Safe Screening for Sparse Regression with the Kullback-Leibler Divergence

Cassio F. Dantas

Emmanuel Soubies and Cédric Févotte

IRIT, Université de Toulouse, CNRS

European Research Council
Established by the European Commission

Institut de Recherche
en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

Outline

Context and Literature

1. Motivation
2. Safe screening : a quick overview

Our contribution

3. Problem definition
4. Safe screening for the Kullback-Leibler divergence
   - Dual problem and optimality conditions
   - Screening rule and Safe region
5. Experimental results

Motivations

- Goal: accelerate the solution of sparse regression problems with the generalized **Kullback-Leibler** (KL) divergence → safe screening.

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

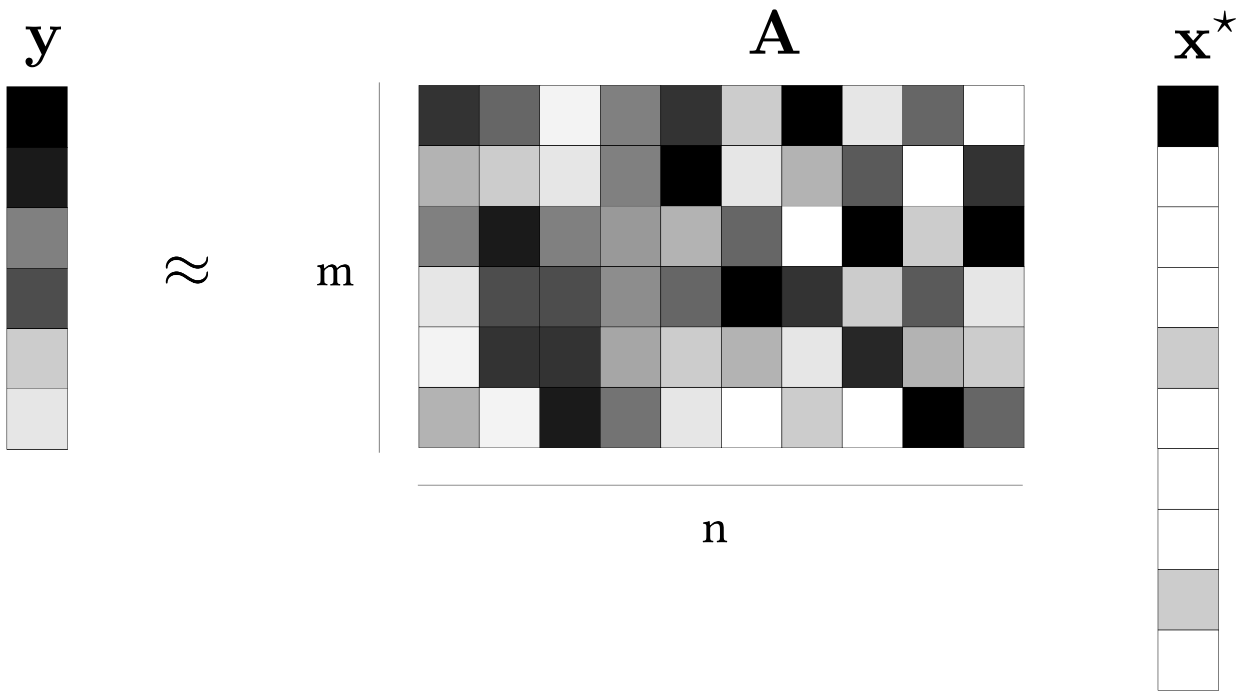
- Maximum likelihood estimation with a **Poisson** observation model.

- Applications:
  - Sparse NMF
  - Count data
    - Text processing: word count
    - Recommendation: view / listening count
  - Medical imaging (Positron emission tomography)

Safe Screening

- Accelerate the solution of **sparse regression problems**.

$$\mathbf{y} \approx \mathbf{A}\mathbf{x}\,, \quad \text{with } \mathbf{x} \text{ sparse}$$

