Faster and still safe: Combining screening techniques and structured dictionaries to accelerate the Lasso

Cássio F. DANTAS, Rémi GRIBONVAL
cassio.fraga-dantas@inria.fr, remi.gribonval@inria.fr

Accelerate the Lasso optimization
by combining two strategies :

1) Safe Screening Rules

2) Fast Structured Dictionaries

inventors for the digital world

Contents

inventors for the digital world

01

Context

Lasso problem

The l1-regularized least squares.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \underbrace{\frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1}_{P(\boldsymbol{\beta})}$$

Denoting :

- $\mathbf{y} \in \mathbb{R}^N$ the observation vector;
- $\mathrm{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_K] \in \mathbb{R}^{N \times K}$ the design matrix (or dictionary);
- $\boldsymbol{\beta} \in \mathbb{R}^{\boldsymbol{K}}$ the sparse representation vector;
- $\lambda > 0$ parameter controlling the sparsity of the solution.

Dual Lasso

Dual formulation of the Lasso problem :

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Delta_{\mathbf{X}}}{\operatorname{argmax}} \underbrace{\frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\left\|\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda}\right\|_2^2}_{D(\boldsymbol{\theta})}$$

Denoting :

- $\boldsymbol{\theta} \in \mathbb{R}^{\boldsymbol{N}}$ the dual variable;
- $\Delta_{\mathbf{X}} = \{\boldsymbol{\theta} \in \mathbb{R}^N : \|\mathbf{X}^T\boldsymbol{\theta}\|_\infty \leq 1\}$ the feasible set;

Motivation

- **Iterative algorithms** are often used to solve the Lasso problem.

- Exemple : ISTA (Iterative Shrinkage-Thresholding Algorithm)

---

while not converged **do**

$$\boldsymbol{\beta}_{t+1} \leftarrow \mathrm{ST}_{\frac{\lambda}{L_t}}\left(\boldsymbol{\beta}_t + \frac{1}{L_t}\mathbf{X}_t^T(\mathbf{y} - \mathbf{X}_t\boldsymbol{\beta}_t)\right)$$

- Two **matrix-vector multiplications** at each iteration.

Quadratic complexity!

Can it be reduced?

02

Fast Structured Dictionaries

Structure ⇒ Acceleration

Accelerate matrix-vector multiplications

Constrain the dictionary matrix to have a certain type of structure.

Examples :

- Kronecker product
- Sparse factors
- Circulant factors
- (...)

Structured Approximation

If the dictionary matrix $\mathbf{X}$ is not structured, find a structured approximation $\tilde{\mathbf{X}}$.

$$\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{E},$$

where $\mathbf{E}$ is the approximation error matrix and $\mathbf{e}_j$ is its j-th column.

Algorithm (high level view)

1) Start Lasso optimization by using the structured $\tilde{\mathbf{X}}$, to take advantage of its reduced multiplication cost.

2) As the algorithm approaches the solution, switch back to the original dictionary $\mathbf{X}$.

while switching criterion not met **do**

$$\boldsymbol{\beta}_{t+1} \leftarrow \mathrm{ST}_{\frac{\lambda}{L_t}}\left(\boldsymbol{\beta}_t + \frac{1}{L_t}\tilde{\mathbf{X}}_t^T(\mathbf{y} - \tilde{\mathbf{X}}_t\boldsymbol{\beta}_t)\right)$$

while not converged **do**

$$\boldsymbol{\beta}_{t+1} \leftarrow \mathrm{ST}_{\frac{\lambda}{L_t}}\left(\boldsymbol{\beta}_t + \frac{1}{L_t}\mathbf{X}_t^T(\mathbf{y} - \mathbf{X}_t\boldsymbol{\beta}_t)\right)$$

Algorithm (high level view)

1) Start Lasso optimization by using the structured $\tilde{\mathbf{X}}$, to take advantage of its reduced multiplication cost.

2) As the algorithm approaches the solution, switch back to the original dictionary $\mathbf{X}$.

while switching criterion not met **do**

$$\boldsymbol{\beta}_{t+1} \leftarrow \mathrm{ST}_{\frac{\lambda}{L_t}}\left(\boldsymbol{\beta}_t + \frac{1}{L_t}\tilde{\mathbf{X}}_t^T(\mathbf{y} - \tilde{\mathbf{X}}_t\boldsymbol{\beta}_t)\right)$$

while not converged **do**

$$\boldsymbol{\beta}_{t+1} \leftarrow \mathrm{ST}_{\frac{\lambda}{L_t}}\left(\boldsymbol{\beta}_t + \frac{1}{L_t}\mathbf{X}_t^T(\mathbf{y} - \mathbf{X}_t\boldsymbol{\beta}_t)\right)$$

03

Safe Screening Rules

Safe Screening

- Rules for identifying inactive dictionary atoms, before solving the problem.

Safe Screening

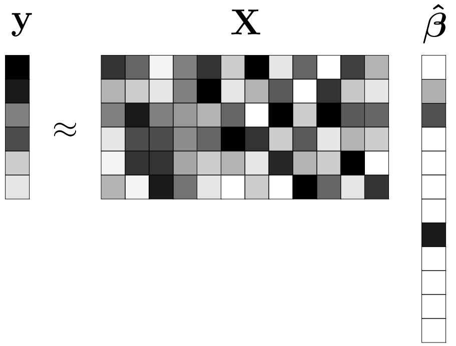
- Rules for identifying inactive dictionary atoms, before solving the problem.

Dictionary columns that will receive zero weight on the reconstruction of the input signal

Safe Screening

- Rules for identifying inactive dictionary atoms, before solving the problem.

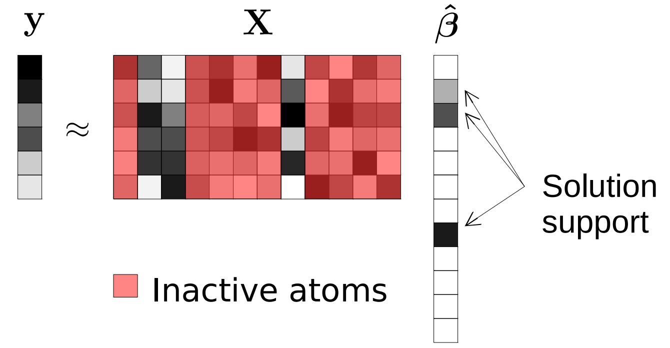
Dictionary columns that will receive zero weight on the reconstruction of the input signal



Safe Screening

- Rules for identifying inactive dictionary atoms, before solving the problem.

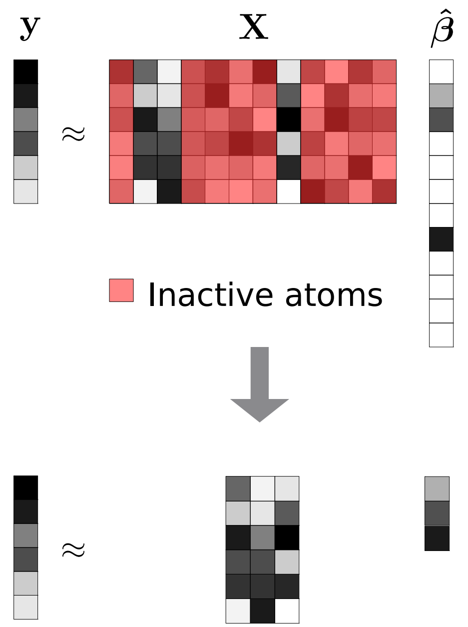
Dictionary columns that will receive zero weight on the reconstruction of the input signal



Safe Screening

- Rules for identifying inactive dictionary atoms, before solving the problem.

Dictionary columns that will receive zero weight on the reconstruction of the input signal



- We can eliminate such atoms.

- Zero risk of false eliminations!

