Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso

Antoine Bonnefoy, Valentin Emiya, Liva Ralaivola, Rémi Gribonval

To cite this version:

Antoine Bonnefoy, Valentin Emiya, Liva Ralaivola, Rémi Gribonval. Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso. IEEE Transactions on Signal Processing, Institute of Electrical and Electronics Engineers, 2015, 63 (19), pp.20. 10.1109/TSP.2015.2447503 . hal-01084986

HAL Id: hal-01084986

https://hal.archives-ouvertes.fr/hal-01084986

Submitted on 25 Nov 2014

Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso

Antoine Bonnefoy, Valentin Emiya, Liva Ralaivola, Rémi Gribonval.

Abstract

Recent computational strategies based on screening tests have been proposed to accelerate algorithms addressing penalized sparse regression problems such as the Lasso. Such approaches build upon the idea that it is worth dedicating some small computational effort to locate inactive atoms and remove them from the dictionary in a preprocessing stage so that the regression algorithm working with a smaller dictionary will then converge faster to the solution of the initial problem. We believe that there is an even more efficient way to screen the dictionary and obtain a greater acceleration: inside each iteration of the regression algorithm, one may take advantage of the algorithm computations to obtain a new screening test for free with increasing screening effects along the iterations. The dictionary is henceforth dynamically screened instead of being screened statically, once and for all, before the first iteration. We formalize this dynamic screening principle in a general algorithmic scheme and apply it by embedding inside a number of first-order algorithms adapted existing screening tests to solve the *Lasso* or new screening tests to solve the *Group-Lasso*. Computational gains are assessed in a large set of experiments on synthetic data as well as real-world sounds and images. They show both the screening efficiency and the gain in terms running times.

Index Terms

Screening test, Dynamic screening, Lasso, Group-Lasso, Iterative Soft Thresholding, Sparsity.

I. Introduction

In this paper, we focus on the numerical solution of optimization problems that consist in minimizing the sum of an ℓ_2-fitting term and a sparsity-inducing regularization term. Such problems are of the form:

$$\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y}) : \tilde{\mathbf{x}} \triangleq \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{D}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\Omega(\mathbf{x}), \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^N$ is an observation; $\mathbf{D} \in \mathbb{R}^{N\times K}$ with $N \leq K$ is a matrix called dictionary; $\Omega : \mathbb{R}^K \to \mathbb{R}_+$ is a convex sparsity-inducing regularization function; and $\lambda > 0$ is a parameter that governs the tradeoff between data fidelity and regularization. Various convex and non-smooth functions Ω may induce the sparsity of the solution $\tilde{\mathbf{x}}$. Here, we consider two instances of problem (1), the *Lasso* [1] and the *Group-Lasso* [2], which differ from each other in the choice of the regularization Ω. A key challenge is to handle (1) when both N and K may be large, which occurs in many real-world applications including denoising [3], inpainting [4] or classification [5]. Algorithms relying on first-order information only, *i.e.* gradient-based procedures [6], [7], [8], [9], are particularly suited to solve these problems, as second-order based methods —*e.g.* using the Hessian— imply too computationally demanding iterations. In the ℓ_2 data-fitting case the gradient relies on the application of the operator $\mathbf{D}$ and its transpose $\mathbf{D}^T$. We use this feature to define *first order* algorithms as those based on the application of $\mathbf{D}$ and $\mathbf{D}^T$. The definition extends to primal-dual algorithm [10], [11].

Accelerating these algorithms remains challenging: even though they provably have fast convergence [12], [6], the multiplications by $\mathbf{D}$ and $\mathbf{D}^T$ in the optimization process are a bottleneck in their computational efficiency, which is thus governed by the dictionary size. We are interested in the general case where no fast transform is associated with $\mathbf{D}$. This occurs for instance with exemplar-based or learned dictionaries. Such accelerations are even more needed during the process of learning high-dimensional dictionaries, which requires solving many problems of the form (1).

This work was supported by Agence Nationale de la Recherche (ANR), project GRETA 12-BS02-004-01. R.G. acknowledges funding by the European Research Council within the PLEASE project under grant ERC-StG-2011-277906.

Algorithm 1
Static screening strategy

```
$\mathbf{D}_0 \leftarrow$ Screen $\mathbf{D}$
loop t
    $\mathbf{x}_{t+1} \leftarrow$Update $\mathbf{x}_t$ using $\mathbf{D}_0$
end loop
```

Algorithm 2
Dynamic screening strategy

```
$\mathbf{D}_0 \leftarrow \mathbf{D}$
loop t
    $\mathbf{x}_{t+1} \leftarrow$ Update $\mathbf{x}_t$ using $\mathbf{D}_t$
    $\mathbf{D}_{t+1} \leftarrow$ Screen $\mathbf{D}_t$ using $\mathbf{x}_{t+1}$
end loop
```

Screening Tests: The convexity of the objective function suggests to use the theory of convex duality [13] to understand and solve such problems. In this context, strategies based on *screening tests* [14], [15], [16], [17], [18], [19] have recently been proposed to reduce the computational cost by considering properties of the dual optimum. Given that the sparsity-inducing regularization Ω entails an optimum $\tilde{\mathbf{x}}$ that may contain many zeros, a screening test is a method aimed at locating a subset of such zeros. It is formally defined as follows:

Definition 1 (Screening test)**.** Given problem $\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y})$ with solution $\tilde{\mathbf{x}}$, a boolean-valued function $T : [1 \dots K] \rightarrow \{0, 1\}$ is a screening test if and only if:

$$\forall i \in [1 \dots K], T(i) = 1 \Rightarrow \tilde{\mathbf{x}}(i) = 0. \tag{2}$$

We assume that solution $\tilde{\mathbf{x}}$ is unique —*e.g.*, it is true with probability one for the *Lasso* if $\mathbf{D}$ is drawn from a continuous distribution [20]. In general, a screening test cannot detect all zeros in $\tilde{\mathbf{x}}$, that is why relation (2) is not an equivalence. An efficient screening test locates many zeros among those of $\tilde{\mathbf{x}}$. From a screening test T a screened dictionary $\mathbf{D}_0 = \mathbf{D}\mathbf{T}$ is defined, where the matrix $\mathbf{T}$ removes from $\mathbf{D}$ the *inactive* atoms corresponding to the zeros located by the screening test T. $\mathbf{T}$ is built by removing, from the $K \times K$ identity matrix, the columns corresponding to screened atoms *i.e.* columns indexed by $i \in [1 \dots K]$ whenever verifies $T(i) = 1$. Property (2) implies that $\tilde{\mathbf{x}}_0$, the solution of problem $\mathcal{P}(\lambda, \Omega, \mathbf{D}_0, \mathbf{y})$, is exactly the same as $\tilde{\mathbf{x}}$ the solution of $\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y})$ up to inserting zeros at the locations of the removed atoms: $\tilde{\mathbf{x}} = \mathbf{T}\tilde{\mathbf{x}}_0$. Any optimization procedure solving problem $\mathcal{P}(\lambda, \Omega, \mathbf{D}_0, \mathbf{y})$ with the screened dictionary $\mathbf{D}_0$ therefore computes the solution of $\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y})$ at a lower computational cost. Algorithm 1 depicts the commonly used strategy [15], [18] to obtain an algorithmic acceleration using a screening test; it rests upon two steps: i) locate some zeros of $\tilde{\mathbf{x}}$ thanks to a screening test and construct the *screened* dictionary $\mathbf{D}_0$ and ii) solve $\mathcal{P}(\lambda, \Omega, \mathbf{D}_0, \mathbf{y})$ using the smaller dictionary $\mathbf{D}_0$.

Dynamic Screening: We propose a new screening principle called *dynamic screening* in order to reduce even more the computational cost of first-order algorithms. We take the aforementioned concept of screening test one step further, and improve existing screening tests by *embedding* them in the iterations of first-order algorithms. We take advantage of the computations made during the optimization procedure to perform a new screening test at each iteration with a *negligible* computational overhead, and we consequently *dynamically* reduce the size of $\mathbf{D}$. For a schematic comparison, the existing *static* screening and the proposed *dynamic* screening are sketched in Algorithms 1 and 2, respectively. One may observe that with the dynamic screening strategy, the dictionary $\mathbf{D}_t$ used at each iteration t is possibly smaller and smaller thanks to successive screenings.

Illustration: We now present a brief illustration of the dynamic screening principle in action, dedicated to the impatient reader who wishes to get a good grasp of the approach without having to enter the mathematics in too much detail. The dynamic screening principle is illustrated in Figure 1 in the particular case of the combined use of ISTA [8] and of a new, dynamic version of the SAFE screening test [15] to solve the *Lasso* problem, which is (1) with $\Omega(\mathbf{x}) = \|\mathbf{x}\|_1$. The screening effect is illustrated through the evolution of the size of the dictionary, *i.e.*, the number of atoms remaining in the screened dictionary $\mathbf{D}_t$. In this example the observation $\mathbf{y}$ and all the $K = 50000$ atoms are vectors drawn uniformly and independently on the unit sphere in dimension $N = 5000$ of $\mathbf{D}$, and we set $\lambda = 0.75 \times \left\|\mathbf{D}^T\mathbf{y}\right\|_\infty$. Here,

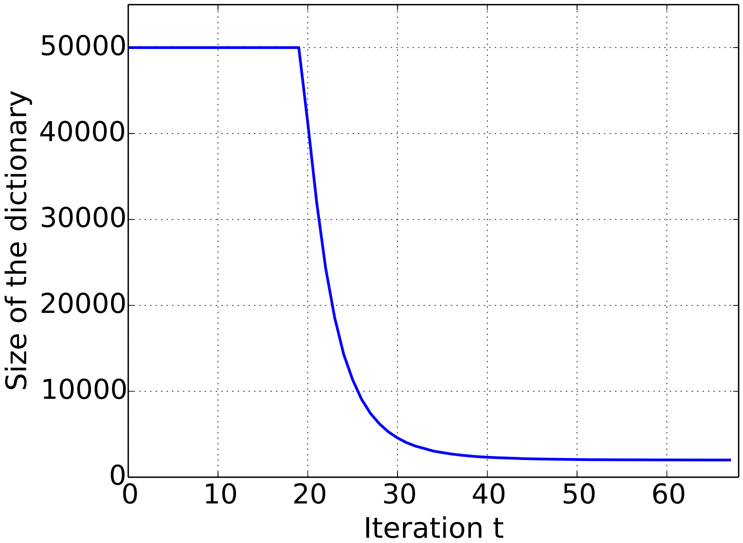

Figure 1: Size of the dictionary $\mathbf{D}_t$ (number of atoms) as a function of the iteration t in a basic dynamic-screening setting, starting with a dictionary with $K = 5000$ atoms.

the actual screening begins to have an effect at iteration $t = 20$. In about ten iterations, the dictionary is dynamically screened down to 5% of its initial size and the computational cost of each iteration gets lower and lower. Then, the last iterations are performed with a reduced computational cost. Consequently, the total running time equals 4.6 seconds while it is 11.8 seconds if no screening is used. One may also observe that the screening test is inefficient in the first 20 iterations. In particular, the dynamic screening at the very first iteration is strictly equivalent to the state-of-the-art (static) screening. Static screening would have been of no use in this case.

Contributions: This paper is an extended version of [21]. Here we propose new screening tests for the *Group-Lasso* and give an unified formulation of the dynamic screening principle for both *Lasso* and *Group-Lasso*, which improves screening tests for both problems. The algorithmic contributions and related theoretical results are introduced in Section II. First, the dynamic screening principle is formalized in a general algorithmic scheme (Algorithm 3) and its convergence is established (Theorem 1). Then, we show how to instantiate this scheme for several first-order algorithms (Section II-B1), as well as for the two considered problems the *Lasso* and *Group-Lasso* (Sections II-B2 and II-B3). We adapt existing tests to make them dynamic for the *Lasso* and propose new screening tests for the *Group-Lasso*. A turnkey instance of the proposed approach is given in Section II-C. Finally, the computational complexity of the dynamic screening scheme is detailed and discussed in Section II-D. In Section III, experiments show how the dynamic screening *principle* significantly reduces the computational cost of first-order optimization algorithms for a large range of problem settings and algorithms. We conclude this paper by a discussion in Section IV. All proofs are given in Appendix.

