Expanding Boundaries of GAP Safe Screening

Cassio F. Dantas

Joint work with: Cédric Févotte and Emmanuel Soubies

IRIT, Université de Toulouse, CNRS

Contributions: Global view

- Goal: expanding the frontiers of GAP safe screening in the context of sparse linear regression problems.
  - Regularity assumptions: global → **local**
  - Non-negativity constraints

- Allows to:
  - Extension to larger class of functions. E.g., β-divergences.
  - Improves upon the existing GAP Safe approach.

Outline

Context and Literature

1. Safe screening : a quick overview

Our contribution

2. Exploiting local properties of the dual function
    - General approach
    - Particular cases
3. Experimental results

Outline

Context and Literature

1. Safe screening : a quick overview

Our contribution

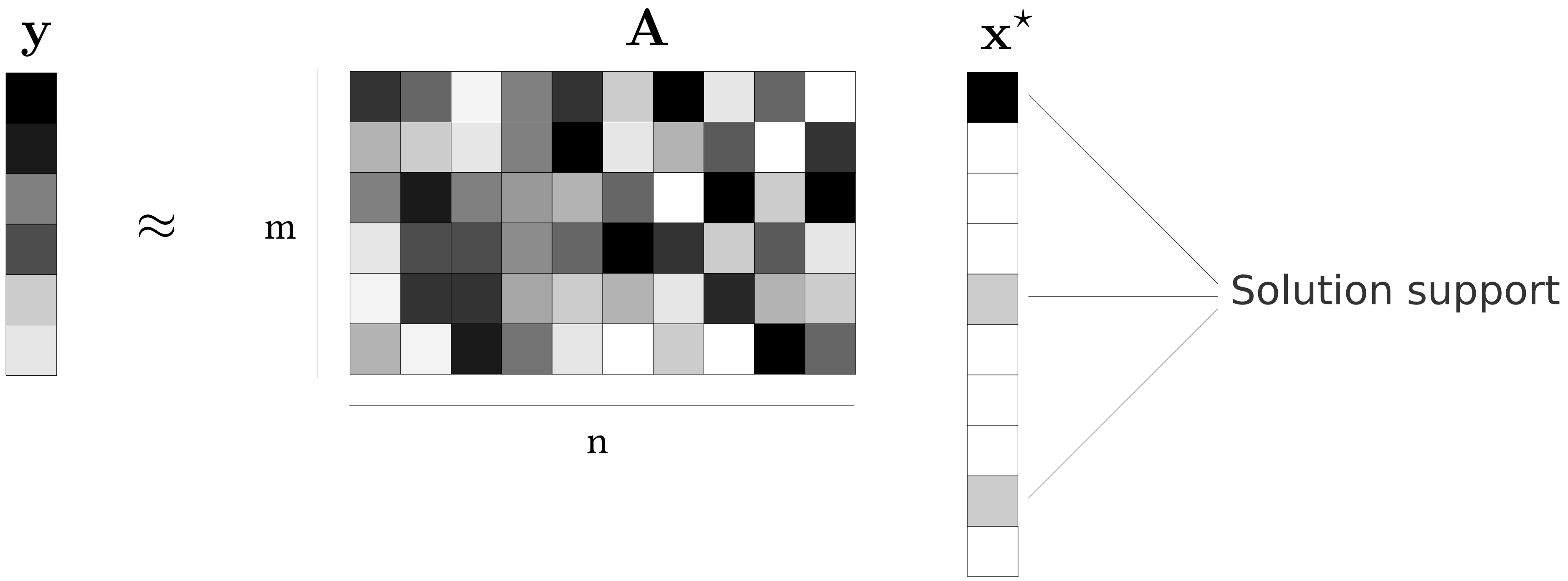
2. Exploiting local properties of the dual function
   - General approach
   - Particular cases
3. Experimental results

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**.

