Screen & Relax

Accélérer la résolution de l'Elastic-Net par identification du support de la solution

Théo Guyard[1,2], Cédric Herzet[2], Clément Elvira[3]
[1] INSA Rennes
[2] INRIA Rennes Bretagne Atlantique
[3] IETR CentraleSupélec

GRETSI – 8 Septembre 2022

General context

Sparse problem

Ingredients of the problem

- A target y

Sparse problem

Ingredients of the problem

- A target y
- A dictionary $\mathrm{A} = \{\mathrm{a}_i\}_{i \in \mathcal{I}}$ made of atoms

Sparse problem

Ingredients of the problem

- A target y
- A dictionary $\mathrm{A} = \{\mathrm{a}_i\}_{i \in \mathcal{I}}$ made of atoms

Objective

- Find a sparse linear combination of atoms that well approximates the target through a given model

Sparse problem

Ingredients of the problem

- A target y
- A dictionary $A = \{a_i\}_{i \in \mathcal{I}}$ made of atoms

Objective

- Find a sparse linear combination of atoms that well approximates the target through a given model

Rough formulation

Problem

Find x sparse such that $y \simeq \text{Model}(Ax)$

Sparse problem

Ingredients of the problem

- A target y
- A dictionary $A = \{a_i\}_{i \in \mathcal{I}}$ made of atoms

Objective

- Find a sparse linear combination of atoms that well approximates the target through a given model

Rough formulation

Problem

Find x sparse such that $y \simeq \mathrm{Model}(Ax)$

Remark : Entries of x weight each atom in the linear combination.

The Elastic-Net problem

Formulation and properties

Target problem

Solve

$$x^\star = \arg\min_x \left\{ \mathrm{P}(x) = \underbrace{\tfrac{1}{2}\|y - Ax\|_2^2}_{f(Ax)} + \underbrace{\lambda(\sigma\|x\|_1 + \tfrac{1-\sigma}{2}\|x\|_2)}_{\lambda g(x)} \right\} \qquad (P)$$

where $\lambda > 0$ and $\sigma \in]0,1[$ are tuning hyperparameters.

Formulation and properties

Target problem

Solve

$$x^\star = \arg\min_x \left\{ \mathrm{P}(x) = \underbrace{\tfrac{1}{2}\|y - Ax\|_2^2}_{f(Ax)} + \underbrace{\lambda(\sigma\|x\|_1 + \tfrac{1-\sigma}{2}\|x\|_2)}_{\lambda g(x)} \right\} \qquad (P)$$

where $\lambda > 0$ and $\sigma \in]0,1[$ are tuning hyperparameters.

Properties of (P)

- Convex non-smooth problem
- Least-squares : Ensures a good reconstruction of the target
- ℓ_1-norm : Enforces sparsity
- ℓ_2-norm : Promotes desirable properties

Formulation and properties

Target problem

Solve

$$\mathrm{x}^\star = \arg\min_{\mathrm{x}} \left\{ \mathrm{P}(\mathrm{x}) = \underbrace{\tfrac{1}{2}\|\mathrm{y} - \mathrm{A}\mathrm{x}\|_2^2}_{f(\mathrm{A}\mathrm{x})} + \underbrace{\lambda\left(\sigma\|\mathrm{x}\|_1 + \tfrac{1-\sigma}{2}\|\mathrm{x}\|_2\right)}_{\lambda g(\mathrm{x})} \right\} \qquad (P)$$

where $\lambda > 0$ and $\sigma \in]0, 1[$ are tuning hyperparameters.

Properties of (P)

- Convex non-smooth problem
- Least-squares : Ensures a good reconstruction of the target
- ℓ_1-norm : Enforces sparsity
- ℓ_2-norm : Promotes desirable properties

Solving (P)

- Broad class of solution methods (gradient-based, pivot-based, ...)
- Acceleration strategies (backtracking, screening tests, ...)

Screening and Relaxing tests

Main idea

In the context of a sparse problem

Main idea

In the context of a sparse problem

- Where are zero and non-zero entries in $x^\star$?

Main idea

In the context of a sparse problem

- Where are zero and non-zero entries in $x^\star$?
- Can we accelerate a given solution method if we knew some ?

Main idea

In the context of a sparse problem

- Where are zero and non-zero entries in $x^\star$?
- Can we accelerate a given solution method if we knew some ?

