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 \(\rightarrow\) local
  - Non-negativity constraints

- Allows to:
  - Extension to larger class of functions. E.g., \(\beta\)-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*.

\[ y \approx Ax, \text{ with } x \text{ sparse} \]

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

![Diagram showing matrix A and solution support](image-url)
Safe Screening

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

\[ y \approx Ax, \text{ with } x \text{ sparse} \]

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

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

\[ y \approx Ax, \text{ with } x \text{ sparse} \]

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

\[
\begin{array}{c}
\text{y} \\
\approx \\
\text{A} \\
\text{x}^* \\
\end{array}
\]
Problem definition

- Primal problem: \( \mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathbb{R}^n} P_\lambda(\mathbf{x}) := F(\mathbf{A}\mathbf{x}) + \lambda \Omega(\mathbf{x}) \)
  
  - \( F : \mathbb{R}^m \rightarrow \mathbb{R} \), 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.
  
  - \( \Omega : \mathbb{R}^n \rightarrow \mathbb{R} \), group-decomposable norm. We set: \( \Omega(\mathbf{x}) = ||\mathbf{x}||_1 \)
  
  - \( \lambda > 0 \) regularization parameter.

- Dual problem: \( \mathbf{\theta}^* = \arg\max_{\mathbf{\theta} \in \Delta_A} D_\lambda(\mathbf{\theta}) := -F^*(-\lambda \mathbf{\theta}) \)
  
  with \( \Delta_A = \{ \mathbf{\theta} \in \mathbb{R}^m \mid ||\mathbf{A}^T \mathbf{\theta}||_{\infty} \leq 1 \} \)

  - Unit ball of the dual norm \( \overline{\Omega} \)
Problem definition

• First-order optimality conditions:

1) \( \lambda \theta^* = -\nabla F(Ax^*) \)  
   \hspace{1cm} \text{(primal-dual link)}

2) \( A^T \theta^* \in \partial \Omega(x^*) \)  
   \hspace{1cm} \text{(subdifferential inclusion)}

\[ \forall j \in \{1, \ldots, n\}, \quad \begin{cases} 
|a_j^T \theta^*| \leq 1, & \text{if } x_j^* = 0 \\
|a_j^T \theta^*| = 1 & \text{if } x_j^* \neq 0
\end{cases} \]
Safe Screening

∀j ∈ {1, . . . , n}, \( \left\{ \begin{array}{l}
|a_j^T \theta^*| \leq 1, \text{ if } x_j^* = 0 \\
|a_j^T \theta^*| = 1 \text{ if } x_j^* \neq 0
\end{array} \right. \)

- Direct consequence: \( |a_j^T \theta^*| < 1 \implies x_j^* = 0 \)

⚠️ In practice, the dual solution \( \theta^* \) is not known.

✔️ Define a safe region \( \mathcal{R} \) which contains \( \theta^* \).

Example: \( \theta \in \mathbb{R}^2 \)

Safe screening rule [El Ghaoui et al. 2012]

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

\[
\max_{\theta \in \mathcal{R}} |a_j^T \theta| < 1 \implies |a_j^T \theta^*| < 1 \implies x_j^* = 0
\]

GAP Safe Screening

GAP Safe sphere [Ndiaye et al. 2017]

Assuming that the dual function $D_\lambda$ is $\alpha$-strongly concave, then for any feasible primal-dual pair $(x, \theta)$

$$\theta^* \in B(\theta, r), \text{ with } r = \sqrt{\frac{2 \text{ Gap}_\lambda(x, \theta)}{\alpha}}$$

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

- Requires global strong concavity of $D_\lambda$
- Can we use local strong concavity instead? Yes, if...