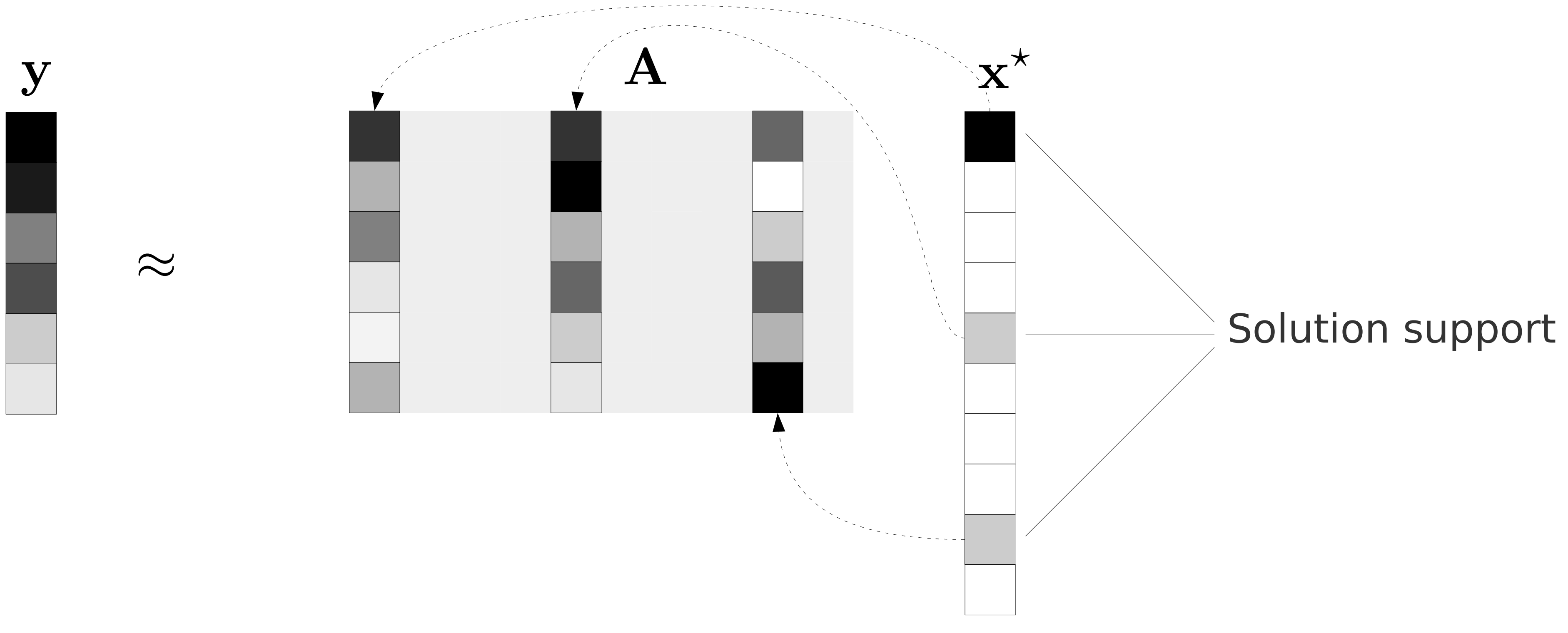


Safe Screening

- 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**.

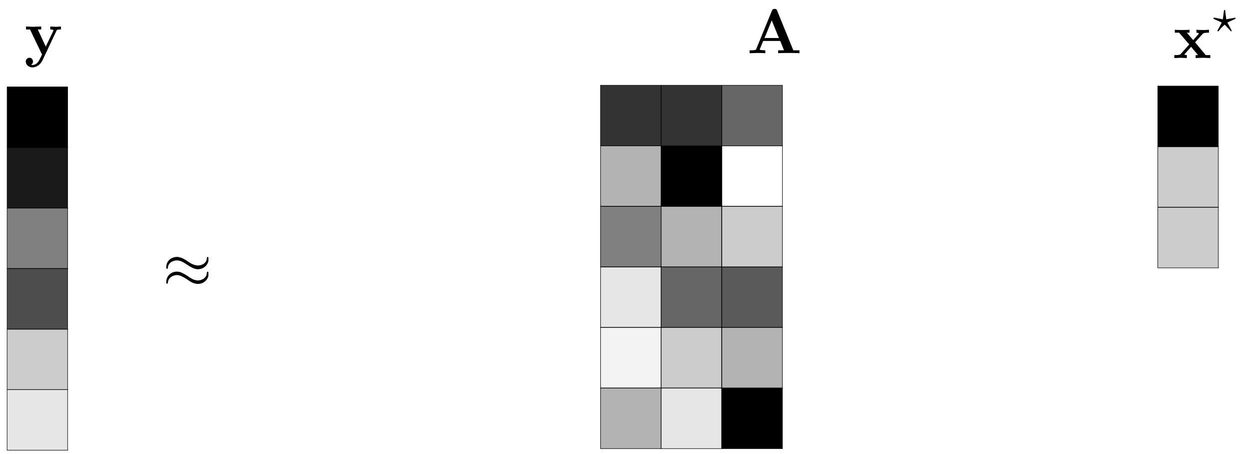


Safe Screening

- 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**.



Problem definition

- Primal problem: $$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} P_\lambda(\mathbf{x}) := F(\mathbf{A}\mathbf{x}) + \lambda\Omega(\mathbf{x})$$
  - $F : \mathbb{R}^m \to \mathbb{R}$, data fidelity function
    - Coordinate-wise separable $F(\mathbf{A}\mathbf{x}) = \sum_{i=1}^m f_i([\mathbf{A}\mathbf{x}]_i)$
    - f_i is proper, lower semi-continuous, convex, differentiable.
  - $\Omega : \mathbb{R}^n \to \mathbb{R}$, group-decomposable norm. We set: $\Omega(\mathbf{x}) = \|\mathbf{x}\|_1$
  - $\lambda > 0$ regularization parameter.

- Dual problem: $$\boldsymbol{\theta}^\star = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Delta_{\mathbf{A}}} D_\lambda(\boldsymbol{\theta}) := -F^*(-\lambda\boldsymbol{\theta})$$
  $$\text{with } \Delta_{\mathbf{A}} = \{\boldsymbol{\theta} \in \mathbb{R}^m \mid \|\mathbf{A}^\mathsf{T}\boldsymbol{\theta}\|_\infty \leq 1\}$$
  Unit ball of the dual norm $\overline{\Omega}$

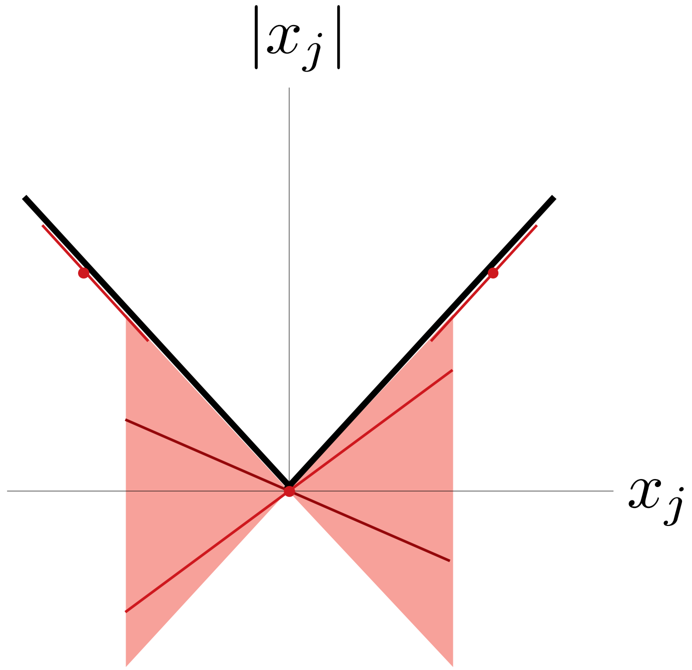
Problem definition

- First-order optimality conditions:

