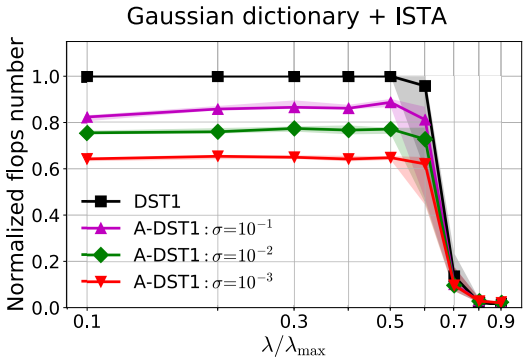


FIGURE 2 – Normalized number of flops as a function of $\lambda/\lambda_{\max}$. Lower values correspond to smaller complexities.

5 Simulation results

In this section, we provide some experiment results on synthetic data as a proof of concept for the proposed approach. We use a $N \times K$ dictionary with columns drawn i.i.d. uniformly on the unit sphere (100×500 in Fig. 1 and 1000×5000 in Fig. 2). We generate unit-norm observations $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ a sparse vector with Gaussian entries and active set determined by a Bernoulli distribution with probability $p = 0.02$. The ISTA algorithm stops when the relative variation of the cost function $\frac{|\mathcal{P}_t - \mathcal{P}_{t-1}|}{\mathcal{P}_t}$ is lower than 10^{-10}. We *mimic* acceleration with a structured approximate dictionary assumed to have RC $= 0.5$, and focus on evaluating the impact of the level of approximation error. For this we generate $\tilde{\mathbf{X}} = \mathbf{X} - \sigma\mathbf{E}$ with $\mathbf{E}$ a matrix with columns drawn i.i.d on the unit sphere, and $\sigma = \|\mathbf{e}_j\| = 10^{-1}, 10^{-2}, 10^{-3}$ respectively. As a reference, with approximation errors around 10^{-2}, accelerations of about 10 times are obtained in [6] for large MEG gain matrices.

Fig. 1 compares the naive and improved switching criteria. The top graphs show the number of remaining atoms across iterations for the screening rule DST1 (which manipulates $\mathbf{X}$) compared to its approximate counterparts (which manipulate $\tilde{\mathbf{X}}$), and the bottom graphs show the associated complexity cost by iteration ($\text{flops}_{\mathbf{X}}$ and $\text{flops}_{\tilde{\mathbf{X}}}$). The cumulated complexity is given by the area below the solid curve and is displayed as a dashed curve. As we can see before switching, the number of screened atoms at a given iteration is smaller for higher approximation errors σ. The improved switching brings considerable complexity savings in that it avoids the switching from being delayed, specially at high approximation errors σ. In particular, for $\sigma = 10^{-1}$, the naive criterion is never met.

Fig. 2 shows the normalized number of flops summed over all iterations ($\frac{\sum_{it} \text{flops}_{\mathbf{X}}}{\sum_{it} \text{flops}_N}$ or $\frac{\sum_{it} \text{flops}_{\tilde{\mathbf{X}}}}{\sum_{it} \text{flops}_N}$) as a function of $\lambda/\lambda_{\max}$. The medians among 10 runs are plotted and the shaded area contains the 25%-to-75% percentiles.

The proposed approach is always advantageous with respect to the conventional dynamic screening, specially for low and intermediate $\lambda/\lambda_{\max}$ values where an acceleration of up to 35% is reached. At high $\lambda/\lambda_{\max}$ values, the screening takes place within very few iterations and the approximate dictionary is quickly replaced by the original one. Finally, for the highest approximation error ($\sigma = 10^{-1}$) the advantage of the proposed approach is considerably mitigated, but it remains consistently better than the conventional screening.

6 Conclusion

We provided means to combine two accelerating strategies for the Lasso problem : screening rules and fast approximate dictionaries. Consistent complexity gains over existing techniques were observed in simulations. Although we restrained ourselves to sphere regions and a specific screening rule (DST1), the same approach can be adapted to other existing rules and region types. Future works include extending the GAP safe rule [4], handling multiple approximate dictionaries with different associated complexity gains as well as using real fast dictionaries (e.g. FAμST [6, 9]).

References

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

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

[3] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. 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] L. L. Magoarou and R. Gribonval, "Flexible multilayer sparse approximations of matrices and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, pp. 688–700, June 2016.

[7] C. Dantas, M. N. da Costa, and R. Lopes, "Learning dictionaries as a sum of kronecker products," *IEEE Signal Processing Letters*, 2017.

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

[9] L. L. Magoarou, R. Gribonval, and A. Gramfort, "FAμST: Speeding up linear transforms for tractable inverse problems," in *23rd European Signal Processing Conference (EUSIPCO)*, Aug 2015.