Screening Test

- Function $\mu(\mathbf{x}_j)$ of the atom $\mathbf{x}_j$

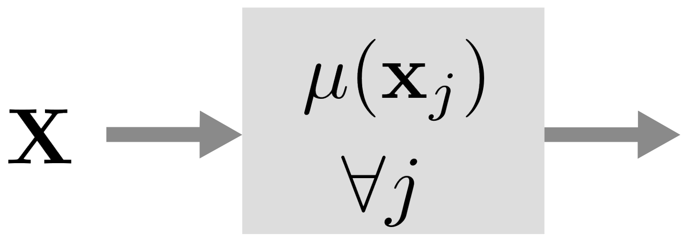
$$\mu(\mathbf{x}_j) < 1 \quad \Longrightarrow \quad \mathbf{x}_j \text{ is surely inactive.}$$

inventors for the digital world

Screening Test

- Function $\mu(\mathbf{x}_j)$ of the atom $\mathbf{x}_j$

$$\mu(\mathbf{x}_j) < 1 \quad \Longrightarrow \quad \mathbf{x}_j \text{ is surely inactive.}$$



Rejection set:

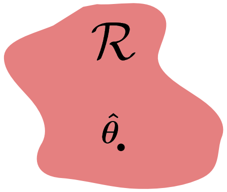
$$\mathcal{A}^c = \{j \in \{1, \ldots, K\} : \mu(\mathbf{x}_j) < 1\}$$

Preserved set:

$$\mathcal{A} = \{j \in \{1, \ldots, K\} : \mu(\mathbf{x}_j) \geq 1\}$$

Screening Test – In practice

Given a region $\mathcal{R}$ (***safe region***) which contains $\hat{\boldsymbol{\theta}}$.

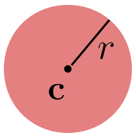


$$\mu_{\mathcal{R}}(\mathbf{x}_j) = \max_{\boldsymbol{\theta} \in \mathcal{R}} |\mathbf{x}_j^T \boldsymbol{\theta}|$$

Sphere test

Safe region is a closed l2-ball with center ***c*** and radius *r* : $\mathcal{R} = B(\mathbf{c}, r)$

$$\mu_{B(\mathbf{c},r)}(\mathbf{x}_j) = |\mathbf{x}_j^T \mathbf{c}| + r\|\mathbf{x}_j\|_2.$$



04

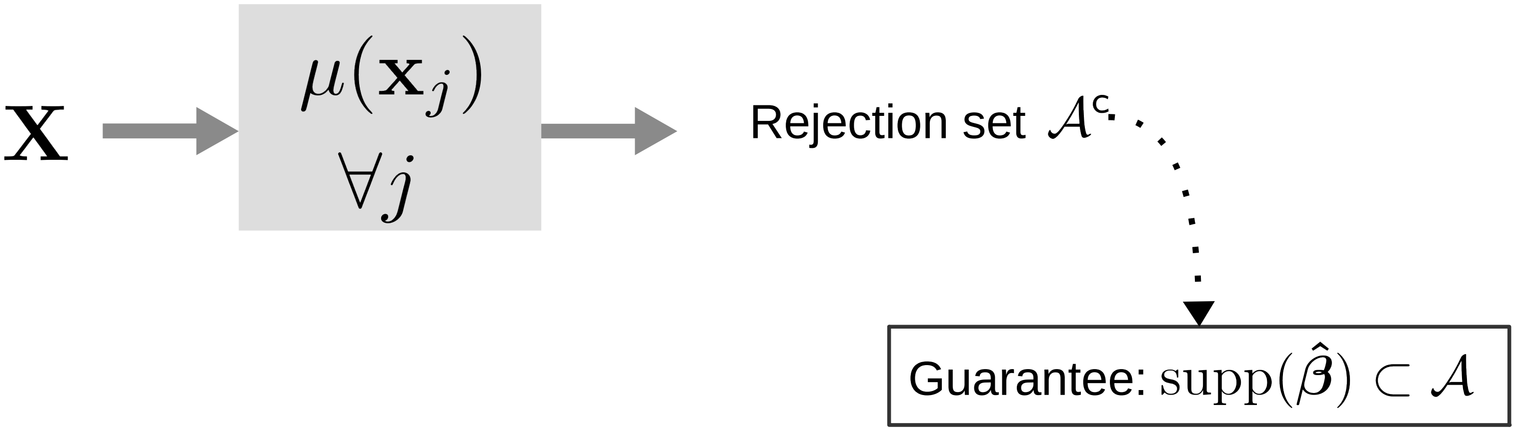
Screening Rules with Approximate Dictionaries

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

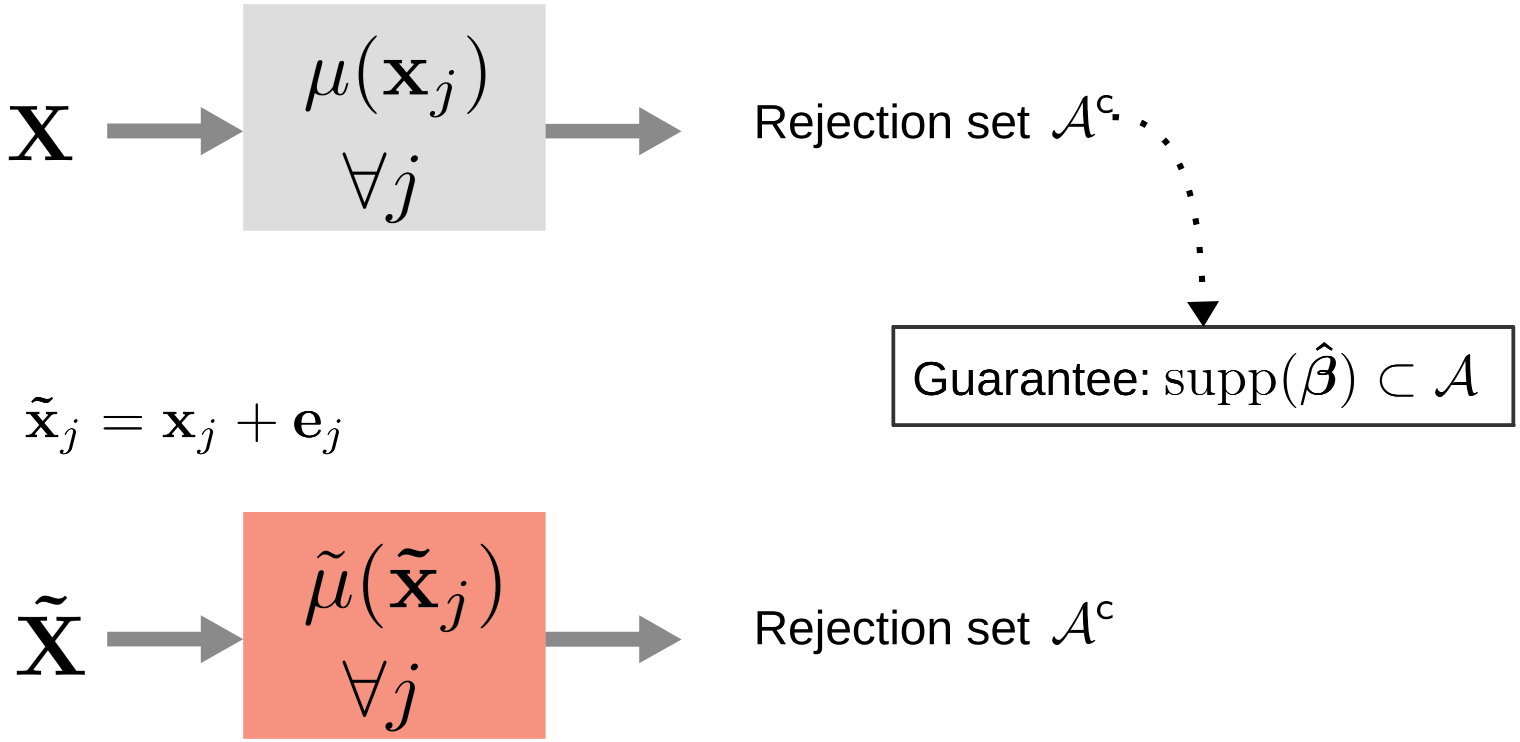
$\mathbf{X}$ → $\mu(\mathbf{x}_j)$ $\forall j$ → Rejection set $\mathcal{A}^c$

