Gap Safe screening rules for faster complex-valued multi-task group Lasso

Mathurin Massias, Joseph Salmon, Alexandre Gramfort
LTCI, Télécom ParisTech, Université Paris-Saclay
46 rue Barrault, 75013, Paris, France
Email: {mathurin.massias, joseph.salmon, alexandre.gramfort}@telecom-paristech.fr

***Abstract*—Linear regression with sparsity-inducing penalties is a popular tool for high-dimensional inverse problems such as source localization. The Group Lasso is a particular choice of penality that considers the $\ell_{2,1}$ norm to promote group (or block) sparsity patterns. Since no general closed-form solution is available for this problem, iterative solvers are needed. This can lead to very slow convergence especially if the problem is ill-conditionned or if the dimension of the problem is particularly large. Safe screening rules [3] (see also [12]) and in particular dynamic ones [1, 2] speed-up the optimization process by progressively discarding regressors identified as irrelevant. In this work, we consider the case where features and observations can be complex-valued, a common case in signal processing when working with time-frequency operators. We derive Gap Safe screening rules in this context and propose a block coordinate descent optimization strategy [11]. In practice, we illustrate significant speed-ups in terms of convergence compared to classical solvers on a neuroscience problem, namely the problem of source localization using magneto and electroencephalography (MEG/EEG).**

I. The complex-valued multi-task group Lasso

In the following, n represents the number of observations (or sensors), q the number of tasks (or time instants), and p the number of features (or variables). Given a matrix of observations $Y \in \mathbb{C}^{n\times q}$, a design matrix (or forward operator) $X \in \mathbb{C}^{n\times p}$ and a set of groups $\mathcal{G}$ (*i.e.,* a partition of $\{1,\dots,p\}$) the complex-valued multi-task group Lasso solves $P_\lambda(\beta)$: $\beta^* \in \underset{\beta\in\mathbb{C}^{p\times q}}{\arg\min}\ \frac{1}{2}\|Y - X\beta\|_F^2 + \lambda\|\beta\|_{F,1}$, where $\lambda > 0$ is a regularization parameter, $\|\beta\|_{F,1} = \sum_{g\in\mathcal{G}} \|\beta_g\|_F$, β_g denotes the sub-matrix of β composed of rows whose indices are in g and $\|A\|_F = (\sum_{i,j}|A_{i,j}|^2)^{\frac{1}{2}}$ denotes the Frobenius norm of A.

For $z, z' \in \mathbb{C}^d$, instead of the Hermitian inner product $\langle z, z'\rangle_{\mathcal{H}} = \sum_{i=1}^d z_i\bar{z}'_i$ we use $\langle z, z'\rangle = \frac{1}{2}\sum_{i=1}^d (z_i\bar{z}'_i + \bar{z}_i z'_i)$. Contrary to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, it is real-valued and enables us to define the Fenchel conjugate of a function $f : \mathbb{C}^d \to \mathbb{R}$ as $f^* : u \mapsto \sup_{z\in\mathbb{C}^d} \langle u, z\rangle - f(u)$. For $X, Y \in \mathbb{C}^{d\times d'}$, we consider $\langle X, Y\rangle = \frac{1}{2}\operatorname{tr}(X^{\mathcal{H}}Y + Y^{\mathcal{H}}X)$. Using this we can prove that $\|\cdot\|_{F,1}^* = \|\cdot\|_{F,\infty}$, and that $\frac{1}{2}\|\cdot\|_F^2$ is its self Fenchel conjugate. The dual problem of $P_\lambda(\beta)$ then reads:

$$\sup_{\theta\in\mathbb{C}^{n\times q}} \inf_{\substack{\beta\in\mathbb{C}^{p\times q}\\ \mu\in\mathbb{C}^{n\times q}}} \frac{1}{2}\|\mu\|_F^2 + \lambda\|\beta\|_{F,1} + \langle\theta, \mu - Y + X\beta\rangle \ .$$