II. Generalized Dynamic Screening

Notation and definitions: $\mathbf{D} \triangleq [\mathbf{d}_1, \ldots, \mathbf{d}_K] \in \mathbb{R}^{N \times K}$ denotes a *dictionary* and $\Gamma \triangleq \{1, \ldots, K\}$ denotes the set of integers indexing the columns, or *atoms*, of $\mathbf{D}$. The i-th component of $\mathbf{x}$ is denoted by $\mathbf{x}(i)$. For a given set $\mathcal{I} \subset \Gamma$, $|\mathcal{I}|$ is the cardinal of $\mathcal{I}$, $\mathcal{I}^c \triangleq \Gamma \backslash \mathcal{I}$ is the complement of $\mathcal{I}$ in Γ and $\mathbf{D}_{[\mathcal{I}]} \triangleq [\mathbf{d}_i]_{i \in \mathcal{I}}$ denotes the sub-dictionary composed of the atoms indexed by elements of $\mathcal{I}$. The notation extends to vectors: $\mathbf{x}_{[\mathcal{I}]} \triangleq [\mathbf{x}(i)]_{i \in \mathcal{I}}$. Given two index sets $\mathcal{I}, \mathcal{J} \subset \Gamma$ and a matrix $\mathbf{M}$, $\mathbf{M}_{[\mathcal{I},\mathcal{J}]} \triangleq [\mathbf{M}(i,j)]_{(i,j) \in \mathcal{I} \times \mathcal{J}}$ denotes the sub-matrix of $\mathbf{M}$ obtained by selecting the rows and columns indexed by elements in $\mathcal{I}$ and $\mathcal{J}$, respectively. We denote primal variables vectors by $\mathbf{x} \in \mathbb{R}^K$ and dual variables vectors by $\boldsymbol{\theta} \in \mathbb{R}^N$. We denote by $[r]_a^b \triangleq \max(\min(r, b), a)$ the projection of r onto the segment $[a, b]$. Without loss of generality, the observation $\mathbf{y}$ and the atoms $\mathbf{d}_i$ are assumed to have unit ℓ_2 norm. For any matrix $\mathbf{M}$, $\|\mathbf{M}\|$ denotes its spectral norm, *i.e.*, its largest singular value.

A. Proposed general algorithm with dynamic screening

Dynamic screening is dedicated to accelerate the computation of the solution of problem (1). It is presented here in a general way.

Let us consider a problem $\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y})$ as defined by eq. (1). First-order algorithms are iterative optimization procedures that may be resorted to solving $\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y})$. Based only on applications of $\mathbf{D}$ and $\mathbf{D}^T$, they build a sequence of iterates $\mathbf{x}_t$ that converges, either in terms of objective values or the iterate itself, to the solution of the problem. In the following, we use the update step function $p(\cdot)$ as a generic notation to refer to any first-order algorithm. The optimization procedure might be formalized as the update

$$(\mathbf{X}_t, \boldsymbol{\theta}_t, \boldsymbol{\alpha}_t) \leftarrow p(\mathbf{X}_{t-1}, \boldsymbol{\theta}_{t-1}, \boldsymbol{\alpha}_{t-1}, \mathbf{D})$$

of several variables. Matrix $\mathbf{X}_t$ is composed of one or several columns that are primal variables and from which one can extract $\mathbf{x}_t$; vector $\boldsymbol{\theta}_t$ is an updated variable of the dual space $\mathbb{R}^N$ —in the sense of convex duality (see II-B2 equation (5) for details); and $\boldsymbol{\alpha}_t$ is a list of updated auxiliary scalars in $\mathbb{R}$.

Algorithm 3 makes explicit the use of the introduced notation $p(\cdot)$ in the general scheme of the dynamic screening principle. The inputs are: the data that characterize problem $\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y})$; the update function $p(\cdot)$ related to the first-order algorithm to be accelerated; an initialization $\mathbf{X}_0$; and a family of screening tests $\{T_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \mathbb{R}^N}$ in the sense[1] of Definition 1. Iteration t begins with an update step at line 4. It is followed by a screening stage in which a screening test $T_{\boldsymbol{\theta}_t}$ is computed using the dual point $\boldsymbol{\theta}_t$ obtained during the update. As shown at line 6, it enables to detect new inactive atoms and to update index set $\mathcal{I}_t$ that gathers all atoms identified as inactive so far. This set is then used to screen the dictionary and the primal variables at lines 7 and 8 using the screening matrix $\mathbf{Id}_{[\mathcal{I}_t^c, \mathcal{I}_{t-1}^c]}$ —obtained by removing columns $\mathcal{I}_{t-1}$ and rows $\mathcal{I}_t$ from the $K \times K$-identity matrix— and its transpose. Thanks to successive screenings, the dimension shared by the primal variables and the dictionary is decreasing and the optimization update can be computed in the reduced dimension $|\mathcal{I}_t^c|$, at a lower computational cost. The acceleration is efficient because lines 6 to 8 have negligible computation impact, as shown in Section II-D and assessed experimentally in Section III.

Algorithm 3 General algorithm with dynamic screening

```
Require: $\mathbf{D}, \mathbf{y}, \lambda, \Omega, \mathbf{X}_0$, screening test $T_{\boldsymbol{\theta}}$ for any $\boldsymbol{\theta} \in \mathbb{R}^N$ and first-order update $p(\cdot)$.
1: $\mathbf{D}_0 \leftarrow \mathbf{D}$, $\mathcal{I}_0 \leftarrow \emptyset, t \leftarrow 1, \bar{\mathbf{X}}_0 \leftarrow \mathbf{X}_0$
2: while stopping criteria on $\mathbf{X}_t$ do
3:     ...................... Optimization Update ......................
4:     $(\mathbf{X}_t, \boldsymbol{\theta}_t, \boldsymbol{\alpha}_t) \leftarrow p(\bar{\mathbf{X}}_{t-1}, \boldsymbol{\theta}_{t-1}, \boldsymbol{\alpha}_{t-1}, \mathbf{D}_{t-1})$
5:     .............................. Screening ...............................
6:     $\mathcal{I}_t \leftarrow \{i \in \Gamma, T_{\boldsymbol{\theta}_t}(i)\} \cup \mathcal{I}_{t-1}$
7:     $\mathbf{D}_t \leftarrow \mathbf{D}_{t-1}\mathbf{Id}_{[\mathcal{I}_{t-1}^c, \mathcal{I}_t^c]}$
8:     $\bar{\mathbf{X}}_t \leftarrow \mathbf{Id}_{[\mathcal{I}_t^c, \mathcal{I}_{t-1}^c]}\mathbf{X}_t$
9:     $t \leftarrow t + 1$
10: end while
11: return $\mathbf{X}_t$
```

Since the optimization update at line 4 is performed in a reduced dimension, the iterates generated by the algorithm with dynamic screening may differ from those of the base first-order algorithm. The following

[1] Algorithm 3 uses notation with screening tests indexed by a dual point $\boldsymbol{\theta}$ however the proposed Algorithm 3 is valid in a more general case for any screening test T that may be designed. We use the notation $T_{\boldsymbol{\theta}}$ since in this paper —and in the literature— screening tests are based on a dual point.

results states that the proposed general algorithm with dynamic screening preserves the convergence to the global minimum of problem (1).

Theorem 1. *Let Ω be a convex and sparsifying regularization function and $p(\cdot)$ the update function of an iterative algorithm. If $p(\cdot)$ is such that for any $\mathbf{D}, \mathbf{y}, \lambda$, the sequence of iterates given by $p(\cdot)$ converges to the solution of $\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y})$, then, for any family of screening tests $\{T_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \mathbb{R}^N}$, the general algorithm with dynamic screening (Algorithm 3) converges to the same global optimum of problem $\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y})$.*

Proof. Since $\forall \boldsymbol{\theta} \in \mathbb{R}^N, T_{\boldsymbol{\theta}}$ is a screening test, problems $\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y})$ and $\mathcal{P}(\lambda, \Omega, \mathbf{D}_t, \mathbf{y})$ for all $t \geq 0$ have the same solution. The sequence $\{\mathcal{I}_t\}_{t \geq 0}$ of located zeros at time t is inclusion-wise non-decreasing and upper bounded by the set of zeros in $\tilde{\mathbf{x}}$ the solution of $\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y})$, so the sequence converges in a finite number of iterations t_0. Then $\forall t \geq t_0, \mathbf{D}_{t_0} = \mathbf{D}_t$ and the existing convergence proofs of the first-order algorithm with update $p(\cdot)$ apply. □

B. Instances of algorithm with dynamic screening

The general form of Algorithm 3 may be instantiated for various algorithms, problems and screening tests. The general form with respect to first-order algorithms is instantiated for a number of possible algorithms in Section II-B1. The algorithm may be applied to solve the *Lasso* and *Group-Lasso* problems through the choice of the regularization Ω. Instances of screening tests $T_{\boldsymbol{\theta}}$ associated to each problem are described in Sections II-B2 and II-B3 respectively. Proofs are postponed to the appendices for readability purposes.

1) First-order algorithm updates: The dynamic screening principle may accelerate many first-order algorithms. Table I specifies how to use Algorithm 3 for several first-order algorithms, namely ISTA [8], TwIST [22], SpaRSA [9], FISTA[6] and Chambolle-Pock [11]. This specification consists in defining the primal variables $\mathbf{X}_t$, the dual variable $\boldsymbol{\theta}_t$, the possible additional variables $\boldsymbol{\alpha}_t$ and the update $p(\cdot)$ used at line 4.

Table I shows two important aspects of first-order algorithms: first, the notation is general and many algorithms may be formulated in this way; second, every $p(\cdot)$ has a computational complexity in $\mathcal{O}(KN)$ per iteration.

Table I makes use of proximal operators $\operatorname{prox}_{\lambda}^{\Omega}(\mathbf{x}) \triangleq \min_{\mathbf{z}} \Omega(\mathbf{z}) + \frac{1}{2\lambda} \|\mathbf{x} - \mathbf{z}\|_2^2$ that handle the non-smoothness of the objective function introduced by the regularization Ω. Beyond the subsequent definitions of the proximal operators for the *Lasso* (see eq. (4)) and the *Group-Lasso* (see eq. (13)), we refer the interested reader to [7] for a full description of proximal methods.

2) Dynamic screening for the Lasso: Let us first recall the *Lasso* problem before giving the screening tests $T_{\boldsymbol{\theta}}$ that may be embedded in the corresponding optimization procedure.

a) The Lasso [1]: The *Lasso* problem uses the ℓ_1-norm penalization to enforce a sparse solution. The *Lasso* is exactly (1) using $\Omega(\mathbf{x}) = \|\mathbf{x}\|_1$:

$$\mathcal{P}^{\textit{Lasso}}(\lambda, \mathbf{D}, \mathbf{y}) : \arg\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{D}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{x}\|_1 . \tag{3}$$

The proximal operator of the ℓ_1-norm is the so-called soft-thresholding operator:

$$\operatorname{prox}_t^{\textit{Lasso}}(\mathbf{x}) \triangleq \operatorname{sign}(\mathbf{x}) \max(|\mathbf{x}| - t, 0). \tag{4}$$

Screening tests [15], [19], [18] rely on the dual formulation of the *Lasso* problem:

$$\tilde{\boldsymbol{\theta}} \triangleq \arg\max_{\boldsymbol{\theta}} \frac{1}{2} \|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2} \left\| \boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda} \right\|_2^2 \tag{5a}$$

$$\text{s.t. } \forall i \in \Gamma, |\boldsymbol{\theta}^T \mathbf{d}_i| \leq 1. \tag{5b}$$

A dual point $\boldsymbol{\theta} \in \mathbb{R}^N$ is said *feasible* for the *Lasso* if it complies with constraints (5b).