$\rightarrow$ Spoiler alert : Yes and yes ! We can leverage duality.

Main idea

In the context of a sparse problem

- Where are zero and non-zero entries in $\mathrm{x}^\star$?
- Can we accelerate a given solution method if we knew some ?

$\rightarrow$ Spoiler alert : Yes and yes ! We can leverage duality.

Primal problem

$$\mathrm{x}^\star = \arg\min_{\mathrm{x}} \mathrm{P}(\mathrm{x}) \qquad (P)$$

$\equiv$

Dual problem

$$\mathrm{u}^\star = \arg\max_{\mathrm{u}} \mathrm{D}(\mathrm{u}) \qquad (D)$$

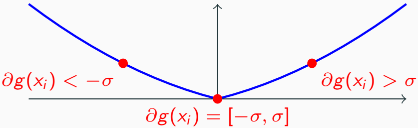
with optimality conditions linking $\mathrm{x}^\star$ and $\mathrm{u}^\star$

Optimality conditions

$$a_i^T u^\star \in \lambda \partial g(x_i^\star)$$

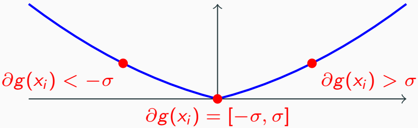
Optimality conditions

$$\mathsf{a}_i^T \mathsf{u}^\star \in \lambda \partial g(x_i^\star)$$



Optimality conditions

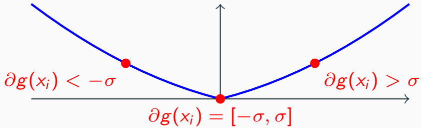
$$\mathrm{a}_i^T \mathrm{u}^\star \in \lambda \partial g(x_i^\star)$$



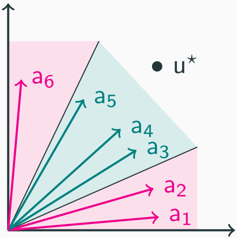
$$|\mathrm{a}_i^T \mathrm{u}^\star| \leq \lambda\sigma \quad \Longleftrightarrow \quad x_i^\star = 0$$
$$|\mathrm{a}_i^T \mathrm{u}^\star| > \lambda\sigma \quad \Longleftrightarrow \quad x_i^\star \neq 0$$

Optimality conditions

$$a_i^T u^\star \in \lambda \partial g(x_i^\star)$$

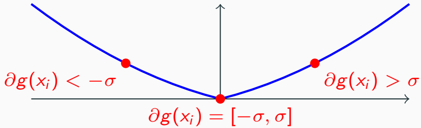


$$|a_i^T u^\star| \le \lambda\sigma \quad \Longleftrightarrow \quad x_i^\star = 0$$
$$|a_i^T u^\star| > \lambda\sigma \quad \Longleftrightarrow \quad x_i^\star \neq 0$$

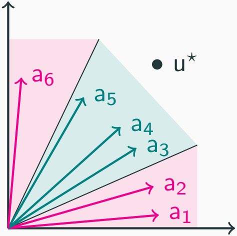


Optimality conditions

$$a_i^T u^\star \in \lambda \partial g(x_i^\star)$$



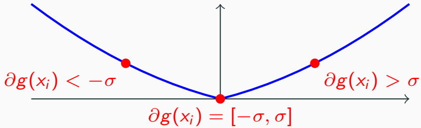
$$|a_i^T u^\star| \leq \lambda\sigma \iff x_i^\star = 0$$
$$|a_i^T u^\star| > \lambda\sigma \iff x_i^\star \neq 0$$



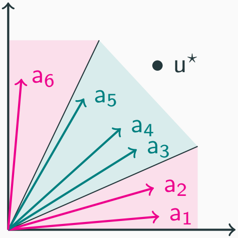
We can identify zeros and non-zeros in $x^\star$

Optimality conditions

$$a_i^T u^\star \in \lambda \partial g(x_i^\star)$$



$$|a_i^T u^\star| \leq \lambda\sigma \iff x_i^\star = 0$$
$$|a_i^T u^\star| > \lambda\sigma \iff x_i^\star \neq 0$$



We can identify zeros and non-zeros in $x^\star$

... but we need $u^\star$!