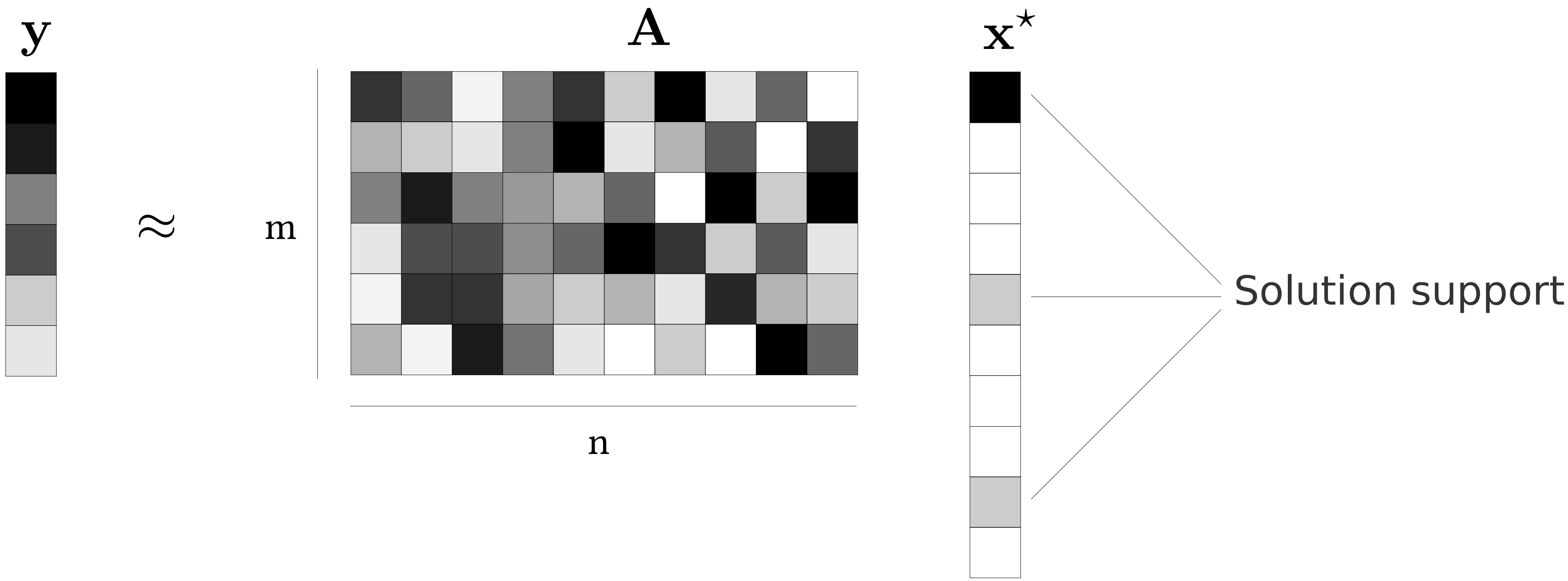


Safe Screening

- Accelerate the solution of **sparse regression problems**.

$$\mathbf{y} \approx \mathbf{A}\mathbf{x}\,, \quad \text{with } \mathbf{x} \text{ sparse}$$

- Core idea: identify and eliminate coordinates not belonging to the **solution support**.

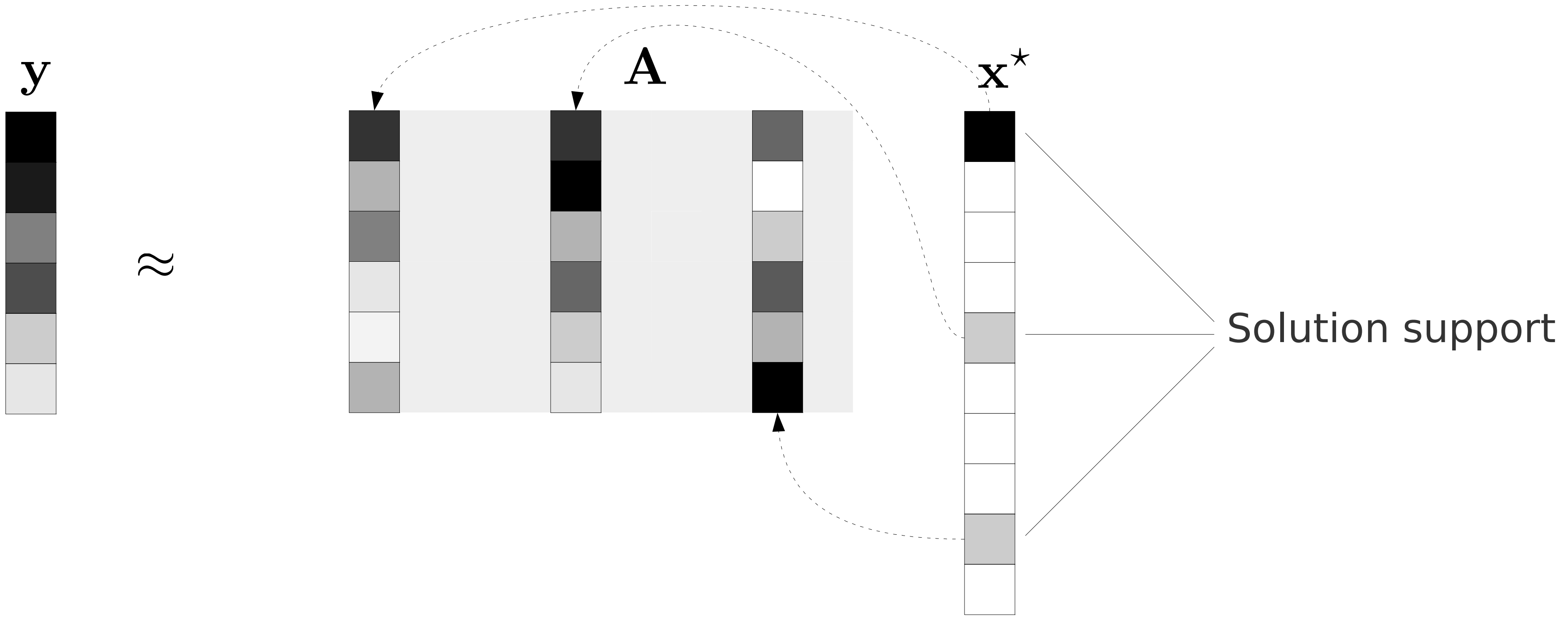


Safe Screening

- Accelerate the solution of **sparse regression problems**.

$$\mathbf{y} \approx \mathbf{A}\mathbf{x}\,, \quad \text{with } \mathbf{x} \text{ sparse}$$

- Core idea: identify and eliminate coordinates not belonging to the **solution support**.