Algorithm	Nature of $\{\mathbf{X}_t, \boldsymbol{\alpha}_t\}$	Optimization *update* $\{\mathbf{X}_t, \boldsymbol{\theta}_t\} \leftarrow p(\mathbf{X}_{t-1}, \boldsymbol{\theta}_{t-1}, \boldsymbol{\alpha}_{t-1}, \mathbf{D})$
ISTA [8]	$\mathbf{X}_t \triangleq \mathbf{x}_t, \boldsymbol{\alpha}_t \triangleq L_t$	$\boldsymbol{\theta}_t \leftarrow \mathbf{D}\mathbf{x}_{t-1} - \mathbf{y}$ $\mathbf{x}_t \leftarrow \mathrm{prox}^{\Omega}_{\frac{\lambda}{L_t}} \left(\mathbf{x}_{t-1} - \frac{1}{L_t}\mathbf{D}^T\boldsymbol{\theta}_t\right)$ L_t is set with the backtracking rule see [6]
TwIST [22]	$\mathbf{X}_t \triangleq [\mathbf{x}_t, \mathbf{x}_{t-1}], \boldsymbol{\alpha}_t \triangleq \emptyset$	$\boldsymbol{\theta}_t \leftarrow \mathbf{D}\mathbf{x}_{t-1} - \mathbf{y}$ $\mathbf{x}_t \leftarrow (1-\alpha)\mathbf{x}_{t-2} + (\alpha - \beta)\mathbf{x}_{t-1} + \beta\, \mathrm{prox}^{\Omega}_{\lambda} \left(\mathbf{x}_{t-1} - \mathbf{D}^T\boldsymbol{\theta}_t\right)$
SpaRSA [9]	$\mathbf{X}_t \triangleq [\mathbf{x}_t], \boldsymbol{\alpha}_t \triangleq L_t$	Same as ISTA except that L_t is set with the Brazilai-Borwein rule [9]
FISTA [6]	$\mathbf{X}_t \triangleq [\mathbf{x}_t, \mathbf{u}_t], \boldsymbol{\alpha}_t \triangleq (l_t, L_t)$	$\boldsymbol{\theta}_t \leftarrow \mathbf{D}\mathbf{u}_{t-1} - \mathbf{y}$ $\mathbf{x}_t \leftarrow \mathrm{prox}^{\Omega}_{\lambda/L_t} \left(\mathbf{u}_{t-1} - \frac{1}{L_t}\mathbf{D}^T\boldsymbol{\theta}_t\right)$ $l_t \leftarrow \frac{1+\sqrt{1+4l_{t-1}^2}}{2}$ $\mathbf{u}_t \leftarrow \mathbf{x}_t + (\frac{l_{t-1}-1}{l_t})(\mathbf{x}_t - \mathbf{x}_{t-1})$ L_t is set with the backtracking rule see [6]
Chambolle-Pock [11]	$\mathbf{X}_t \triangleq [\mathbf{x}_t, \mathbf{u}_t], \boldsymbol{\alpha}_t \triangleq (\tau_t, \sigma_t)$	$\boldsymbol{\theta}_t \leftarrow \frac{1}{1+\sigma_{t-1}}(\boldsymbol{\theta}_{t-1} + \sigma_{t-1}(\mathbf{D}\mathbf{u}_{t-1} - \mathbf{y}))$ $\mathbf{x}_t \leftarrow \mathrm{prox}^{\Omega}_{\lambda\tau_{t-1}} \left(\mathbf{x}_{t-1} - \tau_{t-1}\mathbf{D}^T\boldsymbol{\theta}_t\right)$ $\varphi_t \leftarrow \frac{1}{\sqrt{1+2\gamma\tau_{t-1}}}$ $\tau_t \leftarrow \varphi_t\tau_{t-1}; \sigma_t \leftarrow \frac{\sigma_{t-1}}{\varphi_t}$ $\mathbf{u}_t \leftarrow \mathbf{x}_t + \varphi_t(\mathbf{x}_t - \mathbf{x}_{t-1})$

Table I: *Updates* for first-order algorithms.

From the convex *optimality conditions*, solutions of the *Lasso* problem (3) and its dual (5), $\tilde{\mathbf{x}}$ and $\tilde{\boldsymbol{\theta}}$ respectively, are necessarily linked by:

$$\mathbf{y} = \mathbf{D}\tilde{\mathbf{x}} + \lambda\tilde{\boldsymbol{\theta}} \text{ and } \forall i \in \Gamma, \begin{cases} |\tilde{\boldsymbol{\theta}}^T\mathbf{d}_i| \leq 1 & \text{if } \tilde{\mathbf{x}}(i) = 0, \\ |\tilde{\boldsymbol{\theta}}^T\mathbf{d}_i| = 1 & \text{if } \tilde{\mathbf{x}}(i) \neq 0. \end{cases} \tag{6}$$

We define $\lambda_* \triangleq \left\|\mathbf{D}^T\mathbf{y}\right\|_\infty$. If $\lambda > \lambda_*$, the solution is trivial and is derived from the most simple screening test that may be designed, which screens out *all* atoms. Indeed, starting from the fact that $\tilde{\boldsymbol{\theta}} = \mathbf{y}/\lambda$ is the solution of the dual problem —it is feasible and maximizes (5a)—, we have $\forall i, |\mathbf{d}_i^T\tilde{\boldsymbol{\theta}}| = |\mathbf{y}^T\mathbf{d}_i|/\lambda \leq \lambda_*/\lambda < 1$ so that the optimality conditions impose that $\tilde{\mathbf{x}}(i) = 0$. In other words, we can screen the whole dictionary before entering the optimization procedure. In the following, we will focus on the non-trivial case $\lambda \in]0, \lambda_*]$.

b) Screening tests for the Lasso: The screening tests presented here have been proposed initially in the static perspective in [15], [19], [18]. They use relation (6) to locate some inactive atoms $\mathbf{d}_i$ for which $|\mathbf{d}_i^T\tilde{\boldsymbol{\theta}}| < 1$. The quantity is not directly accessible as the optimal is not known, thus the base concept of screening test is to geometrically construct a region $\mathcal{R}$ that is known to contain the optimal $\tilde{\boldsymbol{\theta}}$ so that the upper-bound $\max_{\boldsymbol{\theta}\in\mathcal{R}} |\mathbf{d}_i^T\boldsymbol{\theta}| \geq |\mathbf{d}_i^T\tilde{\boldsymbol{\theta}}|$ gives a sufficient condition for atom $\mathbf{d}_i$ to be inactive: $\max_{\boldsymbol{\theta}\in\mathcal{R}} |\mathbf{d}_i^T\boldsymbol{\theta}| < 1 \Rightarrow \tilde{\mathbf{x}}(i) = 0$. In particular the previous maximization problem admits a closed form solution when $\mathcal{R}$ is a sphere or a dome.

The SAFE test proposed by L. El Ghaoui et al. in [15] is derived by constructing a sphere from any dual point $\boldsymbol{\theta}$. Xiang et al. in [19], [18] improved it when the particular dual point $\mathbf{y}$ is used. We propose here a homogenized formulation relying on any dual point $\boldsymbol{\theta}$ for each of the three screening tests, generalizing [19], [18] to any dual point, thereby fitting them for use in a dynamic setting. We present these screening tests through the following Lemmata, in order of increasing description complexity and screening efficiency. For more details on the construction of regions $\mathcal{R}$ and the solution of the maximization problem please see the references or proofs in Appendix.

Lemma 2 (The SAFE screening test [15])**.** *For any $\boldsymbol{\theta} \in \mathbb{R}^N$, the following function $T_{\boldsymbol{\theta}}^{\mathrm{SAFE}}$ is a screening test for $\mathcal{P}^{\mathit{Lasso}}(\lambda, \mathbf{D}, \mathbf{y})$:*

$$\begin{aligned} T_{\boldsymbol{\theta}}^{\mathrm{SAFE}} &: \Gamma \to \{0, 1\} \\ & i \mapsto [\![(1 - |\mathbf{d}_i^T \mathbf{c}|) > r_{\boldsymbol{\theta}}]\!] \end{aligned} \tag{7}$$

where $\mathbf{c} \triangleq \frac{\mathbf{y}}{\lambda}$, $r_{\boldsymbol{\theta}} \triangleq \left\| \frac{\mathbf{y}}{\lambda} - \mu\boldsymbol{\theta} \right\|_2$ and $\mu \triangleq \left[\frac{\boldsymbol{\theta}^T\mathbf{y}}{\lambda\|\boldsymbol{\theta}\|_2^2}\right]_{-\|\mathbf{D}^T\boldsymbol{\theta}\|_\infty^{-1}}^{\|\mathbf{D}^T\boldsymbol{\theta}\|_\infty^{-1}}$.

The notation $[\![\mathrm{P}]\!]$ in (7) means that we take the boolean value of the proposition P as the output of the screening test. We also recall that $[r]_a^b \triangleq \max(\min(r, b), a)$ denotes the projection of r onto the segment $[a, b]$. Lemma 2 is exactly El Ghaoui's SAFE test.

The screening test ST3 [18] is a much more efficient test than SAFE, especially when λ_* is high. We extend it in the following Lemma so that it can be used for dynamic screening.

Lemma 3 (The Dynamic ST3: DST3)**.** *For any $\boldsymbol{\theta} \in \mathbb{R}^N$, the following function $T_{\boldsymbol{\theta}}^{\mathrm{DST3}}$ is a screening test for $\mathcal{P}^{\mathit{Lasso}}(\lambda, \mathbf{D}, \mathbf{y})$:*

$$\begin{aligned} T_{\boldsymbol{\theta}}^{\mathrm{DST3}} &: \Gamma \to \{0, 1\} \\ & i \mapsto [\![(1 - |\mathbf{d}_i^T \mathbf{c}|) > r_{\boldsymbol{\theta}}]\!] \end{aligned} \tag{8}$$

where $\mathbf{c} \triangleq \frac{\mathbf{y}}{\lambda} - \left(\frac{\lambda_*}{\lambda} - 1\right)\mathbf{d}_*$, $r_{\boldsymbol{\theta}} \triangleq \sqrt{\left\|\mu\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda}\right\|_2^2 - \left(\frac{\lambda_*}{\lambda} - 1\right)^2}$, $\mu \triangleq \left[\frac{\boldsymbol{\theta}^T\mathbf{y}}{\lambda\|\boldsymbol{\theta}\|_2^2}\right]_{-\|\mathbf{D}^T\boldsymbol{\theta}\|_\infty^{-1}}^{\|\mathbf{D}^T\boldsymbol{\theta}\|_\infty^{-1}}$ and $\mathbf{d}_* \triangleq \arg\max_{\mathbf{d} \in \{\pm\mathbf{d}_i\}_{i=1}^K} \mathbf{d}^T\mathbf{y}$.

When applied with $\boldsymbol{\theta} = \mathbf{y}$ this screening test is exactly the ST3 [18]. Further improvements have been proposed in the Dome test [19] for which we also propose an extended version appropriate for dynamic screening.