Relaxed optimality condition

Let $u^\star \in \mathbb{S}(c, r)$, then

$$|a_i^T c| + r \leq \lambda\sigma \quad \Longrightarrow \quad x_i^\star = 0 \qquad \text{(screening test)}$$

Relaxed optimality condition

Let $u^\star \in \mathbb{S}(c, r)$, then

$$
\begin{aligned}
|a_i^T c| + r \leq \lambda\sigma \quad &\Longrightarrow \quad x_i^\star = 0 \qquad &\text{(screening test)}\\
|a_i^T c| - r > \lambda\sigma \quad &\Longrightarrow \quad x_i^\star \neq 0 \qquad &\text{(relaxing test)}
\end{aligned}
$$

Relaxed optimality condition

Let $u^\star \in \mathbb{S}(c, r)$, then

$$\begin{aligned} |a_i^T c| + r \leq \lambda\sigma \quad &\Longrightarrow \quad x_i^\star = 0 \qquad &\text{(screening test)} \\ |a_i^T c| - r > \lambda\sigma \quad &\Longrightarrow \quad x_i^\star \neq 0 \qquad &\text{(relaxing test)} \end{aligned}$$

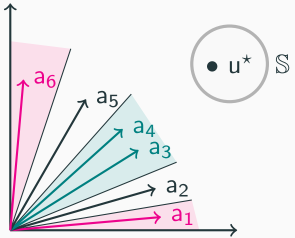
$\rightarrow$ No $u^\star$ needed anymore, but only a "safe region" containing it !

Relaxed optimality condition

Let $u^\star \in \mathbb{S}(c, r)$, then

$$
\begin{aligned}
|a_i^T c| + r \leq \lambda\sigma \quad &\Longrightarrow \quad x_i^\star = 0 \qquad \text{(screening test)}\\
|a_i^T c| - r > \lambda\sigma \quad &\Longrightarrow \quad x_i^\star \neq 0 \qquad \text{(relaxing test)}
\end{aligned}
$$

$\rightarrow$ No $u^\star$ needed anymore, but only a “safe region” containing it !



Dimensionality reduction

Leveraging test results

With screening test

Zero entries identified in $x^\star$ can be discarded from the problem without changing the objective value.

Leveraging test results

With screening test

Zero entries identified in $x^\star$ can be discarded from the problem without changing the objective value.

With relaxing test

Non-zero entries identified in $x^\star$ can be expressed as a linear combination of all the other entries.

Problem reformulation

Notations $(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}})$: Set of zero/non-zero/unclassified entries in $\mathbf{x}^\star$

Problem reformulation

Notations $(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}})$: Set of zero/non-zero/unclassified entries in $\mathsf{x}^\star$

Initial problem

$$\mathsf{x}^\star = \operatorname{argmin}_{\mathsf{x}\in\mathbb{R}^n} \left\{ \mathrm{P}(\mathsf{x}) = \tfrac{1}{2}\|\mathsf{y} - \mathsf{A}\mathsf{x}\|_2^2 + \lambda(\sigma\|\mathsf{x}\|_1 + \tfrac{1-\sigma}{2}\|\mathsf{x}\|_2^2) \right\}$$

n dimensional problem

Problem reformulation

Notations $(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}})$: Set of zero/non-zero/unclassified entries in $\mathsf{x}^\star$

Initial problem

$$\mathsf{x}^\star = \operatorname{argmin}_{\mathsf{x}\in\mathbb{R}^n} \left\{ \mathrm{P}(\mathsf{x}) = \tfrac{1}{2}\|\mathsf{y} - \mathsf{A}\mathsf{x}\|_2^2 + \lambda(\sigma\|\mathsf{x}\|_1 + \tfrac{1-\sigma}{2}\|\mathsf{x}\|_2^2) \right\}$$

n dimensional problem

$$\rightsquigarrow$$

Reduced problem

$$\begin{cases} \mathsf{x}^\star_{\bar{\mathcal{S}}} &= \operatorname{argmin}_{\mathsf{x}\in\mathbb{R}^{|\bar{\mathcal{S}}|}} \left\{ \tilde{\mathrm{P}}(\mathsf{x}) = \tfrac{1}{2}\|\tilde{\mathsf{y}} - \tilde{\mathsf{A}}\mathsf{x}\|_2^2 + \lambda(\sigma\|\mathsf{x}\|_1 + \tfrac{1-\sigma}{2}\|\mathsf{x}\|_{\mathsf{M}}^2) \right\} \\ \mathsf{x}^\star_{\mathcal{S}_1} &= \mathsf{B}\mathsf{x}^\star_{\bar{\mathcal{S}}} + \mathsf{b} \\ \mathsf{x}^\star_{\mathcal{S}_0} &= 0 \end{cases}$$

