



HAL
open science

On the properties of the solution path of the constrained and penalized L2-L0 problems

Junbo Duan, Charles Soussen, David Brie, Jérôme Idier

► To cite this version:

Junbo Duan, Charles Soussen, David Brie, Jérôme Idier. On the properties of the solution path of the constrained and penalized L2-L0 problems. 2009. hal-00361380v2

HAL Id: hal-00361380

<https://hal.science/hal-00361380v2>

Preprint submitted on 23 Feb 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the properties of the solution path of the constrained and penalized L2-L0 problems

Junbo Duan*, Charles Soussen*, David Brie*, Jérôme Idier†

*Centre de Recherche en Automatique de Nancy

†Institut de Recherche en Communication et Cybernétique de Nantes

February 23, 2009

1 Domain of optimization

For $k \leq n$, we define the domain $\mathcal{D}_k \subset \mathbb{R}^n$:

$$\mathcal{D}_k = \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_0 = k\}. \quad (1)$$

Theorem 1 For $k \geq 1$, \mathcal{D}_k is not a closed set, and $\overline{\mathcal{D}_k} = \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_0 \leq k\}$ (denoting by $\overline{}$ the closure operator).

Proof 1 • \mathcal{D}_k is not a closed set: it is easy to find a sequence $\mathbf{x}_j \in \mathcal{D}_k$ ($j \in \mathbb{N}$) whose limit is not in \mathcal{D}_k . For instance, $\mathbf{x}_j = (1/j)\mathbf{e}$, where \mathbf{e} is a given vector in \mathcal{D}_k . \mathbf{x}_j tends towards $\mathbf{0} \notin \mathcal{D}_k$.

- $\overline{\mathcal{D}_k} \subseteq \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_0 \leq k\}$. If $\mathbf{x} \in \overline{\mathcal{D}_k}$, then there exists a sequence $\mathbf{x}_j \in \mathcal{D}_k$ ($j \in \mathbb{N}$) whose limit is equal to \mathbf{x} . Then,

$$\forall \varepsilon > 0, \exists J, j \geq J \Rightarrow \forall i, |\mathbf{x}(i) - \mathbf{x}_j(i)| < \varepsilon.$$

Applying this property with $\varepsilon = \min_{\mathbf{x}(i) \neq 0} |\mathbf{x}(i)|$, we deduce that there exists an iteration J , such that $\forall j \geq J, \forall i, \mathbf{x}(i) \neq 0 \Rightarrow \mathbf{x}_j(i) \neq 0$. In other words, $\|\mathbf{x}\|_0 \leq \|\mathbf{x}_j\|_0 = k$.

- $\{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_0 \leq k\} \subseteq \overline{\mathcal{D}_k}$. Let us show that if \mathbf{x} is such that $\|\mathbf{x}\|_0 \leq k$, then there exists a sequence $\mathbf{x}_j \in \mathcal{D}_k$ whose limit is equal to \mathbf{x} . Given \mathbf{x} , we define \mathbf{x}_j by setting $\mathbf{x}_j(i) = \mathbf{x}(i)$ if $i \in \mathcal{A}(\mathbf{x})$ (support of \mathbf{x}), and by replacing the $k - \|\mathbf{x}\|_0$ first zero valued entries of \mathbf{x} by $1/j$ in \mathbf{x}_j , and setting to 0 the remaining $n - k$ entries $\mathbf{x}_j(i)$. Obviously, $\mathbf{x}_j \in \mathcal{D}_k$ and this sequence tends towards \mathbf{x} .

The consequence of theorem 1 is that

$$\arg \min_{\mathbf{x} \in \mathcal{D}_k} \{\mathcal{E}(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2\}$$

is not always defined, although the minimal value $\min_{\mathbf{x} \in \mathcal{D}_k} \mathcal{E}(\mathbf{x})$ is defined. On the contrary, the set of minimizers

$$\mathcal{X}_c(k) = \arg \min_{\mathbf{x} \in \overline{\mathcal{D}_k}} \mathcal{E}(\mathbf{x}) = \arg \min_{\|\mathbf{x}\|_0 \leq k} \mathcal{E}(\mathbf{x})$$

is properly defined because $\overline{\mathcal{D}_k}$ is a closed set and \mathcal{E} is quadratic and convex (to be completed).

Example 1 Let us consider the minimization of $\|\mathbf{x}\|^2$ over the domain \mathcal{D}_k . For $k \geq 1$, there is no minimizer over \mathcal{D}_k , but the minimal cost $\min_{\mathbf{x} \in \mathcal{D}_k} \|\mathbf{x}\|^2$ is equal to 0. The set of minimizers over $\overline{\mathcal{D}_k}$ is reduced to one vector: $\mathcal{X}_c(k) = \{\mathbf{0}\}$.

Example 2 The set $\mathcal{X}_c(k)$ is not always a singleton. Let us consider the minimization of the 2D cost function $\mathcal{E}(\mathbf{x}) = \mathbf{x}(1)^2$. It is easy to see that $\mathcal{X}_c(0) = \{\mathbf{0}\}$, $\mathcal{X}_c(1) = \{[0, \mathbf{x}(2)]^T, \mathbf{x}(2) \in \mathbb{R}\}$ and $\mathcal{X}_c(2) = \mathcal{X}_c(1)$.

Example 3 Let us consider the minimization of the 2D cost function $\mathcal{E}(\mathbf{x}) = (\mathbf{x}(1) - \alpha)^2$ for a given $\alpha \neq 0$. It is easy to see that $\mathcal{X}_c(0) = \{\mathbf{0}\}$, $\mathcal{X}_c(1) = \{[\alpha, 0]^T\}$ and $\mathcal{X}_c(2) = \{[\alpha, \mathbf{x}(2)]^T, \mathbf{x}(2) \in \mathbb{R}\}$.

Remark 1 Obviously, the sets $\overline{\mathcal{D}_k}$ have a nesting property ($\overline{\mathcal{D}_k} \subset \overline{\mathcal{D}_{k+1}}$), therefore, for all k , we have

$$\forall \mathbf{x}_k \in \mathcal{X}_c(k), \forall \mathbf{x}_{k+1} \in \mathcal{X}_c(k+1), \mathcal{E}(\mathbf{x}_{k+1}) \leq \mathcal{E}(\mathbf{x}_k).$$

Theorem 2 $\mathcal{X}_c(k+1) \cap \overline{\mathcal{D}_k} \subseteq \mathcal{X}_c(k)$.

Proof 2 Let us consider $\mathbf{x}_{k+1} \in \mathcal{X}_c(k+1) \cap \overline{\mathcal{D}_k}$. Since $\overline{\mathcal{D}_k} \subset \overline{\mathcal{D}_{k+1}}$ and \mathbf{x}_{k+1} is a minimizer of \mathcal{E} over $\overline{\mathcal{D}_{k+1}}$, we have $\forall \mathbf{x} \in \overline{\mathcal{D}_k}$, $\mathcal{E}(\mathbf{x}_{k+1}) \leq \mathcal{E}(\mathbf{x})$. As $\mathbf{x}_{k+1} \in \overline{\mathcal{D}_k}$, \mathbf{x}_{k+1} is a minimizer of \mathcal{E} over $\overline{\mathcal{D}_k}$.

2 Working assumptions and notion of constrained solution path

2.1 Unique representation property

We recall the definition of the unique representation property (URP), introduced in [1] in the underdetermined case (when $m \leq n$):

Definition 1 A matrix \mathbf{A} of size $m \times n$ ($m \leq n$) satisfies the URP if and only if any selection of m columns of \mathbf{A} forms a family of linearly independent vectors.

Under the URP assumption, we can solve $\mathbf{y} = \mathbf{A}\mathbf{x}$ by imposing that $\mathbf{x} \in \overline{\mathcal{D}_m}$. The system is then equivalent to $\mathbf{y} = \mathbf{B}\mathbf{z}$ where \mathbf{B} is a matrix of size $m \times m$ extracted from \mathbf{A} , and \mathbf{z} is the corresponding vector extracted from \mathbf{x} , of size $m \times 1$. According to the URP definition, \mathbf{B} is always invertible, and we can find sparse solutions to $\mathbf{y} = \mathbf{A}\mathbf{x}$ with at most m non-zero entries ($\mathbf{z} = \mathbf{B}^{-1}\mathbf{y}$ and then $\mathbf{x} = \{\mathbf{z}, \mathbf{0}\}$ for all the possible extractions \mathbf{B} from \mathbf{A}).

When $m > n$, we adopt the following definition:

Definition 2 A matrix \mathbf{A} of size $m \times n$ ($m > n$) satisfies the URP if and only if it is full rank.

When $m > n$, there is generally no solution to $\mathbf{y} = \mathbf{A}\mathbf{x}$ but the minimizer of $\mathcal{E}(\mathbf{x})$ over \mathbb{R}^n is unique (although not necessarily sparse): $\mathcal{X}_c(n) = \{(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y}\}$.

In the following, we will assume that $\mathbf{y} \neq \mathbf{0}$ and that \mathbf{A} satisfies the URP.

