Screen & Relax : Accelerating the resolution of the Elastic-Net

Paper ID : 2487

Théo Guyard⋆, Cédric Herzet†, Clément Elvira‡

⋆Applied Mathematics Department, INSA Rennes, France | †SIMSMART team, INRIA Rennes-Bretagne Atlantique, France | ‡SCEE team, CentraleSupelec Rennes, France

Objectives

Accelerate the resolution of the Elastic-Net :

- Identification of *zeros* in the optimizer
- Identification of *non-zeros* in the optimizer
- Reduction of the problem dimension
- Reduction of the complexity burden

Introduction

Sparse decomposition aims at finding some approximation of a vector $\mathbf{y} \in \mathbf{R}^m$ as the linear combination of a few columns (dubbed *atoms*) of a dictionary $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_n] \in \mathbf{R}^{m \times n}$. Unfortunately, identifying the sparsest decomposition turns out to be a NP-hard problem. A standard strategy to circumvent this issue consists in approximating this ideal decomposition as the solution of the *Elastic-Net* problem :

$$\min_{\mathbf{x} \in \mathbf{R}^n} P(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 + \tfrac{\gamma}{2}\|\mathbf{x}\|_2^2 \quad (\mathcal{P})$$

where $\lambda > 0$ and $\gamma > 0$. The least-squares term ensures a good approximation of $\mathbf{y}$ by $\mathbf{A}\mathbf{x}$, the ℓ_1-norm enforces sparsity and the ℓ_2-norm provides desirable statistical properties. In the sequel, we note $\mathbf{x}^\star$ the unique minimizer of $(\mathcal{P})$.

Because of its clear practical interest, many contributions of the literature have proposed efficient solving procedures for $(\mathcal{P})$. Of particular interest in this paper is the "*screening*" acceleration technique proposed by El Ghaoui *et al.*. It consists in performing simple tests to identify the *zeros* in $\mathbf{x}^\star$. This knowledge can then be exploited to reduce the dimensionality of the problem by discarding the atoms of the dictionary weighted by zero.

We introduce a dual approach to screening, dubbed "*relaxing*". Our method aims at identifying the position of the *non-zeros* in $\mathbf{x}^\star$. We show that, similarly to screening, this knowledge can be exploited to reduce the dimensionality of the target problem and accelerate its resolution.

Dual problem

Our methodology leverages properties of the *Fenchel dual* problem of $(\mathcal{P})$, given by

$$\max_{\mathbf{u} \in \mathbf{R}^m} D(\mathbf{u}) = \tfrac{1}{2}\|\mathbf{y}\|_2^2 - \tfrac{1}{2}\|\mathbf{y} - \mathbf{u}\|_2^2 - \tfrac{1}{2\gamma}\|[\mathbf{A}^{\mathrm{T}}\mathbf{u} - \lambda]_+\|_2^2 \quad (\mathcal{D})$$

where $[x]_+ \triangleq \max(0, x)$. We note $\mathbf{u}^\star$ the unique maximizer of $(\mathcal{D})$. We also use the following relations

$$\mathbf{u}^\star = \mathbf{y} - \mathbf{A}\mathbf{x}^\star \quad (2)$$

$$\mathbf{x}^\star = \gamma^{-1}[|\mathbf{A}^{\mathrm{T}}\mathbf{u}^\star| - \lambda]_+ \quad (3)$$

deriving from optimality conditions. Combining these relations leads to the following tests.

Screening tests

Goal : Identification of *zeros* in $\mathbf{x}^\star$.

Let $\mathcal{S}(\mathbf{u}, r)$ be a sphere containing $\mathbf{u}^\star$, then

$$\forall i, \quad |\mathbf{A}_i^{\mathrm{T}}\mathbf{u}| + r < \lambda \implies \mathbf{x}_i^\star = 0 \quad (4)$$

Elements that have passed the above test can be discarded safely from $(\mathcal{P})$ as well as the corresponding columns in $\mathbf{A}$.

Relaxing tests

Goal : Identification of *non-zeros* in $\mathbf{x}^\star$.

Let $\mathcal{S}(\mathbf{u}, r)$ be a sphere containing $\mathbf{u}^\star$, then

$$\forall i, \quad |\mathbf{A}_i^{\mathrm{T}}\mathbf{u}| - r > \lambda \implies \mathbf{x}_i^\star \neq 0 \quad (5)$$

Elements that have passed the above test can be expressed as a linear combination of all the other elements of $\mathbf{x}$ in $(\mathcal{P})$.

Screen & Relax strategy

Screening and relaxing tests can be combined and used *within* any iterative method tailored to $(\mathcal{P})$ to reduce its computational burden. At each iteration, one can construct a *safe sphere* $\mathcal{S}(\mathbf{u}, r)$ that is tightened as iterates converge toward the solution. For every element that have passed the test (4), the corresponding index of the problem variable can be discarded. Similarly, for every element that have passed the test (5), the corresponding index of the problem variable can be expressed in function of all the remaining indices. Both of these modifications result in a *problem with a reduced optimization domain* where the problem data are slightly modified. This modification can be handled in an efficient way using rank-one update rules.