Lemma 4 (The Dynamic Dome Test: DDome)**.** *For any $\boldsymbol{\theta} \in \mathbb{R}^N$, the following function $T_{\boldsymbol{\theta}}^{\mathrm{DDome}}$ is a screening test for $\mathcal{P}^{\mathit{Lasso}}(\lambda, \mathbf{D}, \mathbf{y})$:*

$$\begin{aligned} T_{\boldsymbol{\theta}}^{\mathrm{DDome}} &: \Gamma \to \{0, 1\} \\ & i \mapsto [\![Q_{\boldsymbol{\theta}}^l l(\mathbf{d}_*^T\mathbf{d}_i) < \mathbf{x}^T\mathbf{d}_i < Q_{\boldsymbol{\theta}}^u(\mathbf{d}_*^T\mathbf{d}_i)]\!] \end{aligned} \tag{9}$$

where

$$Q_{\boldsymbol{\theta}}^l(t) \triangleq \begin{cases} (\lambda_* - \lambda)t - \lambda + \lambda r_{\boldsymbol{\theta}}\sqrt{1 - t^2}, & \textit{if } t \leq \lambda_* \\ -(\lambda - 1 + \lambda/\lambda_*), & \textit{if } t > \lambda_* \end{cases} \tag{10}$$

$$Q_{\boldsymbol{\theta}}^u(t) \triangleq \begin{cases} (\lambda - 1 + \lambda/\lambda_*), & \textit{if } t < -\lambda_* \\ (\lambda_* - \lambda)t + \lambda - \lambda r_{\boldsymbol{\theta}}\sqrt{1 - t^2}, & \textit{if } t \geq -\lambda_* \end{cases} \tag{11}$$

$r_{\boldsymbol{\theta}} \triangleq \sqrt{\left\|\mu\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda}\right\|_2^2 - \left(\frac{\lambda_*}{\lambda} - 1\right)^2}$, $\mu \triangleq \left[\frac{\boldsymbol{\theta}^T\mathbf{y}}{\lambda\|\boldsymbol{\theta}\|_2^2}\right]_{-\|\mathbf{D}^T\boldsymbol{\theta}\|_\infty^{-1}}^{\|\mathbf{D}^T\boldsymbol{\theta}\|_\infty^{-1}}$ *and* $\mathbf{d}_* \triangleq \arg\max_{\mathbf{d} \in \{\pm\mathbf{d}_i\}_{i=1}^K} \mathbf{d}^T\mathbf{y}$.

When applied with $\boldsymbol{\theta} = \mathbf{y}$ this screening test is exactly the Dome test [19].

Using these Lemmata at dual points $\boldsymbol{\theta}_t$ obtained during iterations allows to progressively reduce the radius $r_{\boldsymbol{\theta}}$ of the considered regions —sphere or dome— and thus improves the screening capacity of the screening tests associated to these regions. The effect of the radius appears clearly in (7,8,10,11). Note that the choice of a new $\boldsymbol{\theta}$, for one of the previous screening tests $T_{\boldsymbol{\theta}}$, only acts on $r_{\boldsymbol{\theta}}$ the radius of the region and not on $\mathbf{c}$ its center.

3) Dynamic screening for the Group-Lasso: The *Lasso* problem embodies the assumption that observation $\mathbf{y}$ may be approximately represented in $\mathbf{D}$ by a sparse vector $\tilde{\mathbf{x}}$. When a particular structure of the data is known, we may additionally assume that, besides sparsity, the representation of $\mathbf{y}$ in $\mathbf{D}$ fits this structure. Inducing the structure into $\tilde{\mathbf{x}}$ is exactly the goal of structured-sparsity regularizers. Among those we focus on the *Group-Lasso* regularization because the group-separability of its objective function (12) particularly fits the screening framework.

a) The Group-Lasso [2]: The *Group-Lasso* is a sparse least-squares problem that assumes some group structure in the sparse solution, in the sense that there are groups of zero coefficients in the solution $\tilde{\mathbf{x}}$. This structure, assumed to be known in advance, is characterized by $\mathcal{G}$, a known partition of Γ, and $w_g > 0$ the weights associated with each group $g \in \mathcal{G}$ (*e.g.* $w_g = \sqrt{|g|}$). Using the group-sparsity inducing regularization $\Omega(\mathbf{x}) \triangleq \sum_{g\in\mathcal{G}} w_g \left\|\mathbf{x}_{[g]}\right\|_2$, the *Group-Lasso* is defined as:

$$\mathcal{P}^{\textit{Group-Lasso}}(\lambda, \mathbf{D}, \mathcal{G}, \mathbf{y}) : \qquad \underset{\mathbf{x}}{\arg\min}\, \frac{1}{2}\left\|\mathbf{D}\mathbf{x} - \mathbf{y}\right\|_2^2 + \lambda \sum_{g\in\mathcal{G}} w_g \left\|\mathbf{x}_{[g]}\right\|_2 . \tag{12}$$

The proximal operator of the group sparsity regularization is the group soft-thresholding:

$$\forall g \in \mathcal{G}, \operatorname{prox}_t^{\textit{Group-Lasso}}(\mathbf{x}_g) \triangleq \max\left(0, \frac{\left\|\mathbf{x}_g\right\|_2 - t w_g}{\left\|\mathbf{x}_g\right\|_2}\right) \mathbf{x}_g. \tag{13}$$

The dual of the *Group-Lasso* problem (12) is (see [17]):

$$\tilde{\boldsymbol{\theta}} \triangleq \underset{\boldsymbol{\theta}}{\arg\max} \frac{1}{2}\left\|\mathbf{y}\right\|_2^2 - \frac{\lambda^2}{2}\left\|\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda}\right\|_2^2 \tag{14a}$$

$$\text{s.t. } \forall g \in \mathcal{G}, \frac{\left\|\mathbf{D}_{[g]}^T \boldsymbol{\theta}\right\|_2}{w_g} \leq 1. \tag{14b}$$

A dual point $\boldsymbol{\theta} \in \mathbb{R}^N$ is said *feasible* for the *Group-Lasso* if it satisfies constraints (14b).
From the convex *optimality conditions*, primal and dual optima are necessarily linked by:

$$\mathbf{y} = \mathbf{D}\tilde{\mathbf{x}} + \lambda\tilde{\boldsymbol{\theta}}, \text{ and } \forall g \in \mathcal{G} \begin{cases} \|\mathbf{D}_{[g]}^T \tilde{\boldsymbol{\theta}}\|_2 \leq w_g & \text{if } \tilde{\mathbf{x}}_{[g]} = \mathbf{0}, \\ \|\mathbf{D}_{[g]}^T \tilde{\boldsymbol{\theta}}\|_2 = w_g & \text{if } \tilde{\mathbf{x}}_{[g]} \neq \mathbf{0}. \end{cases} \tag{15}$$

We now adapt the definition of λ_* to the *Group-Lasso* so that it corresponds to the smallest regularization parameter resulting into a zero solution of (12). Let us define

$$g_* \triangleq \underset{g}{\arg\max} \frac{\left\|\mathbf{D}_{[g]}^T \mathbf{y}\right\|_2}{w_g}, \quad \lambda_* \triangleq \frac{\left\|\mathbf{D}_{[g_*]}^T \mathbf{y}\right\|_2}{w_{g_*}}. \tag{16}$$

As for the *Lasso*, if $\lambda > \lambda_*$ one may screen all the atoms and obtain $\tilde{\mathbf{x}} = \mathbf{0}$. Hence for the *Group-Lasso* setting, we focus on the non-trivial case $\lambda \in]0, \lambda_*]$.

Instances of screening tests for the *Group-Lasso* are presented in the sequel. We extend here the SAFE [15] and the DST3 [18] screening tests to the *Group-Lasso*. To our knowledge there are no published results on this extension to the *Group-Lasso*.

b) Screening tests for the Group-Lasso: As just previously for the *Lasso*, the quantity $\|\mathbf{D}_{[g]}^T\tilde{\boldsymbol{\theta}}\|_2$ in relation (15) is not known except if the problem is solved. Regions $\mathcal{R}$ containing the optimum $\tilde{\boldsymbol{\theta}}$ are considered to use the upper bound $\max_{\boldsymbol{\theta}\in\mathcal{R}} \|\mathbf{D}_{[g]}^T\boldsymbol{\theta}\|_2$ to identify some inactive groups g thanks to relation (15). Please see the proofs in the appendix for details on the construction of the regions and the solution of the maximization problem.

For any index $i \in \Gamma$, we denote by $g(i)$ the unique group $g \in \mathcal{G}$ that contains i. The following Lemma extends the SAFE screening test to the *Group-Lasso*:

Lemma 5 (The Group-SAFE: GSAFE)**.** *For any $\boldsymbol{\theta} \in \mathbb{R}^N$, the following function $T_{\boldsymbol{\theta}}^{\mathrm{GSAFE}}$ is a screening test for $\mathcal{P}^{\text{Group-Lasso}}(\lambda, \mathbf{D}, \mathcal{G}, \mathbf{y})$:*

$$\begin{aligned} T_{\boldsymbol{\theta}}^{\mathrm{GSAFE}} &: \Gamma \to \{0,1\} \\ i &\mapsto \left[\!\!\left[\left(\frac{w_{g(i)}}{\left\|\mathbf{D}_{[g(i)]}\right\|} - \frac{\left\|\mathbf{D}_{[g(i)]}^T\mathbf{c}\right\|_2}{\left\|\mathbf{D}_{[g(i)]}\right\|} \right) > r_{\boldsymbol{\theta}} \right]\!\!\right] \end{aligned} \tag{17}$$

where

$$\mathbf{c} \triangleq \frac{\mathbf{y}}{\lambda}, r_{\boldsymbol{\theta}} \triangleq \left\|\frac{\mathbf{y}}{\lambda} - \mu\boldsymbol{\theta}\right\|_2 \textit{ and } \mu \triangleq \left[\frac{\boldsymbol{\theta}^T\mathbf{y}}{\lambda\|\boldsymbol{\theta}\|_2^2}\right]_{-\min_{g\in\mathcal{G}} \frac{w_g}{\left\|\mathbf{D}_{[g]}^T\boldsymbol{\theta}\right\|_2}}^{\min_{g\in\mathcal{G}} \frac{w_g}{\left\|\mathbf{D}_{[g]}^T\boldsymbol{\theta}\right\|_2}}.$$

The following Lemma extends the screening tests ST3 [18] and DST3 (see Lemma 3) to the *Group-Lasso*.

Lemma 6 (The Dynamic Group ST3: DGST3)**.** *For any $\boldsymbol{\theta} \in \mathbb{R}^N$, the following function $T_{\boldsymbol{\theta}}^{\mathrm{DGST3}}$ is a screening test for $\mathcal{P}^{\text{Group-Lasso}}(\lambda, \mathbf{D}, \mathcal{G}, \mathbf{y})$:*

$$\begin{aligned} T_{\boldsymbol{\theta}}^{\mathrm{DGST3}} &: \Gamma \to \{0,1\} \\ i &\mapsto \left[\!\!\left[\left(\frac{w_{g(i)}}{\left\|\mathbf{D}_{[g(i)]}\right\|} - \frac{\left\|\mathbf{D}_{[g(i)]}^T\mathbf{c}\right\|_2}{\left\|\mathbf{D}_{[g(i)]}\right\|} \right) > r_{\boldsymbol{\theta}} \right]\!\!\right] \end{aligned} \tag{18}$$

where

$$\begin{aligned} \mathbf{c} &\triangleq \left(\mathbf{Id} - \frac{\mathbf{n}\mathbf{n}^T}{\|\mathbf{n}\|_2^2}\right)\frac{\mathbf{y}}{\lambda} + \frac{\mathbf{n}}{\|\mathbf{n}\|_2^2} w_{g_*}^2, \\ \mathbf{n} &\triangleq \mathbf{D}_{[g_*]}\mathbf{D}_{[g_*]}^T\frac{\mathbf{y}}{\lambda_*}, \\ r_{\boldsymbol{\theta}} &\triangleq \sqrt{\left\|\frac{\mathbf{y}}{\lambda} - \mu\boldsymbol{\theta}\right\|_2^2 - \left\|\frac{\mathbf{y}}{\lambda} - \mathbf{c}\right\|_2^2} \textit{ and} \\ \mu &\triangleq \left[\frac{\boldsymbol{\theta}^T\mathbf{y}}{\lambda\|\boldsymbol{\theta}\|_2^2}\right]_{-\min_{g\in\mathcal{G}} \frac{w_g}{\left\|\mathbf{D}_{[g]}^T\boldsymbol{\theta}\right\|_2}}^{\min_{g\in\mathcal{G}} \frac{w_g}{\left\|\mathbf{D}_{[g]}^T\boldsymbol{\theta}\right\|_2}}. \end{aligned}$$

In these two Lemmata, the regions $\mathcal{R}$ used to define the screening tests are spheres and the effect of the radius $r_{\boldsymbol{\theta}}$ on the screening capacity is visible in (17) and (18).