2.2 Cardinality of the set $\mathcal{X}_c(k)$

Theorem 3 For $k \leq \min(m, n)$, the set $\mathcal{X}_c(k)$ is finite under the URP assumption.

Proof 3 Because of the URP assumption, any selection of $k \leq \min(m, n)$ columns of \mathbf{A} yields a matrix \mathbf{B} of size $m \times k$ whose rank is equal to k . Then, the energy reduces to $\mathcal{E}(\mathbf{x}) = \mathcal{E}(\mathbf{z}; \mathbf{0}) = \|\mathbf{y} - \mathbf{B}\mathbf{z}\|^2$ (where $\mathbf{z} \in \mathbb{R}^k$), and there is only one minimizer of $\mathbf{z} \mapsto \mathcal{E}(\mathbf{z}; \mathbf{0})$ over \mathbb{R}^k . Since the number of possible selections of k columns of \mathbf{A} is finite, the set $\mathcal{X}_c(k)$ is finite.

Remark 2 The minimal value of $\mathcal{E}(\mathbf{x})$ (for $\mathbf{x} \in \mathbb{R}^n$) can be reached when minimizing \mathcal{E} over $\overline{\mathcal{D}_{\min(m, n)}}$. Thus, when $m \leq n$, it is not necessary to compute $\mathcal{X}_c(k)$ for $k > m$. According to theorem 3, when $m > n$, all the sets $\mathcal{X}_c(k)$, $k = 0, \dots, n$ are finite.

Theorem 4 When $m \leq n$ and k is such that $m < k \leq n$, the set $\mathcal{X}_c(k)$ is of infinite cardinality.

Proof 4 Given a solution $\mathbf{x}_m \in \mathcal{X}_c(m)$, let $\mathcal{A}(\mathbf{x}_m)$ be the support of \mathbf{x}_m . We consider a support \mathcal{B} of cardinality k such that $\mathcal{A}(\mathbf{x}_m) \subset \mathcal{B} \subseteq \{1, \dots, n\}$, and we extract from \mathbf{A} the matrix \mathbf{B} of size $m \times k$ formed of the columns \mathbf{a}_i of \mathbf{A} ($i \in \mathcal{B}$). Then, let us add to \mathbf{x}_m a vector \mathbf{n} belonging to the null space of \mathbf{B} . Clearly, $\mathbf{x}_m + \mathbf{n} \in \mathcal{X}_c(k)$ because $\|\mathbf{x}_m + \mathbf{n}\|_0 \leq k$ and $\mathcal{E}(\mathbf{x}_m + \mathbf{n}) = \mathcal{E}(\mathbf{x}_m) = \min_{\mathbf{x} \in \mathbb{R}^n} \mathcal{E}(\mathbf{x})$. Since the null space of \mathbf{B} is of dimension $k - m > 0$, $\mathcal{X}_c(k)$ is of infinite cardinality.

As a conclusion, the constrained solution path is defined in the following way, for any case ($m \leq n$ or $m > n$).

Definition 3 *The constrained solution path is the (finite) set*

$$\mathcal{X}_c = \bigcup_{k=0}^{\min(m,n)} \mathcal{X}_c(k).$$

3 Properties of the penalized solution path

3.1 Penalized solution path

For a given $\lambda \geq 0$, we define the set of minimizers of $\mathcal{J}(\mathbf{x}; \lambda) = \mathcal{E}(\mathbf{x}) + \lambda \|\mathbf{x}\|_0$:

$$\mathcal{X}_p(\lambda) = \arg \min_{\mathbf{x} \in \mathbb{R}^n} \{\mathcal{J}(\mathbf{x}; \lambda)\}.$$

By extension, we define $\mathcal{X}_p(+\infty) = \{\mathbf{0}\}$.

Definition 4 *We denote the cardinality of a set $\mathcal{A} \subseteq \{1, \dots, n\}$ by*

$$\|\mathcal{A}\|_0 \triangleq \text{Card}(\mathcal{A}).$$

Definition 5 *We denote by $\mathcal{A}(\mathbf{x}) \subseteq \{1, \dots, n\}$ the support of a vector $\mathbf{x} \in \mathbb{R}^n$.*

Definition 6 *For a given active set \mathcal{A} such that $\|\mathcal{A}\|_0 \leq \min(m, n)$, the corresponding least-square solution is unique (due to the URP assumption). We denote this solution by*

$$\mathbf{x}_{\mathcal{A}} \triangleq \arg \min_{\mathcal{A}(\mathbf{x}) \subseteq \mathcal{A}} \mathcal{E}(\mathbf{x}) \quad (2)$$

and the corresponding least-square cost by

$$\mathcal{E}_{\mathcal{A}} \triangleq \mathcal{E}(\mathbf{x}_{\mathcal{A}}) = \min_{\mathcal{A}(\mathbf{x}) \subseteq \mathcal{A}} \mathcal{E}(\mathbf{x}). \quad (3)$$

Finally, we define the corresponding value of \mathcal{J} by

$$\mathcal{J}_{\mathcal{A}}(\lambda) \triangleq \mathcal{J}(\mathbf{x}_{\mathcal{A}}; \lambda) = \mathcal{E}_{\mathcal{A}} + \lambda \|\mathbf{x}_{\mathcal{A}}\|_0 \quad (4)$$

which is generally different from

$$\min_{\mathcal{A}(\mathbf{x}) \subseteq \mathcal{A}} \mathcal{J}(\mathbf{x}; \lambda).$$

Theorem 5 *If $\lambda > 0$ and $\mathbf{x}_p(\lambda) \in \mathcal{X}_p(\lambda)$, then the support of $\mathbf{x}_p(\lambda)$, denoted by $\mathcal{A} \triangleq \mathcal{A}(\mathbf{x}_p(\lambda))$ for convenience, is such that $\|\mathcal{A}\|_0 \leq \min(m, n)$, and $\mathbf{x}_p(\lambda) = \mathbf{x}_{\mathcal{A}}$.*

Proof 5 — *First, we show that*

$$\mathbf{x}_p(\lambda) \in \arg \min_{\{\mathbf{x} \in \mathbb{R}^n, \mathcal{A}(\mathbf{x}) \subseteq \mathcal{A}\}} \mathcal{E}(\mathbf{x}).$$

Since $\mathbf{x}_p(\lambda)$ is a minimizer of $\mathcal{J}(\mathbf{x}; \lambda)$, the following equivalent inequalities hold for all \mathbf{x} such that $\mathcal{A}(\mathbf{x}) \subseteq \mathcal{A}$:

$$\begin{aligned} \mathcal{J}(\mathbf{x}; \lambda) &\geq \mathcal{J}(\mathbf{x}_p(\lambda); \lambda) \\ \mathcal{E}(\mathbf{x}) + \lambda \|\mathbf{x}\|_0 &\geq \mathcal{E}(\mathbf{x}_p(\lambda)) + \lambda \|\mathcal{A}\|_0 \\ \mathcal{E}(\mathbf{x}) - \mathcal{E}(\mathbf{x}_p(\lambda)) &\geq \lambda (\|\mathcal{A}\|_0 - \|\mathbf{x}\|_0) \geq 0. \end{aligned}$$

We finally deduce that $\mathbf{x}_p(\lambda)$ is a minimizer of \mathcal{E} over the set $\{\mathbf{x} \in \mathbb{R}^n, \mathcal{A}(\mathbf{x}) \subseteq \mathcal{A}\}$.

— *The case where $\|\mathcal{A}\|_0 > \min(m, n)$ never occurs. If it does, remark 2 shows that there exists $\mathbf{x} \in \overline{\mathcal{D}_{\min(m,n)}}$ such that $\mathcal{E}(\mathbf{x}) = \mathcal{E}(\mathbf{x}_p(\lambda))$. Since $\|\mathbf{x}\|_0 \leq \min(m, n) < \|\mathbf{x}_p(\lambda)\|_0 = \|\mathcal{A}\|_0$, $\mathcal{J}(\mathbf{x}; \lambda) < \mathcal{J}(\mathbf{x}_p(\lambda); \lambda)$, which is in contradiction with $\mathbf{x}_p(\lambda) \in \mathcal{X}_p(\lambda)$.*

Finally, $\|\mathcal{A}\|_0 \leq \min(m, n)$ and there is only one minimizer of \mathcal{E} over the set $\{\mathbf{x} \in \mathbb{R}^n, \mathcal{A}(\mathbf{x}) \subseteq \mathcal{A}\}$ (URP assumption), which is $\mathbf{x}_{\mathcal{A}}$.

Corrolary 1 If $\lambda > 0$, the set $\mathcal{X}_p(\lambda)$ is finite and $\mathcal{X}_p(\lambda) \subseteq \overline{\mathcal{D}_{\min(m,n)}}$.

Proof 6 There are at most $\sum_{k=0}^{\min(m,n)} C_n^k$ distinct values $\mathbf{x}_p(\lambda)$ (i.e., $\sum_{k=0}^{\min(m,n)} C_n^k$ sets which are candidate to be a set \mathcal{A} and one optimal \mathbf{x} -value $\mathbf{x}_\mathcal{A}$ per set), which shows that $\mathcal{X}_p(\lambda)$ is a finite set. Additionally, we have seen in theorem 5 that for each solution $\mathbf{x}_p(\lambda)$, $\|\mathbf{x}_p(\lambda)\|_0 = \|\mathcal{A}\|_0 \leq \min(m, n)$.