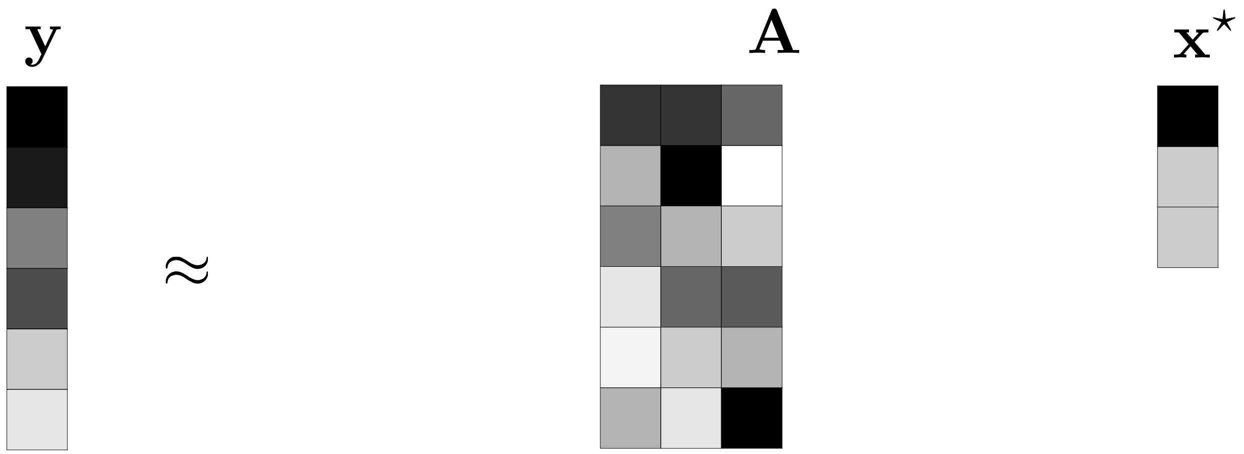


Safe Screening

- Accelerate the solution of **sparse regression problems**.

$$\mathbf{y} \approx \mathbf{A}\mathbf{x}\,, \quad \text{with } \mathbf{x} \text{ sparse}$$

- Core idea: identify and eliminate coordinates not belonging to the **solution support**.



Safe Screening: state-of-the-art

- Initially proposed for the Lasso problem [El Ghaoui et al. 2012]

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{C}} \mathcal{D}(\mathbf{y}, \mathbf{A}\mathbf{x}) + \lambda\Omega(\mathbf{x})$$

$$\downarrow \qquad\qquad \downarrow$$

$$\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

- Extensions:
  - Different regularizations
    - Group-lasso [El Ghaoui et al. 2012], Fused Lasso [Wang et al. 2015], Elastic Net [Fercoq et al. 2015], Sparse-Group Lasso [Wang et al. 2019] ...
  - Different data-fidelity terms
    - Sparse Logistic regression [Wang et al. 2014]
  - Different constraint sets
    - Non-negative Lasso [Wang et al. 2019]
- **KL** divergence case **not previously addressed**!

Outline

Problem definition

- The L1-regularized Kullback-Leibler regression problem:

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n_+} P_\lambda(\mathbf{x}) := \mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{A}\mathbf{x}) + \lambda\|\mathbf{x}\|_1 \qquad \text{(KL-L1)}$$

- Safe Screening for the KL-L1 problem. Technical ingredients:
  - Dual problem
  - First-order optimality conditions
  - Screening rule
  - Safe region

Dual problem

- Primal problem: $$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n_+} P_\lambda(\mathbf{x}) := \mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{Ax}) + \lambda\|\mathbf{x}\|_1 \qquad \text{(KL-L1)}$$

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

  – Dual cost function D_λ and dual feasible set $\mathcal{F}_{\mathbf{A}}$ are obtained by taking the *Fenchel conjugate* of $\mathcal{D}_{\mathrm{KL}}$ and $\|\cdot\|_1 + \mathbb{1}_{\mathbb{R}^n_+}(\cdot)$ respectively.

- First-order optimality conditions:

  1) $\lambda\boldsymbol{\theta}^\star = -\nabla\mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{Ax}^\star)$ (primal-dual link)

  2) $\mathbf{A}^\mathsf{T}\boldsymbol{\theta}^\star \in \partial\|\mathbf{x}^\star\|_1 + \partial\mathbb{1}_{\mathbb{R}^n_+}(\mathbf{x}^\star)$ (subdifferential inclusion)

Optimality conditions

- First-order optimality conditions:

1) $\lambda \boldsymbol{\theta}^\star = -\nabla \mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{A}\mathbf{x}^\star) \quad \Longrightarrow \quad \lambda \boldsymbol{\theta}^\star = \dfrac{\mathbf{y}}{\mathbf{A}\mathbf{x}^\star + \epsilon} - \mathbf{1}$

2) $\mathbf{A}^\mathsf{T} \boldsymbol{\theta}^\star \in \partial \|\mathbf{x}^\star\|_1 + \partial \mathbb{1}_{\mathbb{R}^n_+}(\mathbf{x}^\star) \quad \Longrightarrow \quad \forall j, \begin{cases} \mathbf{a}_j^\mathsf{T} \boldsymbol{\theta}^\star \leq 1, & \text{if } x_j^\star = 0 \\ \mathbf{a}_j^\mathsf{T} \boldsymbol{\theta}^\star = 1 & \text{if } x_j^\star \neq 0 \end{cases}$

Optimality conditions

- First-order optimality conditions:

  1) $\lambda\boldsymbol{\theta}^\star = -\nabla\mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{A}\mathbf{x}^\star) \quad \Longrightarrow \quad \lambda\boldsymbol{\theta}^\star = \dfrac{\mathbf{y}}{\mathbf{A}\mathbf{x}^\star + \epsilon} - \mathbf{1}$