  1) $\lambda \boldsymbol{\theta}^\star = -\nabla F(\mathbf{A}\mathbf{x}^\star)$ (primal-dual link)

  2) $\mathbf{A}^\top \boldsymbol{\theta}^\star \in \partial\Omega(\mathbf{x}^\star)$ (subdifferential inclusion)

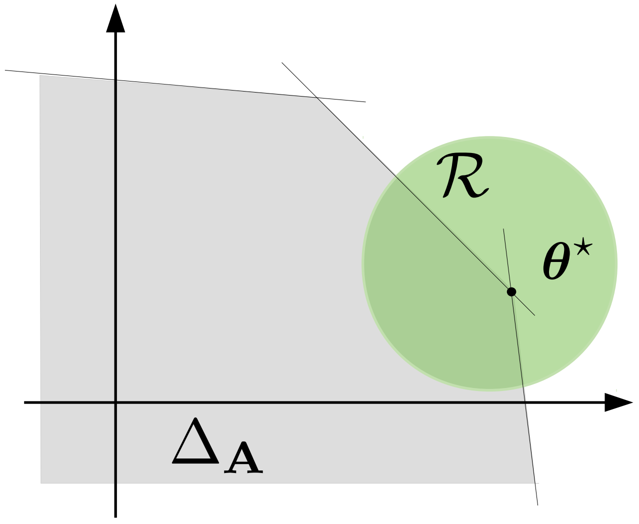
$$\forall j \in \{1, \ldots, n\}, \quad \begin{cases} |\mathbf{a}_j^\top \boldsymbol{\theta}^\star| \leq 1, & \text{if} \quad x_j^\star = 0 \\ |\mathbf{a}_j^\top \boldsymbol{\theta}^\star| = 1 & \text{if} \quad x_j^\star \neq 0 \end{cases}$$



Safe Screening

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

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



- Direct consequence: $|\mathbf{a}_j^T \boldsymbol{\theta}^\star| < 1 \implies x_j^\star = 0$
- 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:

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

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

GAP Safe Screening

GAP Safe sphere [Ndiaye et al. 2017]

Assuming that the dual function D_λ is α-strongly concave, then for any feasible primal-dual pair $(\mathbf{x}, \boldsymbol{\theta})$

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

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

- Requires global strong concavity of D_λ
- Can we use local strong concavity instead? Yes, if...

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

Outline

Context and Literature

1. Safe screening : a quick overview

Our contribution

2. Exploiting local properties of the dual function
    - General approach
    - Particular cases
3. Experimental results

GAP Safe Screening - Revisited

GAP Safe sphere

Assuming that the dual function D_λ is $\alpha_{\mathcal{S}}$-strongly concave, on a subset $\mathcal{S} \in \mathbb{R}^m$ such that $\boldsymbol{\theta}^\star \in \mathcal{S}$, then for any feasible primal-dual pair $(\mathbf{x}, \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \mathcal{S}$

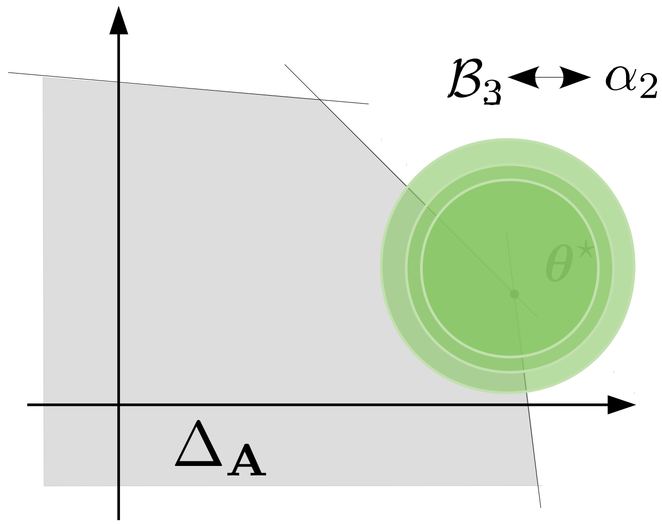
$$\boldsymbol{\theta}^\star \in \mathcal{B}(\boldsymbol{\theta}, r), \text{ with } r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha_{\mathcal{S}}}}$$

Alg. 1) $\mathcal{S} = \mathbb{R}^m$: comes down to standard GAP Safe.

Alg. 2) $\mathcal{S} = \Delta_{\mathbf{A}} \cap \mathcal{S}_0$

Alg. 3) $\mathcal{S} = \mathcal{B}(\boldsymbol{\theta}, r)$: feedback loop between r and α .

$$\begin{tikzcd} \mathcal{B}(\boldsymbol{\theta}, r) \arrow[r] & \alpha_{\mathcal{B}(\boldsymbol{\theta}, r)} \arrow[l, bend left, "r = \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha_{\mathcal{B}(\boldsymbol{\theta}, r)}}}"] \end{tikzcd}$$



Algorithm 0

Algorithm 0 : Iterative solver without screening

Initialize $\mathbf{x} \in \mathbb{R}^n$

Repeat until convergence

  Solver update : $\mathbf{x} \leftarrow \texttt{PrimalUpdate}(\mathbf{x}, \mathbf{A}, \lambda)$

Examples :

- Proximal gradient [Beck, Teboulle, 2009; Harmany et al. 2012]
- Coordinate Descent [Friedman et al. 2010; Hsieh, Dhillon, 2011]
- Majoration-Minimization (Multiplicative Update) [Févotte, Idier 2011]
- (...)

