Safe Screening for Sparse Regression with the Kullback-Leibler Divergence

Cássio F. Dantas, Emmanuel Soubies, Cédric Févotte

IRIT, Université de Toulouse, CNRS, Toulouse, France

3350

Context

Consider the ℓ_1-regularized Kullback-Leibler divergence regression:

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^m_+} P_\lambda(\mathbf{x}) := \mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{A}\mathbf{x}) + \lambda\|\mathbf{x}\|_1 \qquad \text{(KL-}\ell_1\text{)}$$

Goal: given $\mathbf{y}$ and $\mathbf{A}$, find $\mathbf{x}$ sparse and non-negative.

- $\mathbf{y} \in \mathbb{R}^m_+$ is the observation vector,
- $\mathbf{A} \in \mathbb{R}^{m\times n}_+$ is the measurement matrix,
- $\mathbf{x} \in \mathbb{R}^n_+$ is the signal of interest.

Data-fidelity term is the generalized Kullback-Leibler (KL) divergence

$$\mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{z}) = \sum_{i=1}^m y_i \log\left(\frac{y_i}{z_i+\epsilon}\right) - y_i + (z_i+\epsilon). \qquad (1)$$

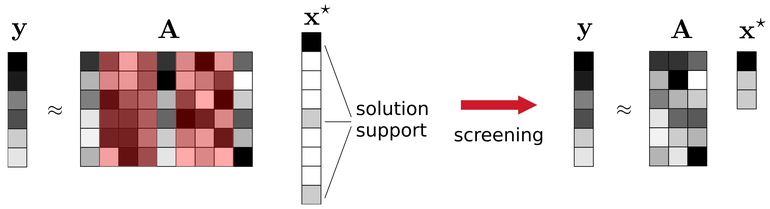
It corresponds to the Poisson negative log-likelihood, up to a constant.

Main objective

What: Accelerate the resolution of the KL-ℓ_1 problem.
How: Deriving a safe screening strategy for the KL divergence.

Safe Screening

Identify and eliminate coordinates not belonging to the solution support.
Zero risk of false identification! [1]



Dual Problem and Optimality Conditions

Dual Problem:

$$\boldsymbol{\theta}^\star = \operatorname*{argmax}_{\boldsymbol{\theta}\in\mathcal{F}_{\mathbf{A}}} D_\lambda(\boldsymbol{\theta}) := \sum_{i=1}^m y_i \log(1+\lambda\theta_i) - \epsilon\lambda\theta_i, \qquad (2)$$
$$\text{with } \mathcal{F}_{\mathbf{A}} = \{\boldsymbol{\theta}\in\mathbb{R}^m \mid \lambda\boldsymbol{\theta} \geq -\mathbf{1},\ \mathbf{A}^\mathsf{T}\boldsymbol{\theta} \leq \mathbf{1}\}$$

First-order optimality conditions:

$$\lambda\boldsymbol{\theta}^\star = \frac{\mathbf{y}}{\mathbf{A}\mathbf{x}^\star+\epsilon} - \mathbf{1} \qquad (3)$$

$$\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star = \begin{cases} 1, & \text{if } x_j^\star > 0, \\ \varrho \leq 1, & \text{if } x_j^\star = 0. \end{cases} \qquad (4)$$

Safe Screening Rule

Consequence of (3): $y_i = 0 \implies \theta_i^\star = -1/\lambda$
Consequence of (4): $\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star < 1 \implies x_j^\star = 0$

Proposition (KL-ℓ_1 Safe Screening Rule). Let $\mathcal{I} = \{i \in [m] : y_i = 0\}$ and $\mathcal{S} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\theta}_\mathcal{I} = -\mathbf{1}/\lambda\}$. Let $\boldsymbol{\theta} \in \mathcal{S}$ and $r > 0$ be such that $\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r)$. Then,

$$\mathbf{a}_j^\mathsf{T}\boldsymbol{\theta} + r\|[\mathbf{a}_j]_\mathcal{I}\|_2 < 1 \quad \implies \quad x_j^\star = 0. \qquad (5)$$

Safe Region

Theorem (KL-ℓ_1 GAP Safe Sphere) Let $(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^n_+ \times (\mathcal{F}_\mathbf{A} \cap \mathcal{S})$ be a primal-dual feasible pair. Then, for

$$r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\bar{\alpha}}} \qquad (6)$$

$$\bar{\alpha} = \lambda^2 \min_{i\in\mathcal{I}^\complement} \frac{y_i}{(1+\max(\|\mathbf{A}\|_1, \lambda)\|\mathbf{a}_i^\dagger\|_1)^2}, \qquad (7)$$

$\mathcal{B}(\boldsymbol{\theta}, r)$ is a safe region, i.e., $\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r)$.