Guarantee: $\operatorname{supp}(\hat{\boldsymbol{\beta}}) \subset \mathcal{A}$

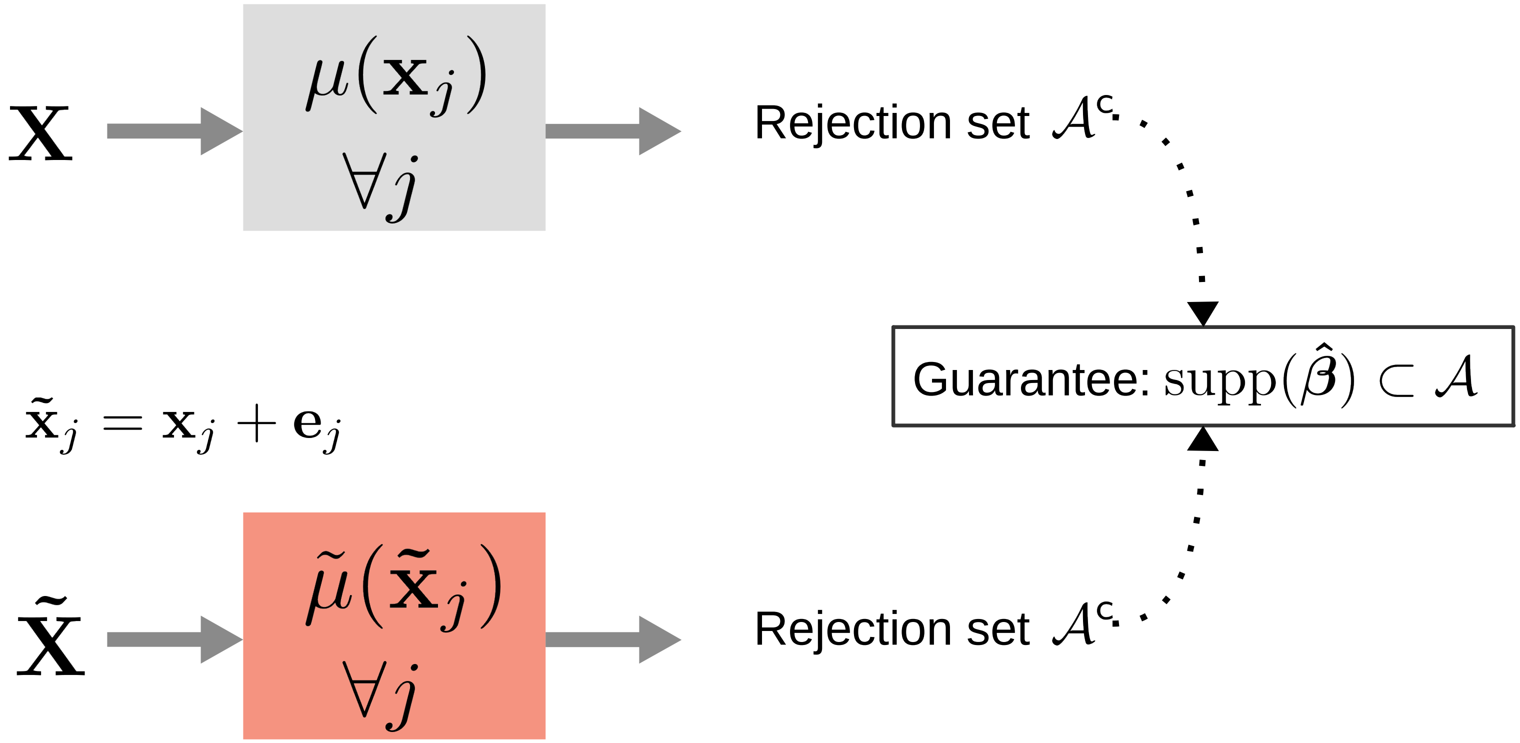
$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$



$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

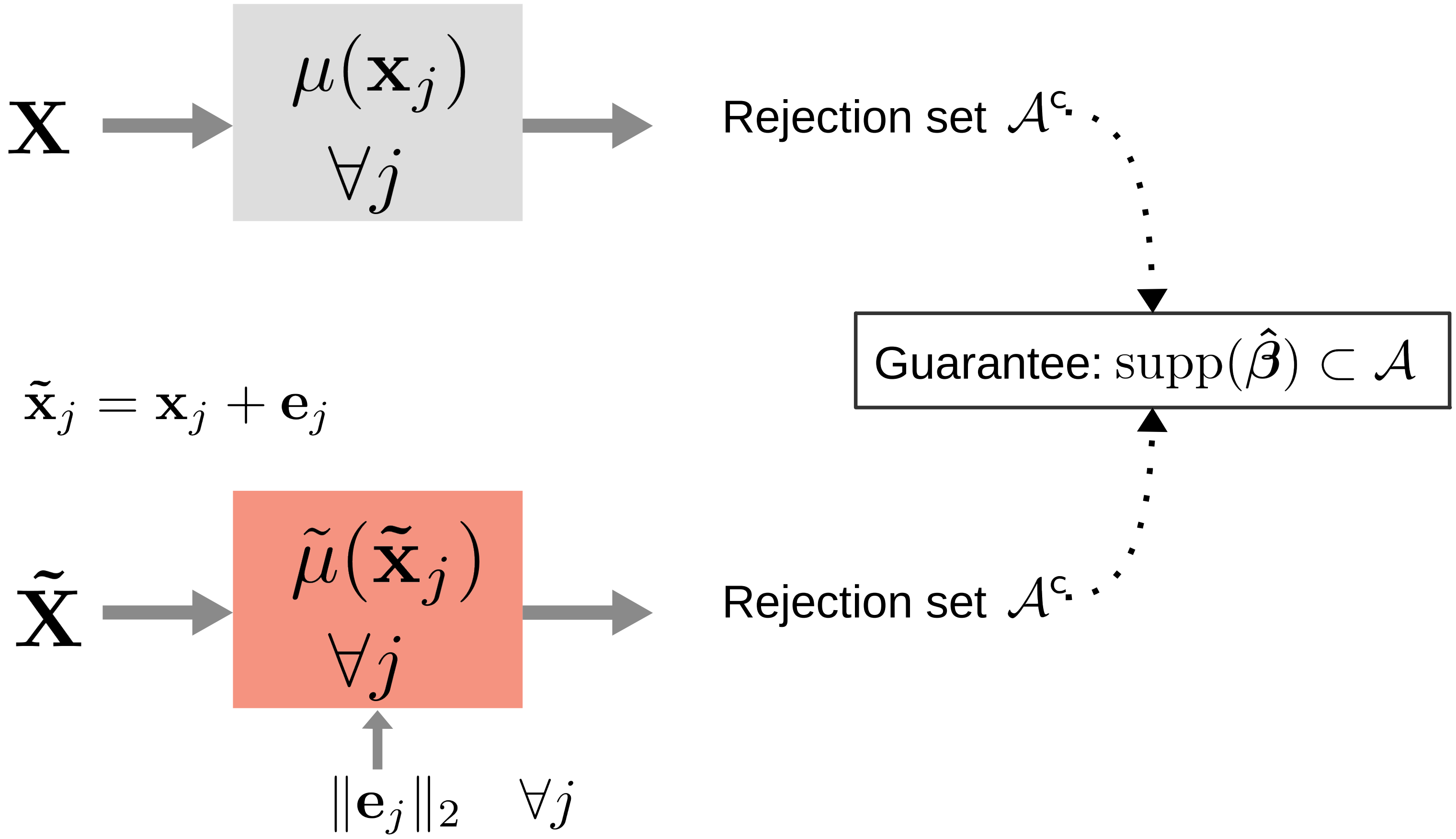


$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$



inventors for the digital world

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$



Extending sphere tests

Suppose a safe sphere $B(\mathbf{c}, r)$ given.

