Screen & Relax for Sparse Support Identification

Théo Guyard[⋆], Cédric Herzet[⋄], Clément Elvira[†]
[⋆]Inria, Rennes, France
[⋄]Ensai, Rennes, France
[†]CentraleSupelec, Rennes, France

SMAI-MODE days
INSA Lyon, France
March 28th, 2024

General Context

Framework

Sparse optimization

- Minimize a loss with a sparse optimizer
- Applications in signal processing, machine learning, statistics, etc...

Framework

Sparse optimization

- Minimize a loss with a sparse optimizer
- Applications in signal processing, machine learning, statistics, etc...

Problem of interest

$$x^\star \in \operatorname{argmin}_x \ f(Ax) + \underbrace{\lambda\|x\|_1 + h(x)}_{g(x)}$$

Framework

Sparse optimization

- Minimize a loss with a sparse optimizer
- Applications in signal processing, machine learning, statistics, etc...

Problem of interest

$$x^\star \in \operatorname{argmin}_x \; f(\mathrm{A}x) + \underbrace{\lambda\|x\|_1 + h(x)}_{g(x)}$$

Working hypotheses

- $f(\cdot)$ and $h(\cdot)$ are proper, closed and convex functions
- $f(\cdot)$ and $h(\cdot)$ are differentiable with Lipschitz-continuous gradient
- $h(\cdot)$ is separable

  $\rightarrow$ With additional mild assumptions ensuring non-degeneracy of the problem

Solving sparse problems

$x^\star \in \mathrm{argmin}_x \ f(Ax) + g(x)$

Solution methods

- Composite objective: smooth + non-smooth-separable
- First-order methods accessing $\nabla f(\cdot)$, $\partial g(\cdot)$, $\mathrm{prox}_{f/g}(\cdot)$, ...
  - Proximal gradient descent
  - Coordinate descent
  - Alternating direction method of multipliers
  - ...

Solving sparse problems

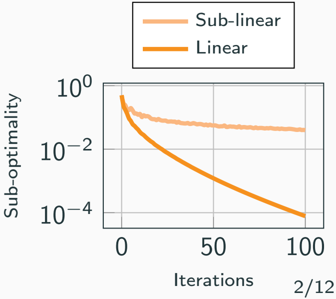
$\mathrm{x}^\star \in \mathrm{argmin}_{\mathrm{x}}\ f(\mathrm{Ax}) + g(\mathrm{x})$

Solution methods

- Composite objective: smooth + non-smooth-separable
- First-order methods accessing $\nabla f(\cdot)$, $\partial g(\cdot)$, $\mathrm{prox}_{f/g}(\cdot)$, ...
  - Proximal gradient descent
  - Coordinate descent
  - Alternating direction method of multipliers
  - ...

Convergence rates

- Sub-linear: $P(\mathrm{x}^{(k)}) - P(\mathrm{x}^\star) \leq C/k^\gamma$
- Linear: $P(\mathrm{x}^{(k)}) - P(\mathrm{x}^\star) \leq Ce^{-\gamma k}$
  - $\rightarrow$ Asymptotically (Peyré et al., 2015)
  - $\rightarrow$ Strong convexity (Aujol et al., 2023)



Sparse structure exploitation

Variable with many useless and few informative entries

Sparse structure exploitation

Variable with many useless and few informative entries

Screening tests

- Identify zeros in $x^\star$
- Dimensionality shrinking
- Computational savings

Sparse structure exploitation

Variable with many useless and few informative entries

Screening tests

- Identify zeros in $x^\star$
- Dimensionality shrinking
- Computational savings

Relaxing tests

- Identify non-zeros in $x^\star$
- Objective smoothing
- Super-linear convergence

Sparse structure exploitation

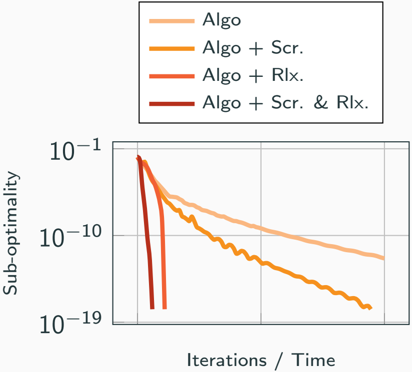
Variable with many useless and few informative entries

Screening tests

- Identify zeros in $x^\star$
- Dimensionality shrinking
- Computational savings

Relaxing tests

- Identify non-zeros in $x^\star$
- Objective smoothing
- Super-linear convergence