  2) $\mathbf{A}^\mathsf{T}\boldsymbol{\theta}^\star \in \partial\|\mathbf{x}^\star\|_1 + \partial\mathbb{1}_{\mathbb{R}^n_+}(\mathbf{x}^\star) \quad \Longrightarrow \quad \forall j, \begin{cases} \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star \leq 1, & \text{if} \quad x_j^\star = 0 \\ \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star = 1 & \text{if} \quad x_j^\star \neq 0 \end{cases}$

- **Consequence 1)** $y_i = 0 \Longrightarrow \theta_i^\star = -\dfrac{1}{\lambda}$

Optimality conditions

- First-order optimality conditions:

  1) $\lambda\boldsymbol{\theta}^\star = -\nabla\mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{A}\mathbf{x}^\star) \quad \Longrightarrow \quad \lambda\boldsymbol{\theta}^\star = \dfrac{\mathbf{y}}{\mathbf{A}\mathbf{x}^\star + \epsilon} - \mathbf{1}$

  2) $\mathbf{A}^\mathsf{T}\boldsymbol{\theta}^\star \in \partial\|\mathbf{x}^\star\|_1 + \partial\mathbb{1}_{\mathbb{R}^n_+}(\mathbf{x}^\star) \quad \Longrightarrow \quad \forall j, \begin{cases} \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star \leq 1, & \text{if} \quad x_j^\star = 0 \\ \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta}^\star = 1 & \text{if} \quad x_j^\star \neq 0 \end{cases}$

- **Consequence 1)** $\quad y_i = 0 \implies \theta_i^\star = -\dfrac{1}{\lambda}$

E.g.: $\qquad \mathbf{y} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \implies \boldsymbol{\theta}^\star = \begin{bmatrix} ? \\ ? \\ -1/\lambda \end{bmatrix}$

Optimality conditions

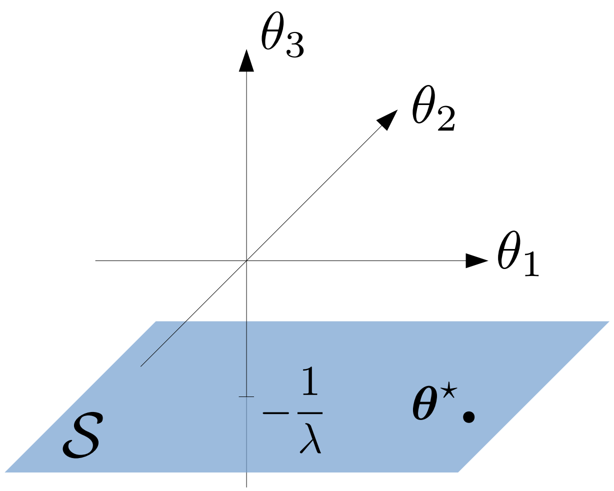
- First-order optimality conditions:

1) $\lambda\boldsymbol{\theta}^\star = -\nabla\mathcal{D}_{\mathrm{KL}}(\mathbf{y} \mid \mathbf{A}\mathbf{x}^\star) \quad \Longrightarrow \quad \lambda\boldsymbol{\theta}^\star = \dfrac{\mathbf{y}}{\mathbf{A}\mathbf{x}^\star + \epsilon} - \mathbf{1}$

2) $\mathbf{A}^\top\boldsymbol{\theta}^\star \in \partial\|\mathbf{x}^\star\|_1 + \partial\mathbb{1}_{\mathbb{R}^n_+}(\mathbf{x}^\star) \quad \Longrightarrow \quad \forall j, \begin{cases} \mathbf{a}_j^\top\boldsymbol{\theta}^\star \leq 1, & \text{if } x_j^\star = 0 \\ \mathbf{a}_j^\top\boldsymbol{\theta}^\star = 1 & \text{if } x_j^\star \neq 0 \end{cases}$

- **Consequence 1)** $y_i = 0 \Longrightarrow \theta_i^\star = -\dfrac{1}{\lambda}$

E.g.: $\mathbf{y} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \Longrightarrow \boldsymbol{\theta}^\star = \begin{bmatrix} ? \\ ? \\ -1/\lambda \end{bmatrix}$



$\boldsymbol{\theta}^\star$ belongs to the hyperplane $\mathcal{S} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \theta_i = -1/\lambda, \ \forall i \text{ s.t. } y_i = 0\}$

Safe Screening rule

2) $$\forall j, \; \begin{cases} \mathbf{a}_j^\top \boldsymbol{\theta}^\star \leq 1, & \text{if } \; x_j^\star = 0 \\ \mathbf{a}_j^\top \boldsymbol{\theta}^\star = 1 & \text{if } \; x_j^\star \neq 0 \end{cases}$$

- **Consequence 2)** $\mathbf{a}_j^\top \boldsymbol{\theta}^\star < 1 \implies x_j^\star = 0$

L. El Ghaoui, V. Viallon, T. Rabbani. *Safe Feature Elimination for the Lasso and Sparse Supervised Learning Problems*. Pacific Journal of Optimization, 2012

Safe Screening rule

2) $$\forall j, \begin{cases} \mathbf{a}_j^\top \boldsymbol{\theta}^\star \leq 1, & \text{if} \;\; x_j^\star = 0 \\ \mathbf{a}_j^\top \boldsymbol{\theta}^\star = 1 & \text{if} \;\; x_j^\star \neq 0 \end{cases}$$