- Adaptation of the existing GAP Safe Sphere [2].
- We use a **local** strong concavity bound $\bar{\alpha}$ on the set $\mathcal{F}_\mathbf{A} \cap \mathcal{S}$ instead of a global bound (nonexistent in our case) as in [2].

Proposed Algorithm

- Usual dynamic safe screening framework.
- Combine with any existing iterative solver for the KL-ℓ_1 problem.

Algorithm 1 : KL-L1 Dynamic GAP Safe Screening

Initialize $\mathbf{x} \in \mathbb{R}^n$, $\mathcal{A} = \{1, \ldots, n\}$, $\bar{\alpha}$ as in (7)

Repeat until convergence

- Primal update : $\mathbf{x}_\mathcal{A} \leftarrow \texttt{PrimalUpdate}(\mathbf{x}_\mathcal{A}, \mathbf{A}_\mathcal{A}, \mathbf{y}, \lambda)$
- Dual update : $\boldsymbol{\theta} \leftarrow \boldsymbol{\Theta}(\mathbf{x}) \in \mathcal{F}_\mathbf{A} \cap \mathcal{S}$
- Safe screening : $r \leftarrow \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\bar{\alpha}}}$
  $\mathcal{A} \leftarrow \{j \in \mathcal{A} \mid \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta} + r\|[\mathbf{a}_j]_\mathcal{I}\|_2 \geq 1\}$
  $\mathbf{x}_{\mathcal{A}^\complement} \leftarrow \mathbf{0}$

Experiments

- **Solvers:** Proximal gradient (SPIRAL), Coordinate descent (CoD), Multiplicative update (MU).
- **Real count datasets:** 20-Newsgroups, NIPS papers (word counts), TasteProfile (song listening counts).
- Input vector $\mathbf{y}$ is a randomly selected column of the dataset. Remaining data forms $\mathbf{A}$.

Table 1: Average speedups (time without/with screening).

	$\lambda/\lambda_{\max}$	10^{-1}		10^{-3}	
	$\varepsilon_{\mathrm{gap}}$	10^{-5}	10^{-7}	10^{-5}	10^{-7}
20 Newsgroups	SPIRAL	1.44	1.59	1.60	1.78
	CoD	2.44	3.42	2.46	3.22
	MU	4.80	7.72	4.49	7.28
NIPS papers	SPIRAL	2.77	3.21	2.26	2.53
	CoD	4.19	5.35	4.12	5.06
	MU	6.71	8.88	5.74	7.31
TasteProfile	SPIRAL	2.54	3.00	2.82	3.21
	CoD	1.75	2.20	2.44	4.22
	MU	2.81	4.11	2.94	4.35

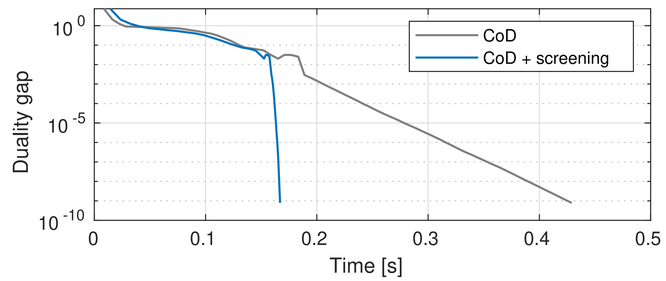


Figure 1: CoD solver convergence vs. time. 20Newsgroups data, $\lambda = 10^{-2}\lambda_{\max}$.

Code available at github.com/cassiofragadantas
Check out our follow-up paper, available here!

References

[1] L. El Ghaoui, V. Viallon, and T. Rabbani, "Safe feature elimination for the lasso and sparse supervised learning problems," *Pacific Journal of Optimization*, 2012.

[2] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon, "Gap safe screening rules for sparsity enforcing penalties," *JMLR*, 2017.

cassio.fraga-dantas@irit.fr