Ultimately, one can reach a point where all the elements have been either screened or relaxed. In this case, the relation (3) allows to recover $\mathbf{x}^\star$ in *closed-form* and up to machine-precision.

Pseudo-code

The following algorithm summarizes our method. Any iterative method tailored to solve $(\mathcal{P})$ can be enhanced with screening and relaxing tests, as soon as it allows to recover a sequence of primal iterates. The dimension of the problem solved is progressively reduced down to zero and ultimately, $\mathbf{x}^\star$ can be recovered exactly.

Algorithm 1: Iterative method for $(\mathcal{P})$ enhanced with a "Screen & Relax" strategy.

Input: Problem data $(\mathbf{A}, \mathbf{y}, \lambda, \gamma)$

while *convergence is not met* **do**
  Update the current iterate $\mathbf{x}^{(t)}$
  Construct a new safe sphere $\mathcal{S}(\mathbf{u}^{(t)}, r^{(t)})$
  Perform the tests (4)-(5) with $\mathcal{S}(\mathbf{u}^{(t)}, r^{(t)})$
  Update the problem data using rank-one rules if new elements have been screening or relaxed
end

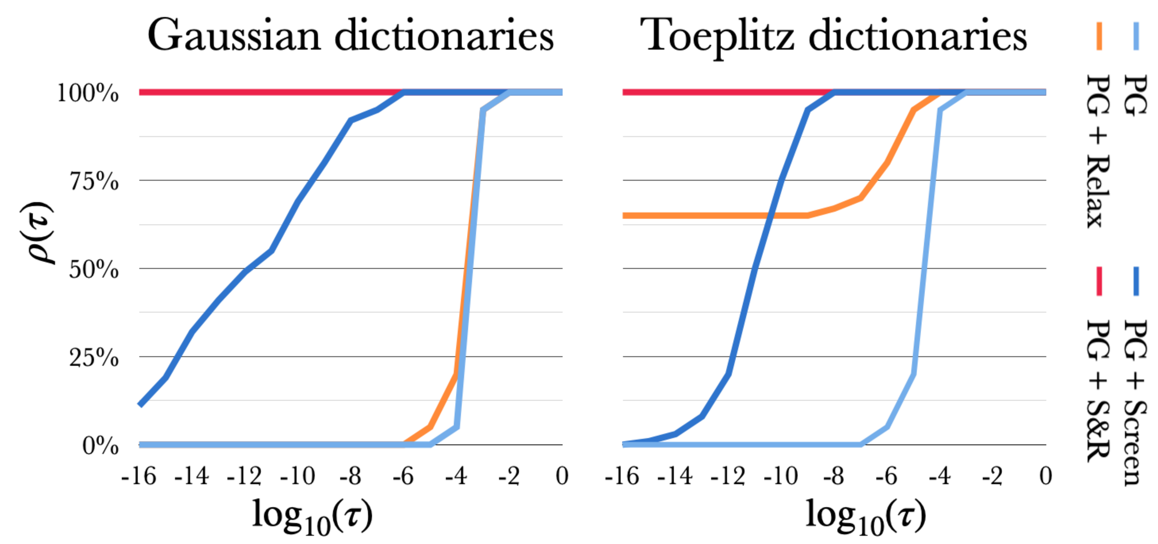
Numerical results

To assess our methodology, we compare four different algorithmic solutions to address $(\mathcal{P})$:

- A Proximal-Gradient (PG) algorithm
- A PG algorithm with additional *screening* tests
- A PG algorithm with additional *relaxing* tests
- A PG algorithm with additional *screening and relaxing* tests

We run a method with a fixed computational budget on 100 different instances of $(\mathcal{P})$. The curve corresponds to the percentage $\rho(\tau)$ of instances for which a duality gap lower than τ is archived.

To generate problem data, we consider the two following setups : 1) The elements of $\mathbf{A}$ are i.i.d. realizations of Gaussian distribution. 2) $\mathbf{A}$ has a Toeplitz structure with shifted versions of a sinc curve. The vector $\mathbf{y}$ is drawn according to an uniform distribution on the m-dimensional sphere. We set $(m, n) = (100, 300)$ and use $(\lambda, \gamma) = (0.5, 0.2) \times \|\mathbf{A}^{\mathrm{T}}\mathbf{y}\|_\infty$. Each problem instance is solved with a budget of 2×10^6 FLOPs for Gaussian dictionaries and 2×10^7 FLOPs for Toeplitz dictionaries.



We can observe that *screening tests* are particularly interesting with *low-correlated dictionaries* such as Gaussian ones. In contrast, *relaxing tests* are particularly interesting with *highly-correlated dictionaries* such as Toeplitz ones. Finally, note that the Screen & Relax methodology archives *machine-precision* ($\tau = 10^{-16}$) in all the instances solved. Indeed, $\mathbf{x}^\star$ is available in closed-form when all the elements of $\mathbf{x}$ have been either screened or relaxed.