- **Consequence 2)** $\mathbf{a}_j^\top \boldsymbol{\theta}^\star < 1 \implies x_j^\star = 0$

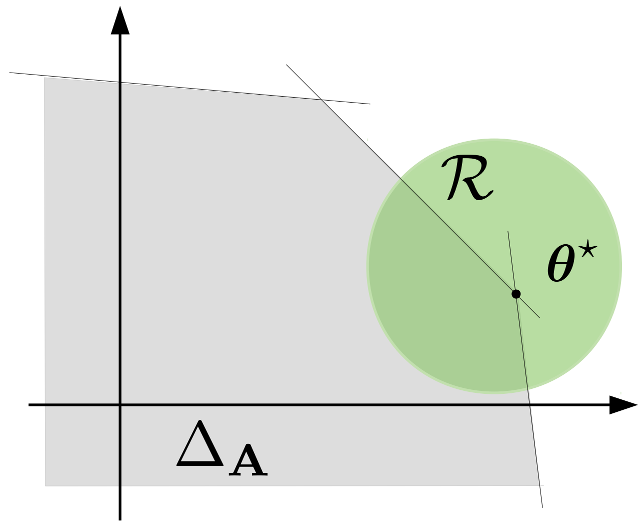
⚠ In practice, the dual solution $\boldsymbol{\theta}^\star$ is not known.

L. El Ghaoui, V. Viallon, T. Rabbani. *Safe Feature Elimination for the Lasso and Sparse Supervised Learning Problems*. Pacific Journal of Optimization, 2012

Safe Screening rule

2) $\forall j, \begin{cases} \mathbf{a}_j^\top \boldsymbol{\theta}^\star \leq 1, & \text{if } x_j^\star = 0 \\ \mathbf{a}_j^\top \boldsymbol{\theta}^\star = 1 & \text{if } x_j^\star \neq 0 \end{cases}$

Example: $\boldsymbol{\theta} \in \mathbb{R}^2$



- **Consequence 2)** $\mathbf{a}_j^\top \boldsymbol{\theta}^\star < 1 \implies x_j^\star = 0$

In practice, the dual solution $\boldsymbol{\theta}^\star$ is not known.

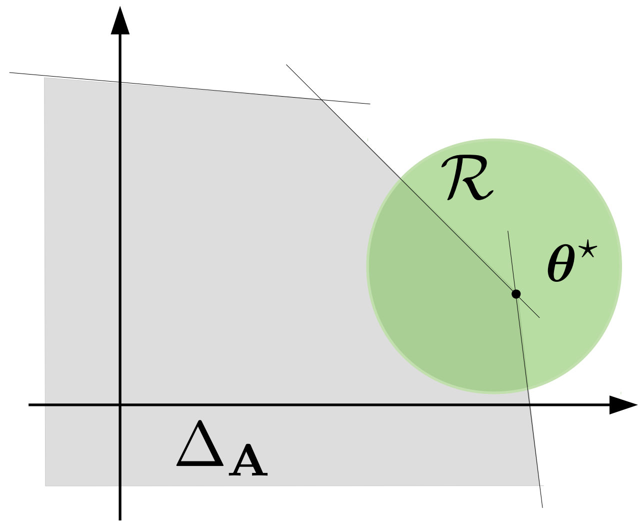
Define a safe region $\mathcal{R}$ which contains $\boldsymbol{\theta}^\star$.

L. El Ghaoui, V. Viallon, T. Rabbani. *Safe Feature Elimination for the Lasso and Sparse Supervised Learning Problems*. Pacific Journal of Optimization, 2012

Safe Screening rule

2) $\forall j, \begin{cases} \mathbf{a}_j^\top \boldsymbol{\theta}^\star \leq 1, & \text{if } x_j^\star = 0 \\ \mathbf{a}_j^\top \boldsymbol{\theta}^\star = 1 & \text{if } x_j^\star \neq 0 \end{cases}$

Example: $\boldsymbol{\theta} \in \mathbb{R}^2$



- **Consequence 2)** $\mathbf{a}_j^\top \boldsymbol{\theta}^\star < 1 \implies x_j^\star = 0$

In practice, the dual solution $\boldsymbol{\theta}^\star$ is not known.

Define a safe region $\mathcal{R}$ which contains $\boldsymbol{\theta}^\star$.

Safe screening rule [El Ghaoui et al. 2012]

Let $\mathcal{R}$ be a safe region, then:

$$\max_{\boldsymbol{\xi} \in \mathcal{R}} \mathbf{a}_j^\top \boldsymbol{\xi} < 1 \implies \mathbf{a}_j^\top \boldsymbol{\theta}^\star < 1 \implies x_j^\star = 0$$

L. El Ghaoui, V. Viallon, T. Rabbani. *Safe Feature Elimination for the Lasso and Sparse Supervised Learning Problems*. Pacific Journal of Optimization, 2012

Safe Screening rule

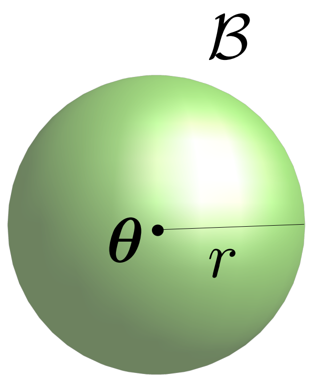
- $\mathcal{R}$ is a sphere $\mathcal{B}(\boldsymbol{\theta}, r)$ with center $\boldsymbol{\theta}$ and radius r

Safe screening rule [El Ghaoui et al. 2012]