Definition 7 The penalized solution path is defined as the union of sets

$$\mathcal{X}_p = \bigcup_{\lambda > 0} \mathcal{X}_p(\lambda).$$

Imposing $\lambda > 0$ (rather than $\lambda \geq 0$) guarantees that $\mathcal{X}_p(\lambda)$ is of finite cardinality for all λ . Moreover, it is easy to see (from theorem 5) that the solution path is of finite cardinality, since all the sets $\mathcal{X}_p(\lambda)$ are included in a common set of cardinality $\sum_{k=0}^{\min(m,n)} C_n^k$: $\{\mathbf{x} \in \mathbb{R}^n, \exists \mathcal{A} \subseteq \{1, \dots, n\}, \|\mathcal{A}\|_0 \leq \min(m, n) \text{ and } \mathbf{x} = \mathbf{x}_\mathcal{A}\}$.

3.2 Piecewise constant property

Theorem 6 The dependence of the set $\mathcal{X}_p(\lambda)$ w.r.t. λ ($\lambda > 0$) is piecewise constant, with a finite number of intervals $(\lambda_i^*, \lambda_{i+1}^*)$: for all i , $\mathcal{X}_p(\lambda)$ is constant for $\lambda \in (\lambda_i^*, \lambda_{i+1}^*)$ and if $\lambda \in (\lambda_i^*, \lambda_{i+1}^*)$, then $\mathcal{X}_p(\lambda) \subseteq \mathcal{X}_p(\lambda_i^*) \cap \mathcal{X}_p(\lambda_{i+1}^*)$.

The minimal cost value $\mathcal{J}(\lambda) \triangleq \min_{\mathbf{x} \in \mathbb{R}^n} \mathcal{J}(\mathbf{x}; \lambda)$ is a continuous and piecewise linear function of λ , and

$$\forall \lambda, \mathcal{J}(\lambda) = \min_{\{\mathcal{A} \subseteq \{1, \dots, n\}, \|\mathcal{A}\|_0 \leq \min(m, n)\}} \mathcal{J}_\mathcal{A}(\lambda). \quad (5)$$

Definition 8 In the following, we will define the values $\lambda = \lambda_i^*$ ($i = 1, \dots, I$) as the **critical values**. These values, together with $\lambda_0^* = 0$ and $\lambda_{I+1}^* = +\infty$, define the piecewise constant domain $\mathcal{X}_p(\lambda)$:

$$0 = \lambda_0^* < \lambda_1^* < \dots < \lambda_I^* < \lambda_{I+1}^* = +\infty. \quad (6)$$

λ_i^* are also the λ -values at which the derivative of \mathcal{J} is changing: at $\lambda = \lambda_i^*$, $\lambda \mapsto \mathcal{J}(\lambda)$ is not differentiable, and \mathcal{J} is linear on each interval $[\lambda_i^*, \lambda_{i+1}^*]$ (see Fig. 1).

Proof 7 — The result (5) can be illustrated geometrically, by considering the affine curves $\lambda \mapsto \mathcal{J}_\mathcal{A}(\lambda)$ for all the possible supports \mathcal{A} such that $\|\mathcal{A}\|_0 \leq \min(m, n)$ (see Fig. 1). Let us prove that (5) holds.

When λ is fixed, let $\mathbf{x}_p(\lambda) \in \mathcal{X}_p(\lambda)$, and let $\mathcal{A} \triangleq \mathcal{A}(\mathbf{x}_p(\lambda))$.

- According to theorem 5, $\|\mathbf{x}_p(\lambda)\|_0 = \|\mathcal{A}\|_0 \leq \min(m, n)$ and $\mathbf{x}_p(\lambda) = \mathbf{x}_\mathcal{A}$. Thus, $\mathcal{E}(\mathbf{x}_p(\lambda)) = \mathcal{E}_\mathcal{A}$ and

$$\mathcal{J}(\mathbf{x}_p(\lambda); \lambda) = \mathcal{J}_\mathcal{A}(\lambda).$$

- $\mathbf{x}_p(\lambda) \in \mathcal{X}_p(\lambda)$ implies that for all $\mathcal{A}' \subseteq \{1, \dots, n\}$ such that $\|\mathcal{A}'\|_0 \leq \min(m, n)$,

$$\mathcal{J}(\mathbf{x}_p(\lambda); \lambda) \leq \mathcal{J}_{\mathcal{A}'}(\lambda) = \mathcal{J}(\mathbf{x}_{\mathcal{A}'}; \lambda).$$

Here, we have shown that (5) holds since $\mathcal{J}(\lambda) = \mathcal{J}(\mathbf{x}_p(\lambda); \lambda)$.

— $\lambda \mapsto \mathcal{J}(\lambda)$ is a continuous and piecewise linear function of λ because of (5). Since the number of affine curves $\lambda \mapsto \mathcal{J}_\mathcal{A}(\lambda)$ is finite, $\lambda \mapsto \mathcal{J}(\lambda)$ is described by a finite set of values $\{(\lambda_i^*, \mathcal{E}_{\mathcal{A}_i}, \|\mathbf{x}_{\mathcal{A}_i}\|_0), i = 0, \dots, I\}$, where $\lambda_0^* = 0 < \lambda_1^* < \dots < \lambda_I^* < \lambda_{I+1}^* = +\infty$. Each value λ_i^* ($i = 1, \dots, I$) corresponds to the intersection between a pair of affine curves (see Fig. 1), and the restriction of \mathcal{J} to a given interval $[\lambda_i^*, \lambda_{i+1}^*]$ is linear:

$$\forall \lambda \in [\lambda_i^*, \lambda_{i+1}^*], \mathcal{J}(\lambda) = \mathcal{E}_{\mathcal{A}_i} + \lambda \|\mathbf{x}_{\mathcal{A}_i}\|_0. \quad (7)$$

In particular, for $i = 0$, we have

$$\forall \lambda \in [0, \lambda_1], \mathcal{J}(\lambda) = \mathcal{E}_{\mathcal{A}_0} + \lambda \|\mathbf{x}_{\mathcal{A}_0}\|_0, \quad (8)$$

where $\mathcal{E}_{\mathcal{A}_0} = \min_{\mathbf{x} \in \mathbb{R}^n} \mathcal{E}(\mathbf{x})$ is the minimal least-square error, and $\|\mathbf{x}_{\mathcal{A}_0}\|_0$ is the minimal L_0 -norm of the minimizers of \mathcal{E} over \mathbb{R}^n . For $i = I$, we have necessarily $\mathbf{x}_{\mathcal{A}_I} = \mathbf{0}$ and $\mathcal{E}_{\mathcal{A}_I} = \|\mathbf{y}\|^2$, thus

$$\forall \lambda \in [\lambda_I^*, +\infty), \mathcal{J}(\lambda) = \|\mathbf{y}\|^2. \quad (9)$$

— For a given interval $[\lambda_i^*, \lambda_{i+1}^*]$, let us show that when $\lambda \in (\lambda_i^*, \lambda_{i+1}^*)$, $\mathcal{X}_p(\lambda)$ is a constant set. For some given λ -value $\in (\lambda_i^*, \lambda_{i+1}^*)$, we consider $\mathbf{x} \in \mathcal{X}_p(\lambda)$, then necessarily, the following equivalent equations hold:

$$\begin{aligned} \mathcal{J}(\lambda) &= \mathcal{J}(\mathbf{x}; \lambda) \\ \mathcal{E}_{\mathcal{A}_i} + \lambda \|\mathbf{x}_{\mathcal{A}_i}\|_0 &= \mathcal{E}(\mathbf{x}) + \lambda \|\mathbf{x}\|_0. \end{aligned}$$

Imagine that $\mathcal{E}_{\mathcal{A}_i} \neq \mathcal{E}(\mathbf{x})$, then, necessarily, the two functions $\mathcal{J}(\lambda') = \mathcal{E}_{\mathcal{A}_i} + \lambda' \|\mathbf{x}_{\mathcal{A}_i}\|_0$ and $\mathcal{J}(\mathbf{x}; \lambda') = \mathcal{E}(\mathbf{x}) + \lambda' \|\mathbf{x}\|_0$ do not coincide for $\lambda' \in [\lambda_i^*, \lambda_{i+1}^*] \setminus \{\lambda\}$. Moreover, $\mathcal{J}(\mathbf{x}; \lambda')$ is strictly lower than $\mathcal{J}(\lambda')$ either for $\lambda' \in [\lambda_i^*, \lambda)$ or for $\lambda' \in (\lambda, \lambda_{i+1}^*]$. This is in contradiction with (7) and the definition of $\mathcal{J}(\lambda')$ in theorem 6.

