Fast Solving of the Group-Lasso via Dynamic Screening

Antoine Bonnefoy, Valentin Emiya, Liva Ralaivola, Rémi Gribonval

Fast Solving of the *Group-Lasso* via Dynamic Screening.

Antoine Bonnefoy*, Valentin Emiya *, Liva Ralaivola *, Rémi Gribonval†.
* Aix-Marseille Université, CNRS UMR 7279 LIF, †Inria.

I. Introduction

The *Group-Lasso* [1] is an optimization problem devoted to finding a *group-sparse approximation* of some signal $\mathbf{y} \in \mathbb{R}^N$ with respect to a predefined dictionary $\mathbf{D} \triangleq [\mathbf{d}_1, \dots, \mathbf{d}_K] \in \mathbb{R}^{N \times K}$ and $\mathcal{G}$ a partition of $[1 \dots K]$, via the minimization of the sum of an ℓ_2-fitting term and a penalization term inducing the group-sparsity. Given $\mathbf{D}, \mathcal{G}$, and $\mathbf{y}$ the *Group-Lasso* problem $\mathcal{P}(\lambda, \mathbf{D}, \mathcal{G}, \mathbf{y})$ writes:

$$\tilde{\mathbf{x}} \triangleq \underset{\mathbf{x} \in \mathbb{R}^K}{\arg\min} \frac{1}{2} \|\mathbf{D}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \sum_{g \in \mathcal{G}} w_g \|\mathbf{x}_{[g]}\|_2, \qquad (1)$$

where the parameter $\lambda > 0$ governs the group-sparsity of $\tilde{\mathbf{x}}$; the sub-vector $\mathbf{x}_{[g]} \triangleq [\mathbf{x}(i)]_{i \in g}^T$ is defined as the concatenation of the coefficients of $\mathbf{x}$ indexed by the elements of g and $\{w_g\}_{g \in \mathcal{G}}$ are the weights associated to each group.

Iterative algorithms such as ISTA and its variants [2], [3], [4], [5] are well suited to handle this problem in real applications for which both N and K may be large. Accelerating these algorithms is yet a key challenge: they remain captive of the dictionary size due to the required multiplications by $\mathbf{D}$ and $\mathbf{D}^T$ over the optimization process.

To overcome this limitation on the *Group-Lasso* problem, we propose to extend the *dynamic screening principle* [6], initially designed for the *Lasso*, to the *Group-Lasso*. The method rests upon the idea of *screening test* [7], [8], [9], [10] to accelerate the computation of the solution of $\mathcal{P}(\lambda, \mathbf{D}, \mathcal{G}, \mathbf{y})$. For the *Lasso*, a screening test aims at locating some of the zeros in the solution $\tilde{\mathbf{x}}$ at low computational cost, in order to construct the reduced or *screened* dictionary $\mathbf{D}_0$, which is dictionary $\mathbf{D}$ trimmed off of its columns that correspond to the located zeros. The solution of the *Lasso* in the reduced dimension is then computed faster than the one of the full *Lasso*. $\tilde{\mathbf{x}}$ is finally reconstructed from the solution of the reduced problem by inserting back the located zeros. Resorting to screening-tests for a *Group-Lasso* problem requires to be able to locate groups $g \in \mathcal{G}$ such that $\tilde{\mathbf{x}}_{[g]} = \mathbf{0}$.

The dynamic screening principle improves previously existing screening tests by taking advantage of the computation made during the optimization procedure to perform a *screening* at each iteration with a negligible computational overhead, and to consequently *dynamically and iteratively* reduce the size of $\mathbf{D}$. Opposing perspectives of the proposed *dynamic* screening and existing *static* screening are schematized in Algorithms 1 and 2.

Algorithm 1
Static screening strategy

$\mathbf{D}_0 \leftarrow$ Screen $\mathbf{D}$
loop k
$\mathbf{x}_{k+1} \leftarrow$Update $\mathbf{x}_k$ using $\mathbf{D}_0$
end loop

Algorithm 2
Dynamic screening strategy

$\mathbf{D}_0 \leftarrow \mathbf{D}$
loop k
$\mathbf{x}_{k+1} \leftarrow$ Update $\mathbf{x}_k$ using $\mathbf{D}_k$
$\mathbf{D}_{k+1} \leftarrow$ Screen $\mathbf{D}_k$ using $\mathbf{x}_{k+1}$
end loop

In the present work we study the extension of the ST3 [9] screening test, to the *Group-Lasso* in a dynamic screening scheme. We call this extension DGST3 for Dynamic-Group-ST3. More details on this work can be found in the journal paper [11] currently under review, which is an extension of [6].