Let $\mathcal{B}(\boldsymbol{\theta}, r)$ be a safe region, then:

$$\max_{\boldsymbol{\xi} \in \mathcal{B}(\boldsymbol{\theta}, r)} \mathbf{a}_j^\top \boldsymbol{\xi} = \mathbf{a}_j^\top \boldsymbol{\theta} + r\|\mathbf{a}_j\|_2 < 1 \quad \Longrightarrow \quad x_j^\star = 0$$

E.g.: $\boldsymbol{\theta} \in \mathbb{R}^3$



Safe Screening rule for KL

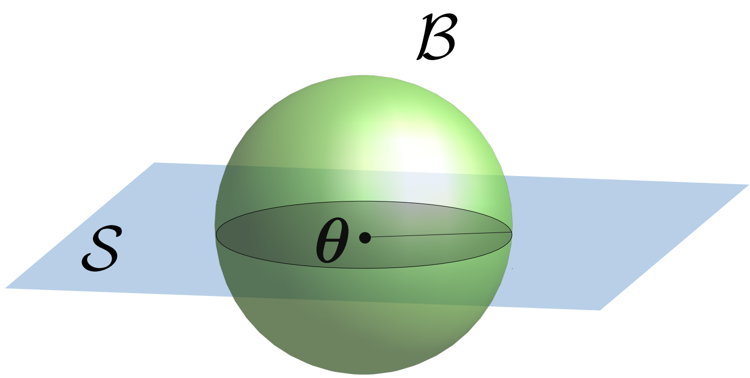
- Improved screening rule for the KL divergence

KL-L1 Safe screening rule [D.S.F. 2021]

Let $\mathcal{B}(\boldsymbol{\theta}, r)$ be a safe region and $\mathcal{S} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \boldsymbol{\theta}_{\mathcal{I}} = -\mathbf{1}/\lambda\}$ with $\mathcal{I} = \{i \in [m] : y_i = 0\}$ and $\boldsymbol{\theta} \in \mathcal{S}$, then:

$$\max_{\boldsymbol{\xi} \in \mathcal{B}(\boldsymbol{\theta}, r) \cap \mathcal{S}} \mathbf{a}_j^\mathsf{T} \boldsymbol{\xi} = \mathbf{a}_j^\mathsf{T} \boldsymbol{\theta} + r \| [\mathbf{a}_j]_{\mathcal{I}^c} \|_2 < 1 \quad \Longrightarrow \quad x_j^\star = 0$$

E.g.: $\boldsymbol{\theta} \in \mathbb{R}^3$



Safe Region

GAP Safe sphere [Ndiaye et al. 2017]

For any feasible primal-dual pair $(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}_+^n \times \mathcal{F}_{\mathbf{A}}$

$$\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r), \text{ with } r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha}}$$

where α is the strong concavity constant of D_λ.

where $\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta}) := P_\lambda(\mathbf{x}) - D_\lambda(\boldsymbol{\theta})$ denotes the duality gap.

E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon. *Gap Safe Screening Rules for Sparsity Enforcing Penalties*. JMLR, 2017.

Safe Region

GAP Safe sphere [Ndiaye et al. 2017]

For any feasible primal-dual pair $(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^n_+ \times \mathcal{F}_{\mathbf{A}}$

$$\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r), \text{ with } r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha}}$$

where α is the strong concavity constant of D_λ.

where $\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta}) := P_\lambda(\mathbf{x}) - D_\lambda(\boldsymbol{\theta})$ denotes the duality gap.

- Requires **global** strong concavity ofD_λ

E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon. *Gap Safe Screening Rules for Sparsity Enforcing Penalties*. JMLR, 2017.

Safe Region

GAP Safe sphere [Ndiaye et al. 2017]

For any feasible primal-dual pair $(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^n_+ \times \mathcal{F}_{\mathbf{A}}$

$$\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r), \text{ with } r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha}}$$

where α is the strong concavity constant of D_λ.

where $\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta}) := P_\lambda(\mathbf{x}) - D_\lambda(\boldsymbol{\theta})$ denotes the duality gap.

- Requires **global** strong concavity ofD_λ
- ISSUE: Not the case for the KL-L1 problem!

E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon. *Gap Safe Screening Rules for Sparsity Enforcing Penalties*. JMLR, 2017.

Safe Region

GAP Safe sphere [Ndiaye et al. 2017]

For any feasible primal-dual pair $(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^n_+ \times \mathcal{F}_{\mathbf{A}}$

$$\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r), \text{ with } r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha}}$$

where α is the strong concavity constant of D_λ.

where $\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta}) := P_\lambda(\mathbf{x}) - D_\lambda(\boldsymbol{\theta})$ denotes the duality gap.

- Requires **global** strong concavity ofD_λ
- ISSUE: Not the case for the KL-L1 problem!
- IDEA: Use **local** strong concavity.

E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon. *Gap Safe Screening Rules for Sparsity Enforcing Penalties*. JMLR, 2017.

Safe Region for KL

KL-L1 GAP Safe sphere [D.S.F. 2021]

For any feasible primal-dual pair $(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^n_+ \times (\mathcal{F}_{\mathbf{A}} \cap \mathcal{S})$

$$\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r), \text{ with } r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\bar{\alpha}}}$$

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

- D_λ is $\bar{\alpha}$-strongly concave on $\mathcal{F}_{\mathbf{A}} \cap \mathcal{S}$
- Note that D_λ is not strongly concave on $\mathcal{F}_{\mathbf{A}}$ only.