We have shown that $\mathcal{E}_{\mathcal{A}_i} = \mathcal{E}(\mathbf{x})$. Since $\mathcal{J}(\lambda) = \mathcal{J}(\mathbf{x}; \lambda)$ and $\lambda > 0$, we deduce that $\|\mathbf{x}_{\mathcal{A}_i}\|_0 = \|\mathbf{x}\|_0$, and that $\forall \lambda' \in [\lambda_i^*, \lambda_{i+1}^*]$, $\mathcal{J}(\lambda') = \mathcal{J}(\mathbf{x}; \lambda')$. Finally, if $\lambda \in (\lambda_i^*, \lambda_{i+1}^*)$ and $\mathbf{x} \in \mathcal{X}_p(\lambda)$, then $\mathbf{x} \in \mathcal{X}_p(\lambda')$ for all $\lambda' \in [\lambda_i^*, \lambda_{i+1}^*]$. $\mathcal{X}_p(\lambda)$ is then a constant set when $\lambda \in (\lambda_i^*, \lambda_{i+1}^*)$, and $\mathcal{X}_p(\lambda) \subseteq \mathcal{X}_p(\lambda_i^*) \cap \mathcal{X}_p(\lambda_{i+1}^*)$.

Lemma 1 The function $\lambda \mapsto \mathcal{J}(\lambda)$ is increasing and concave.

Proof 8 \mathcal{J} is an increasing and concave function as the minimum of a finite set of increasing and concave functions.

Lemma 2 $\mathcal{X}_p(0) \cap \mathcal{X}_p \neq \emptyset$, and if $m \geq n$, then $\mathcal{X}_p(0) \subset \mathcal{X}_p$.

Proof 9 The application of the result of theorem 6: “for all i , if $\lambda \in (\lambda_i^*, \lambda_{i+1}^*)$, then $\mathcal{X}_p(\lambda) \subseteq \mathcal{X}_p(\lambda_i^*)$ ” with $i = 0$ yields

$$\forall \lambda \in (0, \lambda_1^*), \mathcal{X}_p(\lambda) \subseteq \mathcal{X}_p(0).$$

Thus, we always have $\mathcal{X}_p(0) \cap \mathcal{X}_p \neq \emptyset$. For $m \geq n$, $\mathcal{X}_p(0)$ is formed of only one vector, thus $\forall \lambda \in (0, \lambda_1^*)$, $\mathcal{X}_p(\lambda) = \mathcal{X}_p(0)$, and $\mathcal{X}_p(0) \subseteq \mathcal{X}_p$. Since $\mathbf{y} \neq \mathbf{0}$ and \mathbf{A} is full rank, the domain (6) is formed of at least two intervals ($I \geq 1$), thus $\mathcal{X}_p(0) \subset \mathcal{X}_p$.

Theorem 7 For a given λ -value which is distinct from $\lambda_0^*, \lambda_1^*, \dots, \lambda_I^*$, all the elements of $\mathcal{X}_p(\lambda)$ are of same L_0 -norm, which is equal to the derivative of $\mathcal{J}(\lambda)$, and yield the same least-square cost.

Proof 10 Because of theorem 6, for a given value of i , there exists $\mathcal{A}_i \subseteq \{1, \dots, n\}$ such that

$$\forall \lambda' \in [\lambda_i^*, \lambda_{i+1}^*], \mathcal{J}(\lambda') = \mathcal{J}_{\mathcal{A}_i}(\lambda'). \quad (10)$$

Now, let us fix the value of $\lambda \in (\lambda_i^*, \lambda_{i+1}^*)$ and let $\mathbf{x}_p(\lambda) \in \mathcal{X}_p(\lambda)$. Because of theorem 6, $\mathcal{X}_p(\lambda')$ is constant for $\lambda' \in (\lambda_i^*, \lambda_{i+1}^*)$, and $\mathbf{x}_p(\lambda) \in \mathcal{X}_p(\lambda')$ for all $\lambda' \in [\lambda_i^*, \lambda_{i+1}^*]$. (10) implies that

$$\begin{aligned} \forall \lambda' \in [\lambda_i^*, \lambda_{i+1}^*], \mathcal{J}(\mathbf{x}_p(\lambda); \lambda') &= \mathcal{J}_{\mathcal{A}_i}(\lambda') \\ \forall \lambda' \in [\lambda_i^*, \lambda_{i+1}^*], \mathcal{E}(\mathbf{x}_p(\lambda)) + \lambda' \|\mathbf{x}_p(\lambda)\|_0 &= \mathcal{E}_{\mathcal{A}_i} + \lambda' \|\mathbf{x}_{\mathcal{A}_i}\|_0. \end{aligned} \quad (11)$$

Taking the derivative of (11) yields $\|\mathbf{x}_p(\lambda)\|_0 = \|\mathbf{x}_{\mathcal{A}_i}\|_0 = \mathcal{J}'(\lambda)$, and then, due to (11), $\mathcal{E}(\mathbf{x}_p(\lambda)) = \mathcal{E}_{\mathcal{A}_i}$.

Theorem 8 Let $\mathbf{x}_p(\lambda)$ be a sequence such that $\forall \lambda, \mathbf{x}_p(\lambda) \in \mathcal{X}_p(\lambda)$. Then, necessarily, $\|\mathbf{x}_p(\lambda)\|_0$ is a decreasing function of λ , and $\mathcal{E}(\mathbf{x}_p(\lambda))$ is an increasing function of λ .

Proof 11 • Recall that for $i \in \{0, \dots, I\}$, there exists a set \mathcal{A}_i such that if $\lambda \in (\lambda_i^*, \lambda_{i+1}^*)$ and $\mathbf{x}_p(\lambda) \in \mathcal{X}_p(\lambda)$, then $\|\mathbf{x}_p(\lambda)\|_0 = \|\mathbf{x}_{\mathcal{A}_i}\|_0$ (see theorem 7);

- The first result is a direct consequence of theorem 7: $\forall \lambda \notin \{\lambda_0^*, \dots, \lambda_I^*\}$, $\|\mathbf{x}_p(\lambda)\|_0 = \mathcal{J}'(\lambda)$, and of lemma 1: \mathcal{J} is a concave function, thus its derivative (when it is defined) is a decreasing function of λ . At this point, we know that $\lambda \mapsto \|\mathbf{x}_p(\lambda)\|_0$ is piecewise constant on \mathbb{R}_+ , and that its restriction to $\mathbb{R}_+ \setminus \{\lambda_0^*, \dots, \lambda_I^*\}$ is decreasing: $\forall i \in \{1, \dots, I\}$, $\|\mathbf{x}_{\mathcal{A}_{i-1}}\|_0 \geq \|\mathbf{x}_{\mathcal{A}_i}\|_0$. The remaining part is to study the behavior of $\|\mathbf{x}_p(\lambda)\|_0$ at $\lambda = \lambda_i^*$, $i = 0, \dots, I$.

For $i \in \{1, \dots, I\}$, let us show that $\mathbf{x}_p(\lambda_i^*)$ is such that $\|\mathbf{x}_{\mathcal{A}_{i-1}}\|_0 \geq \|\mathbf{x}_p(\lambda_i^*)\|_0 \geq \|\mathbf{x}_{\mathcal{A}_i}\|_0$:

- $\lambda \mapsto \mathcal{J}(\mathbf{x}_p(\lambda_i^*); \lambda)$ and $\lambda \mapsto \mathcal{J}(\lambda)$ coincide at $\lambda = \lambda_i^*$;
- $\mathcal{J}'(\lambda)$ is equal to $\|\mathbf{x}_{\mathcal{A}_{i-1}}\|_0$ when $\lambda \in (\lambda_{i-1}^*, \lambda_i^*)$, and to $\|\mathbf{x}_{\mathcal{A}_i}\|_0$ when $\lambda \in (\lambda_i^*, \lambda_{i+1}^*)$.
- the derivative of $\lambda \mapsto \mathcal{J}(\mathbf{x}_p(\lambda_i^*); \lambda)$ is equal to $\|\mathbf{x}_p(\lambda_i^*)\|_0$.

Due to the definition of $\lambda \mapsto \mathcal{J}(\lambda)$ in theorem 6, the affine function $\lambda \mapsto \mathcal{J}(\mathbf{x}_p(\lambda_i^*); \lambda)$ is necessarily greater or equal to $\lambda \mapsto \mathcal{J}(\lambda)$ for $\lambda \in (\lambda_{i-1}^*, \lambda_i^*)$ and for $\lambda \in (\lambda_i^*, \lambda_{i+1}^*)$. This implies that $\|\mathbf{x}_{\mathcal{A}_{i-1}}\|_0 \geq \|\mathbf{x}_p(\lambda_i^*)\|_0 \geq \|\mathbf{x}_{\mathcal{A}_i}\|_0$.

A similar argument can be given to show that $\|\mathbf{x}_p(\lambda_0^*)\|_0 \geq \|\mathbf{x}_{\mathcal{A}_0}\|_0$.

Finally, we have shown that $\lambda \mapsto \|\mathbf{x}_p(\lambda)\|_0$ is decreasing on \mathbb{R}_+ .