Algorithm 1

Algorithm 1 : Dynamic GAP Safe Screening (DGS) [Ndiaye et al. 2017]

Initialize $\mathbf{x} \in \mathbb{R}^n,\ \mathcal{A} = \{1, \dots, n\},\ \alpha$ global strong concavity bound

Repeat until convergence

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

Dual update : $\boldsymbol{\theta} \leftarrow \boldsymbol{\Theta}\left(\mathbf{x}\right) \in \Delta_{\mathbf{A}}$

Safe screening :

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

$$\mathcal{A} \leftarrow \{j \in \mathcal{A} \mid \max_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}, r)} |\mathbf{a}_j^\top \boldsymbol{\theta}| \geq 1\}$$

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

Algorithm 2

Algorithm 2 : Generalized Dynamic GAP Safe Screening (G-DGS) [D.S.F. 2021]

Initialize $\mathbf{x} \in \mathbb{R}^n,\ \mathcal{A} = \{1, \dots, n\},\ \alpha_{\Delta_\mathbf{A}}$ strong concavity bound on $\mathcal{S} = \Delta_\mathbf{A}$

Repeat until convergence

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

Dual update : $\boldsymbol{\theta} \leftarrow \boldsymbol{\Theta}\left(\mathbf{x}\right) \in \Delta_\mathbf{A}$

Safe screening :

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

$$\mathcal{A} \leftarrow \{j \in \mathcal{A} \mid \max_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}, r)} |\mathbf{a}_j^\mathsf{T} \boldsymbol{\theta}| \geq 1\}$$

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

C. F. Dantas, E. Soubies, C. Févotte. *Expanding Boundaries of Gap Safe Screening*. Submitted to JMLR, 2021.

Algorithm 3

Algorithm 3 : Refined Dynamic GAP Safe Screening (R-DGS) [D.S.F. 2021]

Initialize $\mathbf{x} \in \mathbb{R}^n,\ \mathcal{A} = \{1, \ldots, n\},\ \alpha_{\mathcal{S}}$ strong concavity bound on any valid $\mathcal{S}$

Repeat until convergence

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

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

Safe screening : **Repeat** until $\Delta r < \epsilon_r$

$$r \leftarrow \min\left(r, \sqrt{\frac{2\,\mathrm{Gap}_\lambda(\mathbf{x}, \boldsymbol{\theta})}{\alpha_{\mathcal{S}}}}\right)$$

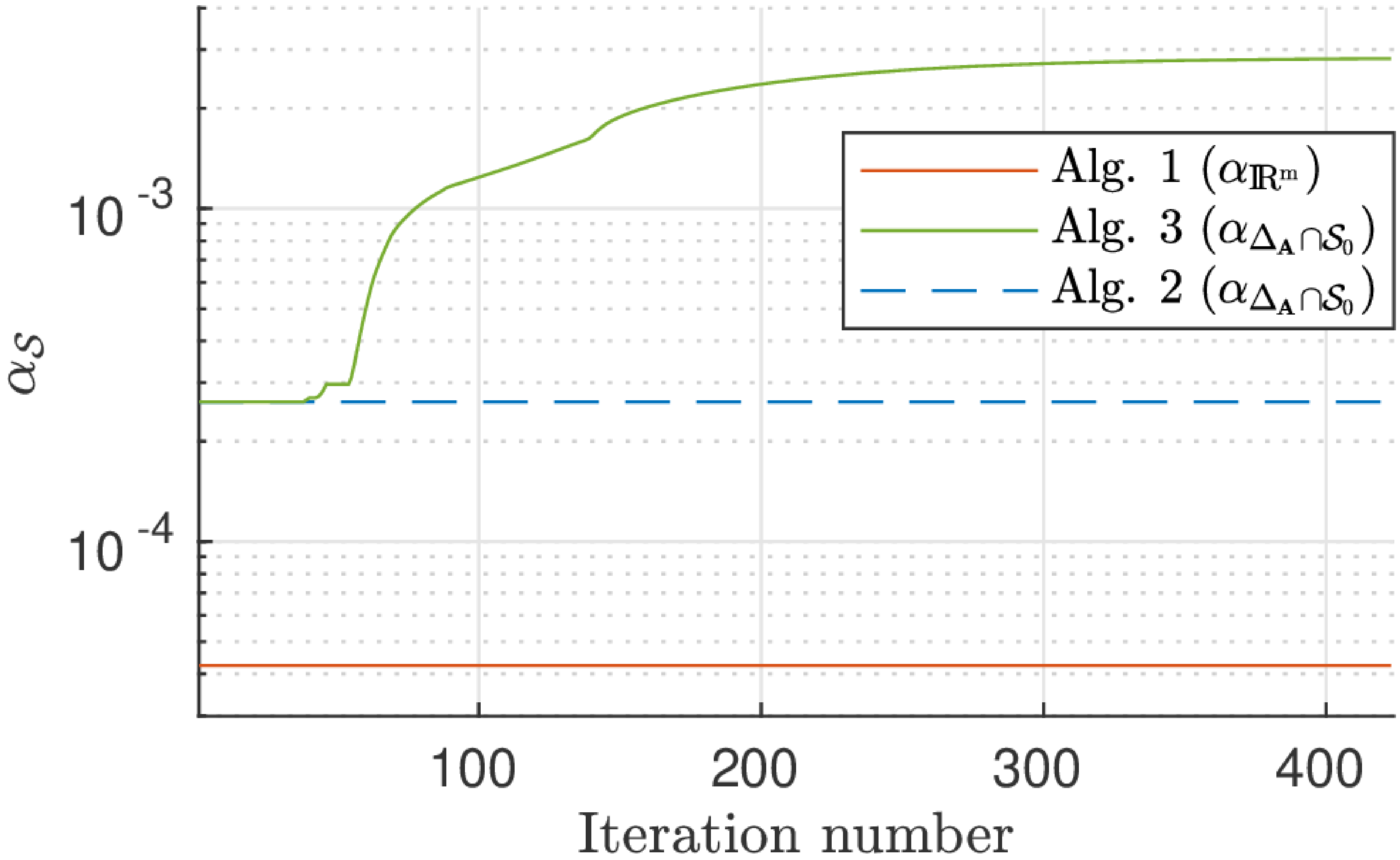
$$\mathcal{S} \leftarrow \mathcal{B}(\boldsymbol{\theta}, r)$$