Sphere test : $$\mu_{B(\mathbf{c},r)}(\mathbf{x}_j) = |\mathbf{x}_j^T \mathbf{c}| + r\|\mathbf{x}_j\|_2.$$

A certain ***« security margin »*** must be added to account for the atom approximation error.

Extending sphere tests

Suppose a safe sphere $B(\mathbf{c}, r)$ given.

Sphere test : $$\mu_{B(\mathbf{c},r)}(\mathbf{x}_j) = |\mathbf{x}_j^T \mathbf{c}| + r\|\mathbf{x}_j\|_2.$$

A certain ***« security margin »*** must be added to account for the atom approximation error.

Sphere test with approximate dictionary :

$$\tilde{\mu}_{B(\mathbf{c},r)}(\tilde{\mathbf{x}}_j) = |\tilde{\mathbf{x}}_j^T \mathbf{c}| + \|\mathbf{e}_j\|_2\|\mathbf{c}\|_2 + r\|\mathbf{x}_j\|_2.$$

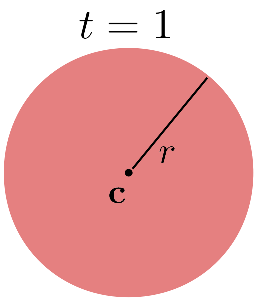
Obtaining a safe sphere

GAP safe sphere :

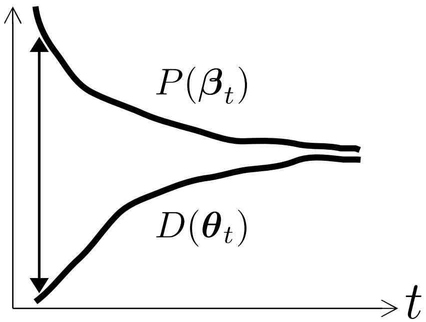
Given a primal-dual estimation $(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)$ at iteration t.

$$\mathbf{c} = \boldsymbol{\theta}_t$$

$$r = \frac{1}{\lambda}\sqrt{2G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)}$$



with $G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t) = P(\boldsymbol{\beta}_t) - D(\boldsymbol{\theta}_t)$ the duality gap at iteration t.



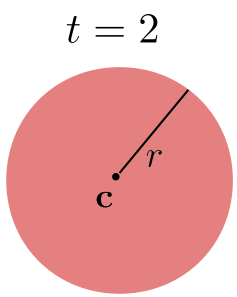
Obtaining a safe sphere

GAP safe sphere :

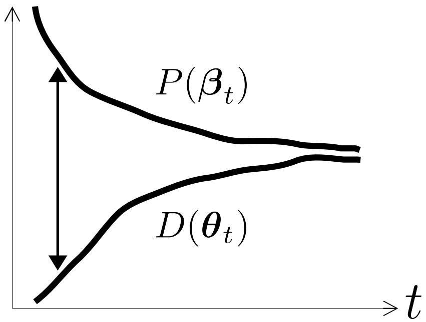
Given a primal-dual estimation $(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)$ at iteration t.

$$\mathbf{c} = \boldsymbol{\theta}_t$$

$$r = \frac{1}{\lambda}\sqrt{2G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)}$$



with $G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t) = P(\boldsymbol{\beta}_t) - D(\boldsymbol{\theta}_t)$ the duality gap at iteration t.



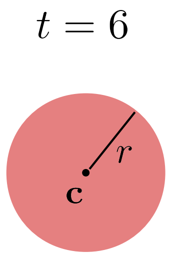
Obtaining a safe sphere

GAP safe sphere :

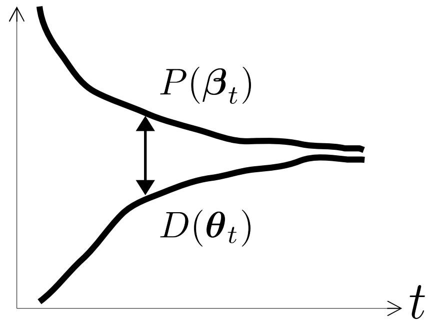
Given a primal-dual estimation $(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)$ at iteration t.

$$\mathbf{c} = \boldsymbol{\theta}_t$$

$$r = \frac{1}{\lambda}\sqrt{2G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)}$$



with $G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t) = P(\boldsymbol{\beta}_t) - D(\boldsymbol{\theta}_t)$ the duality gap at iteration t.



Obtaining a safe sphere

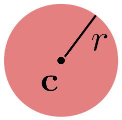
GAP safe sphere :

Given a primal-dual estimation $(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)$ at iteration t .

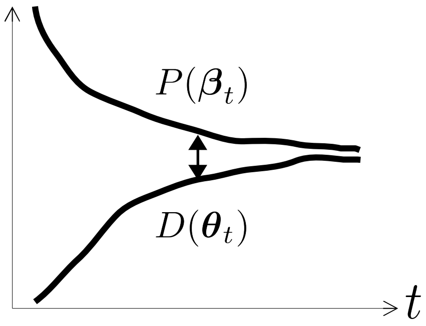
$$\mathbf{c} = \boldsymbol{\theta}_t$$

$$r = \frac{1}{\lambda}\sqrt{2G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)}$$

$t = 10$



with $G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t) = P(\boldsymbol{\beta}_t) - D(\boldsymbol{\theta}_t)$ the duality gap at iteration t .



Obtaining a safe sphere

GAP safe sphere:

Given a primal-dual estimation $(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)$ at iteration t.

$$\mathbf{c} = \boldsymbol{\theta}_t$$

$$r = \frac{1}{\lambda}\sqrt{2G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)}$$

GAP safe sphere with approximate dictionary:

$G(\boldsymbol{\beta}_t, \boldsymbol{\theta}_t)$ cannot be calculated, since $P(\boldsymbol{\beta}_t) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta}_t - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}_t\|_1$ depends on $\mathbf{X}$.