---

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 sphere

Assuming that the dual function $D_\lambda$ is $\alpha_S$-strongly concave, on a subset $S \in \mathbb{R}^m$ such that $\theta^* \in S$, then for any feasible primal-dual pair $(x, \theta)$ with $\theta \in S$

$$\theta^* \in B(\theta, r), \text{ with } r = \sqrt{\frac{2 \text{Gap}_\lambda(x, \theta)}{\alpha_S}}$$

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

Alg. 2) $S = \Delta_A \cap S_0$

Alg. 3) $S = B(\theta, r)$ : feedback loop between $r$ and $\alpha$.

$$B(\theta, r) \xrightarrow{} \alpha_{B(\theta, r)}$$

$$r = \sqrt{\frac{2 \text{Gap}_\lambda(x, \theta)}{\alpha_{B(\theta, r)}}}$$
Algorithm 0: Iterative solver without screening

Initialize \( x \in \mathbb{R}^n \)

Repeat until convergence
  \[
  \text{Solver update: } x \leftarrow \text{PrimalUpdate}(x, A, \lambda)
  \]

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** \( x \in \mathbb{R}^n, \ A = \{1, \ldots, n\}, \ \alpha \) global strong concavity bound

**Repeat** until convergence

- **Primal update:** \( x_A \leftarrow \text{PrimalUpdate}(x_A, A, \lambda) \)
- **Dual update:** \( \theta \leftarrow \Theta (x) \in \Delta_A \)
- **Safe screening:**
  \[
  r \leftarrow \sqrt{\frac{2 \text{Gap}_\lambda(x, \theta)}{\alpha}}
  \]
  \[
  A \leftarrow \{ j \in A \mid \max_{\theta \in B(\theta, r)} |a_j^T \theta| \geq 1 \}
  \]
  \[
  x_{A^c} \leftarrow 0
  \]
Algorithm 2 : Generalized Dynamic GAP Safe Screening (G-DGS) [D.S.F. 2021]

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

Repeat until convergence

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

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

Safe screening : 

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

$$\mathcal{A} \leftarrow \{ j \in \mathcal{A} | \max_{\mathbf{\theta} \in B(\mathbf{\theta}, r)} |a_j^T \mathbf{\theta}| \geq 1 \}$$

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

Algorithm 3

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

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

Repeat until convergence

Primal update: $x_{\mathcal{A}} \leftarrow \text{PrimalUpdate}(x_{\mathcal{A}}, A_{\mathcal{A}}, \lambda)$

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

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

$r \leftarrow \min \left( r, \sqrt{\frac{2 \text{Gap}_\lambda(x, \theta)}{\alpha_S}} \right)$

$S \leftarrow B(\theta, r)$

$\mathcal{A} \leftarrow \{ j \in \mathcal{A} \mid \max_{\theta \in B(\theta, r)} |a_j^T \theta| \geq 1 \}$

$x_{\mathcal{A}^c} \leftarrow 0$

Strong-concavity bound: evolution

- Alg. 1 ($\alpha_{\mathbb{R}^m}$)
- Alg. 2 ($\alpha_{\Delta_{\mathcal{A}} \cap \mathcal{S}_0}$)
- Alg. 3 ($\alpha_{\Delta_{\mathcal{A}} \cap \mathcal{S}_0}$)

Iteration number

$\alpha_s$
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 $F(Ax)$:

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

2) Compute a valid strong concavity bound $\alpha_S$.
   
   Example: $\alpha_{\Delta A}$ 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 $\alpha_B(\theta, r)$
Computing a strong concavity bound $\alpha_S$

Supposing $D_\lambda$ 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 $D_\lambda$ to be twice-differentiable:

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

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

Data-fidelity function $F(Ax) = \sum_i f_i([Ax]_i)$ is coordinate-separable, then, $D_\lambda(\theta) = F^*(-\lambda \theta) = \sum_i f_i^*(-\lambda \theta_i)$ is also coordinate-separable. The Hessian is diagonal with eigenvalues (diagonal entries) given by:

$$\sigma_i(\theta_i) = -\lambda^2 (f_i^*)''(\lambda \theta_i)$$
We distinguished three scenarios regarding the choice of $F(Ax)$:

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.: $\beta$-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 binary 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 $\beta$-divergence minimization.
  - Update step: each coordinate is multiplied by a positive factor.
    \[
    x_j \leftarrow x_j \cdot \frac{a_j^T(y \odot (Ax)^{\beta-2})}{a_j^T(Ax)^{\beta-1} + \lambda} > 0
    \]
  - No real zeros in the solution. Screening solves this issue.
Robustness to initialization of $\alpha$

- 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!


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

Matlab code: github.com/cassiofragadantas
References

Safe Screening


Solvers