- Second result: for a given $i \in \{1, \dots, I\}$, the continuity of \mathcal{J} at $\lambda = \lambda_i^*$ reads $\mathcal{E}_{\mathcal{A}_{i-1}} + \lambda_i^* \|\mathbf{x}_{\mathcal{A}_{i-1}}\|_0 = \mathcal{E}_{\mathcal{A}_i} + \lambda_i^* \|\mathbf{x}_{\mathcal{A}_i}\|_0$. Because $\|\mathbf{x}_{\mathcal{A}_{i-1}}\|_0 \geq \|\mathbf{x}_{\mathcal{A}_i}\|_0$, $\mathcal{E}_{\mathcal{A}_{i-1}} \leq \mathcal{E}_{\mathcal{A}_i}$.

When λ varies from 0 to $+\infty$ and $\lambda \notin \{\lambda_1^*, \dots, \lambda_I^*\}$, $\mathcal{E}(\mathbf{x}_p(\lambda))$ takes sequentially the values $\mathcal{E}_{\mathcal{A}_i}$, $i = 0, \dots, I$. Thus, the restriction of $\lambda \mapsto \mathcal{E}(\mathbf{x}_p(\lambda))$ to $\mathbb{R}_+ \setminus \{\lambda_1^*, \dots, \lambda_I^*\}$ is increasing. With similar arguments than in the first result, we can show that for $i \in \{1, \dots, I\}$, $\mathcal{E}_{\mathcal{A}_{i-1}} \leq \mathcal{E}(\mathbf{x}_p(\lambda_i^*)) \leq \mathcal{E}_{\mathcal{A}_i}$. Finally, $\lambda \mapsto \mathcal{E}(\mathbf{x}_p(\lambda))$ is increasing on \mathbb{R}_+ .

3.3 Cardinality of $\mathcal{X}_p(\lambda)$

It is easy to see that:

- For all $i \in \{1, \dots, I\}$, the cardinality of $\mathcal{X}_p(\lambda_i^*)$ is larger than 2, because at $\lambda = \lambda_i^*$, at least two distinct affine curves $\lambda \mapsto \mathcal{J}_{\mathcal{A}}(\lambda) = \mathcal{E}_{\mathcal{A}} + \lambda \|\mathbf{x}_{\mathcal{A}}\|_0$ intersect (see Fig. 2).
- $\mathcal{X}_p(\lambda)$ is reduced to the unique vector $\mathbf{0}$ for the largest λ -values ($\lambda > \lambda_I^* \Rightarrow \mathcal{X}_p(\lambda) = \{\mathbf{0}\}$).
- For $\lambda = 0$ (the least-square error $\mathcal{E}(\mathbf{x})$ is minimized with no penalty), $\mathcal{X}_p(0)$ is either reduced to the unique vector $(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y}$ when $m \geq n$, or is of infinite cardinality otherwise.

We conclude that at least for $m \geq n$, the cardinality of $\mathcal{X}_p(\lambda)$ is not monotonic w.r.t. λ .

3.4 Relationship between the constrained and the penalized solution paths

Generally, the solution paths \mathcal{X}_c and \mathcal{X}_p do not coincide. This is a consequence of the non-convexity of the L0-norm [2]. However, $\mathcal{X}_p \subseteq \mathcal{X}_c$ is always true (a well-known result in the literature of multi-objective optimization?).

In general, the proposition “ $\forall k, \exists \lambda, \mathcal{X}_c(k) \subseteq \mathcal{X}_p(\lambda)$ ” is false (see Fig. 1).

Theorem 9 *If $\lambda \neq \{\lambda_0^*, \dots, \lambda_I^*\}$, then there exists k such that $\mathcal{X}_p(\lambda) = \mathcal{X}_c(k)$.*

Proof 12 *For a given λ -value, let $\mathbf{x} \in \mathcal{X}_p(\lambda)$, let $\mathcal{A} \triangleq \mathcal{A}(\mathbf{x})$ denote the support of \mathbf{x} and $k_{\mathbf{x}} \triangleq \|\mathbf{x}\|_0 = \|\mathcal{A}\|_0$. According to theorem 5, $\|\mathcal{A}\|_0 \leq \min(m, n)$ and $\mathbf{x} = \mathbf{x}_{\mathcal{A}}$.*

– *Let us show that $\mathbf{x} \in \mathcal{X}_c(k_{\mathbf{x}})$. If $\mathbf{x} \notin \mathcal{X}_c(k_{\mathbf{x}})$, there exists a (minimal) support \mathcal{B} such that $\|\mathcal{B}\|_0 \leq k_{\mathbf{x}}$ and $\mathcal{E}_{\mathcal{B}} < \mathcal{E}(\mathbf{x}) = \mathcal{E}_{\mathcal{A}}$, then $\mathcal{J}(\mathbf{x}_{\mathcal{B}}; \lambda) < \mathcal{J}(\mathbf{x}_{\mathcal{A}}; \lambda)$. This is in contradiction with $\mathbf{x} \in \mathcal{X}_p(\lambda)$.*

At this point, we have shown that

$$\forall \lambda, \forall \mathbf{x} \in \mathcal{X}_p(\lambda), \exists k_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}_c(k_{\mathbf{x}}),$$

or equivalently,

$$\mathcal{X}_p \subseteq \mathcal{X}_c.$$

– *The following of the proof requires the assumption $\lambda \neq \{\lambda_0^*, \dots, \lambda_I^*\}$. We have seen that if \mathbf{x} and $\mathbf{y} \in \mathcal{X}_p(\lambda)$, then $\mathbf{x} \in \mathcal{X}_c(k_{\mathbf{x}})$ and $\mathbf{y} \in \mathcal{X}_c(k_{\mathbf{y}})$. According to theorem 7, all the elements of $\mathcal{X}_p(\lambda)$ are of same L0-norm. Therefore, $k_{\mathbf{y}} = k_{\mathbf{x}}$. At this point, we have shown that*

$$\forall \lambda \neq \{\lambda_0^*, \dots, \lambda_I^*\}, \exists k_{\lambda}, \mathcal{X}_p(\lambda) \subseteq \mathcal{X}_c(k_{\lambda}).$$

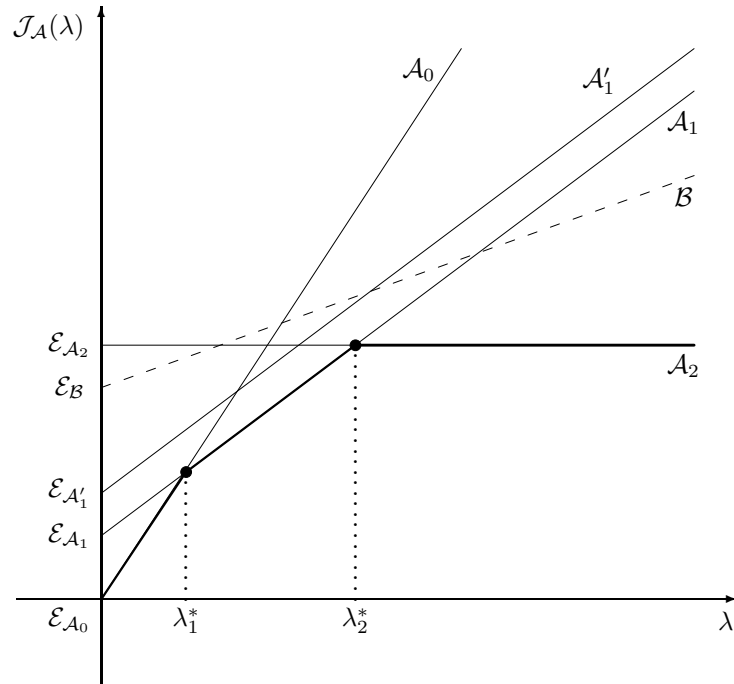


Figure 1: Representation of the affine curves $\lambda \mapsto \mathcal{J}_{\mathcal{A}}(\lambda) = \mathcal{E}_{\mathcal{A}} + \lambda \|\mathbf{x}_{\mathcal{A}}\|_0$ for all the possible supports \mathcal{A} such that $\|\mathcal{A}\|_0 \leq \min(m, n)$. Note that a given affine curve may correspond to several supports \mathcal{A} and \mathcal{B} for which $\forall \lambda, \mathcal{J}_{\mathcal{A}}(\lambda) = \mathcal{J}_{\mathcal{B}}(\lambda)$. The piecewise linear function $\lambda \mapsto \mathcal{J}(\lambda)$ is defined according to (5) and is represented in bold lines. From this illustration, let us comment on the nonequivalence of both solution paths \mathcal{X}_c and \mathcal{X}_p . By following the bold curve representing $\lambda \mapsto \mathcal{J}(\lambda)$, we see that the penalized solution path is described by the active sets \mathcal{A}_0 , \mathcal{A}_1 and \mathcal{A}_2 (and the possible other sets yielding the same three curves $\lambda \mapsto \mathcal{J}_{\mathcal{A}_i}(\lambda)$) for which $\|\mathbf{x}_{\mathcal{A}}\|_0$ is equal to 3, 2 and 0, respectively. \mathcal{A}'_1 is an active set such that $\|\mathbf{x}_{\mathcal{A}'_1}\|_0 = 2$ but $\mathcal{E}_{\mathcal{A}'_1} > \mathcal{E}_{\mathcal{A}_1}$. No active set such that $\|\mathbf{x}_{\mathcal{A}}\|_0 = 1$ is present in \mathcal{X}_p . \mathcal{B} is the active set such that $\|\mathbf{x}_{\mathcal{B}}\|_0 = 1$ whose energy $\mathcal{E}_{\mathcal{B}}$ is the lowest among all the active sets such that $\|\mathbf{x}_{\mathcal{A}}\|_0 \leq 1$, however, $\forall \lambda, \mathcal{J}_{\mathcal{B}}(\lambda) > \mathcal{J}(\lambda)$. Thus, $\mathcal{X}_c(1) = \{\mathcal{B}\} \notin \mathcal{X}_p$. On the contrary, for all $\lambda \neq \{\lambda_0^*, \dots, \lambda_l^*\}$, $\mathcal{X}_p(\lambda) = \mathcal{X}_c(k_\lambda)$, with $k_\lambda = \mathcal{J}'(\lambda) = 3, 2$ or 0 .