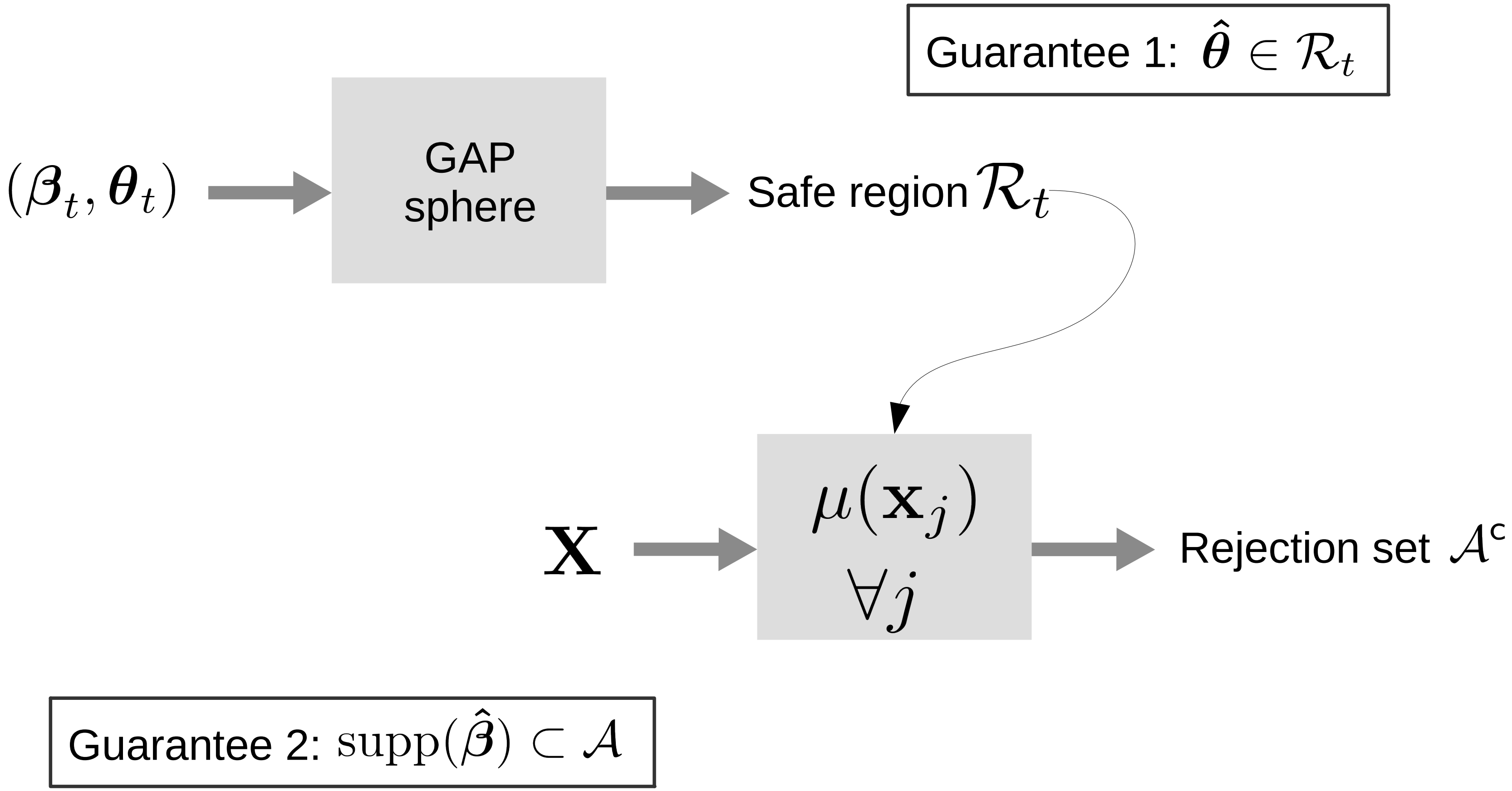
Instead, we use a modified primal $\tilde{P}(\boldsymbol{\beta}_t) = \frac{1}{2}\|\tilde{\mathbf{X}}\boldsymbol{\beta}_t - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}_t\|_1$

$$\tilde{\mathbf{c}} = \tilde{\boldsymbol{\theta}}_t$$

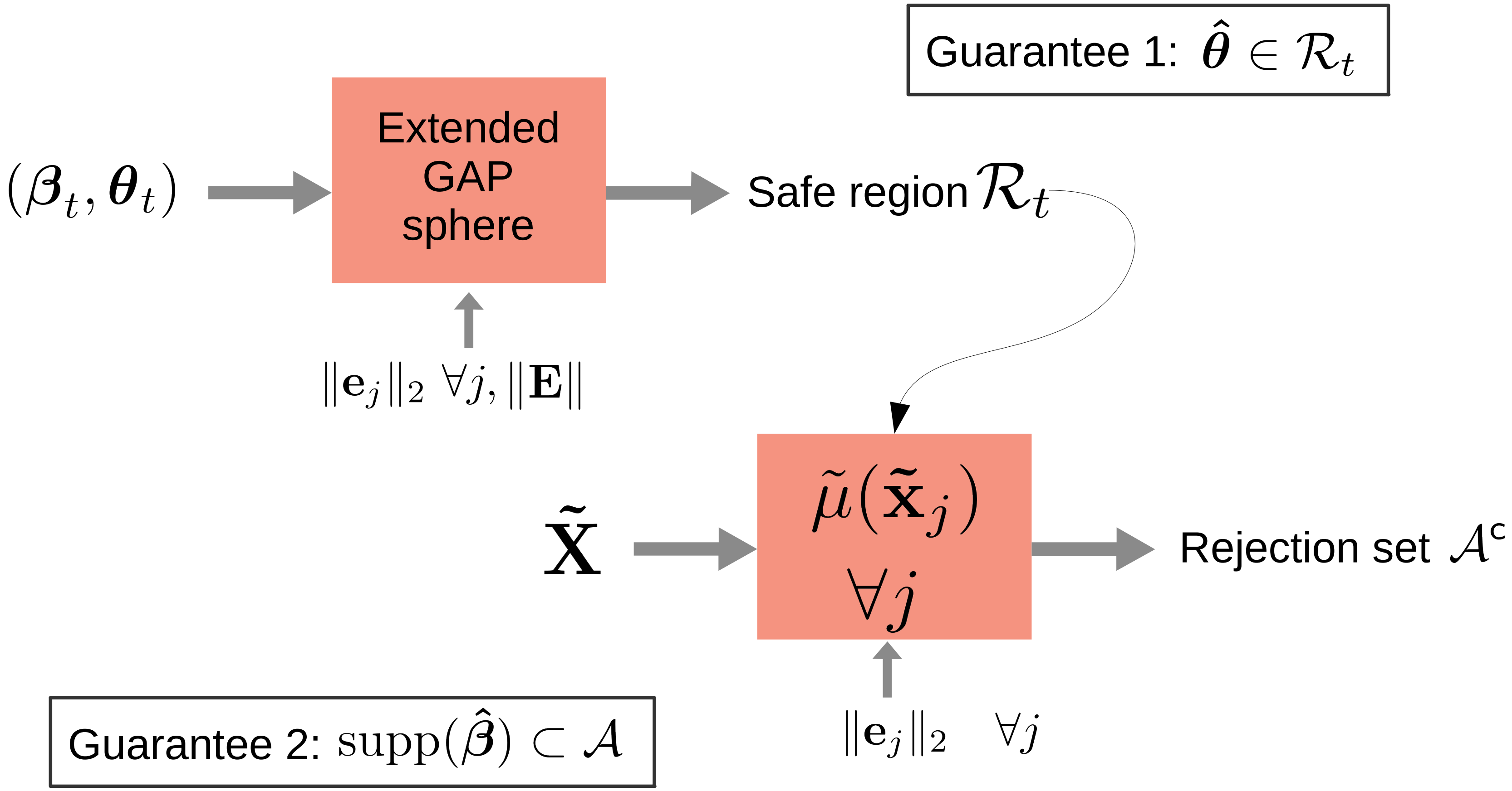
$$\tilde{r} = \frac{1}{\lambda}\sqrt{2\tilde{G}(\boldsymbol{\beta}_t, \tilde{\boldsymbol{\theta}}_t) + 2\delta'}$$

with $\delta' = \|\tilde{\boldsymbol{\rho}}_t\|_2\|\mathbf{E}\|\|\boldsymbol{\beta}_t\|_2 + \frac{1}{2}\|\mathbf{E}\|^2\|\boldsymbol{\beta}_t\|_2^2$