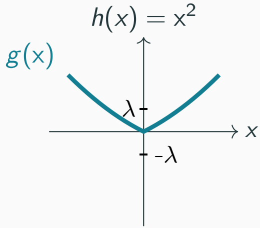


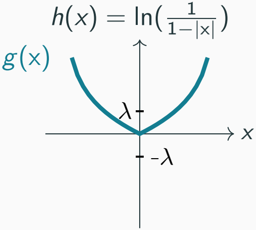
Screening and Relaxing Tests

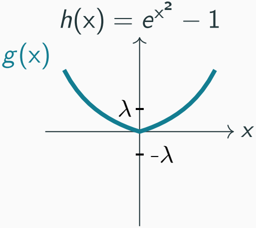
Geometrical idea

Characterize the nullity in $\mathrm{x}^\star$ from $\partial g(\mathrm{x}^\star)$

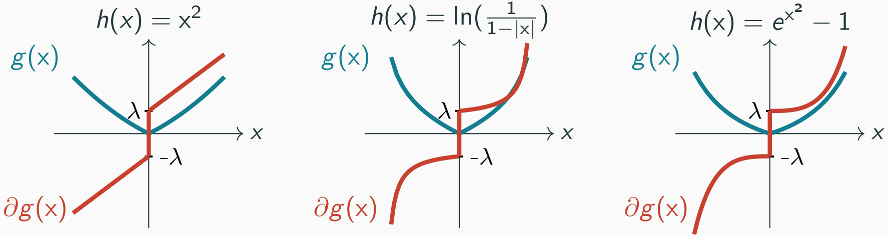
Characterize the nullity in $\mathrm{x}^\star$ from $\partial g(\mathrm{x}^\star)$





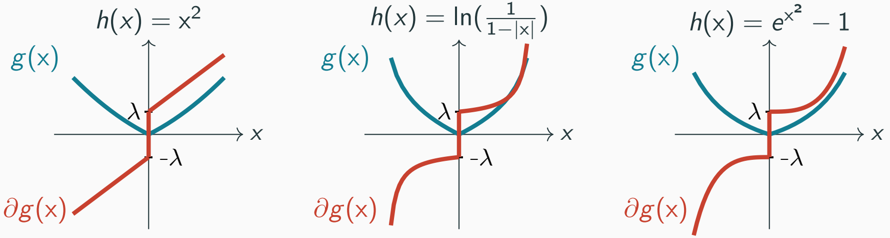


Characterize the nullity in $x^\star$ from $\partial g(x^\star)$



Geometrical idea

Characterize the nullity in $\mathrm{x}^\star$ from $\partial g(\mathrm{x}^\star)$



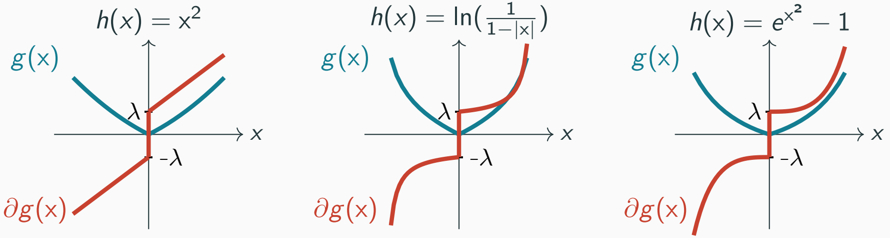
Geometrical screening and relaxing test

$$\partial_i g(\mathrm{x}^\star) \subset [-\lambda, \lambda] \quad \Longleftrightarrow \quad x_i^\star = 0$$
$$\partial_i g(\mathrm{x}^\star) \not\subset [-\lambda, \lambda] \quad \Longleftrightarrow \quad x_i^\star \neq 0$$

Geometrical idea

$g(\mathrm{x}) = \lambda\|\mathrm{x}\|_1 + h(\mathrm{x})$

Characterize the nullity in $\mathrm{x}^\star$ from $\partial g(\mathrm{x}^\star)$



Geometrical screening and relaxing test

$$\partial_i g(\mathrm{x}^\star) \subset [-\lambda, \lambda] \iff x_i^\star = 0$$
$$\partial_i g(\mathrm{x}^\star) \not\subset [-\lambda, \lambda] \iff x_i^\star \neq 0$$