The proposed screening tests have been given for the *Group-Lasso* formulation, but can be readily extended to the *Overlapping Group-Lasso* [23] thanks to the replication trick.

C. A turnkey instance

As a concrete instance of Algorithm 3, we propose to focus on the *Lasso* problem solved by the combined use of ISTA and SAFE. We compare the static screening with the dynamic screening, through implementations given in Algorithms 4 and 5, respectively. The usual ISTA update appears at lines 7 to 9 in Algorithm 4 and lines 5 to 7 in Algorithm 5, where the step size L_t is set using the backtracking strategy as described in [6]. The remaining lines of the algorithm, dedicated to the screening process, are described separately in the following paragraphs.

Algorithm 4 ISTA + *Static* SAFE Screening

```
Require: $\mathbf{D}, \mathbf{y}, \lambda, \mathbf{x}_0 \in \mathbb{R}^K$
 1: .......... Screening ..........
 2: $\mathcal{I} \leftarrow \left\{ i \in \Gamma, |\mathbf{d}_i^T \mathbf{y}| < \lambda - 1 + \frac{\lambda}{\lambda_*} \right\}$
 3: $\mathbf{D}_0 \leftarrow \mathbf{D}_{[\mathcal{I}^c]}$,
 4: $t \leftarrow 1$
 5: while stopping criterion on $\mathbf{x}_t$ do
 6:     ... ISTA update .....
 7:     $\boldsymbol{\theta}_t \leftarrow \mathbf{D}_0 \mathbf{x}_{t-1} - \mathbf{y}$
 8:     $\mathbf{z}_t \leftarrow \mathbf{D}_0^T \boldsymbol{\theta}_t$
 9:     $\mathbf{x}_t \leftarrow \operatorname{prox}_{\lambda/L_t}^{\textit{Lasso}} \left( \mathbf{x}_{t-1} - \frac{1}{L_t} \mathbf{z}_t \right)$
10:     $t \leftarrow t + 1$
11: end while
12: return $\mathbf{x}_t$
```

Algorithm 5 ISTA + *Dynamic* SAFE Screening

```
Require: $\mathbf{D}, \mathbf{y}, \lambda, \mathbf{x}_0 \in \mathbb{R}^K$
 1: $\mathcal{I}_0 \leftarrow \emptyset$, $r_0 \leftarrow +\infty, \mathbf{D}_0 \leftarrow \mathbf{D}$
 2: $t \leftarrow 1$
 3: while stopping criterion on $\mathbf{x}_t$ do
 4:     .... ISTA update .....
 5:     $\boldsymbol{\theta}_t \leftarrow \mathbf{D}_{t-1} \bar{\mathbf{x}}_{t-1} - \mathbf{y}$
 6:     $\mathbf{z}_t \leftarrow \mathbf{D}_{t-1}^T \boldsymbol{\theta}_t$
 7:     $\mathbf{x}_t \leftarrow \operatorname{prox}_{\lambda/L_t}^{\textit{Lasso}} \left( \bar{\mathbf{x}}_{t-1} - \frac{1}{L_t} \mathbf{z}_t \right)$
 8:     ........... Screening ...........
 9:     $\mu_t \leftarrow \left[ \frac{\boldsymbol{\theta}_t^T \mathbf{y}}{\lambda \|\boldsymbol{\theta}_t\|_2^2} \right]_{-\|\mathbf{z}_t\|_\infty^{-1}}^{\|\mathbf{z}_t\|_\infty^{-1}}$
10:     $\mathbf{v}_t \leftarrow \mu_t \boldsymbol{\theta}_t$
11:     $r_t \leftarrow \left\| \frac{\mathbf{y}}{\lambda} - \mathbf{v}_t \right\|_2$
12:     $\mathcal{I}_t \leftarrow \left\{ i \in \Gamma, |\mathbf{d}_i^T \mathbf{y}| < \lambda(1 - r_t) \right\} \cup \mathcal{I}_{t-1}$
13:     $\mathbf{D}_t \leftarrow \mathbf{D}_{t-1} \mathbf{Id}_{[\mathcal{I}_{t-1}^c, \mathcal{I}_t^c]}$
14:     $\bar{\mathbf{x}}_t \leftarrow \mathbf{Id}_{[\mathcal{I}_t^c, \mathcal{I}_{t-1}^c]} \mathbf{x}_t$
15:     $t \leftarrow t + 1$
16: end while
17: return $\mathbf{x}_t$
```

The state-of-the-art static screening shown in Algorithm 4 is the successive use of the SAFE screening test —Lemma 2 with $\boldsymbol{\theta} = \mathbf{y}$ results exactly in lines 2-3— *prior* to the ISTA algorithm. The dictionary is screened once for all, using information from the initially-available data $\mathbf{D}^T\mathbf{y}$ and λ.

The proposed dynamic screening principle is shown in Algorithm 5. The iteration is here composed of two stages: a) the ISTA update (lines 5-7) which is exactly the same as lines 7-9 of Algorithm 4 except that the dictionary changes along iterations and b) the screening step (lines 9-13) which aims at reducing the dictionary size thanks to the information contained in the current iterates $\boldsymbol{\theta}_t$ and $\mathbf{z}_t$. The screening process appears at lines 11-13 where the index sets $\mathcal{I}_t$ of screened atoms form a non-decreasing inclusion-wise sequence (line 12).

D. Computational complexity of the dynamic screening

The screening test introduces only a negligible computational overhead because it mainly relies on the matrix-vector multiplications already performed in the first-order algorithm update. We present now the computational ingredients of the acceleration obtained by dynamic screening.

Algorithm 3 implements the iterative alternation of the update of any first-order algorithm —*e.g.* those from Table I— at line 4, and a screening process at line 6-8. This screening process consists of the pairing of two distinct stages. First the set $\mathcal{I}_t$ of screened atoms is computed at line 6 using one of the Lemmata 2 to 6. Analyzing these Lemmata shows that the expensive computation required to evaluate the screening test $T_{\boldsymbol{\theta}_t}(i)$ for all $i \in \Gamma$ is both due to the products $\mathbf{d}_i^T\mathbf{c}$ for all $i \in \Gamma$ and to the computation of the scalar μ which needs the product $\mathbf{D}_t^T\boldsymbol{\theta}_t$. Thus, determining a set of inactive atoms may cost $\mathcal{O}(KN)$ per iteration. Fortunately, the computation $\mathbf{D}^T\mathbf{c}$ can be done once for all at the beginning of the algorithm. Table I shows that the computation $\mathbf{D}_t^T\boldsymbol{\theta}_t$ is already done by every first-order algorithm. So determining $\mathcal{I}_t$ produces an overhead of $\mathcal{O}(\left|\mathcal{I}_{t-1}^c\right| + N)$ only. Second the proper screening operations reduce the size of the dictionary and the primal variables at lines 7 and 8, with a small computation requirement because matrix $\mathbf{Id}_{\left[\mathcal{I}_t^c,\mathcal{I}_{t-1}^c\right]}$ has only $|\mathcal{I}_t^c|$ non-zero elements. So, finally, the computation overhead entailed by the embedded screening test has complexity $\mathcal{O}(\left|\mathcal{I}_{t-1}^c\right| + N)$ at iteration t which is negligible compared with the complexity $\mathcal{O}(\left|\mathcal{I}_{t-1}^c\right| N)$ for the optimization update $p(\cdot)$. A detailed complexity analysis is given in Section III-A. Finally the total computation cost of the algorithm with dynamic screening may be much smaller than the base first-order algorithm. This is evaluated experimentally in Section III.

III. Experiments

This section is dedicated to experiments made to assess the practical relevance of the proposed dynamic screening principle[2]. More precisely, we aim at providing a rich understanding of its properties beyond what the theory can demonstrate. The questions of interest deal with the computational performance and may be formulated as follows:

- how to measure and evaluate the benefits of dynamic screening?
- what is the efficiency of dynamic screening in terms of the overall acceleration compared to the algorithm without screening or with static screening?
- to which extent does the computational gain depend on problems, algorithms, synthetic and real data, screening tests?

A. How to evaluate: performance measures

Let us first notice from Theorem 1 that whatever strategy is used —no-screening/static screening/dynamic screening—, the algorithms converge to the same optimal $\tilde{\mathbf{x}}$. Consequently, there is no need to evaluate the *quality* of the solution and we shall only focus on *computational aspects*.

The main figure of merit that we use is based on an estimation of the number of floating-point operations (flops) required by the algorithms with no screening (flops_N), with static screening (flops_S) and with dynamic screening (flops_D) for a complete run. We will represent experimental results by the normalized number of flops $\frac{\text{flops}_\text{S}}{\text{flops}_\text{N}}$ and $\frac{\text{flops}_\text{D}}{\text{flops}_\text{N}}$ that reflect the acceleration obtained respectively by the static screening and dynamic screening strategies over the base algorithm with no screening. Computing such quantities requires to experimentally record the number of iterations t_f and for each iteration t, the size of the dictionary $|\mathcal{I}_t^c|$ and the sparsity of the current iterate $\left\|\mathbf{x}_t\right\|_0$. They are defined for the *Lasso* as:

flops_N	$\sum_{t=1}^{t_f}\left[(K + \left\|\mathbf{x}_t\right\|_0)N + 4K + N\right]$
flops_S	$KN + \sum_{t=1}^{t_f}\left[(\left|\mathcal{I}_0^c\right| + \left\|\mathbf{x}_t\right\|_0)N + 4\left|\mathcal{I}_0^c\right| + N\right]$
flops_D	$\sum_{t=1}^{t_f}\left[(\left|\mathcal{I}_t^c\right| + \left\|\mathbf{x}_t\right\|_0)N + 6\left|\mathcal{I}_t^c\right| + 5N\right]$

and for the *Group-Lasso* as:

flops_N	$\sum_{t=1}^{t_f}\left[(K + \left\|\mathbf{x}_t\right\|_0)N + 4K + N + 3\left|\mathcal{G}\right|\right]$
flops_S	$KN + \sum_{t=1}^{t_f}\left[(\left|\mathcal{I}_0^c\right| + \left\|\mathbf{x}_t\right\|_0)N + 4\left|\mathcal{I}_0^c\right| + N + 3\left|\mathcal{G}\right|\right]$
flops_D	$\sum_{t=1}^{t_f}\left[(\left|\mathcal{I}_t^c\right| + \left\|\mathbf{x}_t\right\|_0)N + 7\left|\mathcal{I}_t^c\right| + 5N + 5\left|\mathcal{G}\right|\right]$

Indeed, one update of a first-order algorithm at iteration t requires at least $2\left|\mathcal{I}_t^c\right| N + \left|\mathcal{I}_t^c\right| + N$ to compute the gradient, and the proximal operator of the *Lasso* and *Group-Lasso* need, $3\left|\mathcal{I}_t^c\right|$ and $3\left|\mathcal{I}_t^c\right| + 3\left|\mathcal{G}\right|$ operations, respectively (see Table I, (4) and (13)). The dynamic screening requires the computation of μ: $2N + \left|\mathcal{I}_t^c\right|$ and $2N + 2\left|\mathcal{I}_t^c\right| + 2\left|\mathcal{G}\right|$ for *Lasso* and *Group-Lasso* respectively, (see Lemma 2-6), and $2N$ for

[2] The code in Python and data are released for reproducible research purposes at http://pageperso.lif.univ-mrs.fr/~antoine.bonnefoy.

the computation of the $r_{\boldsymbol{\theta}}$. The screening step is then computed in $|\mathcal{I}_t^c|$ operations. The static screening approach implies a separated initialization of the screening test which requires KN operations.