With the properties of the Fenchel transform, this amounts to solving $D_\lambda(\theta)$: $\max_{\theta\in\Delta_X} \|Y\|_F^2/2 - \lambda^2\|Y/\lambda - \theta\|_F^2/2$, introducing the dual feasible set $\Delta_X = \left\{\theta \in \mathbb{C}^{n\times q} \mid \|X^{\mathcal{H}}\theta\|_{F,\infty} \le 1\right\}$. A popular iterative solver for the multi-task group lasso uses a block coordinate descent (BCD) scheme [11]: at each iteration k, for each group g successively, $P_\lambda(\beta)$ is optimized *w.r.t.* β_g.

II. Gap safe screening rules for complex Lasso

Let β^* be an optimal solution of the complex-valued multi-task group Lasso $P_\lambda(\beta)$ and θ^* be the (unique) dual solution, both linked through $Y = X\beta^* + \lambda\theta^*$. Fermat's rule reads:

$$\forall g, X_g^{\mathcal{H}}\theta^* \in \partial\|\beta_g^*\|_F = \begin{cases} \beta_g^*/\left\|\beta_g^*\right\|_F, & \text{if } \beta_g^* \neq 0 \\ \mathcal{B}_{\|\cdot\|_F}, & \text{otherwise} \end{cases},$$

where $\partial\|\cdot\|_F$ is defined relatively to the real-valued $\langle\cdot,\cdot\rangle$, and $\mathcal{B}_{\|\cdot\|_F}$ stands for the $\|\cdot\|_F$-unit ball. We thus have $\left\|X_g^{\mathcal{H}}\theta^*\right\|_F < 1 \Rightarrow \beta_g^* = 0$. Though, θ^* being unknown, this rule has to be relaxed: given a *safe region* $\mathcal{C}$ containing θ^*, the rule becomes $\sup_{\theta\in\mathcal{C}} \left\|X_g^{\mathcal{H}}\theta\right\|_F < 1 \Rightarrow \beta_g^* = 0$. In a series of work [4, 8, 9] so called *Gap Safe* rule have been proposed. It uses at iteration k the safe ball centered on $\theta_k = (Y - X\beta_k)/(\lambda\alpha_k)$ (α_k chosen so that θ_k is dual feasible), with radius $\sqrt{2(P_\lambda(\beta_k) - D_\lambda(\theta_k))/\lambda^2}$. To ensure that $D_\lambda(\theta_k)$ is increasing, we slightly modify this rule, and update θ_k only when $D_\lambda(\theta_k) > D_\lambda(\theta_{k-1})$. We use $\theta_0 = Y/\lambda_{\max}$, where $\lambda_{\max} = \|X^{\mathcal{H}}Y\|_{F,\infty}$.

III. Numerical experiments

The MEG/EEG inverse problem with $\ell_{2,1}$ regularization leads to a particular case of multi-task group Lasso [10, 5]: in this context, Y is a matrix of sensor measurements (n signals of length q), X is the composition of the forward operator (encoding the electromagnetic dependency between source amplitudes and measurements) and a time-frequency decomposition operator, and β is the complex-valued matrix of time-frequency coefficients [7]. We adopt the *free orientation* setting which leads to the estimation of a vector field: blocks of three consecutives rows of β represent the activity of one source, decomposed over three orthogonal spatial dimensions. Imposing a $\|\cdot\|_{F,1}$ group penalty over these blocks results in a solution where only a few sources are active, which is a desired property from a biological standpoint.

We use data from the MNE dataset [6], with $n = 302$ MEG sensors, 7498 sources (22494 oriented dipoles), and 181 time instants which corresponds to about 300 ms of event related field (ERF) data following an auditory stimulation in the left ear. We decompose the signals over 1518 time-frequency atoms.