$\partial g(\mathrm{x}^\star)$ is not available

Duality to rescue

Characterize the nullity in $x^\star$ from the dual problem

Duality to rescue

Characterize the nullity in $x^\star$ from the dual problem

Dual problem

$$u^\star \in \operatorname{argmax}_u \; -f^\star(-u) - g^\star(A^\top u)$$

$$\boxed{a_i^\top u^\star \in \partial_i g(x^\star)}$$

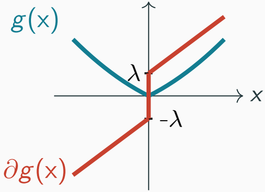
Duality to rescue

Characterize the nullity in $\mathrm{x}^\star$ from the dual problem

Dual problem

$$\mathrm{u}^\star \in \operatorname{argmax}_{\mathrm{u}} \ -f^\star(-\mathrm{u}) - g^\star(\mathrm{A}^\top \mathrm{u})$$

$$\boxed{\mathrm{a}_i^\top \mathrm{u}^\star \in \partial_i g(\mathrm{x}^\star)}$$



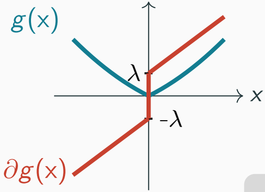
Duality to rescue

Characterize the nullity in $x^\star$ from the dual problem

Dual problem

$$u^\star \in \operatorname{argmax}_u \; -f^\star(-u) - g^\star(A^T u)$$

$$\boxed{a_i^T u^\star \in \partial_i g(x^\star)}$$



Screening and relaxing test

$$|a_i^T u^\star| \le \lambda \quad \Longleftrightarrow \quad x_i^\star = 0$$
$$|a_i^T u^\star| > \lambda \quad \Longleftrightarrow \quad x_i^\star \neq 0$$

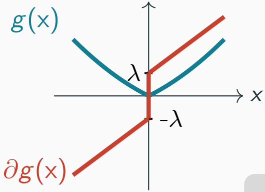
Duality to rescue

Characterize the nullity in $x^\star$ from the dual problem

Dual problem

$$u^\star \in \operatorname{argmax}_u \; -f^\star(-u) - g^\star(A^T u)$$

$$\boxed{a_i^T u^\star \in \partial_i g(x^\star)}$$



Screening and relaxing test

$$|a_i^T u^\star| \leq \lambda \quad \Longleftrightarrow \quad x_i^\star = 0$$
$$|a_i^T u^\star| > \lambda \quad \Longleftrightarrow \quad x_i^\star \neq 0$$

$u^\star$ is not available

Safe regions

Characterize the nullity in $x^\star$ from a safe region

Safe regions

Characterize the nullity in $x^\star$ from a safe region

$$\boxed{\text{Safe region: } u^\star \in \mathcal{R}}$$

Safe screening and relaxing test

$$\max_{u \in \mathcal{R}} |a_i^T u| \leq \lambda \quad \Longrightarrow \quad x_i^\star = 0$$
$$\min_{u \in \mathcal{R}} |a_i^T u| > \lambda \quad \Longrightarrow \quad x_i^\star \neq 0$$

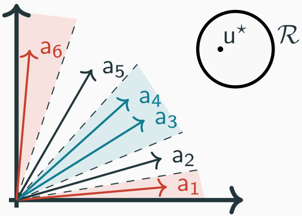
Safe regions

Characterize the nullity in $x^\star$ from a safe region

$$\boxed{\text{Safe region: } u^\star \in \mathcal{R}}$$

Safe screening and relaxing test

$$\max_{u \in \mathcal{R}} |a_i^T u| \leq \lambda \quad \Longrightarrow \quad x_i^\star = 0$$
$$\min_{u \in \mathcal{R}} |a_i^T u| > \lambda \quad \Longrightarrow \quad x_i^\star \neq 0$$



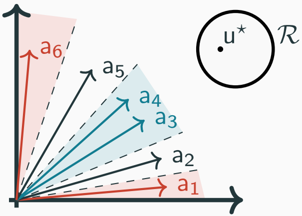
Safe regions

Characterize the nullity in $x^\star$ from a safe region

$$\boxed{\text{Safe region: } u^\star \in \mathcal{R}}$$

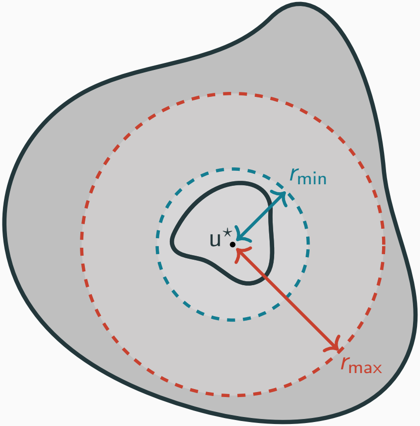
Safe screening and relaxing test

$$\max_{u\in\mathcal{R}} |a_i^T u| \leq \lambda \quad \Longrightarrow \quad x_i^\star = 0$$
$$\min_{u\in\mathcal{R}} |a_i^T u| > \lambda \quad \Longrightarrow \quad x_i^\star \neq 0$$