Note that the primal variable which would be sparse during the optimization procedure which reduce the number of operation required for $\mathbf{D}_t\mathbf{x}_t$ from $|\mathcal{I}_t^c|\,N$ to $\|\mathbf{x}_t\|_0\,N$. Note also that here we do not take into account the time required to compute the matrix norm of the sub-dictionaries corresponding to each group: $\left\|\mathbf{D}_{[g]}\right\|, g \in \mathcal{G}$. These quantities do not depend on the problem $\mathcal{P}(\lambda, \Omega, \mathbf{D}, \mathbf{y})$ but on the dictionary and the groups themselves only, so we consider that they can be computed beforehand for a given dictionary D and a given partition G.

Another option to measure the computational gain consists in actual running times, which we consider as well. The main advantage of this measure would be that it results from actual performance in seconds instead of estimated or asymptotic figure. However, running times depend on the implementation so that it may not be the right measure in the current context of algorithm design. For each screening strategy (no screening/static/dynamic) we measure the running times $t_N/t_S/t_D$. The performance are then represented in terms of normalized running times $\frac{t_D}{t_N}$ and $\frac{t_S}{t_N}$. Eventually, one may wonder whether those measures, flops and times, are somehow equivalent, which will be checked and discussed in Sections III-C and III-D.

B. Data material

1) Synthetic data: For experiments on synthetic data, we used two types of dictionaries that are widely used in the state-of-the-art of sparse estimation and screening tests. The first one is a normalized Gaussian dictionary in which all atoms $\mathbf{d}_i$ are drawn i.i.d. uniformly on the unit sphere, *e.g.*, by normalizing realizations of $\mathcal{N}(\mathbf{0}, \mathbf{Id}_N)$. The second one is the so-called Pnoise introduced in [19], for which all $\mathbf{d}_i$ are drawn i.i.d. as $\mathbf{e}_1 + 0.1\kappa\mathbf{g}$ and normalized, where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{Id}_N)$, $\kappa \sim \mathcal{U}(0,1)$ and $\mathbf{e}_1 \triangleq [1, 0, \ldots, 0]^T \in \mathbb{R}^N$ is the first natural basis vector. We set the data dimension to $N \triangleq 2000$ and the number of atoms to $K \triangleq 10000$.

In the experiments on the *Lasso*, observations $\mathbf{y}$ were drawn i.i.d. from the exact same distribution as the atoms of dictionaries described above. In experiments on the *Group-Lasso*, all groups were built randomly with the same number of atoms in each group. Observations were generated from a Bernouilli-Gaussian distribution: $|\mathcal{G}|$ independent draws of a Bernoulli distribution of parameter $p = 0.05$ were used to determine for each group if it was active or not. Then coefficients of active groups are drawn i.i.d. from a standard Gaussian distribution while they are set to zero in inactive groups. The observation $\mathbf{y}$ was generated as the (l_2-normalized) sum of $\mathbf{Dx}$ and of a Gaussian noise such that the signal-to-noise ratio equals 20dB.

2) Audio data: For experiments on real data we performed the estimation of the sparse representation of audio signals in a redundant Discrete Cosine Transform (DCT) dictionary, which is known to be adapted for audio data. Music and speech recordings were taken from the material of the 2008 Signal Separation Evaluation Campaign [24]. We considered 30 observations $\mathbf{y}$ with length $N = 1024$ and sampling rate 16 kHz and the number of atoms is set to $K \triangleq 10000$.

3) Image data: Experiments on the MNIST database [25] have been performed too. The database is composed of images of $N \triangleq 28 \times 28 = 784$ pixels representing handwritten digits from 0 to 9 and is split into a training set and a testing set. The dictionary $\mathbf{D}$ is composed of $K \triangleq 10000$ vectorized images from the training set, with 1000 randomly-chosen images for each digit. Observations were taken randomly in the test set.

C. Solving the Lasso with several algorithms.

We addressed the *Lasso* problem with four different algorithms from Table I: ISTA, FISTA, SpaRSA and Chambolle-Pock. Algorithms stop at iteration t if either $t > 200$ or the relative variation $\frac{|F(\mathbf{x}_{t-1}) - F(\mathbf{x}_t)|}{F(\mathbf{x}_t)}$ of the objective function $F(\mathbf{x}) \triangleq \frac{1}{2}\|\mathbf{Dx} - \mathbf{y}\|_2^2 + \lambda\Omega(\mathbf{x})$ is lower than 10^{-7}. We used a Pnoise dictionary and three different strategies for each algorithm: no screening, static ST3 screening and dynamic ST3 screening. Algorithms were run for several values of λ to assess the performance for various sparsity levels.

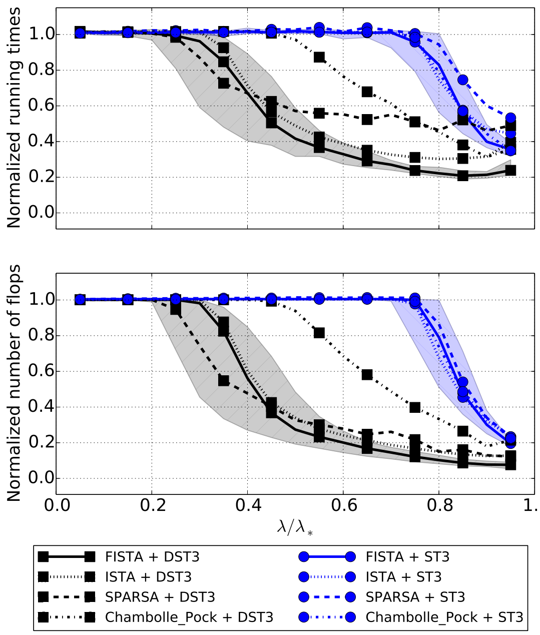

Figure 2: Normalized running times and normalized number of flops for solving the *Lasso* with a Pnoise dictionary.

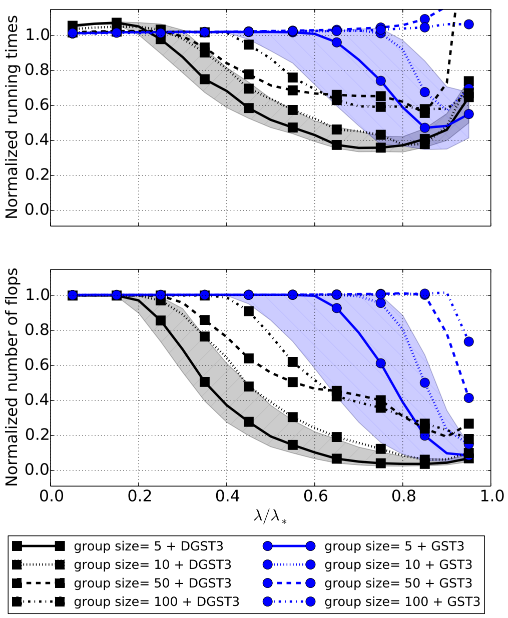

Figure 3: Normalized running times an number of flops for FISTA solving the *Group-Lasso* with different group sizes (5, 10, 50, 100).

Figure 2 shows the normalized running times and normalized number of flops for algorithms with dynamic screening (black squares) and for the corresponding algorithms with static screening (blue circle) as a function of λ/λ_*. Lower values account for faster computation. The medians among 30 runs are plotted and the shaded area contains the 25%-to-75% percentiles for FISTA, in order to illustrate the typical distribution of the values (similar areas are observed for the other algorithms but are not reported for readability).

For all algorithms, the dynamic strategy shows a significant acceleration in a wide range of parameter $\lambda \geq 0.3\lambda_*$. For $\lambda \geq 0.5\lambda_*$, computational savings reach about 75% of the running time and 80% of the number of flops. The static strategy is efficient in a much reduced range $\lambda \geq 0.8\lambda_*$, with lower computational gains. Among all tested algorithms, FISTA has the largest ability to be accelerated, which is really interesting as it is also known to be very efficient in terms of convergence rate. Note that due to the normalization of running times and flops, Figure 2 cannot be used to draw any conclusion on which of ISTA, FISTA, SpaRSA or Chambolle-Pock is the fastest algorithm. Finally, one may observe that the running time and flops measures have similar trends, supporting the idea that only one of them may be used to assess computational performance in a fair way.

D. Solving the Group-Lasso for various group sizes

We addressed the *Group-Lasso* with FISTA using Pnoise data and dictionary as described in III-B with several group sizes. In Figure 3, the median normalized running times and number of flops over 30 runs are plotted, the shaded area representing the 25%-to-75% percentiles when the group size is 5.

The computational gains obtained by the dynamic screening strategy are of the same order than for the *Lasso*, with large savings in a wide range $\lambda \geq 0.3\lambda_*$. One may anticipate that when groups grow larger it is more difficult to locate some inactive groups, Figure 3 confirms this intuition: screening tests become less and less efficient to locate inactive groups and consequently the acceleration is not as efficient as in the *Lasso* problem. As for the *Lasso*, running times and flops have similar trends, even if we observe a larger discrepancy. The discrepancy is due to implementation details. For instance the many loops on groups required for the computation of the screening tests are hard to handle efficiently in python.

E. Comparing screening tests

From the previous experiments, we retained FISTA to solve the *Lasso* on synthetic data with a Pnoise dictionary and a Gaussian dictionary, on real audio data, and on images. Results are reported in Figure 4 for all the proposed screening tests.

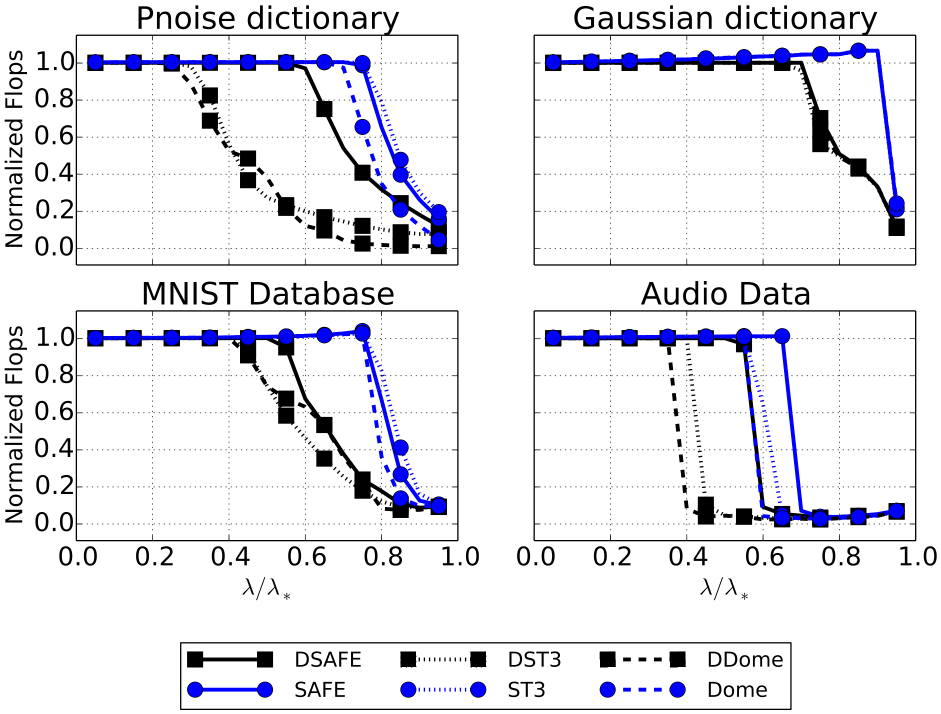

Figure 4: Computational gain of screening strategies for various data and screening tests, on the *Lasso* solved by FISTA.