$$\mathcal{A} \leftarrow \{j \in \mathcal{A} \mid \max_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}, r)} |\mathbf{a}_j^\top \boldsymbol{\theta}| \geq 1\}$$

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

C. F. Dantas, E. Soubies, C. Févotte. *Expanding Boundaries of Gap Safe Screening*. Submitted to JMLR, 2021.

Strong-concavity bound: evolution



Outline

Context and Literature

1. Safe screening : a quick overview

Our contribution

2. Exploiting local properties of the dual function
   - General approach
   - Particular cases
3. Experimental results

Particular cases - Recipe

Applying GAP Safe screening to a convex data-fidelity function $F(\mathbf{Ax})$:

1) Compute the corresponding dual function $D_\lambda = F^*$ (Fenchel conjugate)

2) Compute a valid strong concavity bound $\alpha_{\mathcal{S}}$.

Example : $\alpha_{\Delta_{\mathbf{A}}}$ or $\alpha_{\mathbb{R}^m}$ (if globally strongly concave)

Extra step for Algorithm 3 (refinement approach) :

3) Compute strong concavity bound on a given ball $\alpha_{\mathcal{B}(\boldsymbol{\theta},r)}$

Computing a strong concavity bound $\alpha_{\mathcal{S}}$

Supposing D_λ to be twice-differentiable :

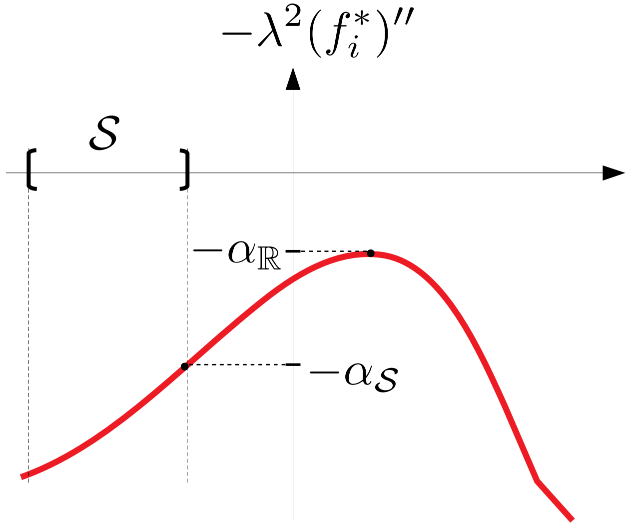
1) Compute the eigenvalues of the Hessian $\nabla^2 D_\lambda$.

2) Upper-bound the largest eigenvalue over the set $\mathcal{S}$.

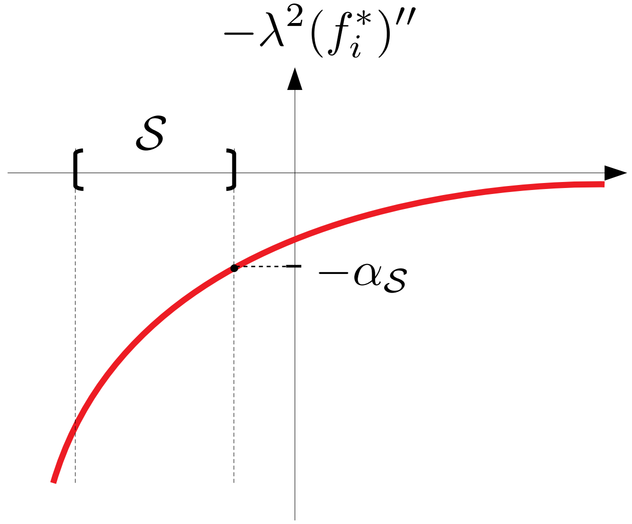
$F = \sum_i f_i \implies D_\lambda = -\sum_i f_i^*$ coordinate-wise separable.

Hessian $\nabla^2 D_\lambda$ is diagonal, with eigenvalues given by $-(f_i^*)''$.

$\alpha > 0$



$-f_i^*$ is globally strongly concave



$-f_i^*$ is *not* globally strongly concave

Computing a strong concavity bound

Supposing D_λ to be twice-differentiable :

$$0 < \alpha_{\mathcal{S}} \leq -\max_{i \in [n]} \sup_{\boldsymbol{\theta} \in \mathcal{S}} \sigma_i(\theta_i),$$

where $\sigma_i\left(\theta_i\right) = -\lambda^2 (f_i^*)''(\lambda\theta_i)$ is i-th eigenvalue of the Hessian $\nabla^2 D_\lambda$.

Data-fidelity function $F(\mathbf{Ax}) = \sum_i f_i([\mathbf{Ax}]_i)$ is coordinate-separable,

then, $D_\lambda(\boldsymbol{\theta}) = F^*(-\lambda\boldsymbol{\theta}) = \sum_i f_i^*(-\lambda\theta_i)$ is also coordinate-separable.