We evaluate the acceleration obtained with dynamic screening for different values of λ (screening is done every 10 pass over all groups with initialization $\beta_0 = 0$). Fig. 1 shows objective convergence for $\lambda = \lambda_{\max}/4$. Screening restricts the BCD steps to a smaller and smaller set of sources, resulting in significant acceleration. As we see in fig. 2, high values of λ result in very sparse solutions, for which screening is well-suited. On the contrary, for a very low λ, few features are discarded and screening does not greatly speed up the convergence.

Acknowledgment

This work was funded by the European Research Council (ERC-YStG-676943).

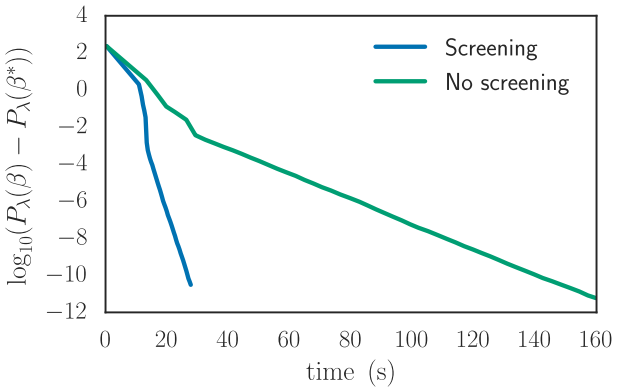

Fig. 1. Convergence toward the optimal objective value for $\lambda = \lambda_{\max}/4$, with (blue) and without screening (green).

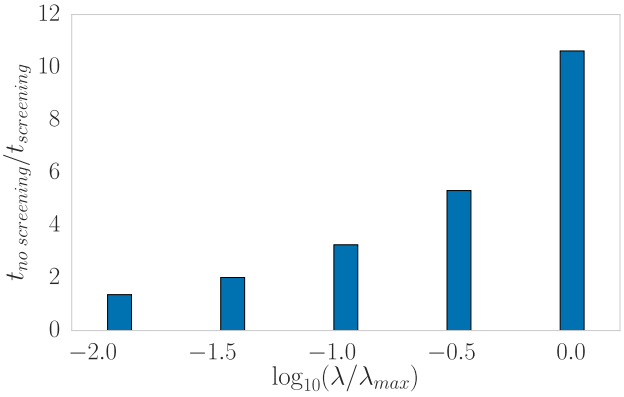

Fig. 2. Ratio between the times to reach a dual gap of 10^{-5} without and with screening, as a function of λ.

References

[1] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval. A dynamic screening principle for the lasso. In *EUSIPCO*, 2014.

[2] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval. Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso. *IEEE Trans. Signal Process.*, 63(19):20, 2015.

[3] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. *J. Pacific Optim.*, 8(4):667–698, 2012.

[4] O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: safer rules for the Lasso. In *ICML*, pages 333–342, 2015.

[5] A. Gramfort, M. Kowalski, and M. S. Hämäläinen. Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Phys. Med. Biol.*, 57(7):1937–1961, Apr. 2012.

[6] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen. MNE software for processing MEG and EEG data. *NeuroImage*, 86:446 – 460, 2014.

[7] A. Gramfort, D. Strohmeier, J. Haueisen, M.S. Hämäläinen, and M. Kowalski. Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations. *NeuroImage*, 70:410 – 422, 2013.

[8] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. GAP safe screening rules for sparse multi-task and multi-class models. *NIPS*, pages 811–819, 2015.

[9] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparsity enforcing penalties. *arXiv preprint: 1611.05780*, 2016.

[10] W. Ou, M. S. Hämaläinen, and P. Golland. A distributed spatio-temporal EEG/MEG inverse solver. *NeuroImage*, 44(3):932–946, Feb. 2009.

[11] Z. Qin, K. Scheinberg, and D. Goldfarb. Efficient block-coordinate descent algorithms for the group lasso. *Mathematical Programming Computation*, 5(2):143–169, 2013.

[12] Z. J. Xiang, Y. Wang, and P. J. Ramadge. Screening tests for lasso problems. *arXiv preprint: 1405.4897*, 2014.