Lasso problem
Limit case with $h(x) = 0$

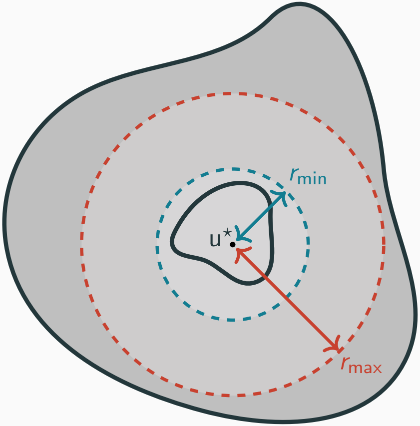
Working regimes



$\mathcal{R} \subset \mathcal{S}(u^\star, r_{\text{min}}) \implies$ all tests passed

$\mathcal{R} \supset \mathcal{S}(u^\star, r_{\text{max}}) \implies$ no tests passed

Working regimes



$\mathcal{R} \subset \mathcal{S}(\mathrm{u}^\star, r_{\min}) \implies$ all tests passed

$\mathcal{R} \supset \mathcal{S}(\mathrm{u}^\star, r_{\max}) \implies$ no tests passed

$\boxed{r_{\min} > 0}$

We know how to construct safe regions with a radius proportional to the (square root of) the duality gap

$\downarrow$

Guaranty to identify all zeros and non-zeros in $\mathrm{x}^\star$ in finite time

Implementation

Reformulation

Initial problem

$$\min_x f(Ax) + \lambda\|x\|_1 + h(x)$$

Reformulation

Initial problem

$$\min_x f(\mathrm{A}\mathrm{x}) + \lambda \|\mathrm{x}\|_1 + h(\mathrm{x})$$

Perfect identif.

$\mathcal{S}_{\mathbf{0}}^\star = \{i \mid x_i^\star = 0\}$

$\mathcal{S}_{\mathbf{1}}^\star = \{i \mid x_i^\star \neq 0\}$

Ideal reformulation

$$\begin{cases} \min_x & f(\mathrm{A}_{\mathcal{S}_{\mathbf{1}}^\star}\mathrm{x}_{\mathcal{S}_{\mathbf{1}}^\star}) + \lambda \boldsymbol{\theta}_{\mathcal{S}_{\mathbf{1}}^\star}^\top \mathrm{x}_{\mathcal{S}_{\mathbf{1}}^\star} + h(\mathrm{x}_{\mathcal{S}_{\mathbf{1}}^\star}) \\ \text{s.t.} & \mathrm{x}_{\mathcal{S}_{\mathbf{0}}^\star} = 0 \end{cases}$$

Initial problem

$$\min_x f(\mathrm{A}\mathrm{x}) + \lambda\|\mathrm{x}\|_1 + h(\mathrm{x})$$

Screen & Relax
$\mathcal{S}_{\mathbf{0}} \subseteq \mathcal{S}_{\mathbf{0}}^\star$
$\mathcal{S}_{\mathbf{1}} \subseteq \mathcal{S}_{\mathbf{1}}^\star$
$\mathcal{S}_{\bullet} = (\mathcal{S}_{\mathbf{0}} \cup \mathcal{S}_{\mathbf{1}})^{\complement}$

Perfect identif.
$\mathcal{S}_{\mathbf{0}}^\star = \{i \mid x_i^\star = 0\}$
$\mathcal{S}_{\mathbf{1}}^\star = \{i \mid x_i^\star \neq 0\}$

Ideal reformulation

$$\begin{cases} \min_x & f(\mathrm{A}_{\mathcal{S}_{\mathbf{1}}^\star}\mathrm{x}_{\mathcal{S}_{\mathbf{1}}^\star}) + \lambda\boldsymbol{\theta}_{\mathcal{S}_{\mathbf{1}}^\star}^{\mathrm{T}}\mathrm{x}_{\mathcal{S}_{\mathbf{1}}^\star} + h(\mathrm{x}_{\mathcal{S}_{\mathbf{1}}^\star}) \\ \text{s.t.} & \mathrm{x}_{\mathcal{S}_{\mathbf{0}}^\star} = 0 \end{cases}$$