Dynamic screening



Extended dynamic screening

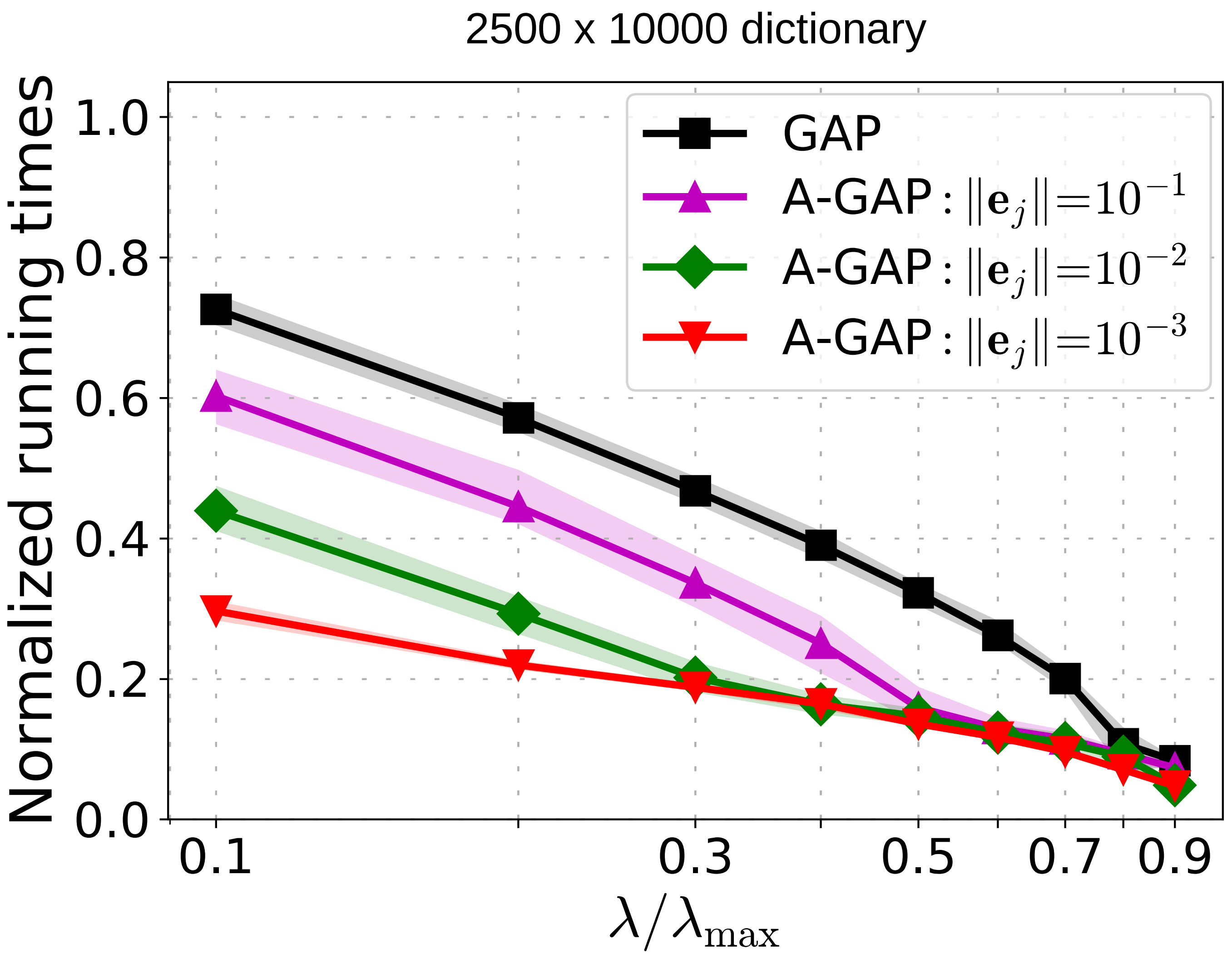


We now have safe screening rules that manipulate an approximate dictionary.

But, what’s the impact of the numerous security margins? Is it still worth it?

05

Results

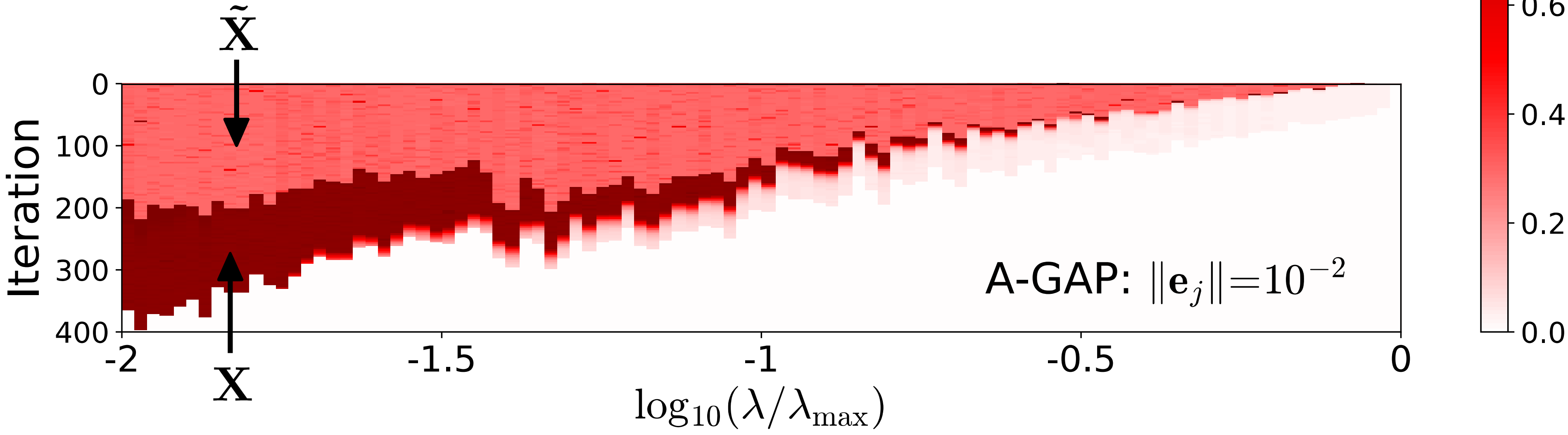

2500 x 10000 dictionary
Normalized running times
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.3
0.5
0.7
0.9
$\lambda/\lambda_{\max}$
GAP
A-GAP : $\|\mathbf{e}_j\| = 10^{-1}$
A-GAP : $\|\mathbf{e}_j\| = 10^{-2}$
A-GAP : $\|\mathbf{e}_j\| = 10^{-3}$

Running times per iteration

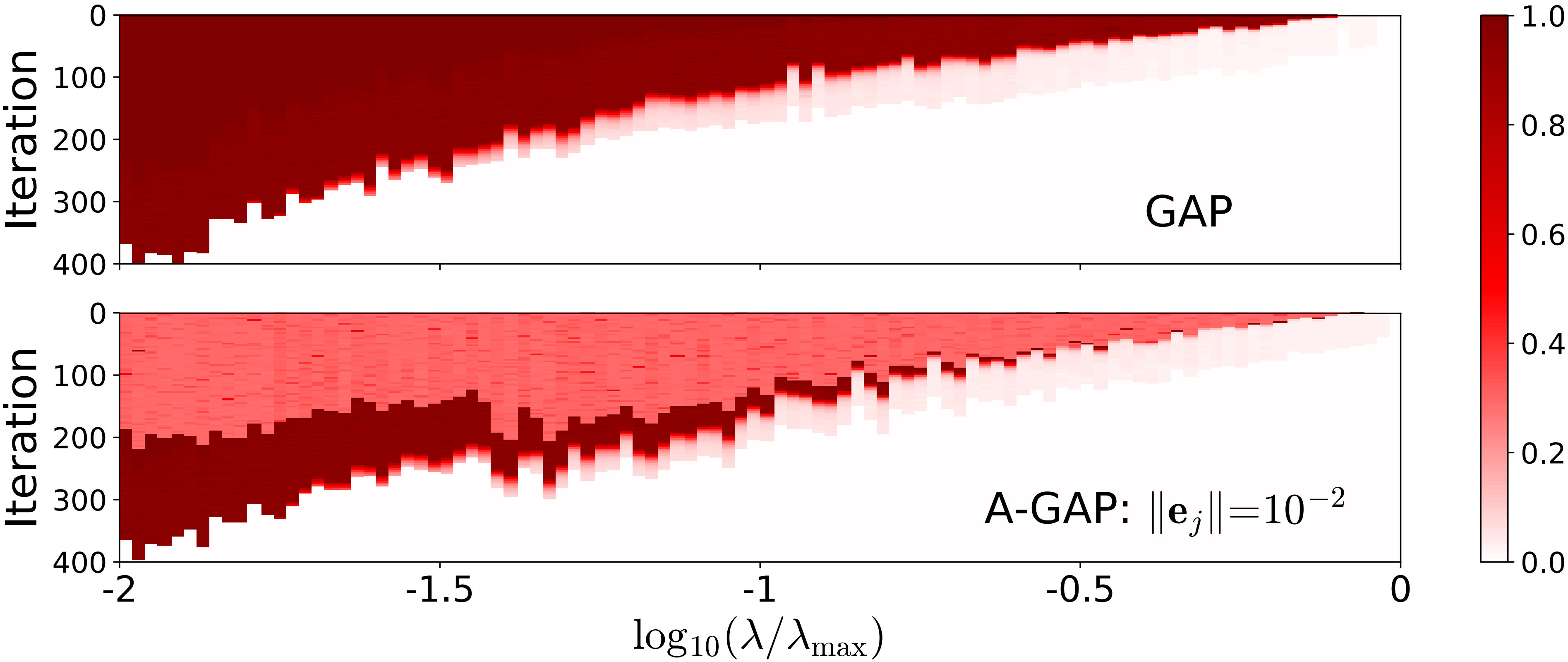
- **Less inactive atoms are identified by the extended screening.**
- **BUT, structured dictionary makes the initial iterations much faster.**