The Hessian is diagonal with eigenvalues (diagonal entries) given by :

$$\sigma_i\left(\theta_i\right) = -\lambda^2 (f_i^*)''(\lambda\theta_i)$$

Notable particular cases

We distinguished three scenarios regarding the choice of $F(\mathbf{Ax})$:

1) Dual function is globally strongly concave + cannot be improved locally.
   E.g.: quadratic distance (Lasso).

2) Dual function is globally strongly concave + can be improved locally.
   E.g.: logistic function.

3) Dual function is only locally (not globally) strongly concave.
   E.g.: β-divergence (with $\beta \in [1, 2)$), Kullback-Leibler divergence ($\beta = 1$).

Outline

Context and Literature

1. Safe screening : a quick overview

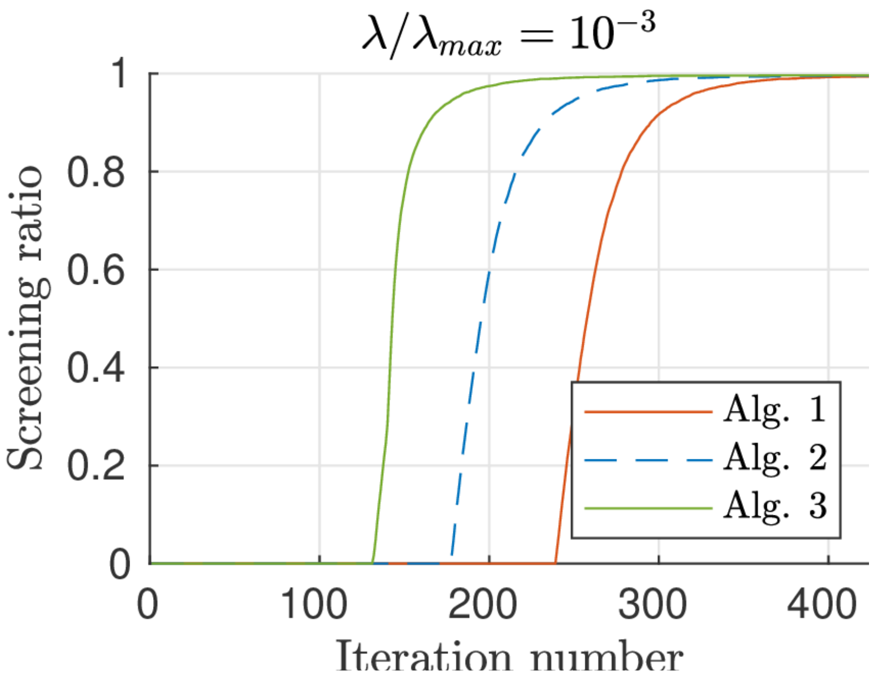
Our contribution

2. Exploiting local properties of the dual function
   - General approach
   - Particular cases
3. Experimental results

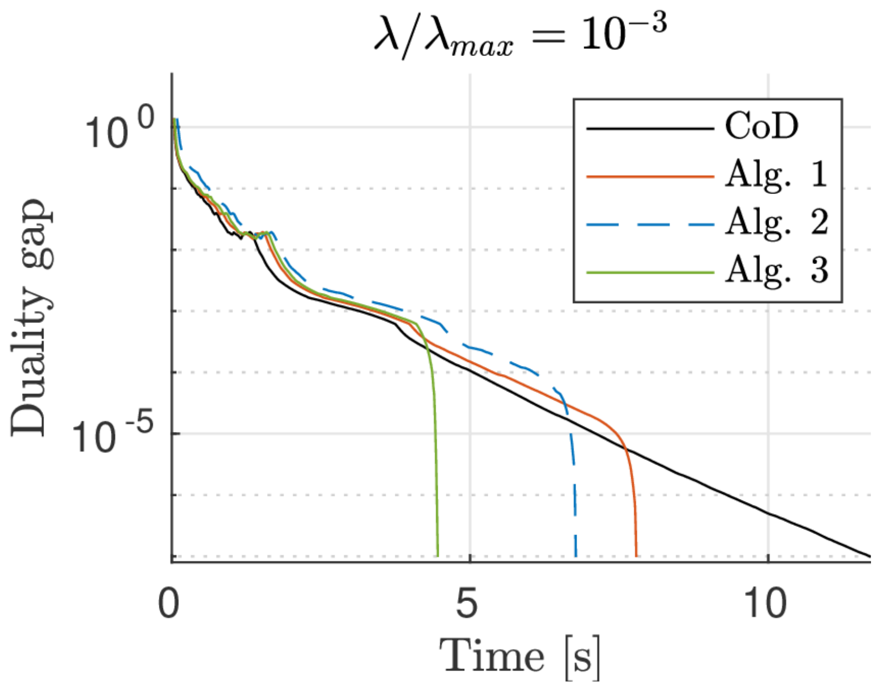
Experiments

- Scenario 2 : Logistic regression
  - Solver : Coordinate descent
  - Dataset : Leukemia bynary classification.