For all kinds of data and all screening tests, dynamic screening again provides a large acceleration on a wide range of λ values and improves the static screening strategy. In the case of the Pnoise dictionary and of audio data, the ST3 and Dome tests bring in important improvement over the SAFE test, in the static and dynamic strategies. Indeed, λ_* is close to 1 in these cases so that the radius $r_{\boldsymbol{\theta}}$ in (7) for SAFE is much larger than in (3) for ST3 and in (4) for Dome, which degrades the screening efficiency of SAFE. This difference is even more visible when the dynamic screening strategy is used. As a counterpart, the Gaussian dictionary have small correlation between atoms. In this dictionary, ST3 and Dome do not improve the performance of SAFE, but the dynamic strategy allows a higher acceleration ratio and for a larger range of parameter λ.

IV. Discussion

We have proposed the dynamic screening principle and shown that this principle is relevant both theoretically and practically. When first-order algorithms are used, dynamic screening induces stronger acceleration on the *Lasso* and *Group-Lasso* solvings than static screening, and in a wider range of λ.

The convergence theorem (Theorem 1) makes very few assumptions on the iterative algorithm, meaning that dynamic screening principle can be applied to many algorithms —*e.g.* second order algorithms.

Conversely, dynamic screening tests may produce different iterates than those of the base algorithm on which it is applied and hence may modify the convergence rate. Can we ensure that the dynamic screening preserves the convergence rate of any first-order algorithm? Answering this question would definitely anchor dynamic screening in a theoretical context.

We presented here algorithms designed to compute the *Lasso* problem for a given λ. Departing from that, one might be willing to compute the whole regularization path as done by the *LARS* algorithm [26]. Thoroughly studying how screening might be combined with *LARS* is another exciting subject that we plan to work on in a near future.

In a recent work [27] Wang et. *al* introduce a way to adapt the static dome test in a continuation strategy. This work relies on exact solutions of successive computation for higher λ parameters. Iterative optimization algorithms do not give exact solutions hence examining how the dome test can be adapted dynamically in an iterative optimization procedure might be of great interest and lead to new approaches.

Given the nice theoretical and practical behavior of Orthogonal Matching Pursuit [28], [29], investigating how it can be paired with dynamic screening is a pressing and exciting matter but poses the problem of dealing with the non-convex ℓ_0 regularization which prevents from using the theory and toolbox of convex optimality.

Lastly, as in [15], we are curious to see how dynamic screening may show up when other than an ℓ_2 fit-to-data is studied: for example, this situation naturally occurs when classification-based losses are considered. As sparsity is often a desired feature for both efficiency (in the prediction phase) and generalization purposes, being able to work out well-founded results allowing dynamic screening is of the utmost importance.

REFERENCES

[1] R. Tibshirani, “Regression shrinkage and selection via the Lasso,” *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[2] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2006.

[3] S. Chen, D. Donoho, and M. Saunders, “Atomic decomposition by Basis Pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[4] M. Elad, J.-L. Starck, P. Querre, and D. Donoho, “Simultaneous cartoon and texture image inpainting using Morphological Component Analysis (MCA),” *Applied and Computational Harmonic Analysis*, vol. 19, no. 3, pp. 340 – 358, 2005.

[5] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation.” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, Feb. 2009.

[6] 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.

[7] P. L. Combettes and V. R. Wajs, “Signal recovery by proximal forward-backward splitting,” *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[8] 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. 1457, pp. 1413–1457, 2004.

[9] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, “Sparse reconstruction by separable approximation,” *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2479–2493, 2009.

[10] H. U. K. J. Arrow, L. Hurwicz, *Studies in linear and non-linear programming, With contributions by Hollis B. Chenery [and others].* Stanford University Press, Stanford, Calif, 1964.

[11] A. Chambolle and T. 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, 2011.

[12] Y. Nesterov, “A method of solving a convex programming problem with convergence rate o (1/k2),” *Soviet Mathematics Doklady*, vol. 27, no. 2, pp. 372–376, 1983.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

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

[15] L. El Ghaoui, V. Viallon, and T. Rabbani, “Safe Feature Elimination in Sparse Supervised Learning,” EECS Department, University of California, Berkeley, Tech. Rep., 2010.

[16] 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, Mar. 2012.

[17] J. Wang, B. Lin, P. Gong, P. Wonka, and J. Ye, “Lasso Screening Rules via Dual Polytope Projection,” *CoRR*, pp. 1–17, 2012.

[18] Z. J. Xiang, H. Xu, and P. J. Ramadge, “Learning sparse representations of high dimensional data on large scale dictionaries,” in *Advances in Neural Information Processing Systems*, vol. 24, 2011, pp. 900–908.

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

[20] R. J. Tibshirani, "The Lasso problem and uniqueness," *Electronic Journal of Statistics*, vol. 7, pp. 1456–1490, 2013.

[21] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval, "A Dynamic Screening Principle for the Lasso," in *Proc. of EUSIPCO*, 2014.

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

[23] L. Jacob, G. Obozinski, and J.-P. Vert, "Group Lasso with overlap and graph Lasso," in *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 433–440.

[24] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation*, Mar. 2009.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, 1998.

[26] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[27] Y. Wang, Z. J. Xiang, and P. L. Ramadge, "Lasso screening with a small regularization parameter," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3342–3346.

[28] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, pp. 3397–3415, 1993.

[29] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

SCREENING TESTS FOR THE *Group-Lasso*

This section is dedicated to the proofs of the screening tests given in the Lemmata 2 to 6.

Proof of Lemmata 2 and 3. Since the *Lasso* is a particular case of the *Group-Lasso*, *i.e.* groups of size one with $w_g = 1$ for all $g \in \mathcal{G}$, Lemmata 2 and 3 are direct corollaries of Lemmata 5 and 6. □

Base concept: Extending what has been proposed in [15], [18] we construct screening tests for *Group-Lasso* using optimality conditions of the *Group-Lasso* (15) jointly with the dual problem (14). These screening tests may locate inactive groups in $\mathcal{G}$. According to (15), groups g such that $\|\mathbf{D}_{[g]}^T\tilde{\boldsymbol{\theta}}\|_2 < w_g$ correspond to inactive groups which can be removed from $\mathbf{D}$. The optimum $\tilde{\boldsymbol{\theta}}$ is not known but we can construct a region $\mathcal{R} \subset \mathbb{R}^N$ that contains $\tilde{\boldsymbol{\theta}}$ so that $\max_{\boldsymbol{\theta}\in\mathcal{R}} \|\mathbf{D}_{[g]}^T\boldsymbol{\theta}\|_2$ gives a sufficient condition to screen groups:

$$\max_{\boldsymbol{\theta}\in\mathcal{R}} \left\|\mathbf{D}_{[g]}^T\boldsymbol{\theta}\right\|_2 < w_g \Rightarrow \left\|\mathbf{D}_{[g]}^T\tilde{\boldsymbol{\theta}}\right\|_2 < w_g \Rightarrow \tilde{\mathbf{x}}_{[g]} = \mathbf{0} \tag{19}$$

There is no general closed-form solution of the maximization problem in (19) that would apply for arbitrary regions $\mathcal{R}$, moreover the quadratic nature of the maximization prevents closed-form solutions even for some simple regions $\mathcal{R}$. We now present the instance of this concept when $\mathcal{R}$ is a sphere. The sphere centered on $\mathbf{c}$ with radius r is denoted by $\mathcal{S}_{\mathbf{c},r}$.

Sphere tests: Consider a sphere $\mathcal{S}_{\mathbf{c},r}$ that contains the dual optimum $\tilde{\boldsymbol{\theta}}$, the screening test associated with this sphere requires to solve $\max_{\boldsymbol{\theta}\in\mathcal{S}_{\mathbf{c},r}} \|\mathbf{D}_{[g]}^T\boldsymbol{\theta}\|_2$ for each group g, which has no closed-form solution. Thus we use the triangle inequality to obtain a closed-form upper-bound on the solution: Lemma 7 provides the corresponding sphere-test.

Lemma 7 (Sphere Test for *Group-Lasso*)**.** *If $r \geq 0$ and $\mathbf{c} \in \mathbb{R}^N$ are such that $\tilde{\boldsymbol{\theta}} \in \mathcal{S}_{\mathbf{c},r}$, then the following function T^{GSPHERE} is a screening test for $\mathcal{P}_{\textit{Group-Lasso}}(\lambda, \mathbf{D}, \mathcal{G}, \mathbf{y})$:*

$$\begin{aligned} T^{\text{GSPHERE}} &: \Gamma \to \{0,1\} \\ i &\mapsto \left[\!\!\left[\left(\frac{w_{g(i)}}{\left\|\mathbf{D}_{[g(i)]}\right\|} - \frac{\left\|\mathbf{D}_{[g(i)]}^T\mathbf{c}\right\|_2}{\left\|\mathbf{D}_{[g(i)]}\right\|} \right) > r \right]\!\!\right] \end{aligned} \tag{20}$$

Proof. Let $i \in \Gamma$ such that $T^{\text{GSPHERE}}(i) = 1$ and $r \geq 0$ and $\mathbf{c} \in \mathbb{R}^N$ are such that $\tilde{\boldsymbol{\theta}} \in \mathcal{S}_{\mathbf{c},r}$. We use the triangle inequality to upper bound $\max_{\boldsymbol{\theta}\in\mathcal{S}_{\mathbf{c},r}} \|\mathbf{D}_{[g]}^T\boldsymbol{\theta}\|_2$:

$$\max_{\boldsymbol{\theta}\in\mathcal{S}_{\mathbf{c},r}} \left\|\mathbf{D}_{[g(i)]}^T\boldsymbol{\theta}\right\|_2 \leq \left\|\mathbf{D}_{[g(i)]}^T\mathbf{c}\right\|_2 + \max_{\boldsymbol{\theta}\in\mathcal{S}_{\mathbf{c},r}} \left\|\mathbf{D}_{[g(i)]}^T(\mathbf{c}-\boldsymbol{\theta})\right\|_2$$
$$\leq \left\|\mathbf{D}_{[g(i)]}^T\mathbf{c}\right\|_2 + r\left\|\mathbf{D}_{[g(i)]}\right\|$$

Which, as $T^{\text{GSPHERE}}(i) = 1$ gives $\left\|\mathbf{D}_{[g(i)]}^T\boldsymbol{\theta}\right\|_2 < w_{g(i)}$. Then using (19) we have that $\tilde{\mathbf{x}}_{[g(i)]} = \mathbf{0}$ and $\tilde{\mathbf{x}}(i) = 0$. □

Dynamic Construction of feasible point using Dual Scaling for the Group-Lasso: Before giving the proof of Lemmata 5 and 6, we need to introduce the dual-scaling strategy that computes from any dual point $\boldsymbol{\theta}$, a feasible dual point that satisfies (14b) by definition and aim at being close to $\mathbf{y}/\lambda$ to obtain an efficient screening test. Proposed by El Ghaoui in [15] for the *Lasso* this method applied to the *Group-Lasso* is given in the following Lemma:

Lemma 8. *Among all feasible scaled versions of $\boldsymbol{\theta}$, the closest to $\mathbf{y}/\lambda$ is $\mathbf{v} = \mu\boldsymbol{\theta}$ where:*

$$\mu \triangleq \left[\frac{\boldsymbol{\theta}^T\mathbf{y}}{\lambda\|\boldsymbol{\theta}\|_2^2}\right]_{-s_{\min}}^{s_{\min}} \quad \textit{with} \quad s_{\min} \triangleq \min_{g\in\mathcal{G}} \frac{w_g}{\left\|\mathbf{D}_{[g(i)]}^T\boldsymbol{\theta}\right\|_2}. \tag{21}$$

Proof. The dual-scaling problem for the *Group-Lasso* is:

$$\mu \triangleq \arg\min_{s\in\mathbb{R}} \left\|s\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda}\right\|_2 \quad \text{s.t. } \forall g \in \mathcal{G}, \left\|\mathbf{D}_{[g]}^T s\boldsymbol{\theta}\right\|_2 < w_g \tag{22}$$

The solution of (22) is the projection onto the feasible segment $[-s_{\min}, s_{\min}] \subset \mathbb{R}$ of the solution of $\arg\min_{s\in\mathbb{R}} \left\|s\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda}\right\|_2$. This solution is given in (21). □

We now prove the SAFE test for *Group-Lasso* using Lemmata 7 and 8 following the same arguments as in [15].