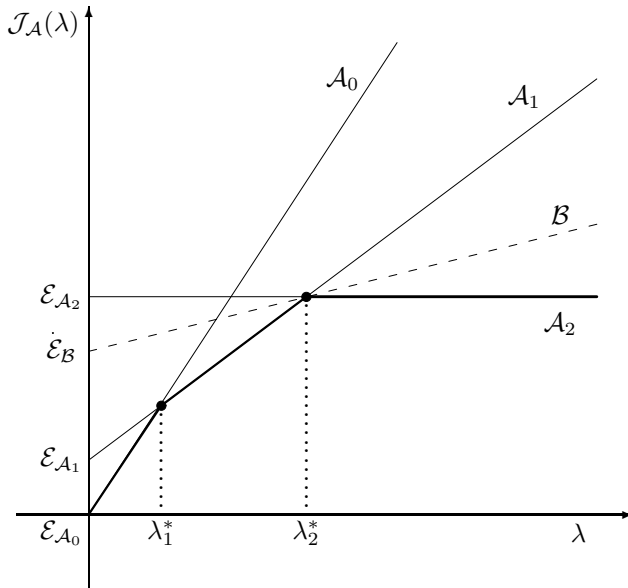


Figure 2: Content of $\mathcal{X}_p(\lambda)$ at a critical λ -value $\lambda = \lambda_i^*$, $i \geq 1$: $\mathcal{X}_p(\lambda_i^*) \subset \mathcal{X}_c$, and $\text{Card}(\mathcal{X}_p(\lambda_i^*)) \geq 2$. On this example, $\mathcal{X}_p(\lambda_2^*) = \mathcal{X}_c(0) \cup \mathcal{X}_c(1) \cup \mathcal{X}_c(2)$ and $\text{Card}(\mathcal{X}_p(\lambda_2^*)) \geq 3$ since \mathbf{x}_{A_1} , \mathbf{x}_{A_2} and $\mathbf{x}_B \in \mathcal{X}_p(\lambda_2^*)$.

— Now, let us prove the reverse inclusion. Given λ , there exists at least one \mathbf{x} such that $\mathbf{x} \in \mathcal{X}_p(\lambda)$, $k_\lambda = \|\mathbf{x}\|_0$ and $\mathbf{x} \in \mathcal{X}_c(k_\lambda)$. For all $\mathbf{y} \in \mathcal{X}_c(k_\lambda)$, we have necessarily $\mathcal{E}(\mathbf{y}) = \mathcal{E}(\mathbf{x})$ and $\|\mathbf{y}\|_0 \leq \|\mathbf{x}\|_0$, thus $\mathcal{J}(\mathbf{y}; \lambda) \leq \mathcal{J}(\mathbf{x}; \lambda)$. Since $\mathbf{x} \in \mathcal{X}_p(\lambda)$, we deduce that $\mathcal{J}(\mathbf{y}; \lambda) = \mathcal{J}(\mathbf{x}; \lambda)$ and that $\mathbf{y} \in \mathcal{X}_p(\lambda)$. This completes the proof, since we have shown that

$$\forall \lambda \neq \{\lambda_0^*, \dots, \lambda_I^*\}, \mathcal{X}_c(k_\lambda) \subseteq \mathcal{X}_p(\lambda).$$

Actually, $k_\lambda = \mathcal{J}'(\lambda)$ according to theorem 7.

3.5 Content of $\mathcal{X}_p(\lambda)$ at critical λ -values

Lemma 3 If $\lambda_{i-1}^* < \lambda < \lambda_i^* < \lambda' < \lambda_{i+1}^*$, then $\mathcal{X}_p(\lambda) \cap \mathcal{X}_p(\lambda') = \emptyset$.

Proof 13 According to theorem 7, all the vectors of $\mathcal{X}_p(\lambda)$ (respectively of $\mathcal{X}_p(\lambda')$) are of same L_0 -norm, which is the derivative of \mathcal{J} at λ (resp. λ'). Thus, if $\mathcal{X}_p(\lambda) \cap \mathcal{X}_p(\lambda') \neq \emptyset$, the derivative of \mathcal{J} is constant on $(\lambda_{i-1}^*, \lambda_{i+1}^*) \setminus \{\lambda_i^*\}$, which is in contradiction with the definition of λ_i^* (critical point, at which the derivative of \mathcal{J} is changing).

Theorem 10 If $\lambda_{i-1}^* < \lambda < \lambda_i^* < \lambda' < \lambda_{i+1}^*$, then $\mathcal{X}_p(\lambda) \cup \mathcal{X}_p(\lambda') \subseteq \mathcal{X}_p(\lambda_i^*)$, thus $\text{Card}(\mathcal{X}_p(\lambda)) + \text{Card}(\mathcal{X}_p(\lambda')) \leq \text{Card}(\mathcal{X}_p(\lambda_i^*))$ (Card denotes the cardinality). If $\lambda \in (\lambda_{i-1}^*, \lambda_{i+1}^*) \setminus \{\lambda_i^*\}$, then $\text{Card}(\mathcal{X}_p(\lambda)) < \text{Card}(\mathcal{X}_p(\lambda_i^*))$.

See illustration in Fig. 2.

Proof 14 • *First result:* according to theorem 6, if $\lambda \in (\lambda_i^*, \lambda_{i+1}^*)$, then $\mathcal{X}_p(\lambda) \subseteq \mathcal{X}_p(\lambda_i^*) \cap \mathcal{X}_p(\lambda_{i+1}^*)$. According to lemma 3, if $\lambda_{i-1}^* < \lambda < \lambda_i^* < \lambda' < \lambda_{i+1}^*$, then $\mathcal{X}_p(\lambda) \cap \mathcal{X}_p(\lambda') = \emptyset$. Thus, $\mathcal{X}_p(\lambda) \cup \mathcal{X}_p(\lambda') \subseteq \mathcal{X}_p(\lambda_i^*)$ and $\text{Card}(\mathcal{X}_p(\lambda)) + \text{Card}(\mathcal{X}_p(\lambda')) \leq \text{Card}(\mathcal{X}_p(\lambda_i^*))$.

- *Second result:* since neither $\mathcal{X}_p(\lambda)$ nor $\mathcal{X}_p(\lambda')$ is empty, their cardinality is larger or equal to 1, then, applying $\text{Card}(\mathcal{X}_p(\lambda)) + \text{Card}(\mathcal{X}_p(\lambda')) \leq \text{Card}(\mathcal{X}_p(\lambda_i^*))$, we deduce that $\text{Card}(\mathcal{X}_p(\lambda)) < \text{Card}(\mathcal{X}_p(\lambda_i^*))$ and $\text{Card}(\mathcal{X}_p(\lambda')) < \text{Card}(\mathcal{X}_p(\lambda_i^*))$.

4 SBR and CSBR algorithms

4.1 SBR iterates and output

Let us consider the SBR algorithm for a given λ -value. An SBR iterate takes the form of:

- an active set \mathcal{A} (for simplicity, we omit the dependence w.r.t. λ);
- the corresponding least-square minimizer $\mathbf{x}_{\mathcal{A}} = \arg \min_{\{\mathbf{x} \in \mathbb{R}^n, \mathcal{A}(\mathbf{x}) \subseteq \mathcal{A}\}} \mathcal{E}(\mathbf{x})$.

$\hat{\mathbf{x}}(\lambda) \triangleq \mathbf{x}_{\mathcal{A}}$ is chosen as the estimator of a minimizer (there may be several) of $\mathcal{J}(\mathbf{x}; \lambda) = \mathcal{E}(\mathbf{x}) + \lambda \|\mathbf{x}\|_0$ over \mathbb{R}^n .

First, recall that for the SBR iterates (and in particular when SBR terminates), $\|\mathbf{x}_{\mathcal{A}}\|_0 = \|\mathcal{A}\|_0$. This property can be guaranteed by including in the SBR loops a small procedure which removes from the active set \mathcal{A} all the indices $i \in \mathcal{A}$ such that $\mathbf{x}_{\mathcal{A}}(i) = 0$ (however, these removals rarely occur in practice). The following remark follows from this property.