Screening performance



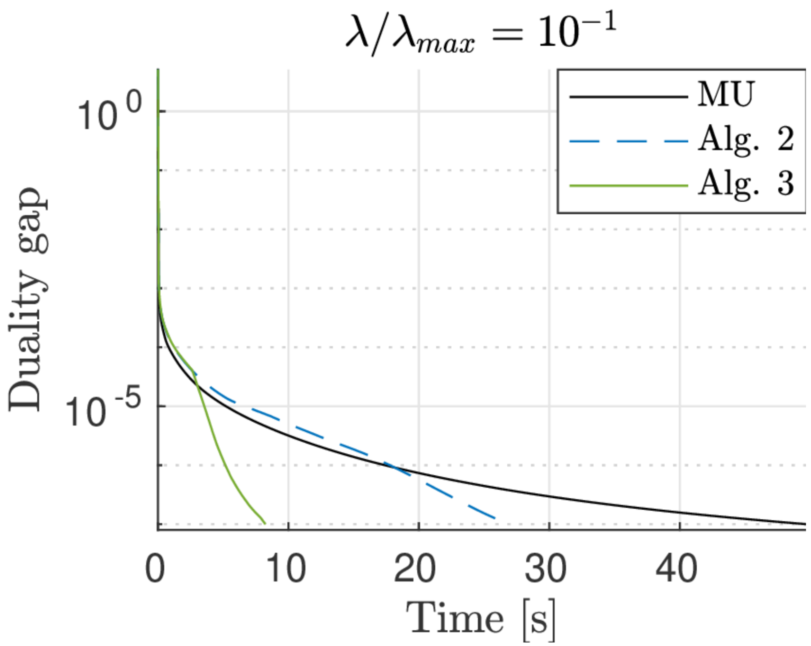
Convergence time



Experiments

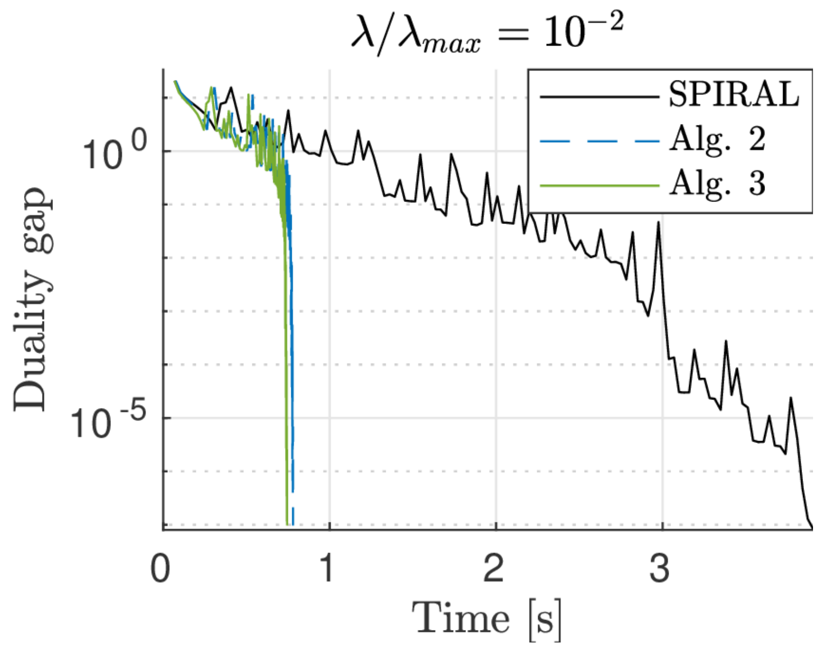
- Scenario 3 (no global strong-concavity)

$(\beta = 1.5)$-divergence



- Solver: multiplicative update
- Dataset: Urban hyperspectral image

Kullback-Leibler divergence



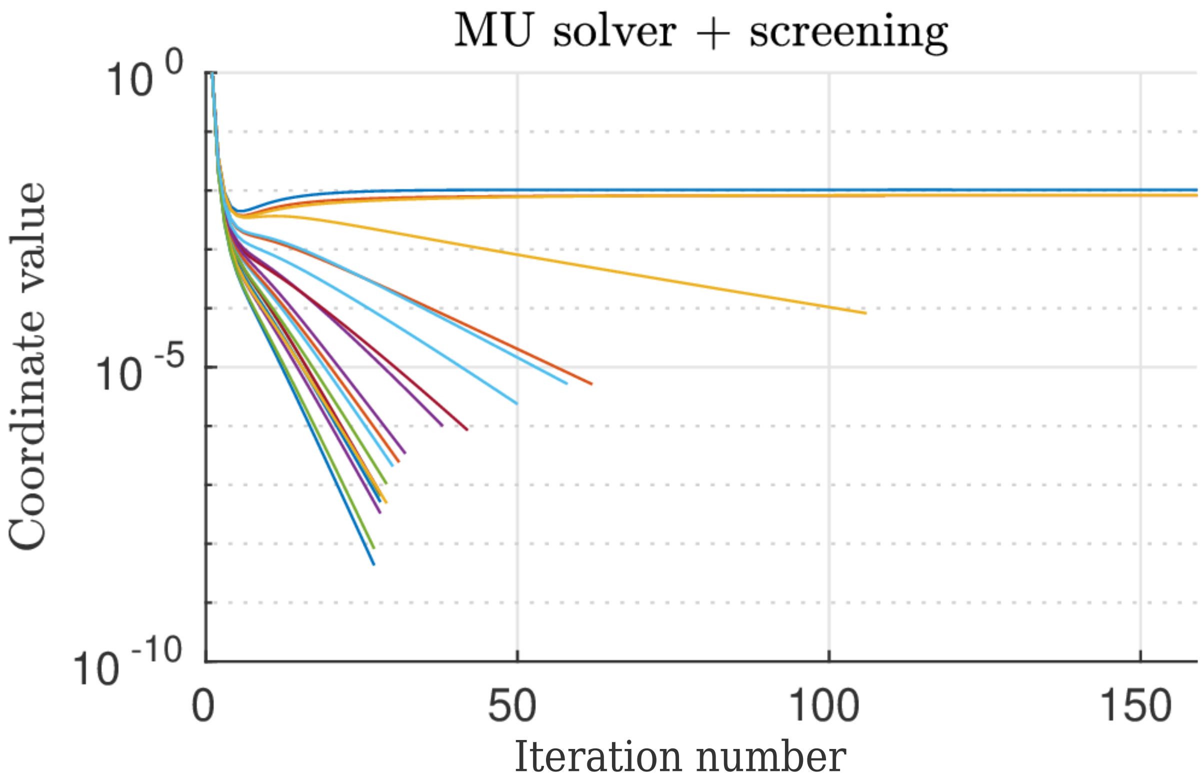
- Solver: proximal gradient
- Dataset: NIPS papers (word count)

Support identification for MU solver

- Popular approach for β-divergence minimization.
  - Update step : each coordinate is multiplied by a positive factor.