Proof of Lemma 5. Let $\boldsymbol{\theta} \in \mathbb{R}^N$ and $\mathbf{v}$ be its feasible scaled version obtained by dual-scaling (Lemma 8). Since $\mathbf{y}/\lambda$ attains the minimum of the unconstrained objective (14a) of the dual problem (14), and since $\mathbf{v}$ complies with all the constraints (14b), the distance between the optimum $\tilde{\boldsymbol{\theta}}$ and $\mathbf{y}/\lambda$ is upper bounded by $\left\|\frac{\mathbf{y}}{\lambda} - \mathbf{v}\right\|_2$, *i.e.*, $\tilde{\boldsymbol{\theta}} \in \mathcal{S}_{\frac{\mathbf{y}}{\lambda}, \left\|\frac{\mathbf{y}}{\lambda}-\mathbf{v}\right\|_2}$. And Lemma 7 concludes the proof. □

Proof of Lemma 6. This proof is illustrated graphically in Figure 5 in 2D. We first construct geometrically objects that are involved in the proof. Recalling that

$$g_* \triangleq \arg\max_g \frac{\left\|\mathbf{D}_{[g]}^T\mathbf{y}\right\|_2}{w_g} \text{ and } \lambda_* \triangleq \frac{\left\|\mathbf{D}_{[g_*]}^T\mathbf{y}\right\|_2}{w_{g_*}}, \tag{23}$$

we define the set of dual point complying with the constraint associated with group g_* as

$$\mathcal{V}_* \triangleq \left\{\boldsymbol{\theta} \in \mathbb{R}^N, \left\|\mathbf{D}_{[g_*]}\boldsymbol{\theta}\right\|_2 \leq w_{g_*}\right\}$$

This set $\mathcal{V}_*$ is the set contained in the ellipsoid:

$$\mathcal{E}_* \triangleq \left\{\boldsymbol{\theta} \in \mathbb{R}^N, \left\|\mathbf{D}_{[g_*]}\boldsymbol{\theta}\right\|_2^2 = w_{g_*}^2\right\}.$$

Point $\mathbf{y}/\lambda_*$ is on the ellipsoid: we define $\mathbf{n}$ as a normal vector to the ellipsoid $\mathcal{E}_*$ at this point. Such a vector is built from the gradient of $f(\boldsymbol{\theta}) \triangleq \frac{1}{2}\left\|\mathbf{D}_{[g_*]}^T\boldsymbol{\theta}\right\|_2^2$ at $\mathbf{y}/\lambda_*$:

$$\mathbf{n} \triangleq \nabla f\left(\frac{\mathbf{y}}{\lambda_*}\right) = \mathbf{D}_{[g_*]}\mathbf{D}_{[g_*]}^T\frac{\mathbf{y}}{\lambda_*} \tag{24}$$

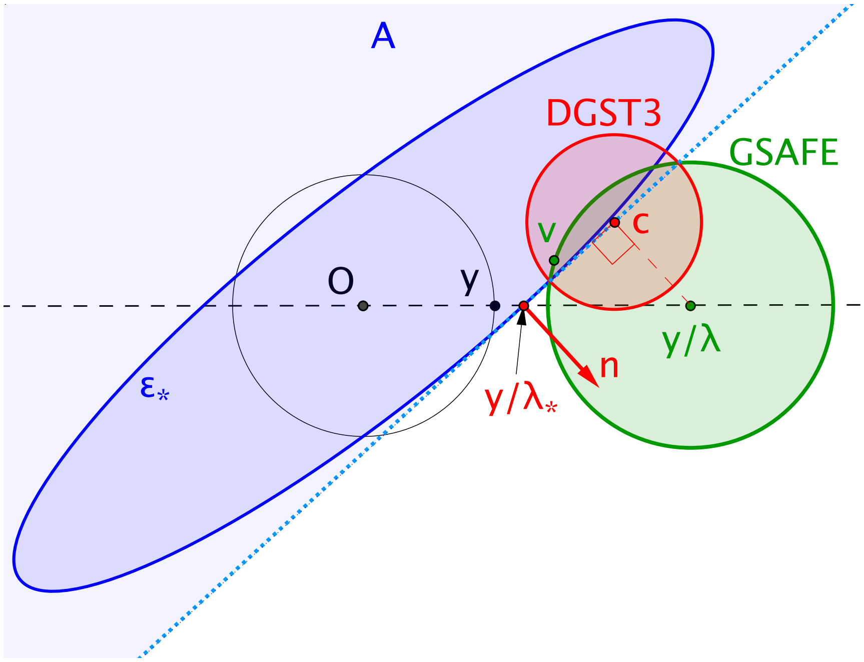

Figure 5: Geometrical illustration of regions associated to the screening tests for the *Group-Lasso*

We denote by $\mathcal{A}$ the half-space defined by the hyperplane tangent to the ellipsoid at $\mathbf{y}/\lambda_*$ that contains $\mathcal{V}_*$: $\mathcal{A} \triangleq \left\{\boldsymbol{\theta} \in \mathbb{R}^N, \boldsymbol{\theta}^T\mathbf{n} \leq w_{g_*}^2\right\}$.

By construction $\mathcal{V}_* \subset \mathcal{A}$. We now construct the new sphere $\mathcal{S}^{\text{DGST3}}$ containing $\tilde{\boldsymbol{\theta}}$. Let $\boldsymbol{\theta} \in \mathbb{R}^N$ and $\mathbf{v}$ the feasible scaled version of $\boldsymbol{\theta}$ obtained by dual-scaling. We know $\tilde{\boldsymbol{\theta}} \in \mathcal{V}_*$ (since $\tilde{\boldsymbol{\theta}}$ is feasible) and $\tilde{\boldsymbol{\theta}} \in \mathcal{S}_{\frac{\mathbf{y}}{\lambda},\left\|\frac{\mathbf{y}}{\lambda}-\mathbf{v}\right\|_2}$ (see proof of Lemma 5), so we have: $\tilde{\boldsymbol{\theta}} \in \mathcal{V}_* \cap \mathcal{S}_{\frac{\mathbf{y}}{\lambda},\left\|\frac{\mathbf{y}}{\lambda}-\mathbf{v}\right\|_2} \subset \mathcal{A} \cap \mathcal{S}_{\frac{\mathbf{y}}{\lambda},\left\|\frac{\mathbf{y}}{\lambda}-\mathbf{v}\right\|_2}$.

Then the new sphere $\mathcal{S}^{\text{DGST3}}$ is defined as the bounding sphere of $\mathcal{A} \cap \mathcal{S}_{\frac{\mathbf{y}}{\lambda},\left\|\frac{\mathbf{y}}{\lambda}-\mathbf{v}\right\|_2}$, its center is the projection of $\dfrac{\mathbf{y}}{\lambda}$ on $\mathcal{H}$ which is given by:

$$\mathbf{c} = \frac{\mathbf{y}}{\lambda} - \left(\frac{\mathbf{n}^T\mathbf{y}}{\lambda} - w_{g_*}^2\right)\frac{\mathbf{n}}{\|\mathbf{n}\|_2^2}$$

and its radius is given by the Pythagoras theorem:

$$r = \sqrt{\left\|\frac{\mathbf{y}}{\lambda} - \mathbf{c}\right\|_2^2 - \left\|\frac{\mathbf{y}}{\lambda} - \mathbf{v}\right\|_2^2}$$

We now formally check that $\tilde{\boldsymbol{\theta}} \in \mathcal{S}_{\mathbf{c},r}$ by showing that $\mathcal{A} \cap \mathcal{S}_{\frac{\mathbf{y}}{\lambda},\left\|\frac{\mathbf{y}}{\lambda}-\mathbf{v}\right\|_2} \subset \mathcal{S}^{\text{DGST3}}$. Let $\boldsymbol{\theta} \in \mathcal{A} \cap \mathcal{S}_{\frac{\mathbf{y}}{\lambda},\left\|\frac{\mathbf{y}}{\lambda}-\mathbf{v}\right\|_2}$, then:

$$\begin{aligned}
\left\|\frac{\mathbf{y}}{\lambda} - \mathbf{v}\right\|_2^2 \geq \left\|\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda}\right\|_2^2 \\
&= \left\|\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda} + \left(\frac{\mathbf{n}^T\mathbf{y}}{\lambda} - w_{g_*}^2 - \frac{\mathbf{n}^T\mathbf{y}}{\lambda} + w_{g_*}^2\right)\frac{\mathbf{n}}{\|\mathbf{n}\|_2^2}\right\|_2^2 \\
&= \|\boldsymbol{\theta} - \mathbf{c}\|_2^2 + \left\|\left(\frac{\mathbf{n}^T\mathbf{y}}{\lambda} - w_{g_*}^2\right)\frac{\mathbf{n}}{\|\mathbf{n}\|_2^2}\right\|_2^2 - 2\frac{w_{g_*}^2}{\|\mathbf{n}\|_2^2}\left(\frac{\lambda_*}{\lambda} - 1\right)\left(\boldsymbol{\theta}^T\mathbf{n} - \frac{\lambda_*}{\lambda}w_{g_*}^2 + \left(\frac{\lambda_*}{\lambda} - 1\right)w_{g_*}^2\right).
\end{aligned}$$

Where the last equality is obtained using the definition of $\mathbf{n}$:

$$\frac{\mathbf{n}^T\mathbf{y}}{\lambda} - w_{g_*}^2 = \frac{\lambda_*}{\lambda}\frac{\mathbf{y}^T\mathbf{n}}{\lambda_*} - w_{g_*}^2 = \left(\frac{\lambda_*}{\lambda} - 1\right) w_{g_*}^2$$

Then, using the definition of $\mathbf{c}$, we have:

$$\left\|\frac{\mathbf{y}}{\lambda} - \mathbf{v}\right\|_2^2 \geq \|\boldsymbol{\theta} - \mathbf{c}\|_2^2 + \left\|\frac{\mathbf{y}}{\lambda} - \mathbf{c}\right\|_2^2 + 2\frac{w_{g_*}^2}{\|\mathbf{n}\|_2^2}\left(\frac{\lambda_*}{\lambda} - 1\right)\left(w_{g_*}^2 - \boldsymbol{\theta}^T\mathbf{n}\right).$$

As $\boldsymbol{\theta}$ is contained in $\mathcal{A}$ we have $0 \leq w_{g_*}^2 - \boldsymbol{\theta}^T\mathbf{n}$ and:

$$\left\|\frac{\mathbf{y}}{\lambda} - \mathbf{v}\right\|_2^2 \geq \|\boldsymbol{\theta} - \mathbf{c}\|_2^2 + \left\|\frac{\mathbf{y}}{\lambda} - \mathbf{c}\right\|_2^2$$

We finally obtain the radius:

$$\|\boldsymbol{\theta} - \mathbf{c}\|_2^2 \leq \left\|\frac{\mathbf{y}}{\lambda} - \mathbf{v}\right\|_2^2 - \left\|\frac{\mathbf{y}}{\lambda} - \mathbf{c}\right\|_2^2 = r^2.$$

Then $\mathcal{A} \cap \mathcal{S}_{\frac{\mathbf{y}}{\lambda}, \left\|\frac{\mathbf{y}}{\lambda} - \mathbf{v}\right\|_2} \subset \mathcal{S}^{\mathrm{DGST3}}$ and $\tilde{\boldsymbol{\theta}} \in \mathcal{S}^{\mathrm{DGST3}}$. Lemma 7 concludes the proof of Lemma 6. □