Proposed Algorithm

Algorithm 1 : KL-L1 Dynamic GAP Safe Screening [D.S.F. 2021]

Initialize $\mathbf{x} \in \mathbb{R}^n,\ \mathcal{A} = \{1, \ldots, n\},\ \bar{\alpha}$ strong concavity bound on $\mathcal{F}_{\mathbf{A}} \cap \mathcal{S}$

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 :

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

$$\mathcal{A} \leftarrow \{j \in \mathcal{A} \mid \mathbf{a}_j^\top \boldsymbol{\theta} + r\|[\mathbf{a}_j]_{\mathcal{I}^c}\|_2 \geq 1\}$$

$$\mathbf{x}_{\mathcal{A}^c} \leftarrow \mathbf{0}$$

Proposed Algorithm

Algorithm 1 : KL-L1 Dynamic GAP Safe Screening [D.S.F. 2021]

Initialize $\mathbf{x} \in \mathbb{R}^n, \ \mathcal{A} = \{1, \ldots, n\}, \ \bar{\alpha}$ strong concavity bound on $\mathcal{F}_{\mathbf{A}} \cap \mathcal{S}$

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 : $\boldsymbol{\theta} \leftarrow \boldsymbol{\Theta}\left(\mathbf{x}\right) \in \mathcal{F}_{\mathbf{A}} \cap \mathcal{S}$

Safe screening :

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

$$\mathcal{A} \leftarrow \{j \in \mathcal{A} \mid \mathbf{a}_j^{\top}\boldsymbol{\theta} + r\|[\mathbf{a}_j]_{\mathcal{I}^c}\|_2 \geq 1\}$$

$$\mathbf{x}_{\mathcal{A}^c} \leftarrow \mathbf{0}$$

Considered solvers:

- Proximal gradient [Harmany et al. 2012]
- Coordinate descent [Hsieh, Dhillon, 2011]
- Majorize-minimize (Multiplicative Update) [Févotte, Idier 2011]

Proposed Algorithm

Algorithm 1 : KL-L1 Dynamic GAP Safe Screening [D.S.F. 2021]

Initialize $\mathbf{x} \in \mathbb{R}^n, \ \mathcal{A} = \{1, \ldots, n\}, \ \bar{\alpha}$ strong concavity bound on $\mathcal{F}_{\mathbf{A}} \cap \mathcal{S}$

Repeat until convergence

Primal update : $\mathbf{x}_{\mathcal{A}} \leftarrow \texttt{PrimalUpdate}(\mathbf{x}_{\mathcal{A}}, \mathbf{A}_{\mathcal{A}}, \mathbf{y}, \lambda)$

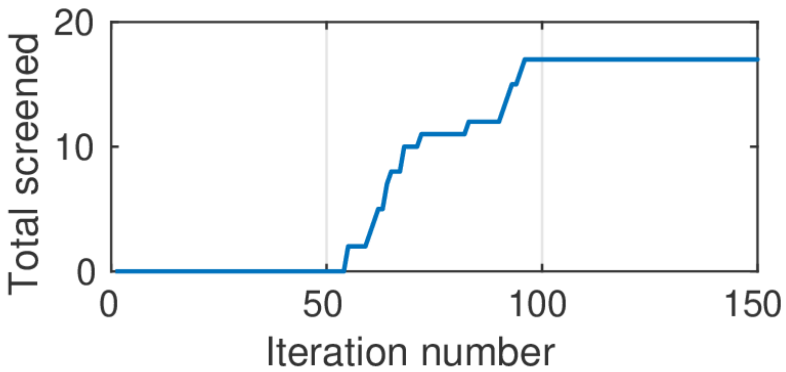
Dual update : $\boldsymbol{\theta} \leftarrow \boldsymbol{\Theta}(\mathbf{x}) \in \mathcal{F}_{\mathbf{A}} \cap \mathcal{S}$

Safe screening :

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

$$\mathcal{A} \leftarrow \{j \in \mathcal{A} \mid \mathbf{a}_j^\mathsf{T}\boldsymbol{\theta} + r\|[\mathbf{a}_j]_{\mathcal{I}^c}\|_2 \geq 1\}$$