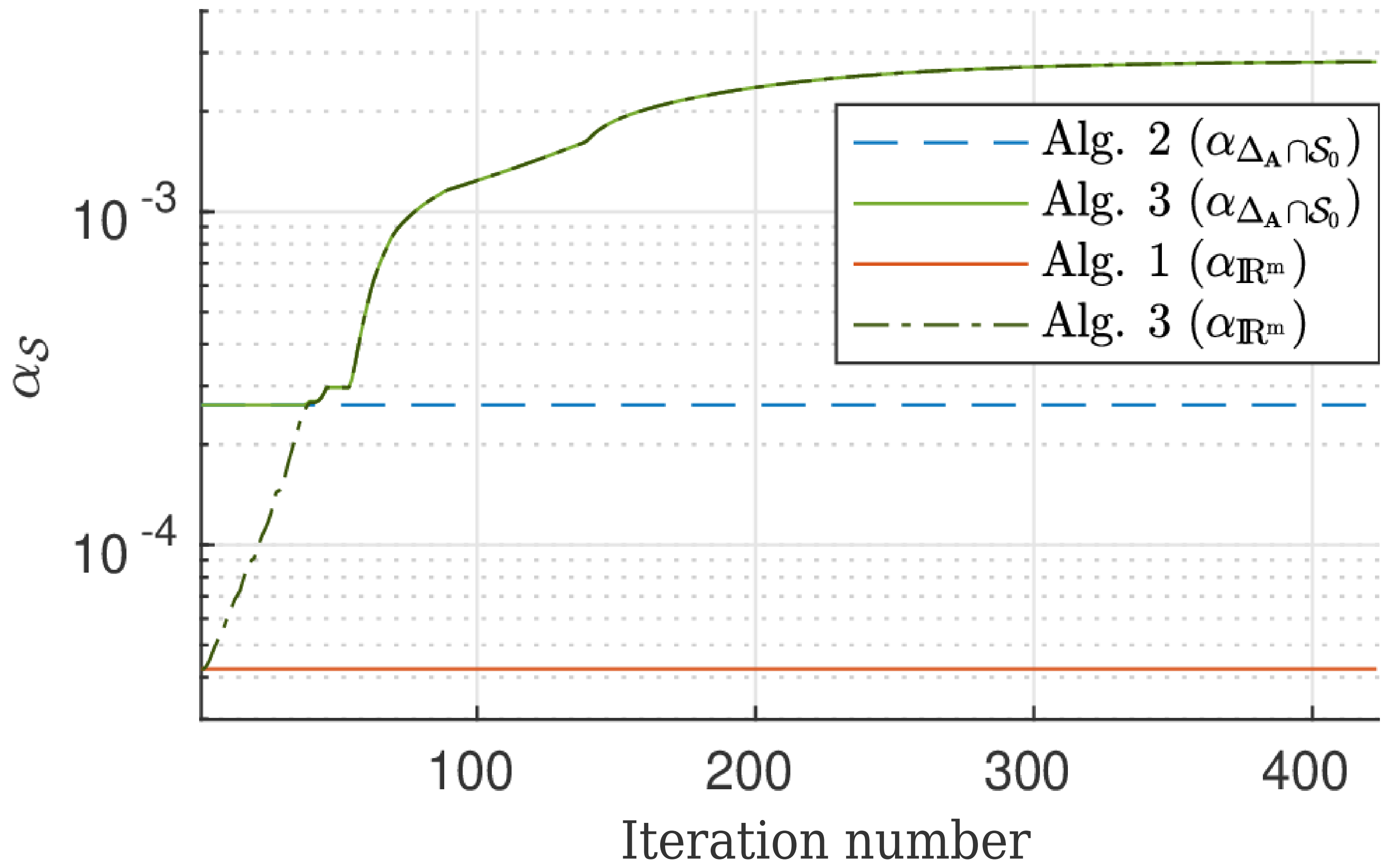
$$x_j \leftarrow x_j \cdot \underbrace{\frac{\mathbf{a}_j^\top(\mathbf{y} \odot (\mathbf{Ax})^{\beta-2})}{\mathbf{a}_j^\top(\mathbf{Ax})^{\beta-1} + \lambda}}_{>0}$$

  - No real zeros in the solution. Screening solves this issue.



Robustness to initialization of α

- Algorithm 3 initialized with global (worse) and local (better) bounds.



Concluding remarks

- GAP Safe extension exploiting **local properties** of the cost function.
  - Expands the class of admissible functions;
  - Potential improvement on previously applicable cases.
- Iterative refinement of the GAP safe sphere.
- Significant improvements both in terms of screening performance and convergence time.

Check out the full paper!

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

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

[3] C. F. Dantas, E. Soubies, C. Févotte. *Expanding Boundaries of GAP Safe Screening*. Submitted to JMLR. Feb 2021.

Solvers

[4] A. Beck, M. Teboulle. A Fast Iterative Shrinkage-thresholding Algorithm for Linear Inverse Problems. SIAM Journal on Imaging Sciences, Jan 2009.

[5] 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, Mar 2012.

[6] J. H. Friedman, T. Hastie, R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 2010.

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

[8] Cédric Févotte and Jérôme Idier. Algorithms for Nonnegative Matrix Factorization with the β-divergence. Neural Computation, Sep 2011.