Remark 3 For a given λ -value, the cost of an SBR iterate $\mathbf{x}_{\mathcal{A}}$ is $\mathcal{J}_{\mathcal{A}}(\lambda) = \mathcal{E}_{\mathcal{A}} + \lambda \|\mathcal{A}\|_0$. Because the cost of SBR iterates only depend on their support and for convenience, we will omit their dependence w.r.t. \mathbf{x} .

Remark 4 SBR terminates after a finite number of iterations. Moreover, a set \mathcal{A} cannot be explored twice while running SBR.

Proof 15 SBR is a descent algorithm and the number of sets \mathcal{A} which are reachable is finite (i.e., the number of subsets of $\{1, \dots, n\}$).

Remark 5 When SBR terminates, the estimate $\hat{\mathbf{x}}(\lambda) = \mathbf{x}_{\mathcal{A}}$ is generally not included in $\mathcal{X}_p(\lambda)$ because SBR is a sub-optimal algorithm.

Remark 6 At the SBR output \mathcal{A} , \mathcal{J} is “locally minimum w.r.t. \mathcal{A} ”: any replacement of \mathcal{A} by $\mathcal{A} \bullet i$ (where $\bullet \triangleq \cup$ or \setminus) does not yield a decrease of the cost $\mathcal{J}_{\mathcal{A}}(\lambda)$). Formally, this property reads:

$$\forall i, \mathcal{J}_{\mathcal{A}}(\lambda) \leq \mathcal{J}_{\mathcal{A} \bullet i}(\lambda), \quad (12)$$

or equivalently,

$$\begin{aligned} \forall i, \mathcal{E}_{\mathcal{A}} + \lambda \|\mathcal{A}\|_0 &\leq \mathcal{E}_{\mathcal{A} \bullet i} + \lambda \|\mathbf{x}_{\mathcal{A} \bullet i}\|_0 \\ \forall i, \mathcal{E}_{\mathcal{A}} - \mathcal{E}_{\mathcal{A} \bullet i} &\leq \lambda (\|\mathbf{x}_{\mathcal{A} \bullet i}\|_0 - \|\mathcal{A}\|_0). \end{aligned} \quad (13)$$

Here, we do not consider the small removal procedure described above for $\mathcal{A} \bullet i$ (update of $\mathcal{A} \bullet i$ by removing the indices corresponding to the zero valued entries of $\mathbf{x}_{\mathcal{A} \bullet i}$), therefore we use $\|\mathbf{x}_{\mathcal{A} \bullet i}\|_0$, which may be lower than $\|\mathcal{A} \bullet i\|_0$.

4.2 Iterative computation of λ in the CSBR algorithm

When $\lambda = \lambda_q > 0$ (q -th call to SBR), let $\mathcal{A} = \mathcal{A}_q$ be the support of the output of SBR(λ_q). Then, (13) holds. For simplicity, we will omit, when possible, the dependence of \mathcal{A} w.r.t. q . When λ_q is replaced by another value $\lambda \leq \lambda_q$ and $\mathcal{A} = \mathcal{A}_q$ is kept fixed, for which λ -values does (13) remain valid?

When $\bullet = \setminus$, both terms on the left- and right-hand sides of the inequality are strictly negative ($\|\mathbf{x}_{\mathcal{A} \setminus i}\|_0 \leq \|\mathcal{A}\|_0 - 1$), while when $\bullet = \cup$, both terms are positive since $\mathcal{E}_{\mathcal{A}} - \mathcal{E}_{\mathcal{A} \cup i} \geq 0$ and (13) holds for $\lambda = \lambda_q$ (this implies that $\|\mathbf{x}_{\mathcal{A} \cup i}\|_0 = \|\mathcal{A}\|_0$ or $\|\mathcal{A}\|_0 + 1$). Therefore, (13) remains valid for $\lambda \neq \lambda_q$ if and only if

$$(0 \leq) \max_{i \notin \mathcal{A} \text{ and } \|\mathbf{x}_{\mathcal{A} \cup i}\|_0 = \|\mathcal{A}\|_0 + 1} (\mathcal{E}_{\mathcal{A}} - \mathcal{E}_{\mathcal{A} \cup i}) \leq \lambda \leq \min_{i \in \mathcal{A}} \left[\frac{\mathcal{E}_{\mathcal{A}} - \mathcal{E}_{\mathcal{A} \setminus i}}{\|\mathbf{x}_{\mathcal{A} \setminus i}\|_0 - \|\mathcal{A}\|_0} \right]. \quad (14)$$

The lower bound of (14) can be simplified to $\max_{i \notin \mathcal{A}} (\mathcal{E}_{\mathcal{A}} - \mathcal{E}_{\mathcal{A} \cup i})$ because if $i \notin \mathcal{A}$ is such that $\|\mathbf{x}_{\mathcal{A} \cup i}\|_0 = \|\mathcal{A}\|_0$, then (13) implies that $\mathcal{E}_{\mathcal{A}} - \mathcal{E}_{\mathcal{A} \cup i} = 0$. Thus, including these indices i in the computation of the lower bound of (14) does not change its value, and (14) simplifies to

$$\max_{i \notin \mathcal{A}} (\mathcal{E}_{\mathcal{A}} - \mathcal{E}_{\mathcal{A} \cup i}) \leq \lambda \leq \min_{i \in \mathcal{A}} \left[\frac{\mathcal{E}_{\mathcal{A}} - \mathcal{E}_{\mathcal{A} \setminus i}}{\|\mathbf{x}_{\mathcal{A} \setminus i}\|_0 - \|\mathcal{A}\|_0} \right]. \quad (15)$$

Given λ_q , the next λ -value $\lambda_{q+1} < \lambda_q$ is found by computing the lower bound of (15).

How to choose λ_{q+1} ? Setting λ_{q+1} to the lower bound of (15) is not judicious, since for this λ -value, \mathcal{J} is still “locally minimum w.r.t. \mathcal{A} ” in the sense of (12). One possibility is to set λ_{q+1} to the lower bound of (15) minus some $\varepsilon > 0$, without guarantee that this value is larger than the “next lower bound” of (15). Another possibility is to sort the values of

$$\tilde{\lambda}_i \triangleq \mathcal{E}_{\mathcal{A}} - \mathcal{E}_{\mathcal{A} \cup i} \geq 0 \quad (16)$$

for all indices $i \notin \mathcal{A}$, and then to **set λ_{q+1} to the mean of the two largest values**. This setting ensures that the inequality $\tilde{\lambda}_i \leq \lambda_{q+1}$ does not hold for one value of $\tilde{\lambda}_i$ only.

- If the number of indices $i \notin \mathcal{A}$ such that $\tilde{\lambda}_i > 0$ is equal to 1, then we set λ_{q+1} to half of the value of $\tilde{\lambda}_i$.
- If all indices $i \notin \mathcal{A}$ are such that $\tilde{\lambda}_i = 0$, then we terminate CSBR.
- If \mathcal{A} is the complete set $\{1, \dots, n\}$, the lower bound of (15) is undefined, and we terminate CSBR.

Remark 7 For a given $i \notin \mathcal{A}$ for which $\|\mathbf{x}_{\mathcal{A} \cup i}\|_0 = \|\mathcal{A}\|_0 + 1$, $\tilde{\lambda}_i$ is the λ -value for which both affine curves $\lambda \mapsto \mathcal{J}_{\mathcal{A}}(\lambda)$ and $\lambda \mapsto \mathcal{J}_{\mathcal{A} \cup i}(\lambda)$ intersect. Similarly, for $i \in \mathcal{A}$, the value

$$\tilde{\lambda}_i \triangleq \frac{\mathcal{E}_{\mathcal{A}} - \mathcal{E}_{\mathcal{A} \setminus i}}{\|\mathbf{x}_{\mathcal{A} \setminus i}\|_0 - \|\mathcal{A}\|_0} \quad (17)$$

is the λ -value for which both affine curves $\lambda \mapsto \mathcal{J}_{\mathcal{A}}(\lambda)$ and $\lambda \mapsto \mathcal{J}_{\mathcal{A} \setminus i}(\lambda)$ intersect.

Proof 16 For $i \notin \mathcal{A}$ and for the λ -value $\tilde{\lambda}_i$, (13) is an equality, then $\mathcal{J}_{\mathcal{A}}(\tilde{\lambda}_i) = \mathcal{J}_{\mathcal{A} \cup i}(\tilde{\lambda}_i)$. Since $\|\mathbf{x}_{\mathcal{A} \cup i}\|_0$ is supposed to be different from $\|\mathcal{A}\|_0$, both affine curves $\lambda \mapsto \mathcal{J}_{\mathcal{A}}(\lambda)$ and $\lambda \mapsto \mathcal{J}_{\mathcal{A} \cup i}(\lambda)$ are not parallel (their slopes are equal to $\|\mathcal{A}\|_0$ and $\|\mathbf{x}_{\mathcal{A} \cup i}\|_0 = \|\mathcal{A}\|_0 + 1$ respectively), and they intersect at $\lambda = \tilde{\lambda}_i$. A similar proof holds in the case where $i \in \mathcal{A}$ and \cup is replaced by \setminus .