Running times per iteration

- **Less inactive atoms are identified by the extended screening.**
- **BUT, structured dictionary makes the initial iterations much faster.**



Running times per iteration



06

Conclusion

Conclusion

- The proposed approach combines screening rules and fast approximate dictionaries.
- It reduces even further the execution time w.r.t screening rules alone.

Potential extensions

- Other region types (e.g. domes)
- Other problems than Lasso (e.g. Group-Lasso, Regularized Logistic Regression)

Thank you!

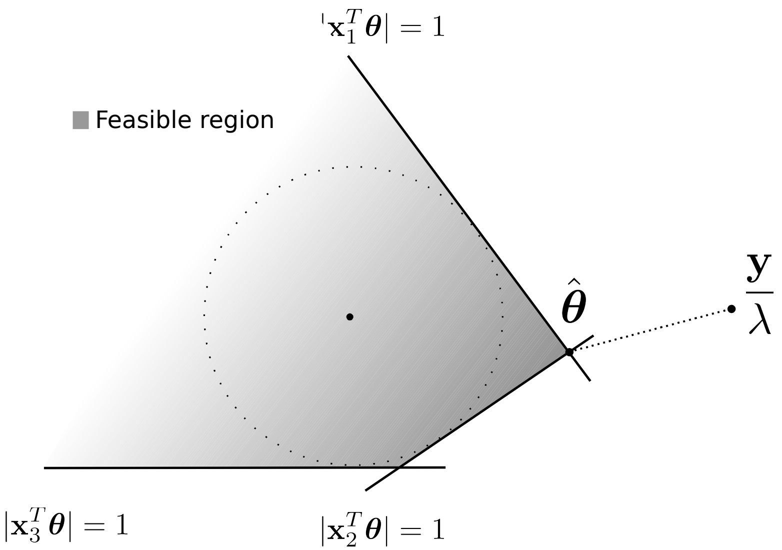
Questions?

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

Screening test – Details

Dual formulation of the Lasso problem :

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \quad \frac{1}{2}\|\mathbf{y}\|_2^2 - \frac{\lambda^2}{2}\left\|\boldsymbol{\theta} - \frac{\mathbf{y}}{\lambda}\right\|_2^2$$
$$\text{s.t.} \qquad \|\mathbf{X}^T\boldsymbol{\theta}\|_\infty \leq 1.$$



Projection problem !

At the dual solution $\hat{\boldsymbol{\theta}}$:

- Constraints on $\mathbf{x}_1$ and $\mathbf{x}_2$ are active, i.e.

$$|\mathbf{x}_1^T\hat{\boldsymbol{\theta}}| = 1, \quad |\mathbf{x}_2^T\hat{\boldsymbol{\theta}}| = 1$$

- Constraints on $\mathbf{x}_3$ is inactive, i.e.

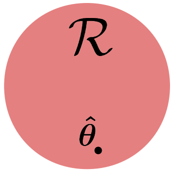
$$|\mathbf{x}_3^T\hat{\boldsymbol{\theta}}| < 1$$

Screening test – Details

- Every dictionary atom for which $|\mathbf{x}_j^T \hat{\boldsymbol{\theta}}| < 1$ is ***inactive***.
- Then, simply calculate $|\mathbf{x}_j^T \hat{\boldsymbol{\theta}}|$ for all j and discard all atoms for which the result is smaller than 1.

Dual solution $\hat{\boldsymbol{\theta}}$ is not known.

Identify a region $\mathcal{R}$ (***safe region***) which contains $\hat{\boldsymbol{\theta}}$.



Sufficient condition : $\forall \quad \boldsymbol{\theta} \in \mathcal{R}, \quad |\mathbf{x}_j^T \boldsymbol{\theta}| < 1 \quad \Longrightarrow \quad \mathbf{x}_j$ is inactive

Swithing criterion

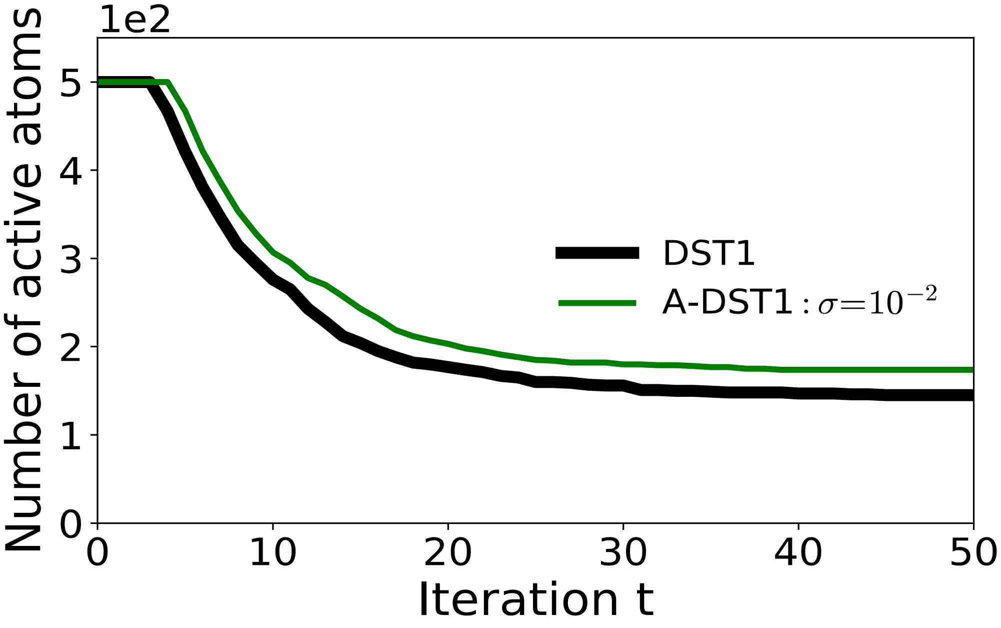
Reasons to switch back from $\tilde{\mathbf{X}}$ to $\mathbf{X}$:

- **Convergence**: to avoid converging to the solution of the approximate problem.
  - The higher the approximation error, the sooner we need to switch.