II. Method

We state here the main Lemma on dynamic screening for the *Group-Lasso*. Proofs, which rely on the dual problem and the optimality condition of the *Group-Lasso*, are given in [11]. We define λ_* as the smallest penalization parameter resulting into a zero solution of (1): $g_* \triangleq \arg\max_g \left\|\mathbf{D}_{[g]}^T \mathbf{y}\right\|_2 w_g^{-1}$, and $\lambda_* \triangleq \left\|\mathbf{D}_{[g_*]}^T \mathbf{y}\right\|_2 w_{g_*}^{-1}$. $\mathbf{D}_{[g]} \triangleq [\mathbf{d}_i]_{i \in g}$ denotes the sub-dictionary indexed by g.

Lemma 1 (The Dynamic Group ST3: DGST3)**.** *Given a problem $\mathcal{P}(\lambda, \mathbf{D}, \mathcal{G}, \mathbf{y})$, and any iterate $\mathbf{x}$ of an iterative algorithm, let us define $\boldsymbol{\theta} = \mathbf{D}\mathbf{x} - \mathbf{y}$, one can screen all groups $g \in \mathcal{G}$ satisfying:*

$$\left(\frac{w_g}{\left\|\mathbf{D}_{[g]}\right\|_2} - \frac{\left\|\mathbf{D}_{[g]}^T \mathbf{c}\right\|_2}{\left\|\mathbf{D}_{[g]}\right\|_2} \right) > r_{\boldsymbol{\theta}}$$

where

$$r_{\boldsymbol{\theta}} \triangleq \sqrt{\left\|\frac{\mathbf{y}}{\lambda} - \mu\boldsymbol{\theta}\right\|_2^2 - \left\|\frac{\mathbf{y}}{\lambda} - \mathbf{c}\right\|_2^2},$$

$$\mu \triangleq \operatorname{sign}(\boldsymbol{\theta}^T \mathbf{y}) \min\left(\frac{|\boldsymbol{\theta}^T \mathbf{y}|}{\lambda \|\boldsymbol{\theta}\|_2^2}, \min_{g \in \mathcal{G}} w_g \left\|\mathbf{D}_{[g]}^T \boldsymbol{\theta}\right\|_2^{-1} \right)$$

$$\mathbf{c} \triangleq \left(\mathbf{Id} - \frac{\mathbf{n}\mathbf{n}^T}{\|\mathbf{n}\|_2^2} \right) \frac{\mathbf{y}}{\lambda} + \frac{\mathbf{n}}{\|\mathbf{n}\|_2^2} w_{g_*}^2, \mathbf{n} \triangleq \mathbf{D}_{[g_*]} \mathbf{D}_{[g_*]}^T \frac{\mathbf{y}}{\lambda_*}.$$

Note that expensive computation appears when computing the matrix-vector products for $\boldsymbol{\theta}, \mu$ and $\mathbf{c}$. $\mathbf{c}$ can be computed once for all the iterations of the algorithm.Actually all the variants of ISTA already compute $\boldsymbol{\theta}$ and $\mathbf{D}^T\boldsymbol{\theta}$ so we have $\boldsymbol{\theta}$ and μ almost for free.

The proposed screening test is given for the *Group-Lasso*, but can be readily extended to the *Overlapping Group-Lasso* [12] thanks to the replication trick.

III. Results

Various experiments are proposed in [11], we present here one of them on synthetic data. For this experiment we used a Pnoise dictionary introduced in [9], for which all $\{\mathbf{d}_i\}_{i=1}^K$ are drawn i.i.d. as $\mathbf{e}_1 + 0.1\mathcal{U}(0,1)\mathcal{N}(\mathbf{0}, \mathbf{Id}_N)$ and normalized, where $\mathbf{e}_1 \triangleq [1, 0, \dots, 0]^T \in \mathbb{R}^N$. We took $N = 2000, K = 10000$. Groups were built randomly with the same number of atoms in each group. Observations $\mathbf{y}$ were generated from a Bernoulli-Gaussian distribution: $\mathbf{x}_\mathbf{y} = [b_g \mathbf{x}_{[g]}^T]^T$, where $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{Id}_K)$ and for all group g, b_g was drawn independently from a Bernoulli distribution of parameter $p = 0.05$. The observation $\mathbf{y} \triangleq \mathbf{D}\mathbf{x}_\mathbf{y}$ was corrupted by additive Gaussian noise to reach a SNR of 20dB and normalized.