Problem reformulation

Notations $(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}})$: Set of zero/non-zero/unclassified entries in $x^\star$

Initial problem

$$x^\star = \operatorname{argmin}_{x \in \mathbb{R}^n} \left\{ \mathrm{P}(x) = \tfrac{1}{2}\|y - Ax\|_2^2 + \lambda(\sigma\|x\|_1 + \tfrac{1-\sigma}{2}\|x\|_2^2) \right\}$$

n dimensional problem

$$\rightsquigarrow$$

Reduced problem

$$\begin{cases} x^\star_{\bar{\mathcal{S}}} & = \operatorname{argmin}_{x \in \mathbb{R}^{|\bar{\mathcal{S}}|}} \left\{ \tilde{\mathrm{P}}(x) = \tfrac{1}{2}\|\tilde{y} - \tilde{A}x\|_2^2 + \lambda(\sigma\|x\|_1 + \tfrac{1-\sigma}{2}\|x\|_M^2) \right\} \\ x^\star_{\mathcal{S}_1} & = Bx^\star_{\bar{\mathcal{S}}} + b \\ x^\star_{\mathcal{S}_0} & = 0 \end{cases}$$

$n - |\mathcal{S}_0| - |\mathcal{S}_1|$ dimensional problem (similar structure)

Some linear algebra operations (negligible computational cost)

Dynamic Screen & Relax principle

Algorithm 1: "Screen & Relax" solving procedure

Input: A, y, λ, σ, $x^{(0)}$

$(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}}) \leftarrow (\emptyset, \emptyset, \{1, \ldots, n\})$

while *convergence criterion is not met* **do**

Dynamic Screen & Relax principle

Algorithm 2: "Screen & Relax" solving procedure

Input: A, y, λ, σ, $x^{(0)}$

$(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}}) \leftarrow (\emptyset, \emptyset, \{1, \dots, n\})$

while *convergence criterion is not met* **do**

    Update the current iterate

Dynamic Screen & Relax principle

Algorithm 3: "Screen & Relax" solving procedure

Input: A, y, λ, σ, $x^{(0)}$

$(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}}) \leftarrow (\emptyset, \emptyset, \{1, \dots, n\})$

while *convergence criterion is not met* **do**

    Update the current iterate

    Update $(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}})$ with screening and relaxing tests

Dynamic Screen & Relax principle

Algorithm 4: "Screen & Relax" solving procedure

Input: A, y, λ, σ, $x^{(0)}$

$(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}}) \leftarrow (\emptyset, \emptyset, \{1, \ldots, n\})$

while *convergence criterion is not met* **do**

- Update the current iterate
- Update $(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}})$ with screening and relaxing tests
- Reduce the problem

Dynamic Screen & Relax principle

Algorithm 5: "Screen & Relax" solving procedure

Input: A, y, λ, σ, $x^{(0)}$

$(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}}) \leftarrow (\emptyset, \emptyset, \{1, \ldots, n\})$

while *convergence criterion is not met* **do**

  Update the current iterate

  Update $(\mathcal{S}_0, \mathcal{S}_1, \bar{\mathcal{S}})$ with screening and relaxing tests

  Reduce the problem

  if $\bar{\mathcal{S}} = \emptyset$ **then**

    The solution is available in closed form

  end

end

Some numerical results

Experimental setup

Synthetic data generation

- Generate the dictionary A randomly
- Generate a k-sparse vector $x^{\dagger}$
- Set $y = Ax^{\star} + \text{noise}$
- Solve (P) with a tailored method

Experimental setup

Synthetic data generation

- Generate the dictionary A randomly
- Generate a k-sparse vector $x^\dagger$
- Set $y = Ax^\star + \text{noise}$
- Solve (P) with a tailored method

Concurrent methods

- Accelerated proximal-gradient algorithm
- Same with screening tests
- Same with relaxing tests
- Same with screening and relaxing tests