$$\mathbf{x}_{\mathcal{A}^c} \leftarrow \mathbf{0}$$



Outline

Experiments

Convergence time

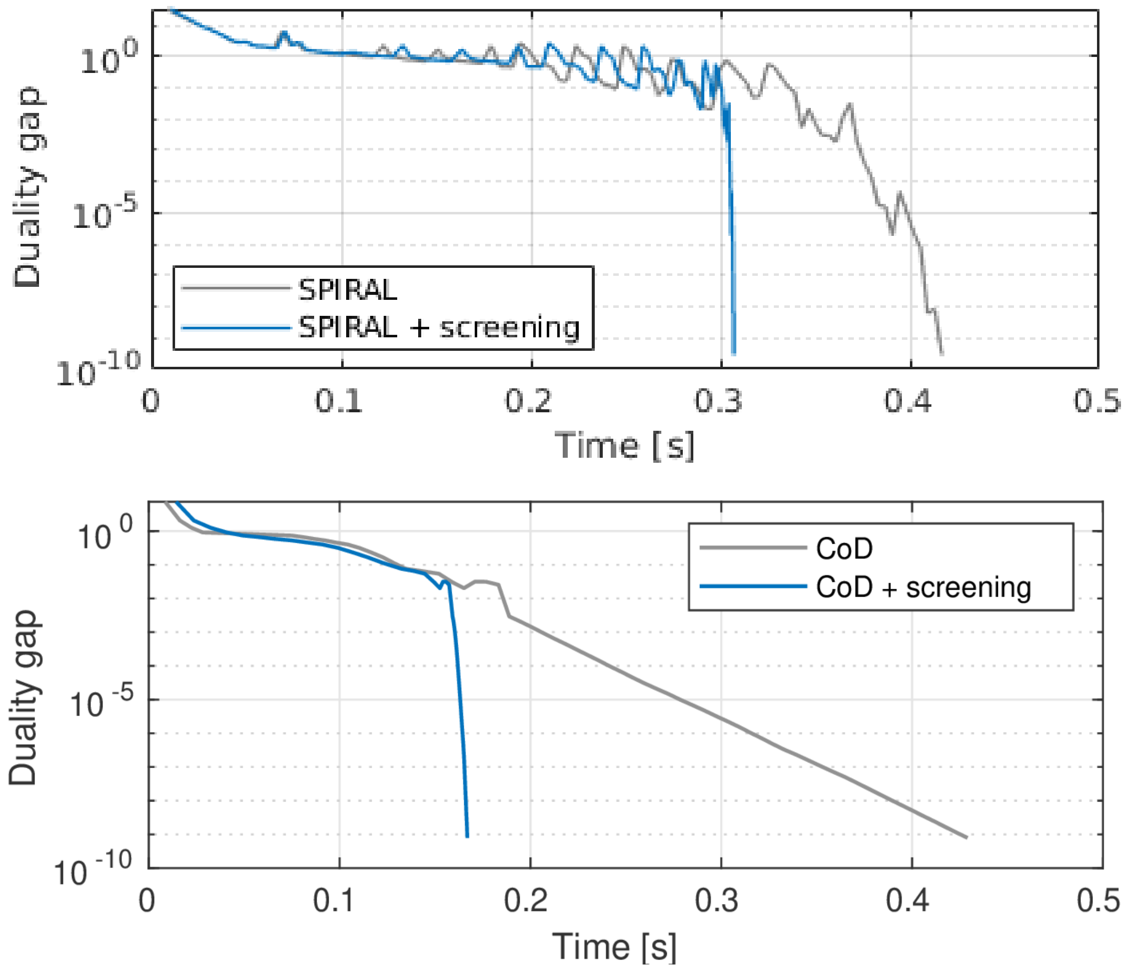


Figure: Convergence vs. Time. 20-Newsgroup data, $\lambda = 10^{-2}\lambda_{\max}$.

Experiments

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

	$\lambda/\lambda_{\max}$	10^{-1}		10^{-3}	
	ε_{gap}	10^{-5}	10^{-7}	10^{-5}	10^{-7}
20 Newsgr.	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

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

$\lambda_{\max} = \max\left(\mathbf{A}^\mathsf{T}(\mathbf{y}-\epsilon)\right)/\epsilon$ is the bound above which $\mathbf{x}^\star = \mathbf{0}$.

Concluding remarks

- Main contribution : safe screening technique for the KL-L1 problem.
  - Improved screening rule for the particular KL case.
  - Adaptation of GAP Safe sphere exploiting **local properties** of the cost function.
- Significant improvements in terms of convergence time.
- Extensions: check our **follow-up paper** (below)!
  - Other group-decomposable regularizations.
  - Other β-divergences as data fidelity.
  - Tighter local strong concavity bound.

C. F. Dantas, E. Soubies, C. Févotte. *Expanding Boundaries of GAP Safe Screening*. 2021.

Available at: hal.archives-ouvertes.fr/hal-03147502

Matlab code: github.com/cassiofragadantas

References

Contact me: cassio.fraga-dantas@irit.fr

Safe Screening

[1] L. El Ghaoui, V. Viallon, T. Rabbani. *Safe Feature Elimination for the Lasso and Sparse Supervised Learning Problems*. Pacific Journal of Optimization, Oct 2012.

[2] O. Fercoq, A. Gramfort, J. Salmon. Mind the Duality Gap : Safer Rules for the Lasso. ICML, 2015.

[3] E. Ndiaye, O. Fercoq, A. Gramfort, J. Salmon. *Gap Safe Screening Rules for Sparsity Enforcing Penalties*. Journal of Machine Learning Research, Nov 2017.

[4] J. Wang, J. Zhou, J. Liu, P. Wonka, J. Ye. *A Safe screening Rule for Sparse Logistic Regression*. NeurIPS, 2014.

[5] J. Wang, W. Fan, J. Ye. *Fused Lasso Screening Rules via the Monotonicity of Subdifferentials*. IEEE Transactions of Pattern Analysis and Machine Intellingence, 2015.

[6] J. Wang, Z. Zhang, J. Ye. *Two-layer Feature Reduction for Sparse-group Lasso via decomposition of Convex Sets*. Journal of Machine Learning Research, 2019.

KL-L1 Solvers

[7] Z. T. Harmany, R. F. Marcia, R. M. Willett. This is SPIRAL-TAP: *Sparse Poisson Intensity Reconstruction Algorithms - Theory and Practice*. IEEE Transactions on Image Processing, 2012.

[8] C. Hsieh, I. S. Dhillon. *Fast Coordinate Descent Methods with Variable Selection for Non-negative Matrix Factorization*. In Proc. ACM SIGKDD, 2011.

[9] C. Févotte, J. Idier. *Algorithms for Nonnegative Matrix Factorization with the β-divergence*. Neural Computation, Sep 2011.