On 30 realization of the Pnoise data for each group-size, we solved the *Group-Lasso* problem for different λ values, with 3 variants of FISTA: (i) the base FISTA as describe in [3], (ii) FISTA with initial GST3 static screening, and (iii) FISTA with DGST3 dynamic screening. Figure 1 shows the normalized flops number of (ii) (circle) and (iii) (black square) with respect to (i), as a function of λ/λ_*. Low values account for fast computation.

Experiments show that the DGST3 significantly reduces the computational cost of the optimization compared to the GST3 in a large range of λ values. The computational saving reaches up to 90% and 80% with respect to the base algorithm and the algorithm with static screening, respectively.

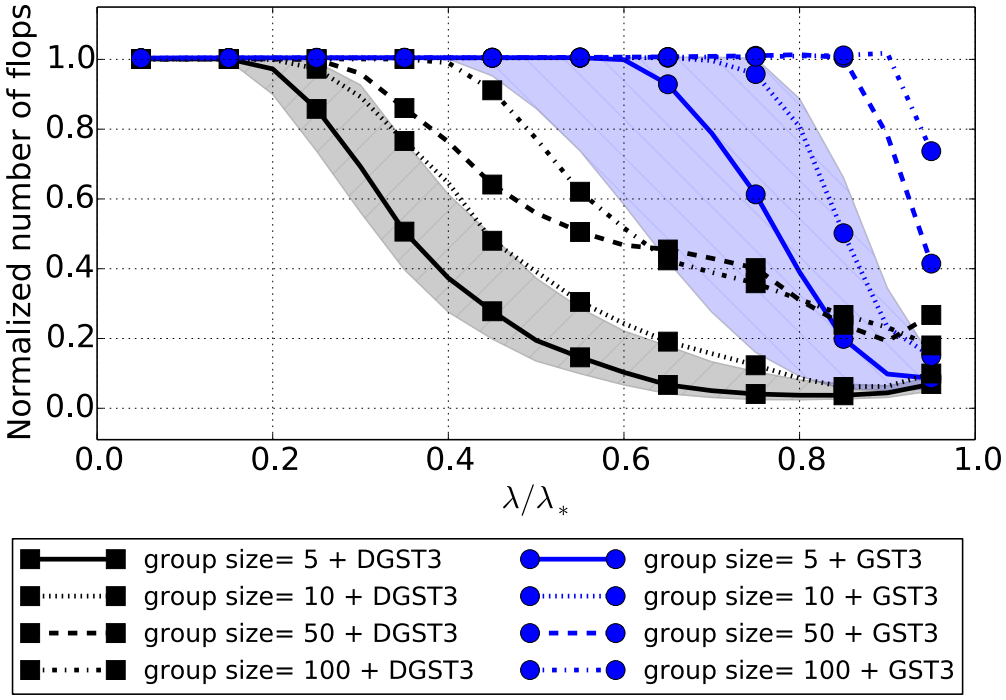

Figure 1: Normalized number of flops on Pnoise dictionary w.r.t. FISTA with no screening, one corresponds to no speed up and low values account for fast computation. Various group sizes are considered. The median over 30 runs are plotted, and for group size = 5, the shaded area contains the 25%-to-75% percentiles.

ACKNOWLEDGMENT

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.

REFERENCES

[1] 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.

[2] 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.

[3] 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. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/080716542

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

[5] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. on Sig. Proc.*, vol. 57, no. 7, pp. 2479–2493, 2009.

[6] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval, "A Dynamic Screening Principle for the Lasso," in *Proc. of EUSIPCO*, 2014. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00880787

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

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

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

[10] 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," *J. of the Royal Stat. Society: Series B*, vol. 74, no. 2, pp. 245–266, 2012. [Online]. Available: http://doi.wiley.com/10.1111/j.1467-9868.2011.01004.x

[11] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval. (2014) Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso. [Online]. Available: http://arxiv.org/abs/1412.4080

[12] 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.