Partial reformulation

$$\begin{cases} \min & f(\mathrm{A}_{\mathcal{S}_{\mathbf{1}}\cup\mathcal{S}_\bullet}\mathrm{x}_{\mathcal{S}_{\mathbf{1}}\cup\mathcal{S}_\bullet}) + \lambda\boldsymbol{\theta}_{\mathcal{S}_{\mathbf{1}}}^{\mathrm{T}}\mathrm{x}_{\mathcal{S}_{\mathbf{1}}} + h(\mathrm{x}_{\mathcal{S}_{\mathbf{1}}}) \\ & \qquad + \lambda\|\mathrm{x}_{\mathcal{S}_\bullet}\|_1 + h(\mathrm{x}_{\mathcal{S}_\bullet}) \\ \text{s.t.} & \mathrm{x}_{\mathcal{S}_{\mathbf{0}}} = 0 \end{cases}$$

Algorithm embedding

Algorithm 1 Screen & Relax

initialize $(\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_\bullet) = (\emptyset, \emptyset, [\![1, n]\!])$
repeat
    // Iterate update
    $\mathrm{x}^k_{\mathcal{S}_\bullet} \leftarrow 1^{st}\mathrm{OrderIteration}(\mathrm{x}^{k-1})$
    $\mathrm{x}^k_{\mathcal{S}_1} \leftarrow 2^{nd}\mathrm{OrderIteration}(\mathrm{x}^{k-1})$
    $\mathrm{x}^k_{\mathcal{S}_0} \leftarrow 0$
    // Problem update
    Construct a new safe region $\mathcal{R}^k$ from x^k
    Update $(\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_\bullet)$ using $\mathcal{R}^k$
until convergence criterion is met

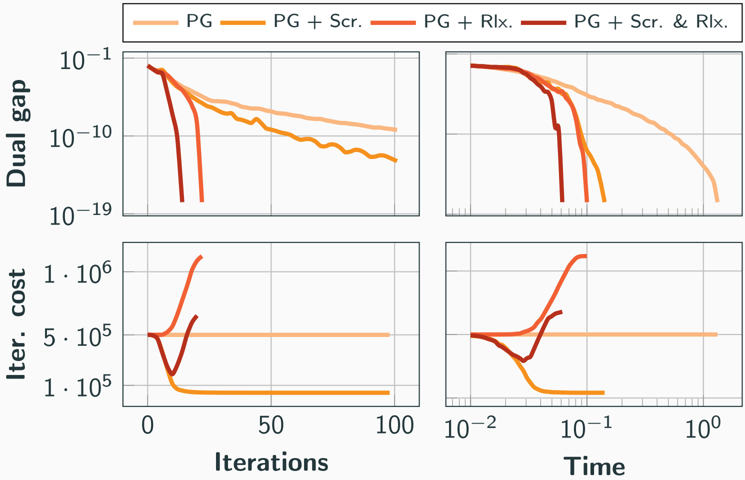
Iteration cost reduction: $n \to n - |\mathcal{S}_0|$

Faster convergence rate: (sub-)linear $\to$ super-linear

Numerics

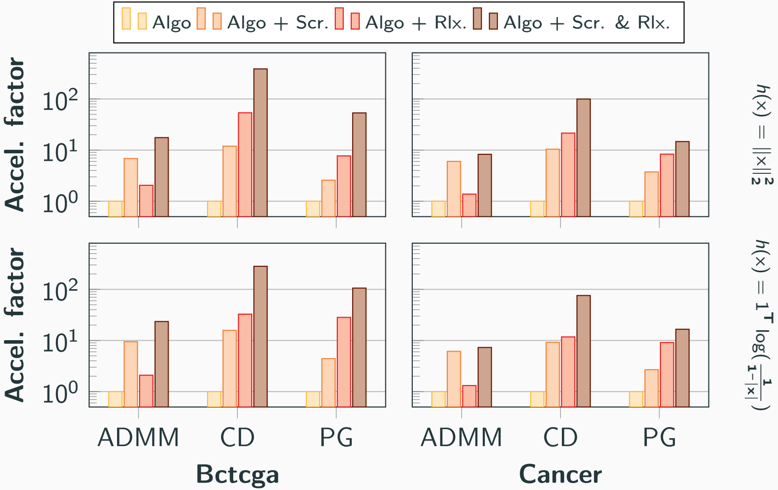
Screen & Relax effects

$$\min_x f(Ax) + \lambda\|x\|_1 + h(x)$$



Regularization path fitting

$$\min_x f(Ax) + \lambda \|x\|_1 + h(x)$$



Regression - 356 × 17,322

Classification - 100 × 12,600

Take-home message

- Structure matters in sparse problems
- Identification of zeros
  - Screening tests
  - Allows for dimensionality reduction
  - Computational savings
- Identification of non-zeros
  - Relaxing tests
  - Allows for objective smoothing
  - Faster convergence rate
- Screen & Relax strategy to benefit from
  - Computational savings
  - Accelerated convergence

Question time

🤓