Remark 8 A set \mathcal{A} of cardinality larger than $\min(m, n)$ cannot be explored.

Proof 17 — SBR: if a set \mathcal{A} of cardinality larger than $\min(m, n)$ is explored, then SBR has earlier explored at least one set \mathcal{B} of cardinality $\min(m, n)$ (recall that the initial solution is $\mathcal{A} = \emptyset$). Due to the URP assumption, $\mathcal{E}_{\mathcal{B}} = \mathcal{E}_{\mathcal{A}} = \min_{\mathbf{x} \in \mathbb{R}^n} \mathcal{E}(\mathbf{x})$ is the optimal least-square cost. Therefore, $\|\mathcal{A}\|_0 > \|\mathcal{B}\|_0 \Rightarrow \forall \lambda > 0, \mathcal{J}_{\mathcal{A}}(\lambda) > \mathcal{J}_{\mathcal{B}}(\lambda)$. This cannot occur because SBR is a descent algorithm.

— CSBR. Recursively, for each $\lambda = \lambda_q$, if the initial set \mathcal{A}_{q-1} (input of $SBR(\lambda_q)$) is of cardinality lower than $\min(m, n)$, then the output \mathcal{A}_q of $SBR(\lambda_q)$ is also of cardinality lower than $\min(m, n)$.

4.3 Termination of CSBR

Remark 9 When CSBR terminates, $\mathcal{E}_{\mathcal{A}} = \mathcal{J}_{\mathcal{A}}(0)$ is locally minimum w.r.t. \mathcal{A} , then $\forall i \notin \mathcal{A}, \mathcal{E}_{\mathcal{A} \cup i} = \mathcal{E}_{\mathcal{A}}$.

Remark 10 CSBR terminates after a finite number of SBR iterations.

Proof 18 • According to remark 4, for a given λ -value λ_q , $SBR(\lambda_q)$ terminates after a finite number of iterations.

- According to remark 7, each value of λ_q is such that there exists $\underline{\mu}_q$ and $\overline{\mu}_q$ such that

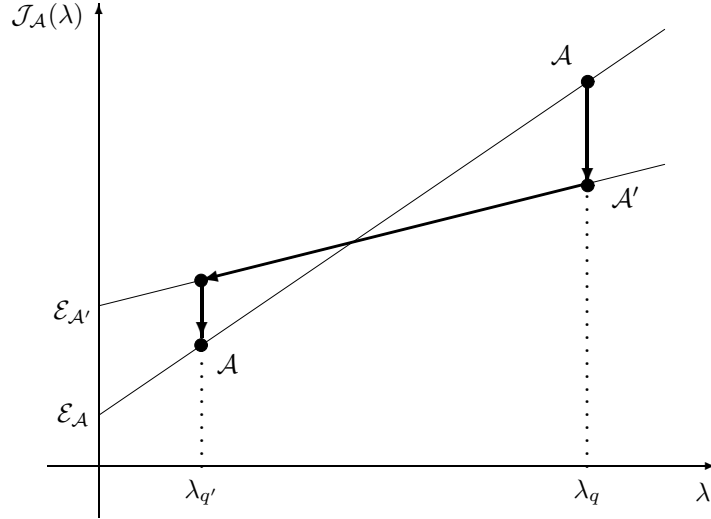


Figure 3: A given set \mathcal{A} may be explored twice during the CSBR procedure, at different λ -values. Here, each vertical line corresponds to one call to SBR (\mathcal{A} and \mathcal{A}' are two SBR iterates at λ_q and $\lambda_{q'}$), *i.e.*, to a fixed λ -value while the plain lines are the affine curves $\lambda \mapsto \mathcal{J}_{\mathcal{A}}(\lambda)$ and $\lambda \mapsto \mathcal{J}_{\mathcal{A}'}(\lambda)$.

- $0 \leq \underline{\mu}_q \leq \lambda_q \leq \overline{\mu}_q$;
- $\lambda_q = (\underline{\mu}_q + \overline{\mu}_q)/2$;
- $\underline{\mu}_q$ and $\overline{\mu}_q$ are critical values for which two affine curves $\lambda \mapsto \mathcal{J}_{\mathcal{A}}(\lambda)$ and $\lambda \mapsto \mathcal{J}_{\mathcal{A}}(\lambda \cup i)$ intersect.

From the recursive construction of the sequence $(\lambda_q, q \geq 0)$, it is clear that $\forall q, \overline{\mu}_q < \lambda_{q-1}$, thus $\forall q, \overline{\mu}_q < \overline{\mu}_{q-1}$. Since each value of $\overline{\mu}_q$ can be associated to a given intersection between two affine curves $\lambda \mapsto \mathcal{J}_{\mathcal{A}}(\lambda)$ and $\lambda \mapsto \mathcal{J}_{\mathcal{B}}(\lambda)$ and the number of possible subsets \mathcal{A} and \mathcal{B} of $\{1, \dots, n\}$ whose cardinality is lower than $\min(m, n)$ is finite, the number of possible values taken by $\overline{\mu}_q$ is also finite. Since the sequence $(\overline{\mu}_q, q \geq 0)$ satisfies $\forall q, \overline{\mu}_q < \overline{\mu}_{q-1}$, we conclude that the number of iterations q at which SBR(λ_q) is run is finite.

Despite remark 10, we cannot claim that a given set \mathcal{A} is never explored twice during the CSBR procedure. In remark 4, we have seen that a given set \mathcal{A} can never be explored twice while running SBR for a given λ -value. However, \mathcal{A} may be explored several times while running CSBR, *i.e.*, once while running SBR at some λ -value λ_q , and another time while running SBR at another λ -value $\lambda_{q'} \leq \lambda_q$ (for $q' > q$). See Fig. 3 for a simple illustration.

Remark 11 When CSBR terminates, the solution $\mathbf{x}_{\mathcal{A}}$ is an unconstrained least-square estimate.

Proof 19 Let us define the residual $\mathbf{r} = \mathbf{y} - \mathbf{A}\mathbf{x}_{\mathcal{A}}$ and the unit vectors $\mathbf{e}_i \in \mathbb{R}^n$ in which all entries are equal to 0 except the i -th entry, equal to 1. Then, $\mathbf{A}\mathbf{e}_i = \mathbf{a}_i$, where \mathbf{a}_i stands for the i -th column of \mathbf{A} .

Firstly, we prove that $\forall i \notin \mathcal{A}, \mathbf{a}_i^T \mathbf{r} = 0$. According to remark 9, \mathcal{A} is such that $\forall i \notin \mathcal{A}, \mathcal{E}_{\mathcal{A} \cup i} = \mathcal{E}_{\mathcal{A}}$. Then, the following inequalities hold:

$$\begin{aligned} \forall i \notin \mathcal{A}, \quad \forall \varepsilon \in \mathbb{R}, \quad \mathcal{E}(\mathbf{x}_{\mathcal{A}} + \varepsilon \mathbf{e}_i) - \mathcal{E}(\mathbf{x}_{\mathcal{A}}) &\geq 0 \\ \forall i \notin \mathcal{A}, \quad \forall \varepsilon \in \mathbb{R}, \quad \|\mathbf{r} - \varepsilon \mathbf{a}_i\|^2 - \|\mathbf{r}\|^2 &\geq 0 \\ \forall i \notin \mathcal{A}, \quad \forall \varepsilon \in \mathbb{R}, \quad \varepsilon^2 \|\mathbf{a}_i\|^2 - 2\varepsilon \mathbf{a}_i^T \mathbf{r} &\geq 0 \\ \forall i \notin \mathcal{A}, \quad \mathbf{a}_i^T \mathbf{r} &= 0. \end{aligned}$$

Secondly, because $\mathbf{x}_{\mathcal{A}}$ is a solution to the constrained problem (2):

$$\begin{aligned} \forall i \in \mathcal{A}, \quad \forall \varepsilon \in \mathbb{R}, \quad \mathcal{E}(\mathbf{x}_{\mathcal{A}} + \varepsilon \mathbf{e}_i) - \mathcal{E}(\mathbf{x}_{\mathcal{A}}) &\geq 0 \\ \forall i \in \mathcal{A}, \quad \mathbf{a}_i^T \mathbf{r} &= 0. \end{aligned}$$

Finally, we deduce that

$$\begin{aligned} \forall i \in \{1, \dots, n\}, \quad \mathbf{a}_i^T \mathbf{r} &= 0 \\ \mathbf{A}^T \mathbf{r} &= \mathbf{0} \\ \mathbf{A}^T \mathbf{A} \mathbf{x}_{\mathcal{A}} - \mathbf{A}^T \mathbf{y} &= \mathbf{0} \\ \nabla \mathcal{E}(\mathbf{x}_{\mathcal{A}}) &= \mathbf{0}. \end{aligned}$$

Since \mathcal{E} is quadratic, we have shown that $\mathbf{x}_{\mathcal{A}}$ is an unconstrained least-square estimate.

References

- [1] I. F. Gorodnitsky and B. D. Rao, “Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm”, *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 600–616, Mar. 1997.
- [2] I. Das and J. E. Dennis, “A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems”, *Structural optimization*, vol. 14, no. 1, pp. 63–69, 2007.