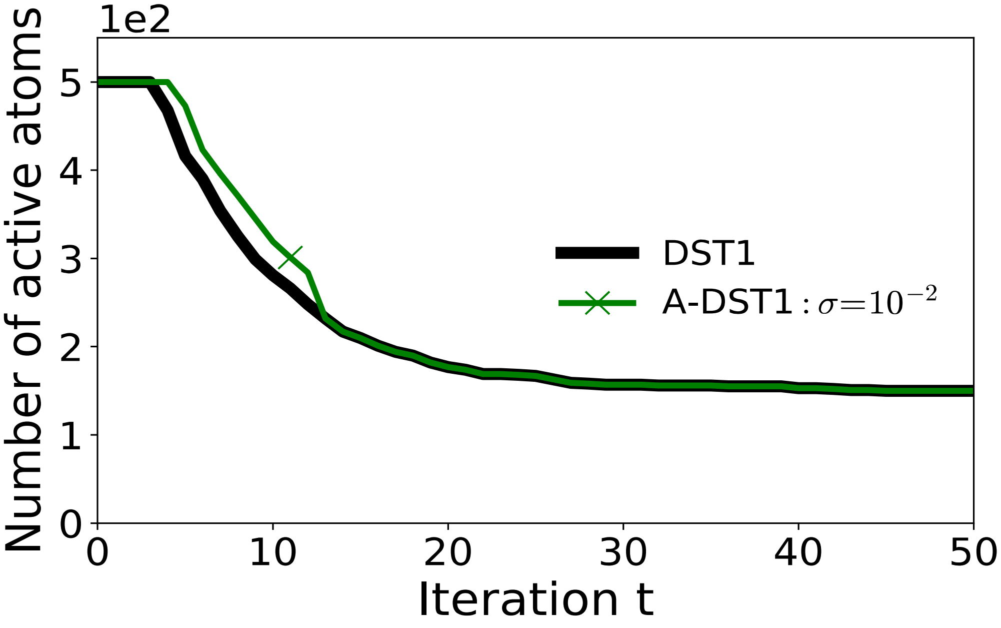
- **Screening ratio**: the number of active atoms may become so small that the use of $\tilde{\mathbf{X}}$ does not pay off anymore.

Comparison

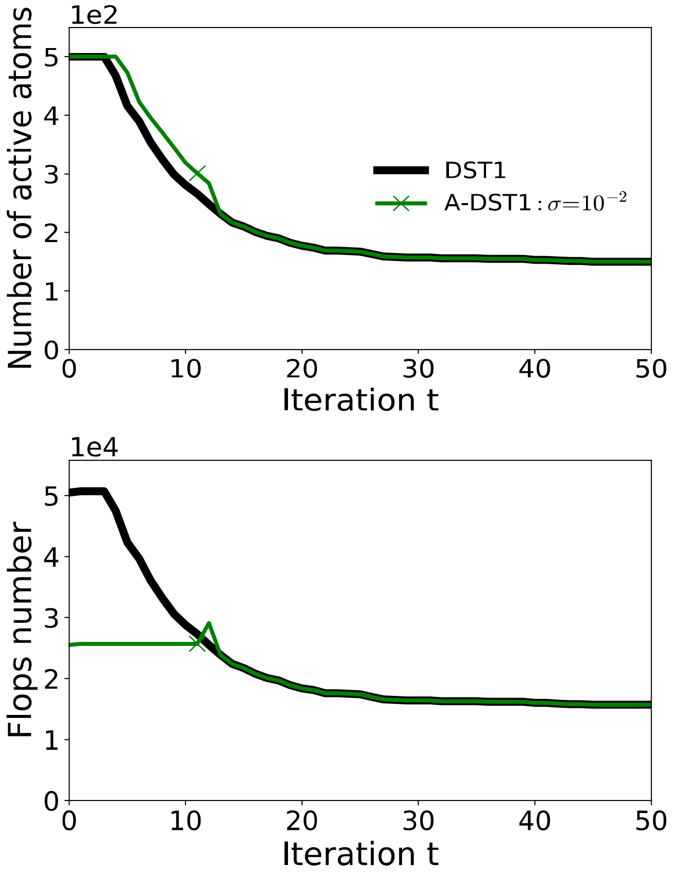
Less inactive atoms are identified by the extended screening.



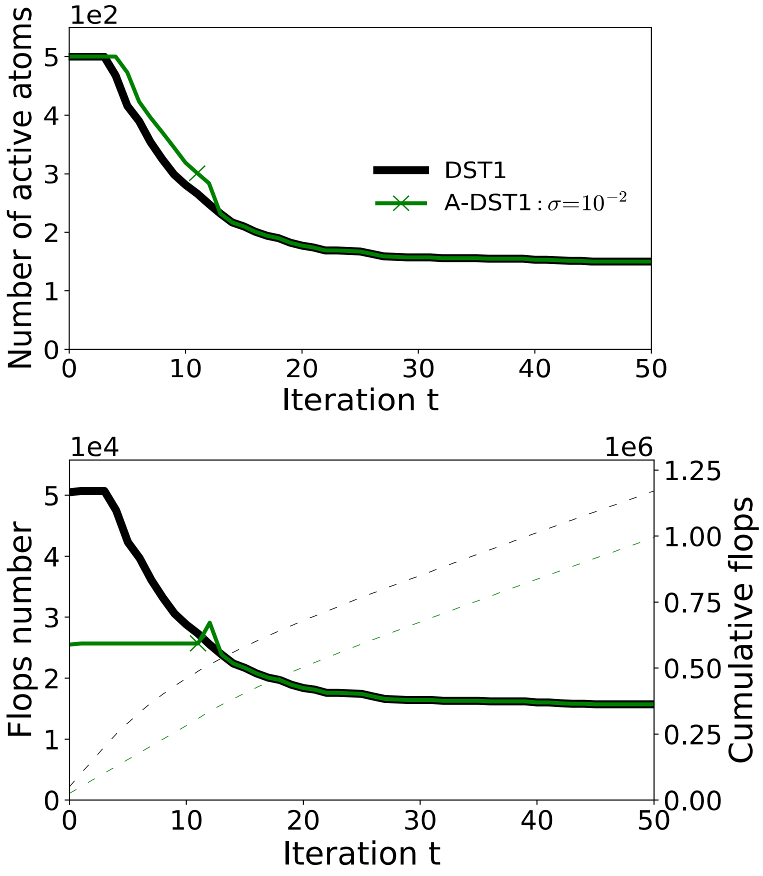
Swithing criterion



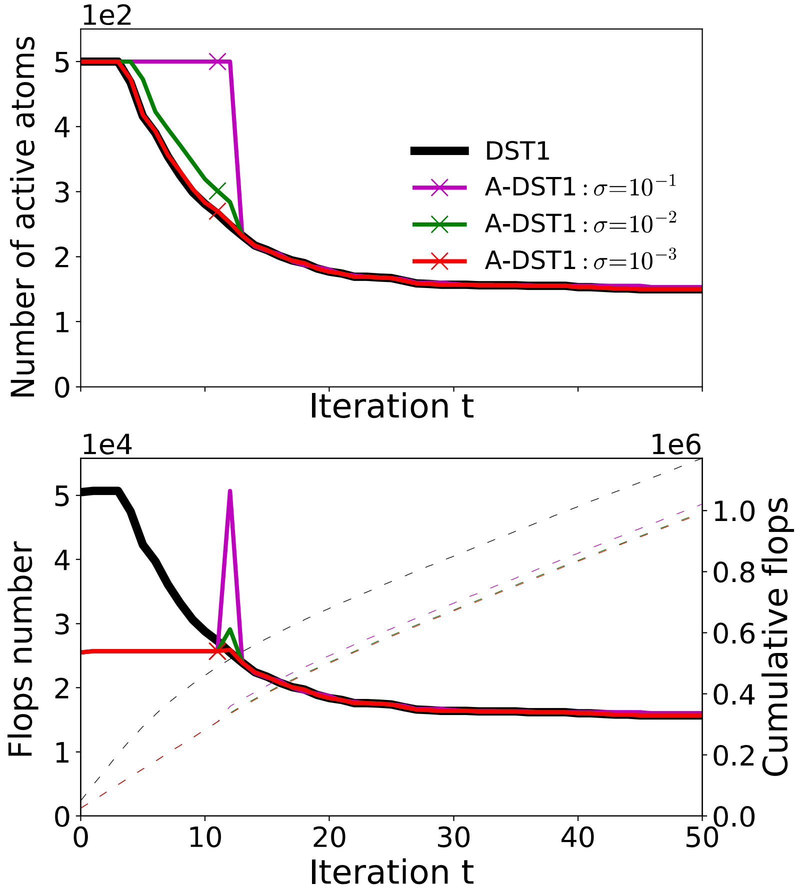
Complexity reduction



Complexity reduction



Impact of the Approximation Error



Inria
inventors for the digital world