Experimental setup

Synthetic data generation

- Generate the dictionary A randomly
- Generate a k-sparse vector $x^{\dagger}$
- Set $y = Ax^{\star} + \text{noise}$
- Solve (P) with a tailored method

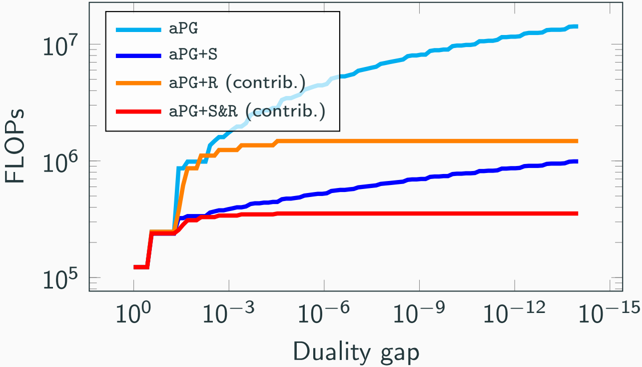
Concurrent methods

- Accelerated proximal-gradient algorithm
- Same with screening tests
- Same with relaxing tests
- Same with screening and relaxing tests

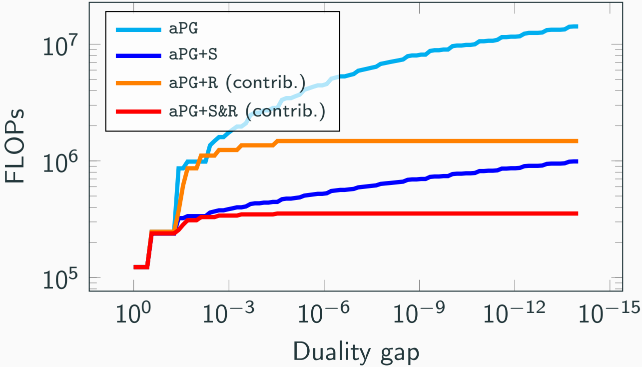
Metrics

- Duality gap : How close is the objective from its optimal value
- FLOPs : Number of linear algebra operations performed

Classical convergence scheme

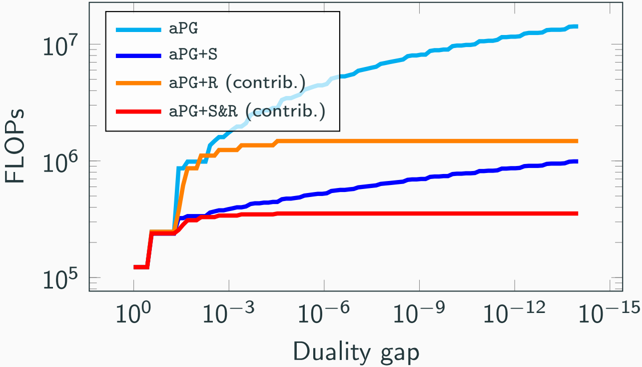


Classical convergence scheme



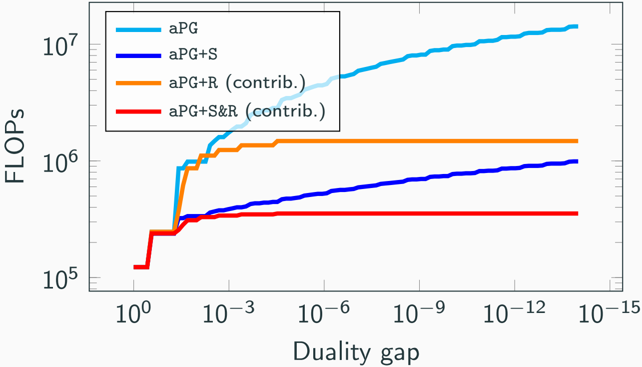
- Complexity reduction with screening and/or relaxing

Classical convergence scheme



- Complexity reduction with screening and/or relaxing
- Convergence to machine precision at some point

Classical convergence scheme



- Complexity reduction with screening and/or relaxing
- Convergence to machine precision at some point
- Gains depend on the sparsity in $x^\star$
    - Too few non-zeros : relaxing has little impact
    - Too many non-zeros : problem updates become binding